Global Hate Speech Classifier

Sarah Gould and Fae Mahmoud and Brandon Moretti
University of Pittsburgh | Human Language Technologies

Abstract

We create a multiclass target identity classifier to identify trends in hate speech across different societies. We use the UC Berkely *Measuring Hate Speech* dataset to train a DistillBERT classifier. (Sachdeva et al., 2022) We test the classifier on English-translated data from three datasets: *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language*, *RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets*, and *BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection* (Mulki et al., 2019), (Assenmacher et al., 2021), (Moon et al., 2020). We also use a zero-shot instruction tuned prompts of Open AI's gpt-4o-mini to classify the data.

1 Introduction

Hate speech is becoming an increasingly prevalent issue. It can be found in corners of the Internet around the world (Persily and Tucker, 2020). However, much of the research surrounding hate speech focuses on the English language and Western societal trends. While there is some research on the targets of hate speech in English, information about target identities in different languages is lacking. For this project we are attempting to understand the global patterns of hate speech, and how the identities-targeted vary across societies and languages. To explore the global trends of targeted identities, we used used two methods to create a multi-class identity target labeler. The goal was to take an input string of text and output a target identity label for that string.

We trained a DistillBERT multi-class classifier on a large English-language hate speech dataset, *Measuring Hate Speech*. This dataset contained multi-label columns for the target identities, which we converted to a multi-class column before using it to fine-tune DistillBERT. DistillBERT had an average Accuracy of 0.86, Precision of 0.81, Recall of 0.79, and a F1 score of 0.80. However, the performance metrics for the various target identities were quite varied. For example, precision was between .53 (Men) and .94 (Jewish), and recall was between .40 (Men) and .98 (Black)

We also used a zero-shot instruction-tuned approach using OpenAi's gpt-4o-mini model for a more modern LLM Approach to the problem. The goal of this was to get an accurate measuring on how well the LLM was able to evaluate which group was being targeted given a text deemed as hate speech compared to our own classifier.

We then used both our DistillBERT classification model and the LLM approach to label hate speech data in Arabic, German, and Korean. We found that for the DistillBERT model the distributions varied widely between the three different languages. Generally the more frequent target identities were identities one was likely to interact with frequently in a specific country, such as White for Germany and Asian for Korean. Each language had 3-4 identities which where frequently targeted while most identities were far less frequently targeted.

We manually evaluated the Arabic, German, and Korean datasets by randomly collecting 100 samples of each and assessing whether the predicted label matched the identity the English-translated text appeared to be targeting. The manual evaluation for the DistillBERT model showed 47 percent accuracy for the Arabic dataset. It showed 53 percent accuracy for the German dataset, and it showed 51 percent accuracy for the Korean Dataset.

The LLM modle showed howed 79 percent accuracy for the Arabic dataset. It showed 78 percent accuracy for the German dataset, and it showed 70 percent accuracy for the Korean Dataset.

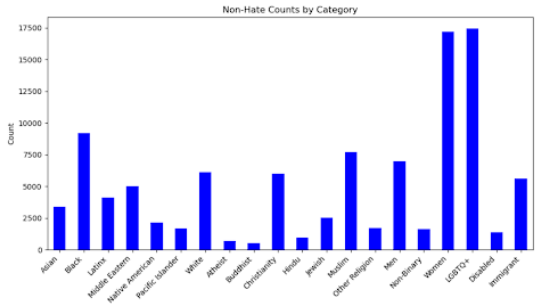


Figure 1: Distribution of Target-Identities in *Measuring Hate Speech*

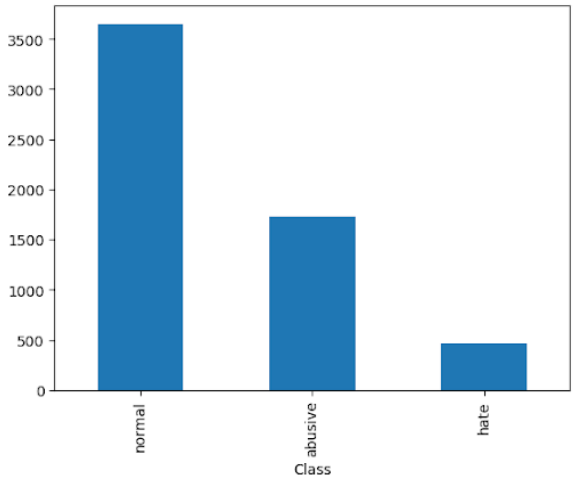


Figure 2: Distribution of abusive, normal, and hate comments for *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language*

