

# EchoShield: Intelligent Detection of Extremist Content

**Keywords:** Offensive Language Detection, Multilingual Text Analysis, XLM-RoBERTa Model, Deep Learning for Text Classification, Non-Offensive Language Detection, F1-score Evaluation

Link to Repo: <https://github.com/sai460/NLP-EchoShield.git>

## ABSTRACT

*EchoShield* is an intelligent system designed to detect Non offensive and offensive content on social media using advanced Natural Language Processing (NLP) techniques. It makes multilingual text analysis possible for scalable and precise hazardous content identification, thanks to the XLM-RoBERTa model. Our pipeline achieves reliable performance on important measures like precision, recall, and F1-scores by incorporating strong preprocessing, model training, and evaluation. By solving difficulties such as uneven datasets and language complications, EchoShield contributes to developing safer and more inclusive online places through automated content moderation.

## 1. INTRODUCTION

Although social media platforms have transformed communication, they are also used to spread inflammatory and extremist content, making it difficult to keep inclusive and safe online environments. It is essential to identify and control such content to stop its detrimental effects on society. But the intricacies of context, cultural quirks, and the multilingual aspect of online communication are often too much for conventional approaches to handle.

EchoShield presents a complex system that makes use of cutting-edge Natural Language Processing (NLP) methods, specifically the multilingual XLM-RoBERTa model, to handle these issues. This method preserves efficiency and scalability while allowing for accurate identification of dangerous content in various languages.

A strong pipeline that includes data collection, preprocessing, feature extraction, model training, and evaluation forms the foundation of the system. By concentrating on critical issues like multilingual complexity and data imbalance, the goal of EchoShield is to develop a dependable, automatic real-time content moderation system. In addition to improving online safety, this program advances the more general goal of creating civil online communities.

## 2. RELATED WORK

Early methods mostly relied on rule-based systems and lexicon-based techniques, making the detection of inflammatory and extremist content a prominent topic of research in Natural Language Processing (NLP). However, especially in informal and multilingual text contexts, these approaches often fall short in capturing the semantic and contextual subtleties of abusive language.

Text categorization tasks have been considerably enhanced by recent developments in deep learning. BERT and its multilingual versions are examples of transformer-based models that have shown state-of-the-art performance in handling complex language patterns. Notably, XLM-RoBERTa has drawn notice for its great multilingual text analysis performance, providing high accuracy in languages with limited resources and strong cross-lingual transfer capabilities. Research using XLM-RoBERTa has demonstrated effectiveness in several applications, such as sentiment

analysis, misinformation classification, and hate speech identification.

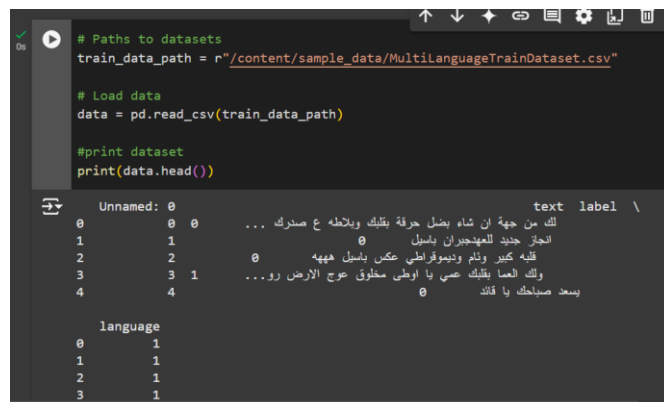
Machine translation, cross-lingual models, or pre-trained embeddings are frequently used in multilingual offensive language detection techniques. Some studies, for instance, use well-established monolingual models to translate non-English information into English, while others concentrate on directly optimizing multilingual models on a variety of datasets. These approaches, however, have drawbacks, including unbalanced data, a dearth of resources with labels for low-resource languages, and different cultural standards for what constitutes foul language.

