Multi-Output Hate Speech Modeling: Fine-Tuning Multilingual Transformer Architectures with Multi-Task Hierarchical Models and LIME Explainable AI

Bella Davies, Nadaa Moharram, Ishani Cheshire
bella_davies@berkeley.edu, nadaa_moharram@berkeley.edu, ishacheshi@berkeley.edu
University of California, Berkeley

Abstract

Hate speech detection across multiple languages and social contexts remains a complex challenge. We propose a multi-label, multi-output framework using fine-tuned transformer models (mBERT, XLM-RoBERTa, RemBERT) for English, French, and Arabic. We enhance the MLMA dataset with curated non-hate examples, and use a hierarchical method where there is an explicit dependency structure between the tasks; the output of one task may be used as input for another. Our system predicts five dimensions of each comment: hate presence, directness, target attribute, targeted group, and sentiment. Fine-tuning yields strong gains over baseline performance, evaluated using per-label and per-language F1 scores. Test predictions are explored using explainable AI method LIME (Local Interpretable Model-Agnostic Explanations), providing detailed explanations of predictions to increase model transparency.

Content Warning: This paper contains hate speech examples and offensive language from the labeled datasets.

1 Introduction

Detecting harmful content is critical for user safety and moderation on social media platforms. Online hate speech can spiral into real violence, as when unmoderated hate-speech on Facebook precipitated genocide in Myanmar (Ortutay, 2022). A robust hate speech detection model would be able to help analyze social behavior—for example, do women in fitness on instagram receive more hate comments than men? Most systems are optimized for English, (as Facebook's was in Myanmar), leaving other low resource or morphological complex languages like Arabic less protected. We propose a Multilingual Hate Speech Detection system to classify hate speech, directness, topic of hate speech, hate-targeted groups, and sentiment in social media comments across multiple languages (English, Arabic and French). Additionally, hate speech detection models using transformer architectures are typically used to model hate speech topics with limited classes. We extend this research for a multi-class, multi-label classification task, which includes a mix of multi-class outputs where each observation has exactly one label (hate speech, directness, target, group), and multi-label outputs where each observation has one or more labels (sentiment). Employing LIME explanations on test predictions enables increased transparency for model decision making for each label and language. Addressing a larger multi-class hate-speech classification problem is a step towards building more inclusive, global NLP systems.

2 Background

Jahan and Oussalah (2023) complete a systematic review of hate-speech detection literature, noting that definitions of hate speech differ across platform and context—e.g. which specific categories are protected—but generally agree that hate speech targets specific groups or individuals on the basis of their connection to specific groups. The most prominent categories of hate speech in the surveyed datasets were "general hate," "abusive," and "cyberbullying," which contrary to the earlier definitions, do not appear to target specific identity groups. When compared to non- deep-learning methods, CNNs improved on the baseline by 7% (Jahan and Oussalah, 2023). BERT variants like XLMRoBERTa and mBERT outperform the single pretrained BERT, while HateBERT, a model trained on offensive Reddit comments, outperforms them in turn (Jahan and Oussalah, 2023). Finally, Jahan and Oussalah (2023) note that English tends to dominate in hate speech databases, representing 26 of 69 surveyed datasets.

Srikissoon and Marivate (2023) aim to investigate multi-class hate speech detection using target-generic datasets. Hate speech varies topically - the nature, language and intensity of hate differ when directed at groups based on race, religion, gender, sexual orientation etc. They assess the performance of mBERT and XLM-RoBERTA on high and low resource languages, with limited sample sizes and class imbalance (Srikissoon and Marivate, 2023). The authors aim to improve the detection of hate speech targeting specific identity groups across multiple languages, finding that mBERT performs well on gender targeted hate speech (Srikissoon and Marivate, 2023).

Wong (2024) aims to use pretrained language models to detect anti-LGBTQ+ hate speech in English comments, and tests the model by comparing model performance across national-variety of English speaking countries. The cross lingual transformer based model Xlm-RoBERTa is used as the baseline model, with each model trained for 8 epochs using the AdamW optimizer. Wong (2024) uses the macro-average and micro-average F1 score to evaluate the models. The RandomOverSampler is used to upsample minority classes. Wong (2024) obtains a 83% average macro-F1 score and 94% average weighted F1-score for English anti-LGBTQ+ hate speech detection. Wong (2024) notes some limitations like sociocultural language differences across the national varieties of English, pitfalls of LLMs for specific tokens, and class imbalance.

Siddiqui et al. (2024) demonstrate the effectiveness of transformer-based models like BERT and XLM-RoBERTA for both binary and fine-grained classification tasks. Siddiqui et al. (2024) train their model to predict the five categories of Disability, Gender, Nationality, Race, and Religion, for the three languages of English, Urdu, and Sindhi, and achieve an 86% weighted F1 score. Additionally Siddiqui et al. (2024) incorporate Local Interpretable Model-Agnostic Explanations (LIME) for enhancing interpretability of the predicted labels. This work underscores the potential of multilingual transformer models and explainable AI for building more accurate and socially aware hate speech detection systems, motivating the need for our extended approach, which scales to many outputs.

In Chung et al. (2020), RemBERT, or Rebalanced mBERT, is an updated version of mBERT which improves performance in low-resource languages and increases generalizability by decoupling input and output embeddings. This reduces input embedding size without significant loss in performance, while simultaneously increasing output embedding size to great gains and preventing the model from overly specializing to the pre-training task (Chung et al., 2020). Furthermore, training data was rebalanced to de-emphasize high-resource languages and add 10 more languages. These updates significantly improved RemBERT's ability to be fine-tuned for a task in one language (e.g. English) and extrapolating that task to other languages.

3 Methods

3.1 Data Collection

Human-annotated datasets in English, French, and Arabic are used for our multi-output hate speech detection. The primary dataset, MLMA Ousidhoum et al. (2019), contains 18,661 Twitter hate comments across four labels: (a) directness (direct/indirect), (b) target attribute (e.g., gender, religion), (c) targeted group (17 options), and (d) sentiment (multi-label: anger, fear, etc.). Directness, target, and group are single-label outputs, while sentiment is multi-label (Ousidhoum et al., 2019). To incorporate non-hate examples (hatespeech = 0, other labels = "normal"), we augment MLMA with filtered samples from three additional English datasets: 1) Davidson et al. (2017) with 2,753 tweets not labeled as hate or offensive; 2) Sachdeva et al. (2022) with 3,678 comments from YouTube, Reddit, and Twitter with neutral sentiment and negative hate scores; 3) Mandl et al. (2019) with 3,591 non-hate tweets.

We used non-hate Arabic tweets collected between August 12,2020 and October 4, 2020 from the 'So Hateful dataset' Zaghouani et al. (2024). We took a random sample of 5K tweets, less than 1% of the original dataset. The tweets were pre-processed to remove duplicates and near-duplicates, as well as short tweets with less than 5 Arabic words and long tweets with more than 80 Arabic words. The tweets cover a variety of topics and themes, reflecting the

diverse interests and perspectives of Arabic Twitter users during the specified time frame Zaghouani et al. (2024).