2 Data

We used four datasets for this project. We used one - *Measuring Hate Speech* - to train our baseline hate speech classifier. The three additional datasets were used to apply the trained classifier to text translated to English from the original languages. The three non-English data sets are in Arabic, German, and Korean. We chose these in an attempt to use data from a variety of languages and geographic areas. In addition, we attempted to choose languages with which at least one of the team members had some experience. We are differentiating areas based on the language used. We found the datasets using `https://hatespeechdata.com/`

2.1 *Measuring Hate Speech*

This is the main dataset we are using to train our dataset. It consists of 39,565 social media comments collected from Instagram, YouTube, and Reddit, and annotated by a team of 7,912 crowd-sourced annotators. The dataset has 131 columns. We will be using 38. This is a mult-labled dataset, meaning that there were many columns each with a binary value indicating whether the text sample fit the specific target identity or not.

2.1.1 Preprocessing *Measuring Hate Speech*

As this is the dataset we used to fine-tune our BERT classifier, it required more intensive preprocessing then our other datasets. We first loaded the 38 relevant columns. Then, since we are attempting to classify the target identity we dropped the all of the non-hate columns from the dataset. We then condensed columns with similar meanings, using guidelines from Yoder et. al. (Yoder et al., 2022). Figure 1 shows the distribution of target identites before low frequency identities were dropped.

We then removed low-frequency (less than 1000 rows) targets from the dataset. At this point we dropped all rows with multiple labeled identities, or rows which had no labeled target identity after the high-frequency examples were removed.

Then, using the `pd.from_dummies()` command we converted the multilabel classification into a multirow classification. In essence, we transformed the many columns of boolean data into a single column of strings - each of which corresponds to the label of a column.

2.2 *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language*

This is our Arabic dataset. It consists of 5851 rows, and each row is a tweet and a classification of that tweet. The focus of the dataset is Syrian/Lebanese political tweets. The dataset has two Columns. The first is the Arabic text we will be classifying The second is the classification of whether the corresponding tweet is “abusive”, “normal” or, “hate”. Figure 2 shows the distribution of abusive, normal, and hate comments. Preprocessing of the data consisted of removing null values, dropping the normal (not abusive or hate) data, and translating the text into English.

2.3 *RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets*

This is our German dataset. It consists of comments on German news sites and their respective rating. The comments were assessed both by a moderator and crowd-sourced annotators. The classification includes the number of crowd-sourced annotators

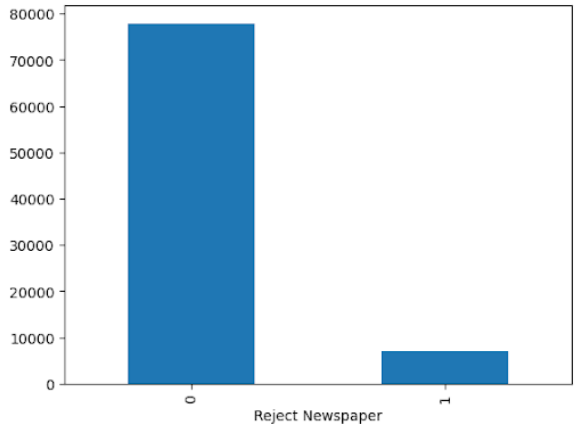


Figure 3: Distribution of Rejected or Not Rejected for *RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets*

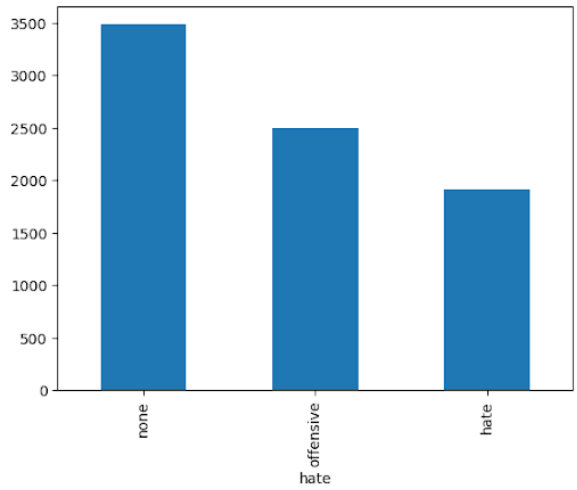


Figure 4: Distribution of Offensive, Hateful, or Normal for *BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection*

who rejected the comment for each specific reason (racisim, threeat, advertising, etc.), whether the moderator rated the comment as offensive, and whether the comment was rejected. There are 85000 rows, which each correspond to a single comment and the classification of that comment. This dataset consists of 14 columns, of which we used 10. Figure 3 shows the distribution of Rejected or Not Rejected comments. Preprocessing of the data consisted of removing null values, dropping the Not Rejected data, and translating the text into English.