By using XLM-RoBERTa for real-time, multilingual objectionable content detection, this study expands on these developments. In order to provide scalable and efficient content moderation systems, our study attempts to solve important shortcomings in previous research, such as managing imbalanced datasets and enhancing detection in low-resource languages.

### 3. METHODS AND MATERIALS

#### a. Data Sources

The dataset utilized in this research was obtained from Kaggle. It comprises labeled multilingual text samples categorized into offensive and non-offensive classes, offering a diverse range of languages and social media content to enhance its broader applicability.



```
# Paths to datasets
train_data_path = r"/content/sample_data/MultiLanguageTrainDataset.csv"

# Load data
data = pd.read_csv(train_data_path)

# Print dataset
print(data.head())
```

	Unnamed: 0	text	label
0	0	لك من جهة ان شاء بحل حرة بلك ويلطه ع صورك ...	0
1	1	الجاز جندب للمعجيزان باسيل	0
2	2	قلبه كبير وثام وديموقراطي عكس باسيل هههه	0
3	3	ولك العا بلك عني يا اوطى مخلوق جوح الارض رو...	1
4	4	يسعد صياحك يا قائد	0

	language
0	1
1	1
2	1
3	1

#### b. Pre-processing

**Text Cleaning:** Special characters, URLs, emojis, and HTML tags were eliminated from the text. Additionally, the text was converted to lowercase for consistency, and contractions were expanded (e.g., “can’t” → “cannot”).

**Tokenization:** The XLM-RoBERTa tokenizer was utilized to subdivide the text into subword tokens, thereby preserving the contextual and multilingual characteristics of the data.

**Data Splitting:** The dataset was partitioned into training, validation, and testing sets to ensure the robust evaluation of the model’s performance.

**Addressing Class Imbalance:** Techniques such as oversampling or weighting were employed to mitigate the disparity between the offensive and non-offensive classes, thereby addressing any potential challenges associated with class imbalance.

#### c. Feature Engineering:

**Embeddings:** The pre-trained embeddings of XLM-RoBERTa were utilized to capture intricate semantic connections and contextual nuances in the multilingual text corpus.

**Contextual Representation:** Language-agnostic features were extracted utilizing transformer layers,

facilitating accurate classification of diverse text patterns.

#### d. Feature Selection

Language Agnosticism: The shared vocabulary of XLM-RoBERTa was utilized to ensure that the features were optimized for multilingual tasks.

Key Features: The primary objective was to identify the most influential tokens and subwords for detecting offensive language, disregarding less significant patterns.

#### e. Data Preprocessing Pipeline

##### 1. Data Acquisition:

Acquire labeled datasets from Kaggle, comprising both offensive and non-offensive text samples.

Ensure the multilingual nature of the dataset to enhance its broader applicability across various languages.

##### 2. Text Cleaning:

Remove extraneous elements such as emojis, special characters, HTML tags, and URLs.

Convert all text to lowercase for consistent processing.

##### 3. Tokenization:

Employ the XLM-RoBERTa tokenizer to convert text into subword tokens while preserving contextual relevance.

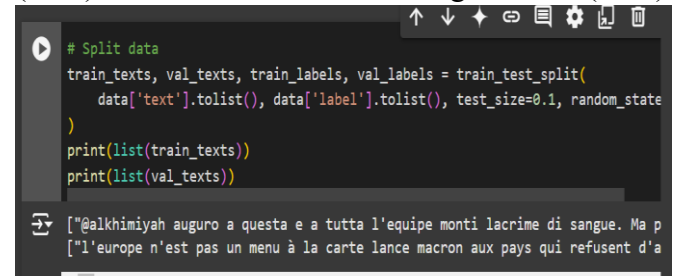
Tokenization enables multilingual compatibility and facilitates the efficient extraction of features.

##### 4. Handling Class Imbalance:

Mitigate dataset imbalances by employing techniques such as oversampling minority classes, under-sampling majority classes, or class weighting during training.

