Groupe 9 - Hate speech detection

Mohamed Hammani[1], **Mohamed Amine Mbarki**[1], **Nabil El-Yazidi**[1]
[1]Department of Computer Science, University of Sherbrooke
{hamm1306, mbam5045, elyn6015}@usherbrooke.ca

1 Abstract

Hate speech detection is essential for creating safe online environments by identifying harmful content and classifying its types, such as race, gender, or religion. Our project aims to enhance hate speech detection performance by proposing innovative approaches. After reviewing related work, we identified limitations in contextual analysis, inspiring our contributions. We developed a novel Retrieval and Classification (RAC) method, inspired by RAG, to improve contextual understanding. Additionally, we explored embeddings from advanced language models, feeding them into classifiers like XGBoost, LSTM, and a hard-voting ensemble. We also pioneered a LangChain template integration with Falcon-7B for enhanced classification. To further enhance transparency, we integrated SHAP and LIME, enabling detailed analysis of feature importance and individual prediction explanations. These original approaches outperform baseline models, as shown by qualitative analysis highlighting challenges with ambiguous texts. Our work advances hate speech detection through innovative methods and proposes future exploration of advanced transformers and a Django interface for practical deployment.

2 Introduction

Hate speech detection is a crucial yet challenging task that plays a vital role in maintaining respectful and inclusive online spaces. With the increasing volume of user-generated content, platforms face difficulties in identifying hateful messages that often appear in subtle, implicit, or context-dependent forms. This makes the task particularly complex, as hate speech can vary widely in tone, target (e.g., race, gender, religion), and linguistic style. Moreover, manual moderation is not scalable, highlighting the need for automated, intelligent systems.

Motivated by these challenges, our project aims to enhance hate speech detection performance by proposing innovative approaches. To this end, we conducted an in-depth analysis of prior research, highlighting the shortcomings of weak supervision and manual annotation methods. Weak supervision, while scalable, often generates noisy labels due to context-free heuristics, leading to poor generalization across datasets (Jin et al., 2023). Manual annotation, though nuanced, is resource-intensive and prone to subjective biases from annotators' backgrounds, risking misclassification or disproportionate targeting of marginalized groups (Hettiachchi et al., 2023).We explored advanced preprocessing and data augmentation techniques, leveraging embeddings from state-of-the-art language models to capture contextual nuances. These embeddings were integrated into classifiers such as XGBoost, LSTM, and a hard-voting ensemble. Our original contributions include a novel Retrieval and Classification (RAC) method, inspired by Retrieval-Augmented Generation, to enhance contextual analysis, and the use of Falcon-7B for embedding generation and a LangChain template integration for classification. Figure 3 illustrates our system's input and output (binary or type classification).In summary, our goal is to deliver a robust solution for hate speech detection, harnessing advancements in language modeling and combining innovative techniques to achieve precise and reliable results.

3 Related Work

In the study by (Malik et al., 2023), hate speech detection was approached through a comparative evaluation of 14 shallow and deep learning models using TF-IDF, GloVe, and transformer-based embeddings. Three benchmark English datasets—Davidson, Founta, and TSA—were used for training and evaluation. Preprocessing included vectorization via TF-IDF or pre-trained embeddings such as GloVe and BERT. Models like SVM, XGBoost, CNN, Bi-LSTM, and MLP were evaluated based on macro and weighted F1 scores. Transformer-based models, particularly BERT and ELECTRA, outperformed others, reaching macro F1 scores up to 0.90 and weighted F1 scores up to 0.98 on the TSA dataset. However, limitations remain, notably performance drops in cross-domain generalization and the higher computational cost of transformer models.

In the research by (Al-Hassan and Al-Dossari, 2019), hate speech detection in social networks was studied from multiple perspectives including supervised, semi-supervised, and deep learning approaches. Preprocessing techniques such as tokenization, POS tagging, n-grams, and word embeddings (Word2Vec, AraVec) were employed to prepare English and multilingual datasets for classification. Models like SVM, CNN, and GRU-based RNNs were applied on datasets such as Twitter, Facebook, and YouTube. In particular, deep models such as CNN and RNN demonstrated promising results, with F1-scores reaching up to 0.93 (Badjatiya et al., 2017). However, the authors highlight ongoing challenges, including the ambiguity of hate definitions across cultures, the lack of annotated data in low-resource languages, and limitations in classifier generalization across domains.