2.4 *BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection*

This is our Korean data set. It consists of 7897 human annotated rows. Each row is a comment found in NAVER entertainment news. Each row has the text, the classification of whether it contains Gender Bias, the classification of whether it contains bias in general, whether it is Offensive, Hateful, or neither, and the title of the article it was listed under. Figure 4 shows the distribution of Offensive, Hate, or None language in this dataset. Preprocessing of the data consisted of removing null values, dropping the None (not Offensive or Hateful) data, and translating the text into English.

2.5 Translation

We translated the non-English data using Google Sheets. Each data-set was first imported into Google Sheets. Then a column was created next to the original language text field, and the following command was used: `=GOOGLETRANSLATE(text, [source_language, target_language])`. Some challenges associated with this method is that Google Sheets would crash if one attempted to translate too many lines at once.

3 Methods

We used a BERT Classifier for our standard method, and instruction-tuned zero shot prompts for our modern method.

3.1 BERT Classifier

BERT is a very powerful classification model that is used for many classification tasks such as the task we are performing in this project. Our BERT classifier model utilized a multi-class, single-label structure. More specifically, we used the DistilBERT model as it is a more compact version of BERT that runs quicker.

By renaming the "multi-class" column to "label", and further fitting a label encoder to the column, we successfully prepared our dataset for preprocessing and training.

Much of our DistillBERT classifier was inspired by Omoniyi's article (Omoniyi, 2024).

3.1.1 Training DistilBERT

Prior to training, we must preprocess our data to be fed into the DistilBERT model. First, we split our data into train/test subdivisions using a 90/10 ratio. By defining a preprocess() function, the "text" column of our data table can be tokenized using Transformer's AutoTokenizer.

The labels are then preserved with tokenization and added to the tokenized dataframe as a column.

The model can then be initialized, making sure to pass the appropriate arguments being `num_labels=14`, as that corresponds to the amount of demographic targets we are training on and `problem_type="single_label_classification"`, corresponding to the multi-class, single-label classification problem that we are working with. Our training arguments are then set as follows:

```
training_args = TrainingArguments(
   output_dir="hatespeech_classifier",
    learning_rate=4e-5,
    per_device_train_batch_size=16,
    per_device_eval_batch_size=16,
    num_train_epochs=5,
    weight_decay=0.01,
    warmup_ratio=0.1,
    eval_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True,
    report_to="none",
)
```

With the training arguments set, the trainer object can be created with the model, the train/test tokenized data, training arguments, and the metric compute function. The metric compute function generates the extracts the predictions and labels and applies an argmax on the predictions. The labels, predictions, and class names (which is just a dictionary containing each demographic in our model) are then passed into `classification_report` so the metrics can be evaluated. After that, all that is left is to run `trainer.train()`.

3.1.2 Using DistillBERT

Now that we have a trained model, it is time to apply the model to other datasets. Using the three datasets that have already been pre-processed, the model will assign a predicted label to each text. The `predict_category()` function was defined to take in a string of text and output the predicted label.

The input string is first tokenized similar to how it is in the train model preprocessing. The model is then set to evaluation mode, and then the main logic of the prediction process is ran in `torch.no_grad()` mode to speed up the calculation process. Following that, the input is sent to the model, the logits are extracted from the model output, and the argmax of the logits give us our predicted target.

Epoch	Training Loss	Validation Loss	Accuracy
1	0.513500	0.516823	0.853406
2	0.418800	0.476644	0.864284
3	0.343900	0.478549	0.866615
4	0.285400	0.529267	0.866097
5	0.228100	0.556913	0.864802

Figure 6: Model Accuracy

Now, for each dataset, every row of text is passed into the `predict_category()` function and appended to an array that can then paired with its assosscciated text input to construct the data frame that is then outputted to a .csv file.

3.2 LLM

For our LLM, we used a zero-shot instruction-tuned classification approach using OpenAi's gpt-4o-mini model. gpt-4o-mini was trained on Microsoft Azure AI supercomputers. Below you will find the prompt we used for the model.