The French non-hate data was collected to address how hate can vary greatly across cultural contexts and is designed to uncover cultural bias in hate speech datasets Tonneau et al. (2024). A survey was conducted of 75 hate speech datasets across 8 languages (including Arabic, English and French) then analyzed trends in language representation over time, confirming English-language bias Tonneau et al. (2024). Twitter was the main platform for data collection. In this case, we extracted 5K normal / non-hate french tweets to add to our hate detection dataset.

3.2 Data Cleaning

With the five multilingual non-hate datasets, the hate speech label set to 0, and directness, target, group, and sentiment labels set to "normal". Labels in the combined dataset are encoded to numerical values (hatespeech 0-1, directness 0-3, target 0-6, group 0-16, and each sentiment (x8) 0-1). These encodings are converted to 36 binary dummy encoded variables prior to modeling, and these are the model outputs. Text is cleaned to remove usernames, URLs, unicode emoji encodings, and other unwanted symbols to reduce noise and ambiguity during modeling. After dropping 179 duplicates, the final, cleaned dataset contains 38,504 observations.

3.3 Train Test Split

The cleaned dataset is shuffled before splitting due to datasets being concatenated by language. To obtain data splits with similar proportions for the imbalanced multiclass dataset, we utilize skmultilearn.model_selection.iterative_train_test_split as opposed to sklearn.model_selection.train_test_split, which is random. Our final balanced splits are 80% train with 30,702 observations, and 10% validation and 10% test, each with 3,901 observations.

3.4 Data Exploration

The training dataset is evenly distributed among the three languages for hate speech and non-hate speech labelled text. As seen in Figure 1, for each language there are similar proportions for the number of observations that are labeled hate speech vs non-hate speech. However, directness, target, group, and sentiment classes display high imbalance (Figure 2). Furthermore, as seen in Figure 3, certain hate speech targets are only present in English, and others have very few labels in French or Arabic. Certain labels, especially sentiment labels, had some correlation, including the sentiment labels of "anger" and "disgust" (correlation = 0.43), two emotions we would expect to be correlated together (Figure 4). As they still provide subtly different insight on the nature of the hate comment, we maintained their separation. The strongest correlation (0.9) was between "directness" and "normal sentiment," which is logical as 'direct' statements are less likely to contain undercurrents like disgust (Figure 4).

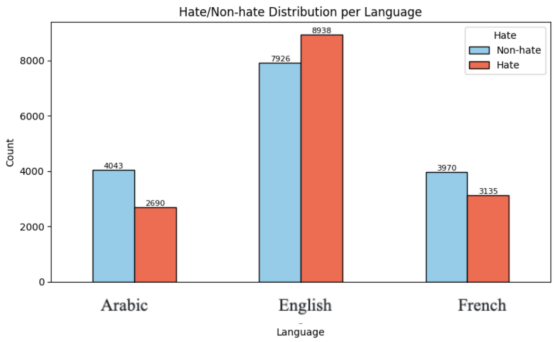

Figure 1: Training Data Distribution of Hate Speech Samples Per Language

3.5 Evaluation Metrics

To evaluate the model, metrics best suited for multiclass learning with imbalanced classes are observed. Each model is evaluated using F1-score, ROC AUC, precision, and recall, each on a micro-average level to capture the performance for each individual label. Model performance is examined at three levels: a) the overall model, b) each label, and c) each label and each language. This level of evaluation and analysis serve to provide detailed insights into the performance of multilingual transformer architectures across different languages and complex prediction tasks.

3.6 Modeling

3.6.1 Baseline Models

TF -IDF and Logistic regression is used as a non-neural baseline to evaluate how well traditional methods perform on a single class. The non-neural baseline predicts the "Target" class with 85% validation F1 score. Term Frequency-Inverse Document Frequency (TFIDF) converts raw text to numeric vectors, and logistic regression classifies by optimizing weight to obtain probabilities for each label. A second baseline trained with the RemBERT transformer architecture and predicts the single binary "Hatespeech" class with 95% validation F1-score. This second baseline model highlights the difference in complexity between this simple binary task as opposed to our multi-output classification task, as a third baseline for the multi-output classification task using mBERT results in 60% validation F1-score.

3.6.2 Transformer Architectures

To improve upon the baseline models, the BERT transformer architectures variations of mBERT (google-bert/bert-base-multilingual-cased), XLM-RoBERTa (FacebookAI/xlm-roberta-base) , and RemBERT (google/rembert) are used for the multi-class task for Hatespeech, Directness, Target, Group, and Sentiment labels. mBERT has been shown to be effective for multilingual hate speech detection in Srikissoon and Marivate (2023) and Siddiqui et al. (2024). In fine-tuning mBERT, we use the distilled version Distil-mBERT (distilbert/distilbert-base-multilingual-cased), which is a smaller, faster, and more efficient version of mBERT retaining 97% of BERT's performance with fewer layers and parameters. XLM-RoBERTa shows performance increases for multilingual hate speech detection in Wong (2024), and Chung et al. (2020) show that RemBERT increases performance for low-resource languages.

3.6.3 Fine-Tuning & Hyperparameters

Multiclass baselines are trained using the final classification layer from each of the three variations of the BERT transformer architectures. Models are trained to optimize the best micro F1-score for all outputs. Binary cross entropy with logits loss function (torch.nn.BCEWithLogitsLoss) is used for all outputs in the multi-class prediction task, and encourages the model to learn shared representations which benefit all classes. The models use the default AdamW optimizer (torch.optim.AdamW) provided with the Huggingface Trainer module, as this is an effective optimizer for large-scale models. Prior to fine-tuning, baseline models for each transformer architecture do not capture positive predictions for minority classes. To improve performance, fine-tuning techniques are implemented, like unfreezing layers and optimizing hyperparameters like learning rate, batch size, and epochs. In fine-tuning iterations, we experiment with batch sizes between 8 and 256, learning rates between 2e-5 and 6e-5, and models trained with 3-5 epochs (Table 1).

Model	Eval F1-Score	Eval Precision	Eval Recall
Baseline mBERT	0.604	0.744	0.495
Fine-Tuned mBERT	0.762	0.818	0.495
Hierarchical mBERT	0.741	0.837	0.664
Baseline RemBERT	0.508	0.652	0.416
Fine-Tuned RemBERT	0.752	0.802	0.709
Hierarchical RemBERT	0.765	0.835	0.705
Baseline XLM-RoBERTa	0.815	0.669	0.704
Hierarchical XLM-RoBERTa	**0.849**	0.849	0.849

Table 1: Validation Metrics for mBERT, RemBERT, and XLM-RoBERTa Fine-Tuning Iterations

3.6.4 Multi-Task Hierarchical Model

In addition to these hyperparameter optimizations, we also experiment with multi-task hierarchical modeling. Due to the nature of the data collection described earlier, the multi-task approach of predicting hate speech prior to predicting the other labels is a reasonable approach. The hierarchical model architecture enforces a dependency structure with 5 classification heads: hate speech is predicted first from the pooled transformer output, then directness is predicted using both the pooled output and the hate speech prediction, target is predicted using the pooled output and directness prediction, and group is predicted using the pooled output and target prediction. Sentiment is predicted independently from the pooled output. Each head is implemented as a linear layer, and the outputs of earlier heads are concatenated with the pooled output to inform subsequent predictions, reflecting the causal and nested relationships among the tasks. This design leverages strong signals from simpler tasks to improve performance on more complex, imbalanced tasks (Table 1).