In the survey by (Nascimento et al., 2023), the authors conducted a comprehensive review of hate speech detection methods on social media, analyzing 83 peer-reviewed papers published since 2015. The study outlines a typical detection pipeline involving data collection via APIs, preprocessing (e.g., lowercasing, stemming, emoji normalization), feature extraction, and classification. Several types of features were discussed, including n-grams, TF-IDF, word embeddings (e.g., Word2Vec, GloVe, BERT), and sentiment scores. Classifiers such as SVM, Random Forest, CNN, and LSTM were evaluated, with some deep learning ensembles achieving up to 98% accuracy on Twitter datasets. Despite these results, challenges remain, such as class imbalance, annotation subjectivity, and the limitations of out-of-vocabulary words in pretrained embeddings. The paper also emphasizes the need for contextual understanding and metadata to improve detection robustness.

To overcome the limitations of prior studies, we adopted a comparative approach to evaluate multiple language and classification models, selecting those with optimal performance for hate speech detection. We assessed embedding models, including transformer-based and generative architectures for their contextual richness. Similarly, we compared classifiers and used a hard-voting ensemble to maximize accuracy. Best practices, including task-specific pre-trained models, data augmentation (Kant, 2023), and contextual data integration, were employed to enhance robustness. Our novel contributions include a Retrieval and Classification (RAC) method, inspired by Retrieval-Augmented Generation, and a LangChain template with Falcon-7B, significantly improving detection performance.

4 Methodology

4.1 Data

The dataset used in this study is the Ethos corpus, a publicly available benchmark specifically curated for hate speech detection tasks (Ioannis Mollas, 2022). It was chosen due to its widespread use in the literature, its high annotation quality, and its relevance for both binary and multi-label classification tasks, making it a standard reference in hate speech research.

For Task A (binary classification), the dataset contains 998 instances, with hate speech accounting for approximately 35.9% of the considered examples and 7.5% excluded from evaluation, highlighting a real-world class imbalance skewed toward the non-hate class, which represents the remaining 56.6%.

For Task B (multi-label classification), each instance can belong to one or more of the following hate categories: *violence*, *directed vs. generalized*, *religion*, *race*, *gender*, *sexual orientation*, *national origin*, and *disability*. The class distribution is imbalanced, as shown in figure below :

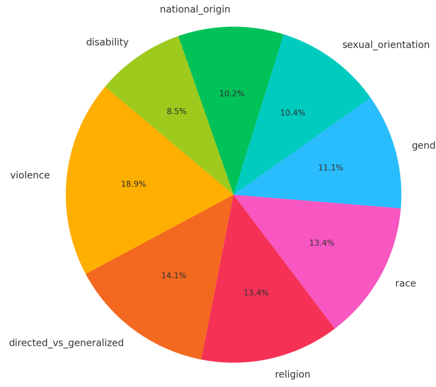


Figure 1: Distribution of hate speech categories in the Ethos dataset (Task B)

This imbalance reflects real-world patterns in online hate speech, which tend to target certain groups more frequently. It also poses a valuable challenge for evaluating the robustness and fairness of hate speech detection models.

4.2 Data preprocessing

We preprocessed the data by removing special characters, emojis, hashtags, URLs, and correcting grammatical errors and out-of-vocabulary words, ensuring suitability for model training. To address limited labeled samples and class imbalance, we employed transformer-based models, including GPT-2, T5, and Pegasus Paraphraser (Kant, 2023), to generate additional training samples. Pegasus, in particular, produced high-quality samples that closely mirrored the original data's linguistic patterns, effectively enriching the dataset. Contextual data was incorporated to enhance model understanding. These steps prepared the data effectively for our hate speech detection model, minimizing biases from text characteristics.