3.2.1 LLM

```
OA_PROMPT = """
The following text is hate speech, identify the target from the following set of labels, answer in only one or two words.
Here is a list of possible target identities:
    Race: asian, black, latinx, Middle Eastern, White;
    Religion:Christian, Jewish, Muslim;
    Origin: Immigrant;
    Gender: Transgender;
    Sexuality: Bisexual, Gay, Lesbian, other;
    Disability: Disabled.

Text: {input}
"""
```

Figure 5: Prompt for gpt-4o-mini

It is OpenAi's most advanced system, and according to them, it typically produces safer and more user-friendly responses compared to their other models. We created a prompt that asked the model to identify the target of hate speech and provided a list of possible target identities. This included subsections of race, gender, religion, origin, sexuality and disability. Not all of our datasets were from English speaking countries, so we translated our datasets all to English and then ran them through the LLM.

4 Results

4.1 DistillBERT

We evaluated our model by investigating the metric scores across the entire model as well as for each target demographic. Each prediction dataset's results are visualized as well for interpretation.

	precision	recall	f1-score	support
Asian	0.93	0.91	0.92	231
Black	0.91	0.98	0.94	903
Latinx	0.88	0.84	0.86	61
Middle Eastern	0.91	0.87	0.89	167
White	0.74	0.83	0.78	226
Christian	0.93	0.91	0.92	228
Jewish	0.94	0.85	0.89	575
Muslim	0.76	0.79	0.77	128
Men	0.53	0.40	0.46	103
Women	0.59	0.41	0.49	99
LGB	0.88	0.91	0.89	233
Transgender	0.67	0.79	0.73	47
Disabled	0.83	0.69	0.75	166
Immigrant	0.84	0.89	0.87	694
accuracy			0.86	3861
macro avg	0.81	0.79	0.80	3861
weighted avg	0.86	0.86	0.86	3861

Figure 7: Metrics by Class

4.1.1 Performance

Figure 6 shows the general performance of the model. Using 5 epochs, we can observe how the metrics change across each epoch. Training loss decreased for each epoch, while validation loss decreased only for the 2nd epoch and increased for each following epochs. Accuracy increased until the 3rd epoch where it peaked before decreasing ever so slightly in the successing epochs.

We the nlooked at the multi-class perfomance of the model. In Figure 7, metric scores (precision, recall, f1-score, support) are broken down by each target class. This allows us to observe which target demographics were the most easy or difficult for the model to accurately identify. Both sex classes appear to have been the most difficult for the model accurately identify. On the other hand, the model fared much better for minority ethnicity classes. The Black demographic had a near perfect recall score. Contrary to the minority ethnicity classes, the White class had significantly poorer metrics in comparison. When looking at the macro average scores, we can observe that the men/women classes may be responsible for dragging down the score. If these two are disregarded, the model performed well overall across the other demographic classes.

4.1.2 Classification Results

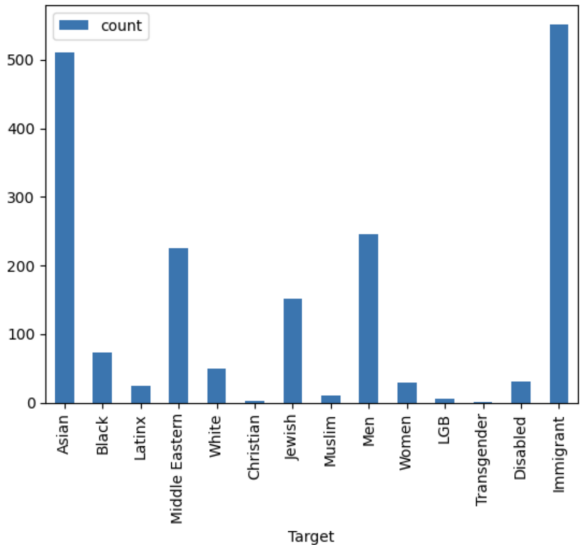


Figure 8: Distribution of Predicted Target Demographics for *BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection*