Model	Overall	Hate Speech	Directness	Target	Group	Sentiment
Overall Metrics	**0.855**	**0.955**	**0.848**	**0.817**	**0.749**	**0.648**
English Metrics	**0.847**	0.994	0.940	0.873	0.761	0.671
Arabic Metrics	**0.913**	0.984	0.814	0.826	0.863	0.721
French Metrics	**0.820**	0.840	0.665	0.680	0.619	0.445

Table 2: XLM-RoBERTa Final Model Test F1 Scores

4 Results and Discussion

Fine-tuning experiments and multi-task hierarchical model architectures described above show 15-25% improvements in validation F-1 scores from the baseline multi-task models and increase the performance for minority classes (Table 1). Overall, we observed that REMBert displayed performance improvements over mBERT, and XLM-RoBERTa displayed performance improvements over both mBERT and REMBert (Table 1). The best-performing model was the fine-tuned, multi-task hierarchical XLM-RoBERTa architecture with 84.9% validation F1-score, and 85.5% test F1-score. This final model uses 12 transformer layers and 125M trainable parameters, with 2e-5 learning rate and batch size of 32, trained for 3 epochs. Table 2 shows the final model F1-scores on the test dataset, by language and class.

The binary hate speech class performs the best, with 95.51% test F1-score (Table 2). English test predictions of hate speech fall at 99.39% test F1-score, followed by Arabic test observations at 98.44% test F1-score and French test observations at 83.96% test F1-score (Table 2). English and French performance tends to decrease across the further dependent tasks, which is expected due to the high dispersion and sparsity across labels for observations in these labels. Arabic differs from this in that performance increases for further dependent tasks, which may be due to the observed lower dispersion among the labels (Table 2, Figure 6). Overall, Arabic comments performed best, contrary to expectations as Arabic had the fewest examples in our dataset. This may be due to the quality of the Arabic dataset, or less dispersion in the data across Target and Group (Figure 3).

In general, the model struggled to distinguish between indirect vs. direct bigotry (Figure 5, 6.) In Arabic and French, the Target class tended to most often mis-classify 'gender,' 'origin,' or 'normal,' as 'other'—this is expected, as 'other' by its catch-all nature is a vague category. English comments tended to have more accurate sentiment classifications than Arabic and French, highlighted by the model's tendency to predict "no sentiment" far more often in Arabic and French (Figure 6). While fine-tuned models achieve higher accuracy scores, performance still falters for minority classes (Figure 5, 6). Some class imbalance is present in our dataset due to the way in which we combined multiple hate speech datasets from different sources. To rectify this class imbalance, further methods can be explored such as a lower prediction threshold, varying model weights for minority classes, or implementing layer-wise learning rates and learning rate schedulers.

5 Explainable AI

To better understand the model behavior of our final model, we incorporate Local Interpretable Model-Agnostic Explanations (LIME) to explain the model predictions on the test dataset, specifically for sparse classes and minority labels, as seen in Siddiqui et al. (2024). For each language and each label, we examine a sample of true positives, true negatives, false positives, and false negatives.[1]

Figure 8 displays correctly classified English comments for the Target class, where the model identifies key words for their corresponding classes. The text with the phrase "retarded chinese" is correctly labeled as Target: Origin, the phrase "dirty jew" is correctly labeled as Target: Religion, and "dyke" is correctly labeled as Target: Sexual Orientation (Figure 8).[1] Figure 9 displays correctly classified English comments for the Group class, where "retard" is correctly classified as Group: Special Needs, and "cunt" is correctly classified as Group: Women (Figure 9). Figure 10 shows examples for minority classes incorrectly classified despite key words being identified for the labels, and Figure 11 shows examples where

[1]Hate speech examples from the labeled dataset.

the true labels are ambiguous and/or incorrect.

Figures 12 and 13 show a sample of predictions for Arabic comments. The text "هههههههه شريرة .. هاد احلى انتقام" is a false positive, identifying it as hate speech when it isn't.[1] The key word here is "شريرة" which means evil, so the model classifies it as hatespeech; however it isn't taking the context into account. This sentence translates to "hahaha evil, this is the best revenge." In contrast a false negative example where a hate speech comment misclassified as Target: Origin, "كلب وسخ ببلدنا ويهيننا لكن فيه رجال يفركون خشمه كلاب عرب الشمال الله يعز المملكة" translates to "A dirty dog in our country insults us, but there are men who rub his nose, Northern Arab dogs. May God honor the Kingdom" which seems to be an insult towards Arabs.[1]

Figure 14 and 15 show predictions for French comments. Hate sentiments are correctly captured, such as the comment "ces des enfants bordel ta 6 de qi c incroyable comment t'es attardé" which translates to "They're children, damn it, your IQ is 6, it's incredible how retarded you are".[1] The word "retarded" causes the model to classify the comment as hate speech. However, it fails with more nuanced issues that potentially need current political context such as this example "le vrai danger de l'islam c'est la gauche je peux te le garantir" that translates to "The real danger of Islam is the left, I can guarantee it."[1]

6 Conclusion

As online spaces become increasingly central to everyday life as means of communication and socialization, hate speech detection is critical to maintain the safety and civility of online interactions. However, the task of hatespeech identification remains extremely complex, with some languages being relatively under-moderated, as well as definitions of hate speech varying based on cultural context. We propose a multi-lingual model which is able to capture that complexity, categorizing comments from three languages (English, Arabic, and French) in five dimensions (hate presence, directness, target attribute, targeted group, and sentiment).

Our final fine-tuned XLM-RoBERTa model achieves this with an F1 score of 0.85 on the test dataset, a significant improvement (25-35%) over other models' baselines. It also maintains relative fairness with similar overall performance across all three languages (0.82-0.91), although performance within the individual five categorization dimensions does vary by language. We explicate the test predictions of our model using the explainable AI method LIME (Local Interpretable Model-Agnostic Explanations), which highlighted which words or phrases in each comment were more influential in classifying the comment as hate-speech or not. In the future, we hope to explore improving the performance in the five dimensions, perhaps by rectifying class imbalances or further augmenting our training dataset. We also hope this may improve the fairness of our model by ensuring that the performance within the five dimensions individually becomes roughly equal across the three languages. The code for this paper can be found here.

