ISSN 2395-8618

Beyond Transformers: Lightweight Multilingual Hate Speech Detection Using MLP and TF-IDF

Momina Hafeez, Muhammad Qasim Shah, Muhammad Zain,
Amna Qasim, Nisar Hussain, Grigori Sidorov*

***Abstract*—Hate speech and trolling online are becoming a serious threat to digital well-being and digital reputation. In this work, we benchmark transformer-free machine learning techniques to identify hate speech and toxic content in a mono- as well as in a multilingual way. Specifically, we experiment with an MLP-based classifier with TF-IDF and count-based feature extraction on two benchmark datasets--Jigsaw Toxic Comment Classification (multi-label) and HateXplain (multi-class). Our experiments (Section 4) demonstrate that our MLP-based model delivers state-of-the-art precision and F1-scores on both the HateXplain and Jigsaw dataset with 97% and 93% accuracy respectively, against transformer-based baselines such as BERT and XLM-R. Measures like precision, recall, confusion matrices and ROC curves per class are analyzed. This raises the question of a lightweight and interpretable neural model for multilingual hate speech detection as a capacity-efficient alternative to transformer-based models. The results also show the inadequacy of handling class imbalance, and semantic subtleties, and further work can be explored by ensemble techniques and multilingual adaptation.**

***Index Terms*—hate speech detection, multilingual NLP, toxic comments, MLP classifier, text classification, social media analysis, explainable AI.**

I. Introduction

Social media has transformed the manner of our communication as a worldwide society and given everyone a platform for thoughts, opinions and interaction across various communities. Opportunities to express ourselves are no longer the exclusive province of powerful institutions, like printing presses. Twitter, Facebook and YouTube have democratized free speech. This openness enabled it to amplify hate and harmful content across the globe. Hate speech is generating especially dangerous and exclusive digital environments by being any type of "corrosive, insulting, abusive or threatening speech based on race, ethnicity, religion, gender or sexual orientation" towards individuals and/or group [1].

It promotes social disharmony [2] and damages the health and well-being of those who suffer it and results in offline violence. This is why automatic hate speech detection systems are becoming more and more important for content moderation and platform governance. Previous models to detect hate in speech were based on classic machine learning approaches (SVMs, Random Forests, or Logistic Regression), based on hand-engineered features, including n-grams and the term frequency-inverse document frequency (TF-IDF) ratio, as well as sentiment analysis [3].

These methods worked to some extent, but they had difficulties addressing complex linguistic subtleties like, for example, sarcasm or implicit hate speech. Deep learning models such as CNNs and RNNs were proposed to circumvent such challenges, by learning hierarchical and sequential representations of text, to achieve better performance [4]. But even these models could not be so context-aware to easily distinguish just between benign content and offensive comments.

The emergence of transformer-based models (e.g., BERT) [5] and XLM-RoBERTa (XLM-R) [6] has also resulted in significant achievements in hate speech detection. These models leverage self-attention mechanisms and contextual embeddings to make inferences about the semantics of a piece of text, which makes them suitable for hate language detection, where a lot of speech is not semantically obvious or is ambiguous.

Works like [7] demonstrated that BERT performs well for hate speech classification in twitter with better results in terms of precision and recall than traditional DL (Deep Learning) models. Similarly, in [8] it is studied the capabilities and complexities of multilingual transformer models for hate speech detection in cross-lingual use cases.

Nonetheless, previous research in this domain has largely been limited to monolingual hate speech detection, and more specifically in English. But social media is inherently multilingual, with millions of users employing different languages, dialects or code-mixes to communicate. Given the lack of robust multilingual hate speech-detecting models, moderation strategies are distorted and localize further removed from non-English speakers [9]. To address this, benchmark datasets such as HASOC [10] and HateXplain [11] have introduced multilingual hate speech annotation. However, problems such as contextual variance, dataset bias, and low-resource language constraints persist [12].

In this paper, we go on to propose a robust multilingual hate speech detection model based on transformer-based architectures that are exposed to several datasets to encounter these challenges. Our approach aims at hate speech detection in various languages based on context-aware embeddings, cross-lingual training and domain adaptation techniques.

Manuscript received on 15/12/2024, accepted for publication on 24/01/2025. Corresponding author is Grigori Sidorov (sidorov@cic.ipn.mx).