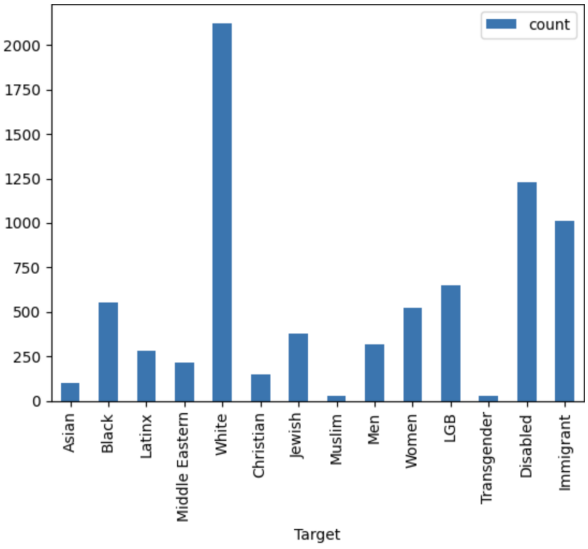


Figure 9: Distribution of Predicted Target Demographics for *RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets*

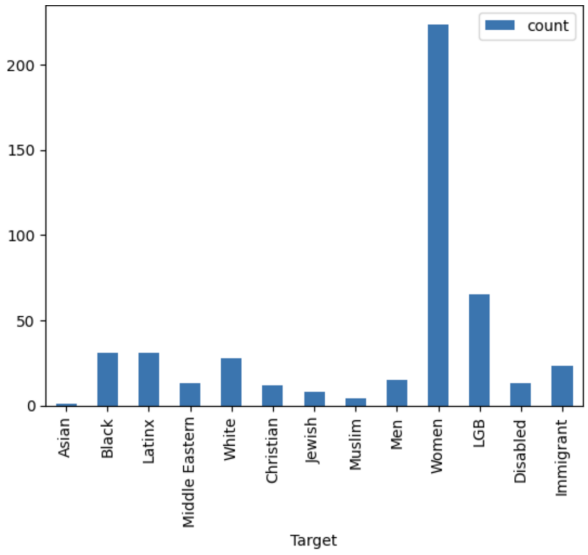
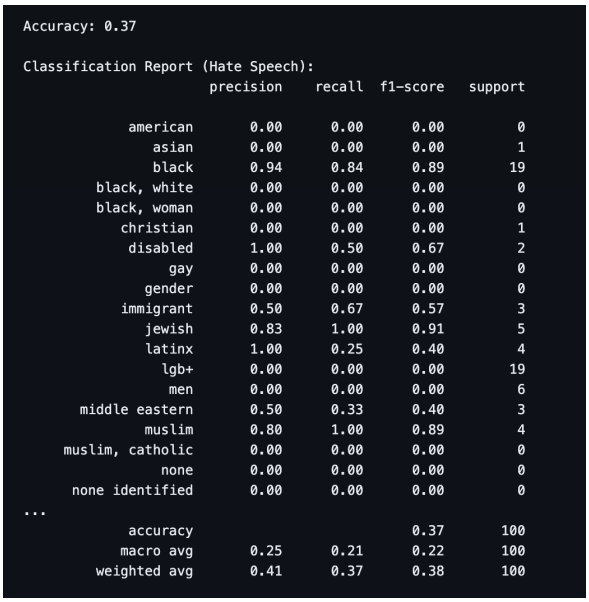


Figure 10: Distribution of Predicted Target Demographics for *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language*



```
Accuracy: 0.37

Classification Report (Hate Speech):
                     precision    recall  f1-score   support

           american       0.00      0.00      0.00         0
              asian       0.00      0.00      0.00         1
              black       0.94      0.84      0.89        19
        black, white       0.00      0.00      0.00         0
        black, woman       0.00      0.00      0.00         0
          christian       0.00      0.00      0.00         1
           disabled       1.00      0.50      0.67         2
                gay       0.00      0.00      0.00         0
             gender       0.00      0.00      0.00         0
          immigrant       0.50      0.67      0.57         3
             jewish       0.83      1.00      0.91         5
             latinx       1.00      0.25      0.40         4
               lgb+       0.00      0.00      0.00        19
                men       0.00      0.00      0.00         6
     middle eastern       0.50      0.33      0.40         3
             muslim       0.80      1.00      0.89         4
  muslim, catholic        0.00      0.00      0.00         0
               none       0.00      0.00      0.00         0
    none identified       0.00      0.00      0.00         0
...
           accuracy                           0.37       100
          macro avg       0.25      0.21      0.22       100
       weighted avg       0.41      0.37      0.38       100
```

Figure 11: Distribution of Predicted Target Demographics for *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language*

After creating the new datasets containing the texts along with their respective predicted target demographic, the results were graphed for easy interpretation.