References

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the 11th International AAAI Conference on Web and Social Media (ICWSM)*. Association for the Advancement of Artificial Intelligence.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. *Proceedings of the 11th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Barbara Ortutay. 2022. Amnesty report finds facebook amplified hate ahead of rohingya massacre in myanmar. PBS News.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqui Song, and Dit-Yan Yeung. 2019. Multilingual and mutli-aspect hate speech analysis. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4675–4684. Association for Computational Linguistics.

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. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP*. European Language Resources Association.

Jawaid Ahmed Siddiqui, Siti Sophiayati Yuhaniz, Ghulam Mujtaba Shaikh, Safdar Ali Seomro, and Zafar Ali Mahar. 2024. Fine-grained multilingual hate speech detection using explainable ai and transformers. *IEEE*, 12.

Trishanta Srikissoon and Vukosi Marivate. 2023. Combating hate: How multilingual transformers can help detect topical hate speech. *EPiC Series in Computing, Proceedings of Society 5.0 Conference 2023*, 93:203–215.

Manuel Tonneau, Diyu Liu, Samuel Fraiberger, Ralph Schroeder, and Scott A. Hale. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH)*, pages 283–311. Association for Computational Linguistics.

Sidney G.-J. Wong. 2024. Sociocultural considerations in monitoring anti-lgbtq+ content on social media.

Wajdi Zaghouani, Hamdy Mubarak, and Md. Rafiul Biswas. 2024. So hateful! building a multi-label hate speech annotated arabic dataset. *Proceedings of the 2024 Joint International Conference on Computational Liguistics, Language Resources and Evaluation (LREC-COLING 2024)*. European Language Resource Association (ELRA) and the International Committee on Computational Linguistics (ICCL).

Appendix

Figure 1: Training Data Distribution of Hate vs. Non-Hate by Language

Language	Non-Hate	Hate	Total
Arabic	4,043 (60%)	2,690 (40%)	6,733 (22%)
English	7,926 (47%)	8,938 (53%)	16,864 (55%)
French	3,970 (55%)	3,135 (45%)	7,105 (23%)
Total	15,939 (52%)	14,763 (48%)	30,702 (100%)