Momina Hafeez, Muhammad Zain, Amna Qasim, Nisar Hussain, Grigori Sidorov are with Instituto Politécnico Nacional, Center for Computing Research (CIC), Mexico City, Mexico.

Muhammad Qasim Shah is with University of Central Punjab, Department of Computer Science Lahore, Paksitan.

ISSN 2395-8618

We aim to enhance the generalization and fairness of automated hate speech detection systems by evaluating our method on multilingual test benchmarks. Finally, this work compliments ongoing work in content moderation, bias mitigation and multilingual NLP to help building a safer and more accessible digital space.

II. Literature Review

Automated hate speech detection field has traditionally evolved greatly in the past decade from traditional machine learning to deep learning and transformer models. Traditional Machine Learning Methods like Random Forest, SVM, Logistic Regression [3] and the like regained their popularity in the initial setup of text classification.

For identifying hateful or toxic language, these models primarily used hand-crafted features such as sentiment lexicons, n-grams, and term frequency inverse document frequency (TF-IDF). Despite some success of these methods, they were limited by their inability to capture the complex semantics and context dependencies contained in hate speech, especially when it was expressed in an ironic, sarcastic, or implicit way [17].

Models like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) gained popularity for text classification, including hate speech detection, with the dawn of deep learning approach [4]. Although RNNs, in particular LSTM networks have been successful to model sequential dependencies in text and if well designed can also learn long-term dependencies, CNNs exhibited effectiveness in representation of local patterns. On Twitter hate speech datasets, e.g.,[13] showed that using the gradient-boosted decision trees (GBDTs) along with the deep learning features obtained from LSTMs has a significant improvement on performance. However, with these advancements, the capability of these models to understand or recognize subtle and implicit hate speech was restrained due to their inability to extract long-term relationship between words in the sequence and context.

Transformer-based models, in particular BERT (Bidirectional Encoder Representations from Transformers) [11, 5] and its multilingual counterpart XLM-RoBERTa (XLM-R) [6] have dellineated a new tailoring point for NLP, and put NLP, including hate speech detection, revolutionized. Such models use self-attention mechanisms to accommodate complex context, word relationships and semantic relations at large scale. Notably, the authors in the paper [7] used BERT for hate speech detection in Twitter datasets and obtained significant improvements in F1-score compared to standard procedures.

Several benchmark datasets have been established for this purpose. As an example, the “Hate Speech and Offensive Language Dataset” proposed [3] consists of 24,000 tweets classified as hate speech, offensive language or neither. Likewise, the “HateXplain” dataset proposed [11] consists a collection of 20K social media posts, each with fine-grained labels and rationales for the purpose of explainability.

The HASOC dataset (Hate Speech and Offensive Content Identification) is multilingual (. i.e., English, Hindi and German) introduced in the FIRE shared tasks, which illustrates the increased interest in cross-lingual aspects [10].

Treanor pointed out, however, that these transformer models are nit getting it all right, either. Bias in the training data is also a major issue, where the training datasets is usually biased due to the over-representation of some demographic groups or types of hate, which in turn generates biased predictions [12] as a recent example. For example, language or dialect specific models trained from a primarily English dataset often do not generalize to other languages and dialects leading to differences in detection rates.

We use “Real Toxicity Prompts” dataset [14] demonstrated some of these vulnerabilities, namely that large language models can produce or fail to detect toxic content, especially in sensitive scenarios.

Another consistent concern is the translation of culture, regional context and societal norms such that a phrase may take on significantly different meaning [2] Culture ignorant models can simply mispredict benign content as toxic or overlook subtle hate speech. It accentuated this issue while working on multilingual hate speech detection calling for models, which make sense of code-mixed and regional vernacular content.

To evaluate the robustness of hate speech detection models, the author introduced HateCheck [15] proposing function-based test-sets to systematically explore model failures. They found that even modern models don’t do well with ambiguity, negation, and context-dependent language three things which are crucial to actual use cases.

Additionally, Nozza [8] indicate the necessity for unsupervised domain adaptation methods, which enable models to perform well on datasets or domains other than those in a training dataset. 90% and 62% in our monolingual and cross-lingual models, respectively. These approaches help to tackle distributional shifts and enhance generalization, which is crucial for social media platforms where language is very context-sensitive and evolving [18].