Figure 8 shows the counts for each target demographic when using the *BEEP! Korean Corpus of Online News Comments for Toxic Speech Detection* dataset. Immigrants and Asians are the notable targets from this particular dataset, with much higher frequencies than the rest.

Figure 9 shows the counts for each target demographic when using the *RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets* dataset. The most notable target in this dataset were white people.

Figure 10 shows the counts for each target demographic when using the *L-HSAB: A Levantine Twitter Dataset for Hate Speech and Abusive Language* dataset. The most notable target in this dataset are women.

4.2 LLM

4.2.1 Performance against Measuring Hate Speech Data

Figure 11 shows the results of the LLM gpt-4o-mini classifier against the Measuring Hate Speech data set. It had difficulty identifying lgb+ identities, with an accuracy reading of zero percent, and .50 percent for middle eastern and immigrant identities. It has an overall accuracy reading of .37.

4.3 Comparing Classification Accuracy between DistillBERT and LLM Approach

To assess the accuracy of the classifiers for the non-English data, we manually evaluated the predictions by randomly collecting 100 samples of each and assessing whether the predicted label matched the identity the English-translated text appeared to be targeting. We did this for both the DistillBERT predictions and the LLM predictions. (Although it would have been trivial to use the LLM method on the non-translated data, for purposes of a fair comnparison we decided to use the same inputs for BERT and the LLM).

4.3.1 Arabic

Figure 12 shows the accuracy of the BERT predictions on the Arabic dataset.

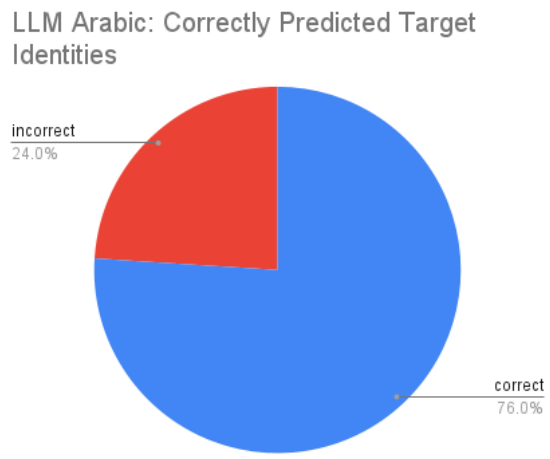


Figure 13: Manual Evaluation Arabic: LLM

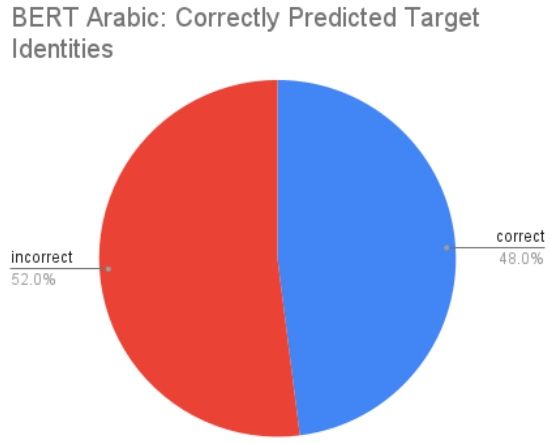


Figure 12: Manual Evaluation Arabic: BERT

Figure 13 shows the accuracy of the LLM predictions on the Arabic dataset.

4.3.2 German

Figure 14 shows the accuracy of the BERT predictions on the German dataset.