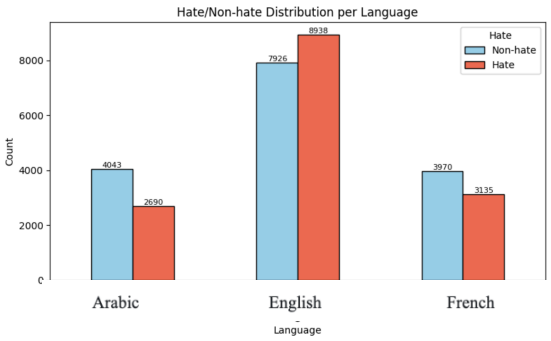

Figure 2: Training Data Distribution of Imbalanced Classes

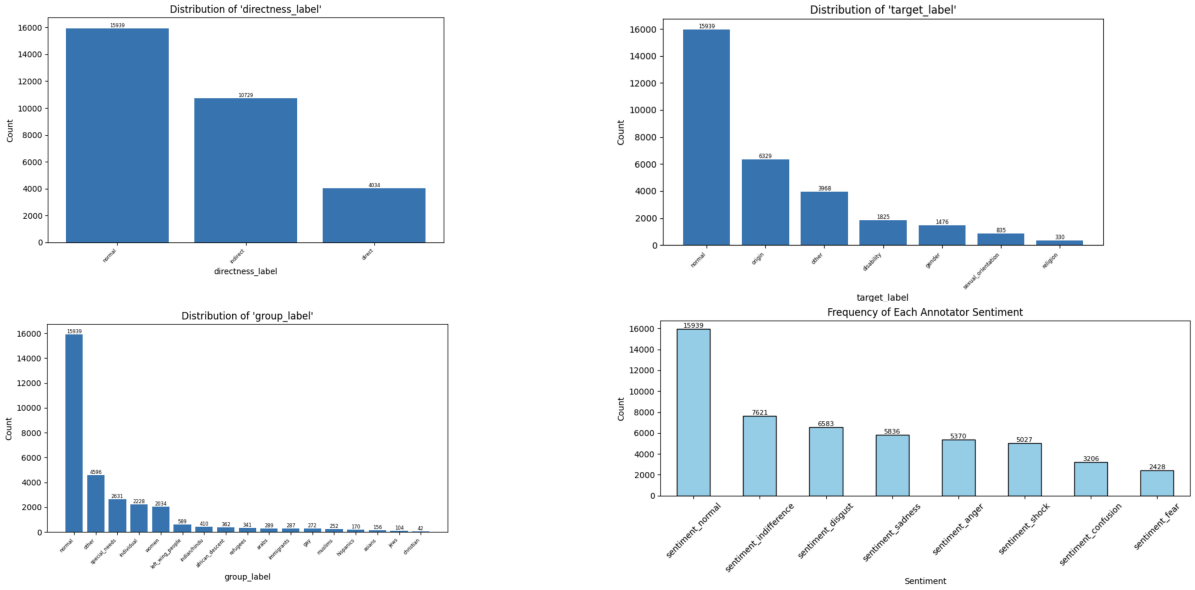

Figure 3: Training Data Distribution of Target and Group by Language

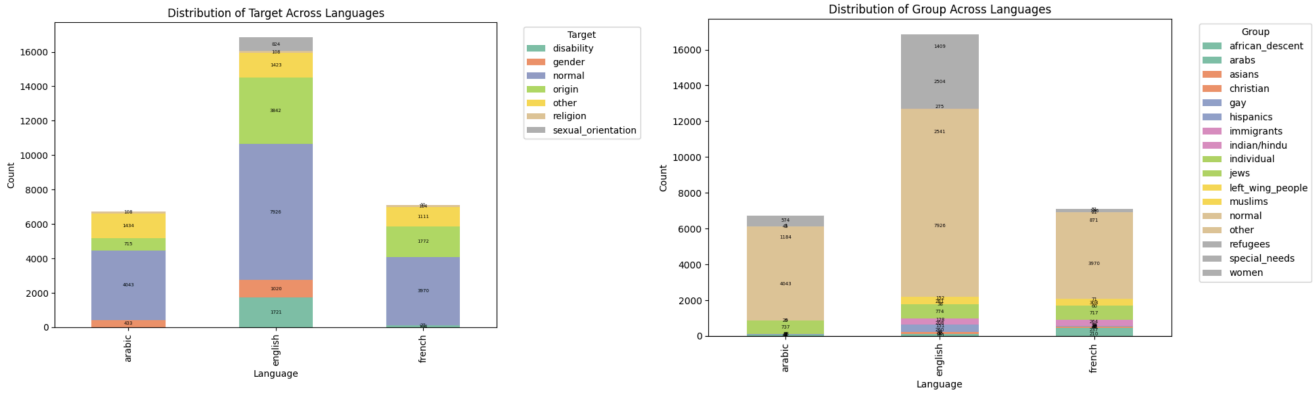

Figure 4: Correlation Matrix of Training Data

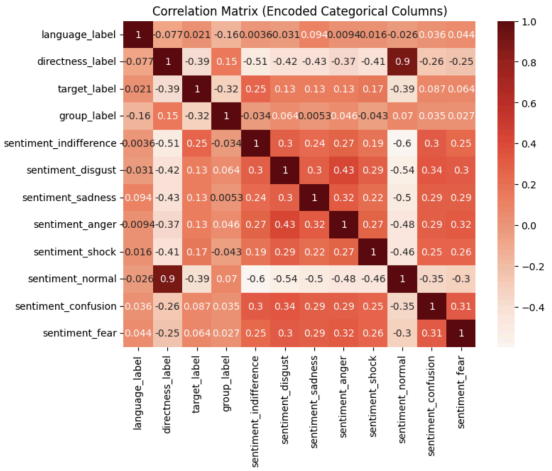

Table 1: Validation Metrics for Training

Model	Eval F1	Eval Precision	Eval Recall
Baseline mBERT	**0.604**	0.774	0.495
Fine-Tuned mBERT	**0.762**	0.818	0.713
Hierarchical mBERT	**0.741**	0.837	0.664
Baseline RemBERT	**0.508**	0.652	0.416
Fine-Tuned RemBERT	**0.752**	0.802	0.709
Hierarchical RemBERT	**0.765**	0.835	0.705
Baseline XLM-RoBERTa	**0.815**	0.669	0.704
Hierarchical XLM-RoBERTa	**0.849**	0.849	0.849

Table 2: Hierarchical XLM-RoBERTa Test Dataset F1-Scores

Model	Overall	Hate Speech	Directness	Target	Group	Sentiment
Overall Model Metrics	**0.855**	**0.955**	**0.848**	**0.817**	**0.749**	**0.648**
English Metrics	**0.847**	0.994	0.940	0.873	0.761	0.671
Arabic Metrics	**0.913**	0.984	0.814	0.826	0.863	0.721
French Metrics	**0.820**	0.840	0.665	0.680	0.619	0.445