4.3 Use of pre-trained embeddings

To generate contextual embeddings for our hate speech detection tasks, we conducted a comparative study of embedding models, including Word2Vec, BERT, RoBERTa, DeBERTa, GPT-3, and Falcon-7B, following a comprehensive literature review. Transformer-based and generative models (BERT, RoBERTa, DeBERTa, GPT-3, Falcon-7B) were selected for their superior ability to capture semantic and contextual nuances, critical for distinguishing subtle hate speech. Similarly, we evaluated classifiers such as SVM, CNN, LSTM, XGBoost, and KNN and a hard-voting ensemble to leverage their complementary strengths. The embeddings, representing comments as high-dimensional vectors, were fed into these classifiers to predict binary (hate/non-hate) and multi-class (e.g., race, gender) labels. This approach enhances classification performance by combining advanced embeddings with robust ensemble learning, addressing limitations of traditional feature engineering.

4.4 Evaluation Method

To assess the performance of our hate speech detection models, we employed multiple evaluation metrics, including **Accuracy**, **Precision**, **Recall**, and **F1-score**, the latter being particularly effective for imbalanced classes. We generated detailed **classification reports** and **confusion matrices** to analyze per-class performance and error distribution. For deep learning models, we tracked **training and validation accuracy and loss** across epochs to monitor learning progress and detect overfitting. Additionally, **macro** and **weighted averages** of the metrics were computed to provide a global evaluation, accounting for class imbalance. To enhance interpretability, we applied **SHAP** (SHapley Additive exPlanations)(Lundberg and Lee, 2017) to identify influential text features driving model predictions and **LIME** (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016) to explain individual predictions by highlighting impactful text segments . These tools provided insights into model behavior and feature contributions.

5 Results

5.1 Task A (Binary classification)

A comparison with the baseline results reported in the ETHOS dataset paper (Mollas et al., 2021) reveals a substantial performance gain achieved by our model. In their binary classification experiments, the best-performing model was **DistilBERT**, which achieved an accuracy of **80.36%** and a macro F1-score of **79.92%**. In contrast, our approach, which combines **GPT-3 embeddings** with an **LSTM classifier**, reaches an accuracy of **97%**, significantly outperforming all baselines. Additionally, our second model—based on our custom **Retrieval and Classification (RaC)** approach—also achieves **97% accuracy**. These results suggest that leveraging large-scale pretrained language models like GPT-3, along with sequence-aware classifiers, provides a notable advantage in capturing nuanced hate speech patterns within the ETHOS dataset.

Our hard-voting ensemble, which aggregates majority class predictions from classifiers including XGBoost, LSTM, SVM, CNN, and KNN, achieved **92% accuracy** with BERT embeddings, **88% accuracy** with RoBERTa embeddings, and **97% accuracy** with GPT embeddings. These results outperformed all individual models trained with the corresponding embeddings, demonstrating that majority voting effectively harnesses diverse classifier strengths to enhance classification performance.

Furthermore, we explored **Falcon-7B** for embedding generation, using a subset of the dataset due to limited GPU resources. This approach yielded a **75% accuracy** on 60 test samples, surpassing our CNN and KNN models with DeBERTa embeddings. We also integrated Falcon-7B with a **LangChain template** (Keivalya Pandya, 2023), assigning a specific role to the agent. The agent provided detailed explanations for its classification decisions, analyzing text semantics and articulating why one class was chosen over another, demonstrating human-like reasoning in understanding hate speech nuances.

5.2 Task B (Multi class classification)

For multiclass classification on the ETHOS dataset, targeting categories such as race, gender, and religion, we evaluated several models using F1-score and accuracy, as summarized in Table 1. The SVM model achieved the highest accuracy of **86.17%** with an F1-score of **0.87**, narrowly outperforming the Neural Network, which recorded an F1-score of **0.8713** and an accuracy of **86.46%**. The CNN and Hard Voting ensemble followed with accuracies of **80.40%** (F1-score: 0.81) and **82.73%** (F1-score: 0.81), respectively, while KNN lagged at **77.23%** accuracy (F1-score: 0.77). These results indicate that SVM and Neural Network models effectively capture the nuanced patterns across multiple hate speech categories, with SVM slightly edging out in overall accuracy, likely due to its robustness in high-dimensional spaces with embeddings.