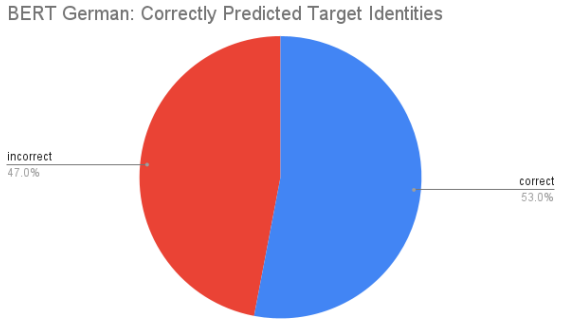


Figure 14: Manual Evaluation German: BERT

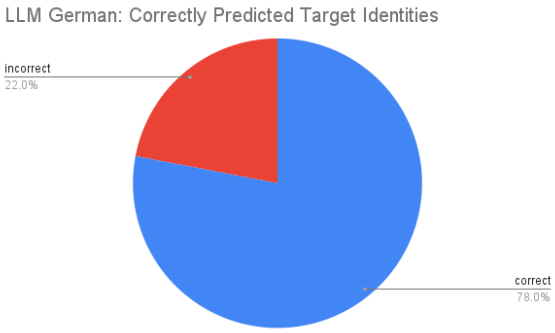


Figure 15: Manual Evaluation German: LLM

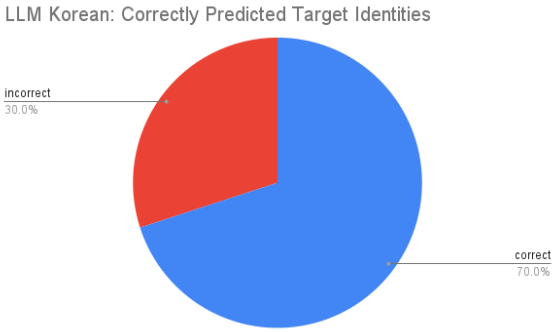


Figure 17: Manual Evaluation Korean: LLM

Figure 15 shows the accuracy of the LLM predictions on the German dataset.

4.3.3 Korean

Figure 16 shows the accuracy of the BERT predictions on the Korean dataset.

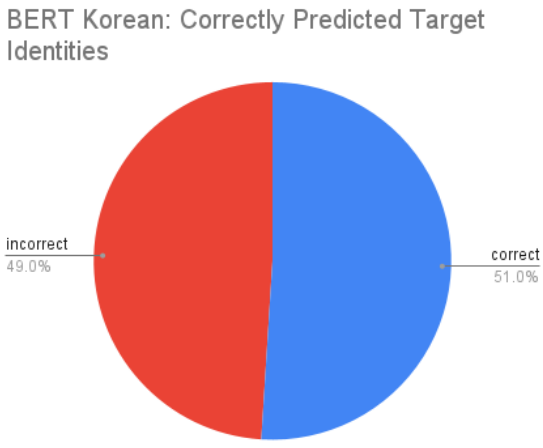


Figure 16: Manual Evaluation Korean: BERT

Figure 17 shows the accuracy of the LLM predictions on the Korean dataset.

4.3.4 Comparison

In general the LLM did significantly better at predicting the same samples. Some of this may be due to the fact that the LLM was trained on far more data than our fine-tuned task. Another major factor is likely the fact that the LLM would frequently identify when No hate was found within the data. Despite the fact that we dropped the normal data from the non-English language sets, a significant amount of the data points had no clear target identity associated with them. The LLM was able to recognize that, while BERT was not.

5 Discussion

The results of this classifier carry a significant weight. It brings to light the hate speech that many marginalized groups experience on a daily basis. Often, hate speech is subtle, which causes it to be overlooked or downplayed. This classifier has the potential to expose harmful language that targets and oppresses marginalized communities. It also highlights the unique struggles faced by different groups. Marginalized communities are often subjected to different stereotypes and forms of oppression due to cultural and social biases. For example, people who come from an Asian background are more likely to be stereotyped as smart and gifted, while those who are African American are deemed as dumb and aggressive. By spotting these patterns in language, the classifier makes it easier for people to recognize and call out hate speech—even when it's not incredibly obvious, like with microaggressions or stereotypes. For example, out classifier even began labeling every text that had the word "idiot" for one of our datasets as targeting women, since that happened to be the highest insult to women in the data set. Additionally, the classifier reveals which groups are most frequently targeted by speakers of different languages. Certain cultures may harbor biases against others due to historical conflicts or societal norms, and this tool can help uncover those underlying patterns.

5.1 Implications of Classifier Results

The fact that the data is translated from the other languages to English may be a cause of some of the inaccuracy as some information is inevitably lost in translation.

Something interesting about these results is that generally the most frequently targeted identities they are associated with identities that are more common in the region the language is spoken, for example. 'White' is the most common in German. Similarly 'Asian' is the second most common for Korean. Some of this is likely due to the topic of the data sets themselves. For example, the Arabic data set had an explicitly political context, so many of the comments reference the Middle East. Similarly, the German data set, being taken from News websites, also frequently mentions political issues. The data set frequently referenced other political parties and was offensive language targeted at other people of the same race, but of different political parties.