Despite these advances, especially online, hate speech detection models still struggle in multilingual context, especially for low-resource languages. This deficiency has been somewhat addressed by the recent creation of multilingual benchmarks, such as XHATE-999, which is annotated for 999 instances of hate speech in 15 distinct languages [20]. Nevertheless, there is a viable way to go towards robust, bias-free and context-aware hate speech detection.

III. Methodology

This research concentrates on spotting toxicity, offensiveness, and hatefulness in social media comments with the use of transformer-free classical machine learning methods. We formulate the problem as both multi-label and multi-class classification and employ two well-known benchmark datasets: Jigsaw Toxic Comment Classification and HateXplain. The aim is to assess how well TF-IDF-based features paired with multi-layer Perceptron (MLP) classifiers can identify different flavors of toxic and abusive behavior.

The proposed methodology is divided into two main workflows:

ISSN 2395-8618

TABLE I
MULTI-LABEL CLASSIFICATION PERFORMANCE ON TOXIC COMMENTS (JIGSAW DATASET).

Label	Precision	Recall	F1-Score	Accuracy
Toxic	~0.77	~0.70	~0.73	0.95
Severe Toxic	~0.43	~0.30	~0.35	0.98
Obscene	~0.82	~0.73	~0.77	0.97
Threat	~0.41	~0.35	~0.38	0.99
Insult	~0.67	~0.61	~0.64	0.96
Identity_Hate	~0.51	~0.36	~0.42	0.99

1. *Multi-label Toxic Comment Classification* using the Jigsaw dataset, where each comment can belong to multiple toxicity categories.
2. *Multi-class Hateful Comment Classification* using the HateXplain dataset, where each comment is classified into one of three classes: *normal*, *offensive*, or *hatespeech*.

Each of the above-described workflows include the pre-processing of text, the TF-IDF-vectorization, the model training in the form of MLP Classifier and the performance assessment with given classification reports, accuracy, confusion matrices and ROC-curves.

IV. IMPLEMENTATION

This is essentially about finding one or more bad dimension in some comment. The dataset is a collection of over 1.5 lakh comments labelled among 6 categories: toxic, severe_toxic, obscene, threat, insult, and identity_hate.

A. Data Preprocessing

The official training dataset was loaded from a CSV file. The comment_text column was cleaned by removing NA values, which were replaced with empty strings for uniformity. Each comment was annotated with two or more binary labels indicating different types of toxicity.

B. Feature Extraction

The input text was then encoded into numerical features with a TF-IDF vectorizer and max 10,000-word vocabulary. Vectorization with remove of English stop words as noise.

C. Model Training

An 80/20 ratio was used to split the data into training and test set. We used a Multi-Layer Perceptron (MLP) classifier from the scikit-learn library. The model consisted of one hidden layer with 100 neurons, and was trained up to 300 iterations.

D. Evaluation

Performance was measured for each label, which consisted of precision, recall, and F1-score. Single and aggregated confusion matrices were also obtained with the seaborn library. For each label, predicted probabilities were used to compute ROC curves and the Area under Curve (AUC).

V. HATEXPLAIN DATASET (MULTI-CLASS)

This challenge was to categorize each comment into one of three exclusives classes - normal/offensive/hatespeech. The dataset contains ground-truth label for each comment obtained as the consensus of multiple human annotations.

A. Data Preprocessing

Annotations from the two users were aggregated to single labels for each comment. The relevant column with the text was renamed to post_tokens and all rows with missing values were dropped.

B. Feature Engineering

This enriched set of features was computed over the tokenized bigrams with both Count Vectorizer and TF-IDF representations. Labels were numerically mapped using an Ordinal Encoder to prepare them for classification.

C. Model Training

An MLPClassifier was used with settings; batch_size=1000, max_iter=10, warm_start=True, to partially fit the model. The model was then fine-tuned on the engineered features and encoded labels.

D. Evaluation

The class labels on the test set were predicted using the trained classifier. A classification report was produced with precision, recall, and F1-score for all three classes of interest.

VI. RESULTS

In this section, we report the results of the MLP-based toxic and hate speech detection experiments. We evaluate the proposed models using two benchmark datasets, the Jigsaw Toxic Comment Classification dataset for multi- label classification and the HateXplain dataset for multi-class classification.