To further analyze the SVM's performance, we plotted the F1-scores per class, as shown in Figure 4. The figure reveals that most classes achieve F1-scores above 0.8, demonstrating strong and consistent performance across categories such as race, gender, and religion. Classes 1, 2, 3, 5, and 6 scored around 0.9, reflecting excellent precision and recall balance. However, Class 4 exhibited a slightly lower F1-score of approximately 0.77, indicating potential challenges in detecting this category, possibly due to underrepresented samples or overlapping features with other classes. Classes 0 and 7, with F1-scores of around 0.85 and 0.82, respectively, also performed well but showed minor variability. Overall, the SVM model's balanced performance across classes underscores its suitability for multiclass hate speech detection .

6 Conclusion and Perspectives

Our project successfully developed a robust framework for hate speech detection on the ETHOS dataset, achieving strong performance in both binary and multiclass classification tasks. For Task A, our approach leveraging **GPT-3 embeddings** with an LSTM classifier and a hard-voting ensemble attained an impressive **97% accuracy**, significantly outperforming the baseline DistilBERT model at 80.36%. The innovative Retrieval and Classification (RaC) method also matched this performance at **97% accuracy**, demonstrating the efficacy of combining retrieval-augmented strategies with advanced embeddings. In Task B, targeting categories like race, gender, and religion, our SVM model achieved the highest accuracy of **86.17%** with an F1-score of 0.87, showing consistent performance across most classes, as evidenced by F1-scores largely above 0.8. The integration of **Falcon-7B** with a LangChain template provided valuable semantic explanations, while SHAP and LIME enhanced model interpretability, offering insights into feature importance and prediction rationale. These advancements pave the way for implementing real-time hate speech monitoring systems, effectively addressing both the presence and type of hate.

However, we faced limitations. Computational constraints, particularly limited GPU capacity, restricted model complexity, as seen with **Falcon-7B** embedding generation (**75% accuracy** on 60 samples). Large embedding tensors saturated memory, and class imbalance in the ETHOS dataset hindered generalization, notably for Task B Class 4 (F1-score: 0.75), while smaller models were prioritized due to size constraints.

Looking ahead, several perspectives can enhance our framework. Optimizing embeddings through dimensionality reduction techniques like PCA could mitigate memory constraints while preserving semantic information, enabling the use of larger datasets. Fine-tuning parameters, such as the number of retrieved documents in RaC, could further improve performance, especially for nuanced hate speech detection (e.g., distinguishing overlapping categories like religion and violence). Expanding the classification scope to detect a broader range of hate speech types, potentially incorporating advanced LLMs like LLaMA, could address class imbalance and improve generalization. Finally, enhancing the LangChain template to provide more detailed, context-aware explanations could further improve interpretability, making the system more transparent for real-world deployment.

References

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus.

Danushka Hettiachchi, Vethum Wickramasinghe, Thilina Anuradha, Indigo Hughes, Lachlan Kennedy, and Jorge Goncalves. 2023. Uncovering political bias in emotion inference models: Implications for social media analysis.

Stamatis Karlos Grigorios Tsoumakas Ioannis Mollas, Zoe Chrysopoulou. 2022. Ethos: a multi-label hate speech detection dataset.

Zhijing Jin, Xingyu Lu, Francesco Barbieri, Radu Soricut, Yonatan Bisk, and Yulia Tsvetkov. 2023. Weakly supervised learning for improving hate speech detection.

Utkarsh Kant. 2023. Paraphrase with transformer models.

Mehfuza Holia Keivalya Pandya. 2023. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *arXiv preprint*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*.

Jitendra Singh Malik, Hezhe Qiao, Guansong Pang, and Anton van den Hengel. 2023. Deep learning for hate speech detection: A comparative study. *arXiv preprint*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2021. Ethos: An online hate speech detection dataset. *arXiv preprint*.

Francimaria R. S. Nascimento, George D. C. Cavalcanti, and Márjory Da Costa-Abreu. 2023. Exploring automatic hate speech detection on social media: A focus on content-based analysis.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

7 Team Members' Contribution Throughout the Project

- **Mohamed**: Focused on model interpretability by implementing **LIME** and **SHAP**, providing local and global explanations for the model predictions. He also contributed to robustness analysis and interpretability benchmarking and created the web app pipeline .

- **Mohamed Amine**: Led the implementation of **Retrieval-Augmented Classification (RAC)** and handled embedding generation using **Falcon-7B**. He also performed evaluation across multiple classifiers, contributed to model optimization and integrated our best model in the web app.