Another reason for this could be that that some of the target identity labels are based off of names. For example, in the Korean data set when a Korean name was used that strongly influenced that seemed to strongly influence the classifiers (both LLM and BERT) giving a target identity of 'Asian' whether or not there are any other aspects to mentioned. Similarly if a group was mentioned, the classifiers often identified that group, whether or not the hate was directed towards that target identity.

6 Future Work

There are a few areas for future work on this project. The first is expanding the languages labeled. Due to time constraints we focused on three languages for this project, but a future goal would be to have a truly global perspective to see how hate speech trends differ across cultures. Part of this goal could be to expand the classifier to a truly multilingual classifier that could evaluate data in several languages, rather than using the English-translated text.

Another area to expand the project would be to fully manually label the test data from the non-English data sets. This would allow us to more fully assess the performance of the data, and could assist with the goal of creating a multi-class target identity classifier for non-English datasets. Part of this would be to ideally work with some native language speakers to properly label the text in their own label. This would enable us to more accurately evaluate the model as we would avoid the information lost due to the translation of the text. In addition, this would allow us to have a better understanding of the culture context involved with the hate speech.

7 Limitations

Three big limitations of our classifier were misinformation, data-set accuracy, and tone. Due to our limited data sets, there is the possibility of misinformation that can lead to misclassification. There is also the issue of nuance when it comes to misinformation. Different cultures consider different things offensive, so something offensive in English may not be seen as offensive in Arabic. It makes it even more complicated when you consider humor. Something regarded as offensive humor in one language may be considered innocent in another. Different languages express humor in such different ways that it would be considerably difficult for a classifier to accurately recognize it and accurately classify texts that are meant to be humorous as hateful or not.

Another huge issue was the accuracy of the date set. The data sets we used were not super up-to-date, and countries can change and become more progressive or conservative rather quickly. A data set that is not continuously updated may not accurately represent current social trends or the latest forms of hate speech. Certain terms that were once deemed acceptable may now be considered offensive. New terms of slang could emerge that carry hateful connotations. If the classifier is fed outdated data, it will not be able to accurately represent and keep up with the changing times. The last big issue is that it is incredibly different to differentiate tone in text. The meaning of a phrase can change drastically depending on the tone one uses. A classifier that can't determine tone may label non-hateful speech as hateful.

Manual evaluation also had some severe limitations. Because we did not have the necessary expertise with the languages, we evaluated the labels using the Google translated English, which likely imperfectly translated the meaning of the statements. In addition, we lack the cultural context to truly understand the meaning of some of the statements to assess the target identity. Also, the lack of context to the original post made it difficult to properly assess the target of the text of a comment on a post, as often the text referenced the original post, or comments made by other users.

8 Ethical Issues

This project is a hate speech class identifier, and as such, it matches offensive language with specific marginalized groups. This classifier brings up a few ethical issues. The biggest one is creating and reinforcing biases. Typically with all machine learning systems, there is a risk of bias. If the model is trained on a biased dataset, it may reinforce the existing stereotypes. It may portray certain language-speaking countries as hateful and aggressive when that simply isn't the case. Especially when it comes to the limitations our classifier and data sets already face, if we feed it in a data set that overly portrays, for example, french-speaking regions as super hateful and aggressive against a certain group, the classifier is going to return that French-speaking regions are extremely biased against that group, when that might not be the case.

Another issue is determining whether the speech is hate or criticism. A sentence may be expressing genuine criticism regarding another marginalized group without it being hateful or aggressive. On the other hand, subtle forms of hate can disguise itself as criticism. The classifier may have an issue distinguishing between the two. This can create issues by over-marking text as hate and suppressing valid expression.

9 Group Member Task Breakdown

- Sarah Gould: Data Processing, Text Translation, Manual Evaluation.
- Fae Mahmoud: LLM Classifier and performance metrics
- Brandon Moretti: BERT Classifier and performance metrics

References

Dennis Assenmacher, Marco Niemann, Kilian Müller, Moritz Seiler, Dennis Riehle, Heike Trautmann, and Heike Trautmann. 2021. Rp-mod & rp-crowd: Moderator- and crowd-annotated german news comment datasets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

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.

Temidayo Omoniyi. 2024. Distilbert for multiclass text classification using transformers.

N. Persily and J.A. Tucker. 2020. *Social Media and Democracy*. SSRC Anxieties of Democracy. Cambridge University Press.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How hate speech varies by target identity: A computational analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

A Github Repo

Contact Brandon Moretti to access the private repository which contains the codebase for this project.

`https://github.com/brandonmoretti/MichaelProject`