Figure 5: Test Prediction Confusion Matrices

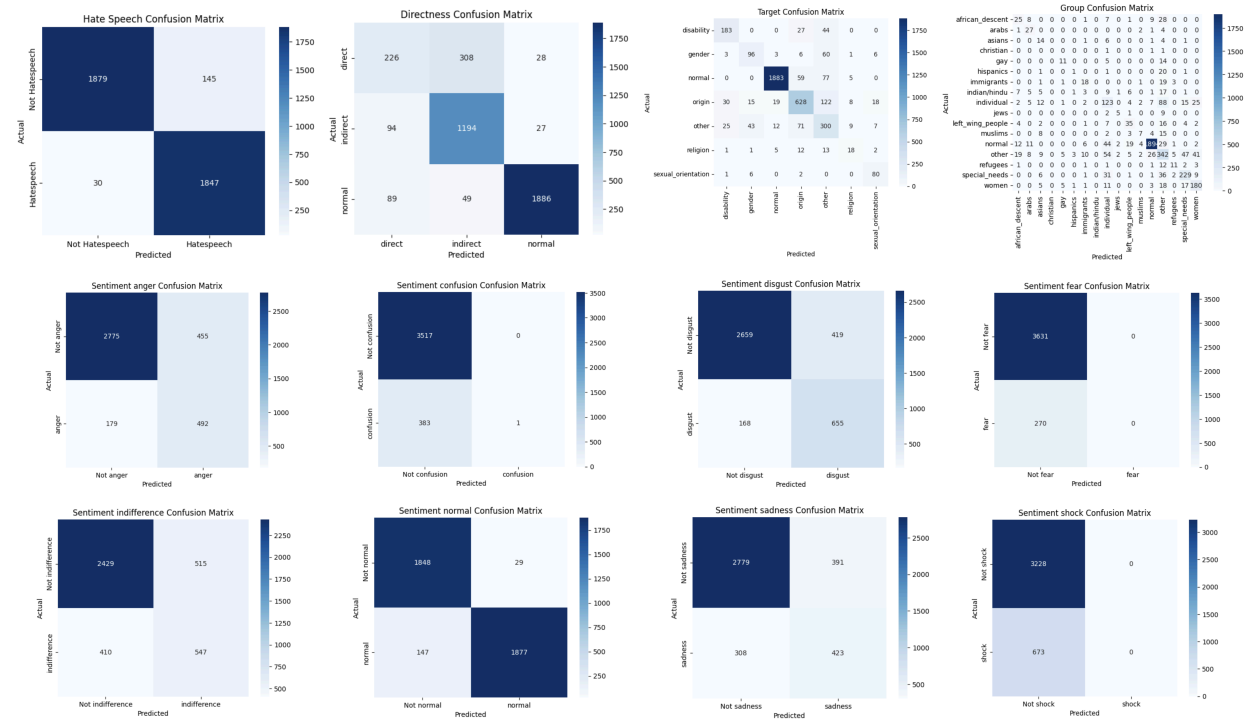

Figure 6: Test Prediction Confusion Matrices Per Language

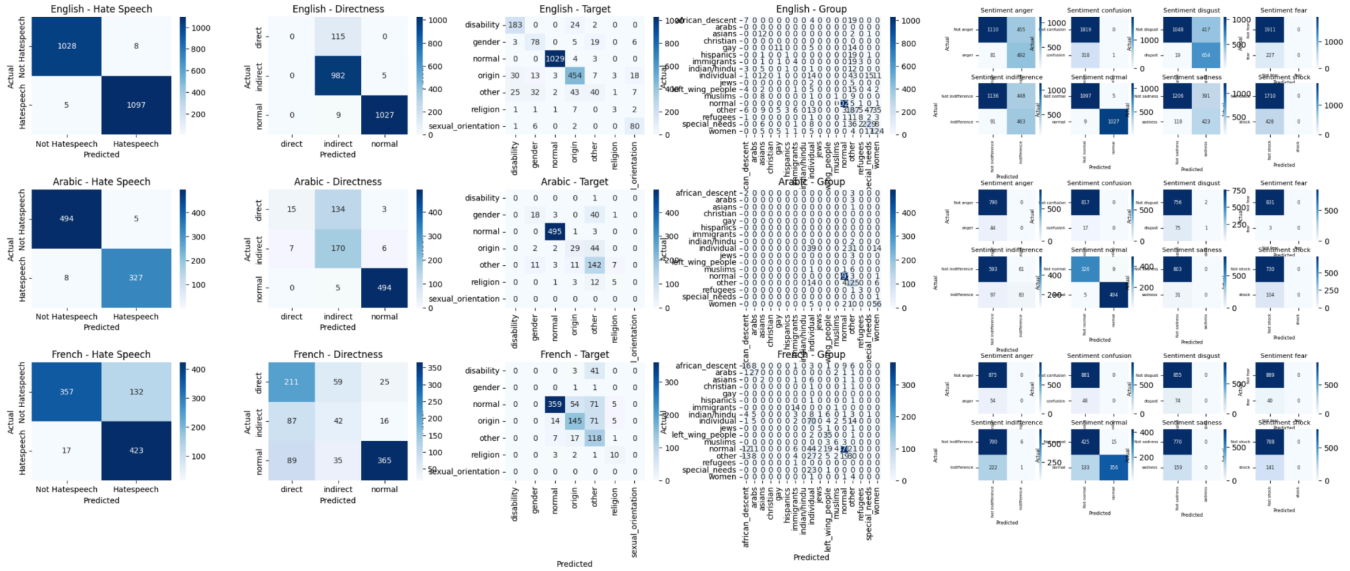

Figure 7: English Hate Speech Predictions

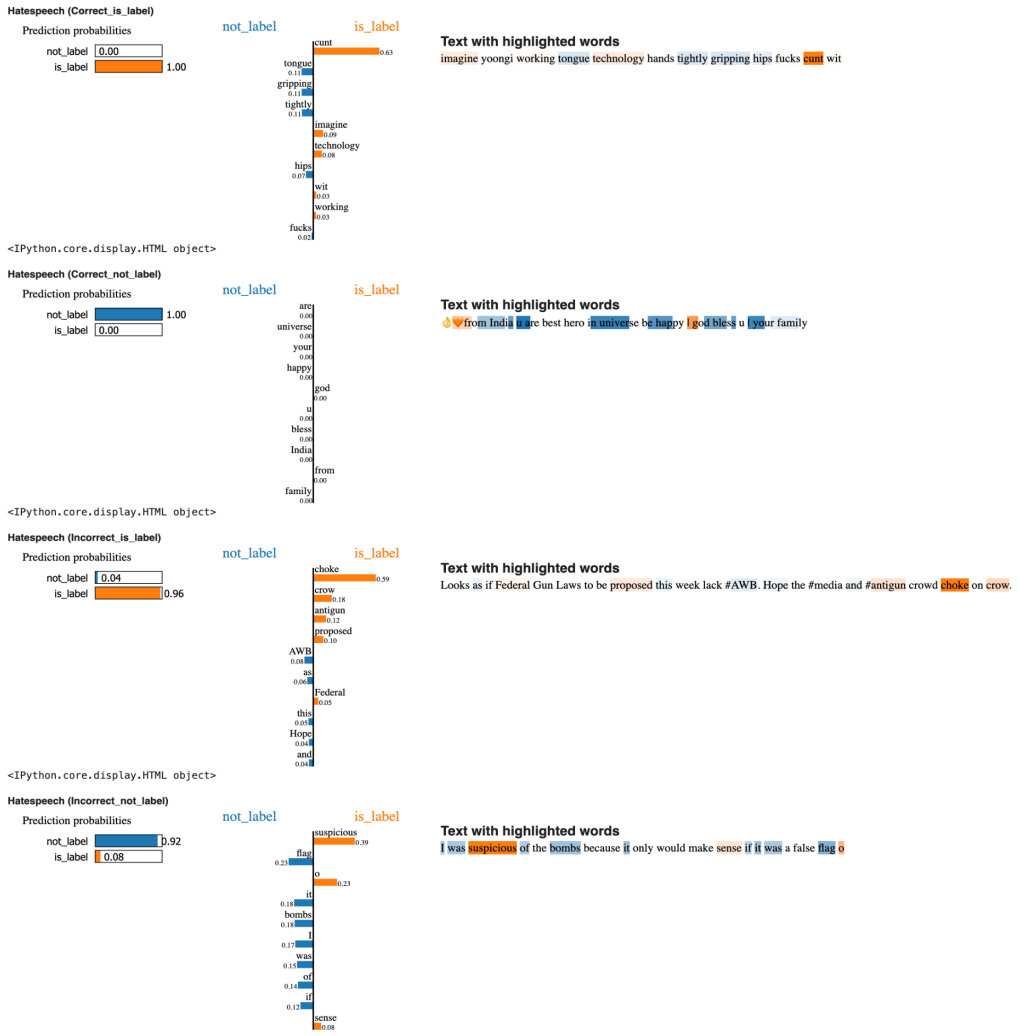

Figure 8: Correctly Classified English “Target” Predictions with LIME

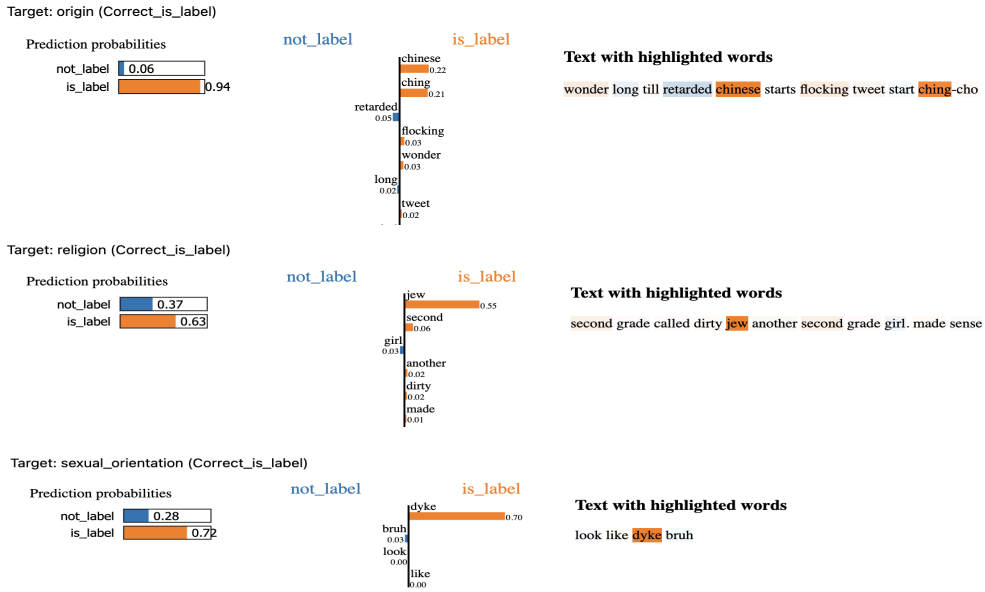

Figure 9: Correctly Classified English “Group” Predictions with LIME

Figure 10: Incorrectly Classified English Predictions with LIME

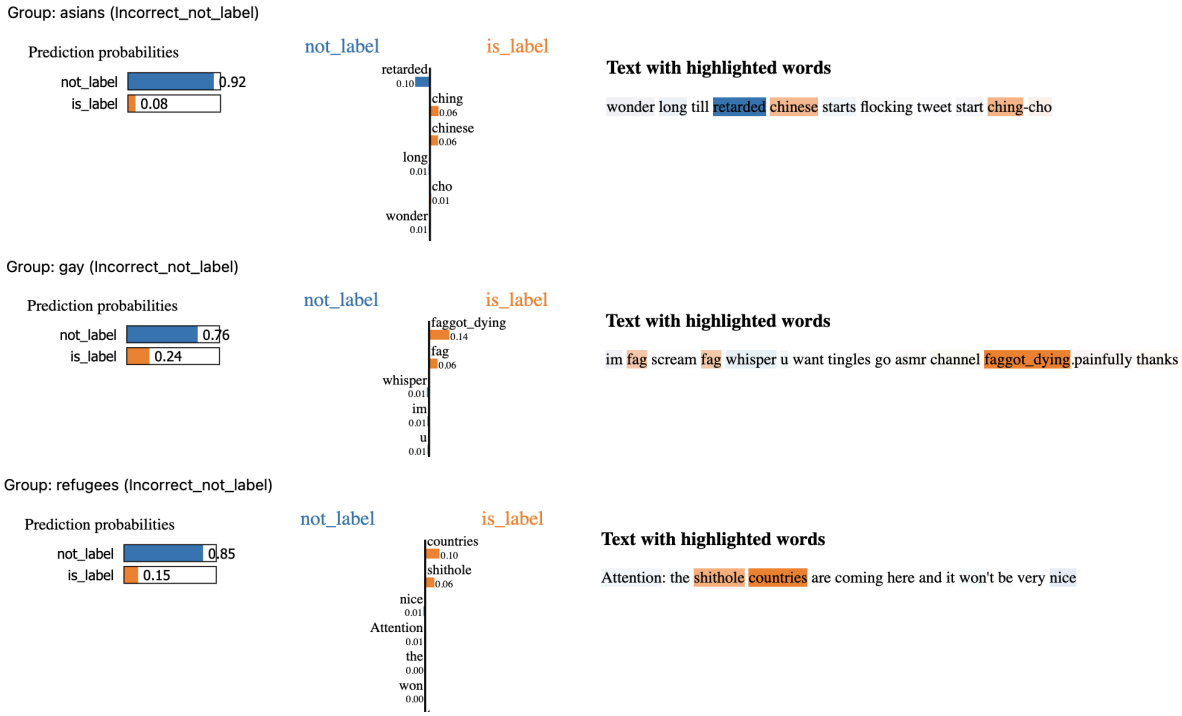

Figure 11: Incorrect/Ambiguous True Labels for English Predictions with LIME

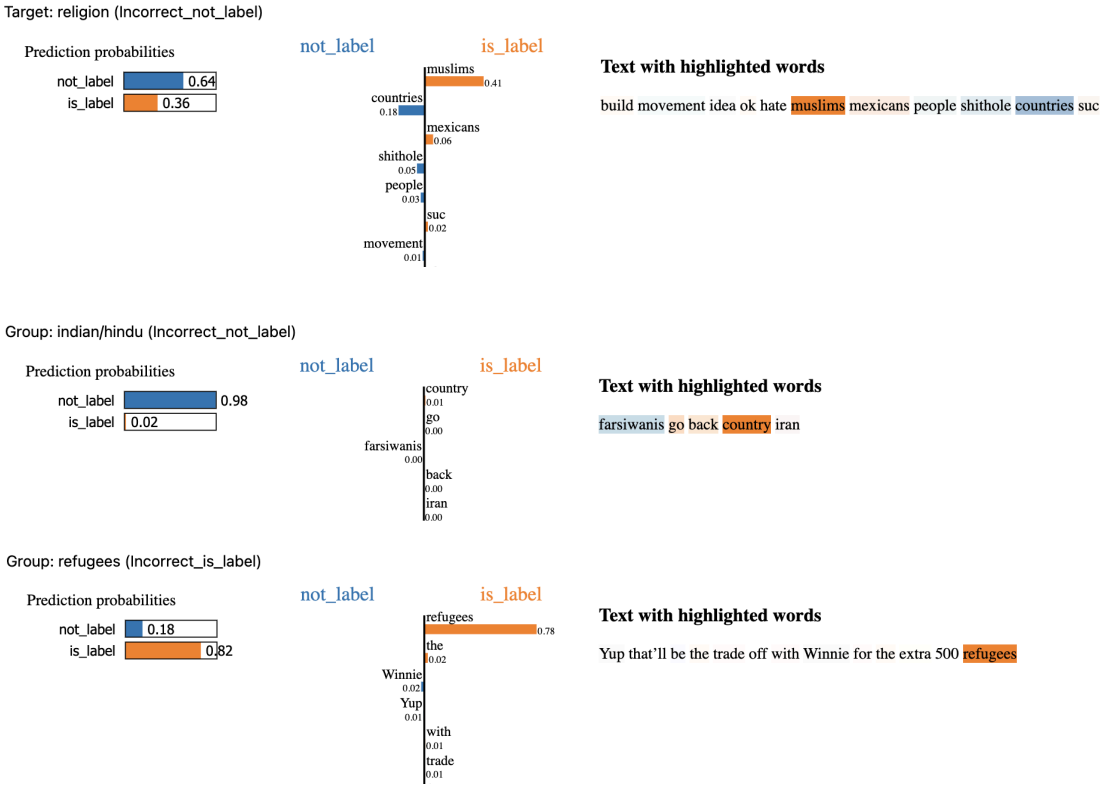

Figure 12: Arabic Hate Speech Predictions

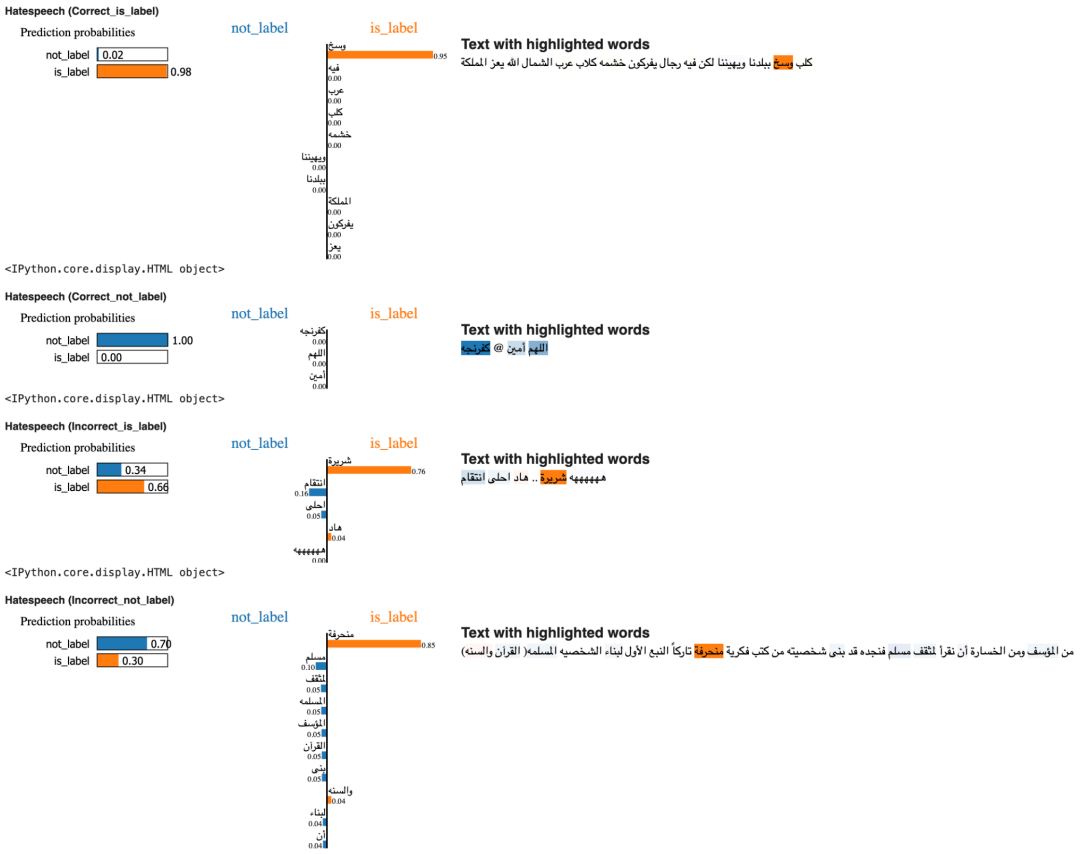

Figure 13: Arabic Target Hate Speech Predictions

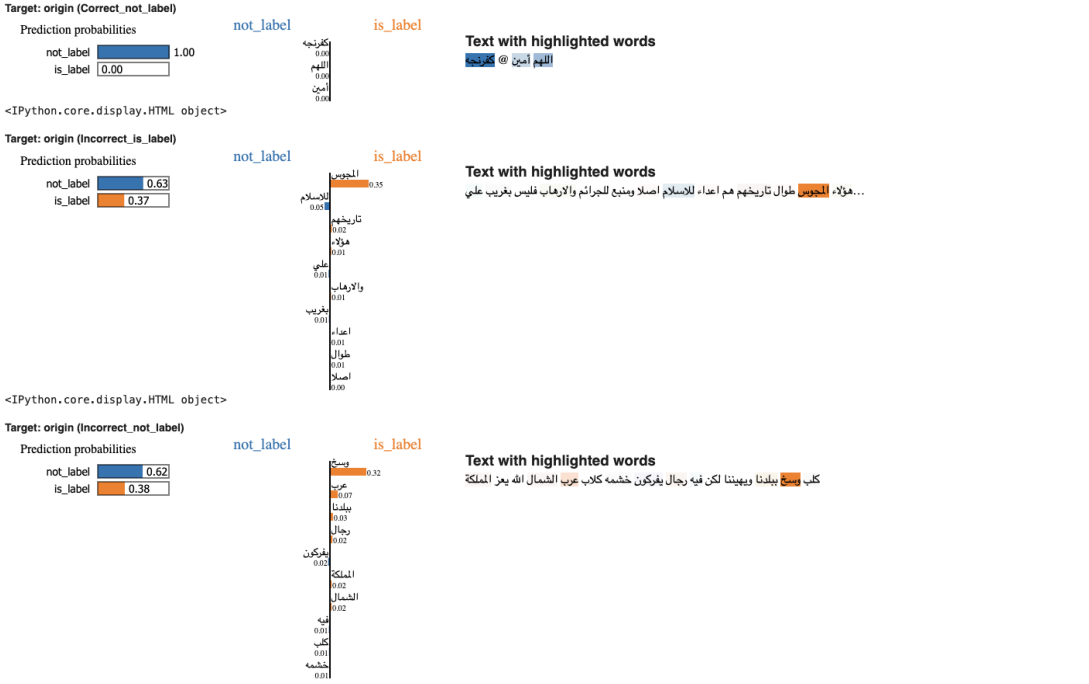

Figure 14: French LIME Explanations

Staying in Oran!! Yout are a dirty Arab.

```
Text: eh bin tu restes à oran!!! sale arabe que tu es...
LIME Explanation (top features):
arabe: −0.628
sale: 0.006
es: 0.003
à: 0.003
oran: 0.002
tu: 0.002
que: −0.001
bin: −0.001
eh: −0.001
restes: −0.001
```