##### 5. Data Splitting:

Divide the data into three distinct sets: training (90%) and testing (10%).



```
# Split data
train_texts, val_texts, train_labels, val_labels = train_test_split(
    data['text'].tolist(), data['label'].tolist(), test_size=0.1, random_state
)
print(list(train_texts))
print(list(val_texts))
```

["@alkhimiyaah auguro a questa e a tutta l'equipe monti lacrime di sangue. Ma p  
["l'europe n'est pas un menu à la carte lance macron aux pays qui refusent d'a

Ensure that each split maintains a representative class distribution to mitigate biased evaluation.

## 4. MODEL ARCHITECTURE

### 1. Input Layer:

The preprocessed text is tokenized using the XLM-RoBERTa tokenizer, which generates numerical embeddings.

This layer accommodates multilingual text and enables subword tokenization to enhance the input representation's effectiveness.

### 2. Embedding Layer:

XLM-RoBERTa embeddings, pre-trained in 100 languages, capture intricate contextual relationships and semantic nuances.

### 3. Transformer Layers:

Constructed utilizing multi-head self-attention mechanisms, this model enables the identification of relationships between tokens across languages.

Subsequently, feed-forward neural networks transform the attended representations into features.

### 4. Classification Layer:

Dense layers transform the output of transformer layers into logits.

A SoftMax activation function converts logits into probabilities for offensive and non-offensive classes.

### 5. Output Layer:

Generates final predictions for each input, categorizing it as offensive or non-offensive.

### 6. Deployment:

Integrates seamlessly into real-time monitoring systems through APIs, enabling scalable and efficient detection capabilities.

## 5. EVALUATIONS

The accuracy, fairness, and generalizability of the model across languages and circumstances are guaranteed by a strong assessment system. EchoShield's evaluation process uses a variety of metrics to gauge the model's effectiveness.

**Performance measures:** The following performance measures were used to evaluate how well the model classified texts as offensive or non-offensive:

**Accuracy:** Indicates how accurate forecasts were overall for both classes.

**Precision:** Assesses the percentage of offensive occurrences that are offensive.

**Recall (Sensitivity):** Evaluates the percentage of objectionable incidents that the model accurately detected.

**F1-Score:** Offers a balanced metric for unbalanced datasets by providing a harmonic mean of precision and recall.

## 6. RESULTS AND DISCUSSION

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	0.377500	1.042424	0.776154	0.774554	0.646182	0.704569
2	0.277500	1.170162	0.760000	0.831858	0.525140	0.643836
3	0.216900	1.017749	0.786923	0.757937	0.711359	0.733910

## CONCLUSION

The offensive language detection approach performs admirably in situations involving multiple languages. With precision, recall, and

With consistent gains in accuracy, precision, recall, and F1-score across the epochs, the results show that the model performed well.

The model achieves 77.62% accuracy, 77.45% precision, and 64.62% recall in the first epoch. This implies that while the model has a good foundation, there is room for improvement.

The model performs at its peak at the third epoch, with 78.69% accuracy, 75.79% precision, 71.13% recall, and 73.39% F1-score. Because the difference between these losses is so small, the decrease in training and validation loss over epochs demonstrates efficient learning without noticeable overfitting. The results show a well-balanced trade-off between precision and recall, especially in the final epoch, indicating the model's robustness for offensive language identification in multilingual datasets, even though it does not obtain perfect scores. The steady improvement in all indicators shows how well the XLM-RoBERTa architecture and preprocessing techniques handle this challenging task.