- **Nabil**: Designed and integrated the **LangChain**-based classification template for hate speech detection. He also contributed to the experimentation pipeline, including preprocessing, the implementation of advanced transformer-based models (BERT variants and GPT), and also created the web app architecture.

8 Annex

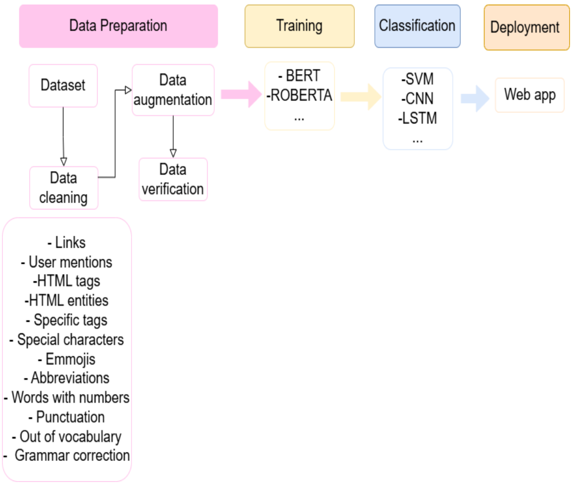


Figure 2: Methodology

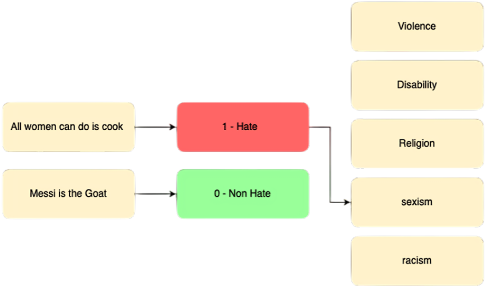


Figure 3: Example of Hate Speech Detection and Type Classification

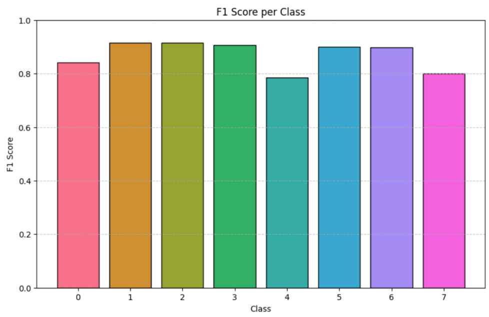
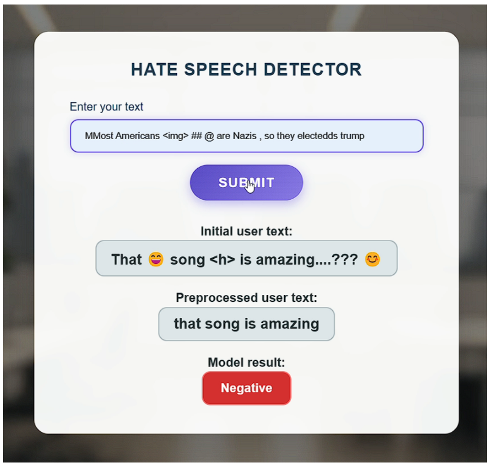


Figure 4: F1-scores per class for the SVM model in multiclass hate speech detection, showing strong performance across most categories.


HATE SPEECH DETECTOR

Enter your text

MMost Americans <img> ## @ are Nazis , so they electedds trump

Initial user text:

That 😄 song <h> is amazing....??? 😊

Preprocessed user text:

that song is amazing

Model result:

Negative

Figure 5: Frontend View of the Deployed Hate Speech Detection System

Model	F1-Score	Accuracy
Neural Network	**0.8713**	**86.46%**
SVM	0.8702	86.17%
CNN	0.8113	80.40%
Hard Voting	0.8100	82.73%
KNN	0.7734	77.23%

Table 1: Summary of the best results obtained for the Multiclass Classification task.

Model	F1-Score	Accuracy
Hard Voting	**0.96**	97.11%
LSTM	0.95	**97.29%**
KNN	0.95	96.93%
CNN	0.95	96.56%
XGBoost	0.91	93.31%
SVM	0.88	93.13%

Table 2: Summary of the best results obtained for the Binary Detection task with GPT embeddings.