Thank you for the insult

```
Class 'Hate' – incorrect predictions (showing up to 2):

Text: merci pour l...insulte.n...
LIME Explanation (top features):
l: 0.284
merci: 0.257
insulte: −0.225
n: −0.121
pour: −0.000
```

he jerks off to sneaky's cosplay, this retard

```
Class 'origin' – incorrect predictions (showing up to 2):

Text: bouahahhaa il se branle sur les cosplay de sneaky cet attardé...
LIME Explanation (top features):
attardé: 0.214
les: 0.013
bouahahhaa: −0.005
il: −0.003
branle: 0.002
sur: −0.002
de: 0.002
cet: −0.001
sneaky: −0.001
cosplay: 0.000
```

Task: Group

Return to your dirty Arab country (oops)

```
Class 'arabs' – correct predictions (showing up to 2):

Text: retourne dans ton pays sale arabe (oups)...
LIME Explanation (top features):
arabe: 0.576
pays: 0.008
dans: −0.007
retourne: 0.005
sale: −0.003
ton: 0.001
oups: 0.001
```

Figure 15: French Hate Speech Predictions

Contributions

Bella

- Background: Siddiqui et al. (2024), Wong et al. (2024)
- Data Collection: Ousidhoum et al. (2019) multilingual hate speech dataset (MLMA), English datasets Davidson et al. (2017), Sachdeva et al. (2022), Mandl et al. (2019)