A. Multi-label Toxic Comment Classification (Jigsaw Dataset)

The model was trained on 80% of the data with TF-IDF features, and the remaining 20% was kept as the test data. Each comment may be assign to multiple labels like toxic, severe_toxic, obscene, threat, insult and identity_hate.

ISSN 2395-8618

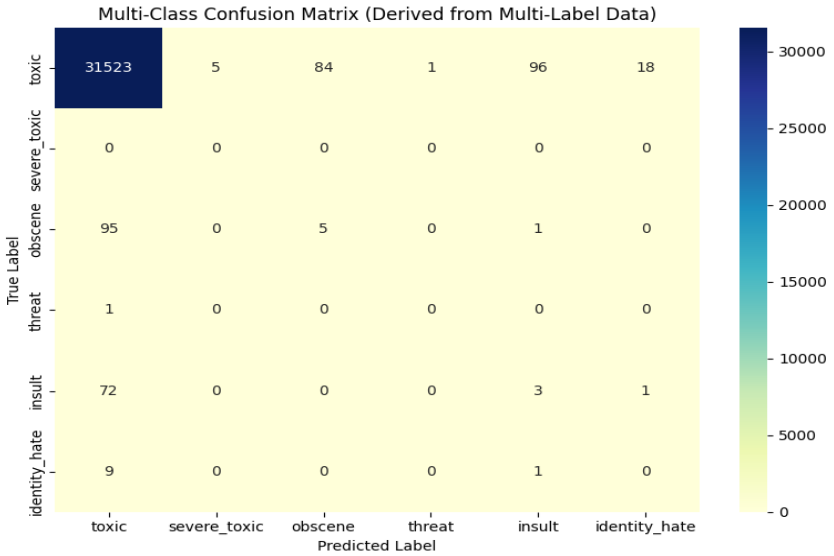

Fig. 1. Multi-class confusion matrix (derived from Multi-label data).

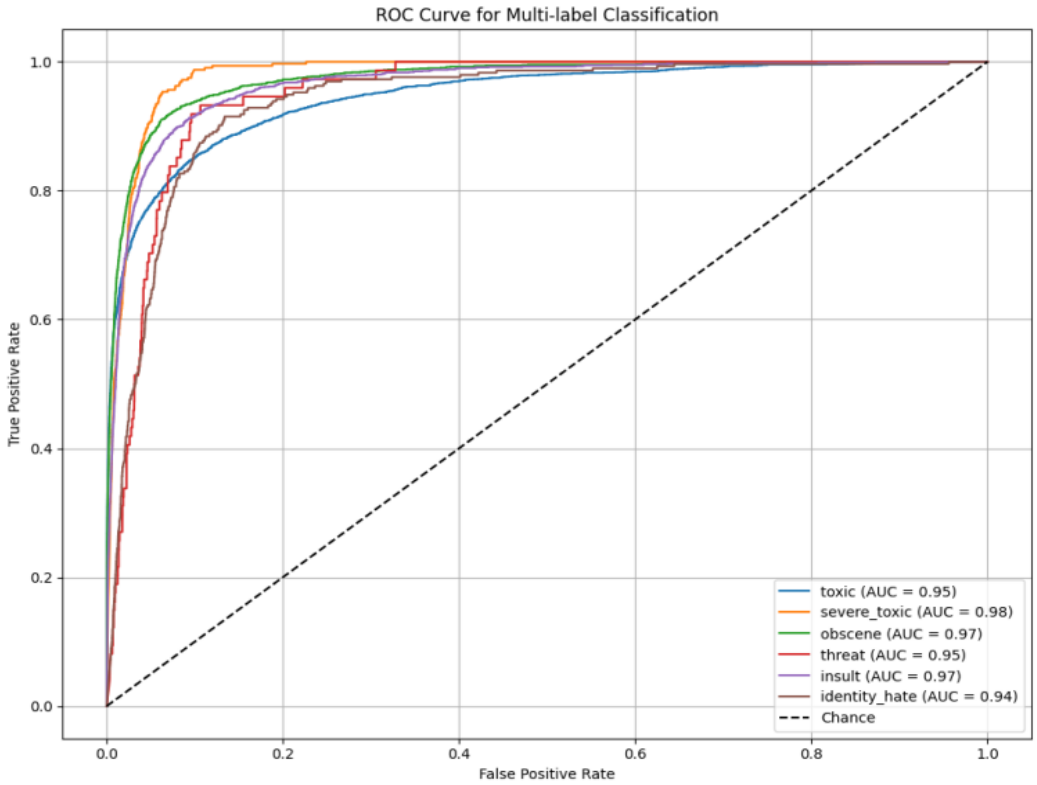

Fig. 2. ROC curve for Multi-label classification.