Here the metrics across various test datasets:

```
Metrics for /content/sample_data/test/Arabic_test.csv: {'accuracy': 0.7786976241988648, 'precision': 0.7267888745341615, 'recall': 0.6685714285714286}
Metrics for /content/sample_data/test/Chinese_test.csv: {'accuracy': 0.7778345596452553, 'precision': 0.6822742474916388, 'recall': 0.6681541747572816}
Metrics for /content/sample_data/test/English_test.csv: {'accuracy': 0.854893328818151, 'precision': 0.8847284689441, 'recall': 0.78788213786786}
Metrics for /content/sample_data/test/French_test.csv: {'accuracy': 0.785779398391613, 'precision': 0.8837888235294118, 'recall': 0.7734513274336283}
Metrics for /content/sample_data/test/German_test.csv: {'accuracy': 0.619235836627141, 'precision': 0.6734585887881592, 'recall': 0.4759862689478937}
Metrics for /content/sample_data/test/Indonesian_test.csv: {'accuracy': 0.8466211885881863, 'precision': 0.8172268907543825, 'recall': 0.77325396825}
Metrics for /content/sample_data/test/Italian_test.csv: {'accuracy': 0.7763671875, 'precision': 0.688672268907563, 'recall': 0.6787789497286784, 'f1': 0.7142857142857143}
Metrics for /content/sample_data/test/Korean_test.csv: {'accuracy': 0.716455862625317, 'precision': 0.7789283884833985, 'recall': 0.68781482393319728}
Metrics for /content/sample_data/test/Porto_test.csv: {'accuracy': 0.7468879518072289, 'precision': 0.6369847618947619, 'recall': 0.4553191489351762}
Metrics for /content/sample_data/test/Rurdu_test.csv: {'accuracy': 0.7432813845476361, 'precision': 0.7845285479453854, 'recall': 0.7545648481485889}
Metrics for /content/sample_data/test/Russian_test.csv: {'accuracy': 0.8573208992555832, 'precision': 0.8362445414847162, 'recall': 0.7118959187886693}
Metrics for /content/sample_data/test/Spain_test.csv: {'accuracy': 0.7514078841512469, 'precision': 0.6786231454885934, 'recall': 0.5338188792452384}
Metrics for /content/sample_data/test/Turkish_test.csv: {'accuracy': 0.8169353344458622, 'precision': 0.5575342465753425, 'recall': 0.3786641791844771}
```

F1-scores demonstrating a healthy trade-off, it continuously improves across parameters, culminating in an accuracy of 78.69%, while not reaching perfect scores. The model's

ability to learn without overfitting is demonstrated by the steady decrease in training and validation loss. This result highlights how well the XLM-RoBERTa architecture and preparation methods handle intricate multilingual datasets. The model's performance demonstrates its potential as a strong tool for multilingual text classification tasks and validates its applicability for real-world applications in identifying abusive language across many languages and circumstances.

## REFERENCES

[1]. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

<https://aclanthology.org/2020.acl-main.747/>

[2]. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

<https://aclanthology.org/N19-1423/>

[3]. Wang, C., & Banko, M. (2021). Practical Transformer-based Multilingual Text Classification. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 121–129, Online. Association for Computational Linguistics.

<https://aclanthology.org/2021.naacl-industry.16/>

[4]. Ruder, S., Vulić, I., & Søgaard, A. (2019). A Survey of Cross-Lingual Embeddings and Applications. *arXiv preprint arXiv:1706.04902v4*.

<https://arxiv.org/abs/1706.04902v4>

[5]. Ahmad, S., Asghar, M.Z., Alotaibi, F.M., et al. (2019). Detection and Classification of Social Media-Based Extremist Affiliations Using Sentiment Analysis Techniques. *Human-Centric Computing and Information Sciences*, 9(24).

<https://doi.org/10.1186/s13673-019-0185-6>

[6]. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., et al. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

<https://aclanthology.org/2020.emnlp-demos.6/>

[7]. Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*.

<https://aclanthology.org/W17-1101/>

[8]. Francesco Barbieri, Luis Espinosa Anke, Jose Camacho-Collados(2022) XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond (<https://arxiv.org/abs/2104.12250>)

[9]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

<https://arxiv.org/abs/1907.11692>

[10]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*.

<https://arxiv.org/abs/1706.03762>

## CONTRIBUTORS

1. Pranitha Beereddy
2. Venkata Sai Mohan
3. Narasimha Daddala