- Data Cleaning: Encode labels, clean text, combine datasets, drop duplicates, dataset splits
- EDA on final combined training dataset: Distributions of classes, distribution of text lengths, correlation and co-occurrence matrices, language differences per class
- Modeling: Baseline mBERT, Fine Tuned mBERT, Hierarchical mBERT with LIME
- Test Predictions and Metrics Per Language, English/French/Arabic LIME Visualizations

Nadaa

- Introduction
- Background: Srikissoon and Marivate. (2023), Tonneau et al. (2024)
- Data Collection: Arabic dataset-Zaghouani et al. (2024), French dataset-Tonneau (2024)
- Data Cleaning: Combine datasets, shuffle combined dataset
- Modeling: Baseline classification model, XLM-RoBERTa baseline and fine-tuning Single prediction, XLM-RoBERTA multi output prediction, hierarchical XLM-RoBERTa
- LIME Analysis: French/Arabic_LIME_Explanations.ipynb

Ishani

- Background: Jahan and Oussalah. (2023), Chung et al. (2020)
- EDA on MLMA dataset: Distributions of classes
- Modeling: RemBERT baseline, RemBERT fine-tuning, RemBERT hierarchical model
- Results (~half)
- Conclusion [entire]
- Methods (part of modeling)
- Introduction (motivation)