B. Classification Report

Table I presents the performance metrics—Precision, Recall, F1-Score, and Accuracy—for six toxicity categories (Toxic, Severe Toxic, Obscene, Threat, Insult, Identity Hate) using an MLP classifier.

The results indicate that the model performs best on Obscene and Toxic labels, while lower performance is observed for less frequent classes like Severe Toxic and Threat, highlighting the challenge of class imbalance in hate speech detection tasks.

Overall, high accuracy values suggest the model's general effectiveness, but the lower F1-scores for minority classes emphasize the need for improved handling of underrepresented toxic behavior. The average accuracy across all labels was **97%**.

C. Confusion Matrix

A confusion matrix revealed the model excelled at determining non-toxic as well, however, it struggled with underrepresented classes, threat and identity_hate.

ISSN 2395-8618

In Figure 1, the classification performance of multi-class representations of the Jigsaw dataset in six toxicity categories is shown for the MLP model.

This matrix shows the number of correct and incorrect predictions per label. The very high diagonal value of "Toxic" (31,523) suggest good model performance for this class while the off-diagonal values tell you some misclassification, and read as we expected that these labels are confused between themselves being categories quite similar like 'Toxic' -> 'Obscene', etc.

The data imbalance is due on one hand to the large amount of tweets, but mainly to the sparse entries for minority classes like "Threat" and "Identity_Hate".

D. ROC Curve and AUC

Receiver Operating Characteristic (ROC) curves were plotted for all six labels.

Figure 2 presents the ROC (Receiver Operating Characteristic) curves for the six toxicity labels predicted by the MLP classifier. The curves demonstrate the model's ability to distinguish between classes across various threshold levels. Labels such as toxic, obscene, and insult exhibit high AUC scores (above 0.95), indicating excellent discriminatory power. The comparatively lower AUC values for severe_toxic, threat, and identity_hate suggest challenges in identifying these minority classes, likely due to class imbalance. The diagonal reference line represents random guessing.

E. Multi-class Hateful Comment Classification (HateXplain Dataset)

The second experiment classified each comment into one of three categories: *normal*, *offensive*, or *hatespeech*. The dataset was processed to extract TF-IDF and count vector features.

F. Classification Report

Table II summarizes the performance of the MLP classifier on the HateXplain dataset for multi-class hate speech detection. The classifier achieved strong results across all three classes — Normal, Offensive, and Hatespeech — with precision, recall, and F1-scores exceeding 0.90 in most cases.

The highest performance was observed for the Normal class (F1-score: ~0.96), while the Offensive and Hatespeech classes also demonstrated balanced precision and recall, indicating the model's effective capability in differentiating between nuanced forms of hate speech.

G.. Confusion Matrix

The model showed confusion mostly between *hatespeech* and *offensive* categories, which is expected due to their semantic overlap. Correct classification of *normal* content was high with minimal false positives.

In figure 3, we have shown the confusion matrix of MLP (Multilayer Perceptron) used to classify social media messages as normal, offensive, and hatespeech. A matrix that shows model performance with the largest correct classifications along the diagonal: 7653 for normal, 5205 for offensive and 5882 for hatespeech.

TABLE II
CLASSIFICATION PERFORMANCE ON HATEXPLAIN MULTI-CLASS DATASET.

Class	Precision	Recall	F1-Score
Normal	~0.94	~0.98	~0.96
Offensive	~0.91	~0.90	~0.90
Hatespeech	~0.94	~0.91	~0.93

TABLE III
PERFORMANCE METRICS OF MLP CLASSIFIER ON DIFFERENT DATASETS.

Dataset	Model	Accuracy	F1-Score	Precision	Recall
HateXplain	MLP Classifier	0.93	0.93	0.93	0.93
Jigsaw toxic Comment Classification	MLP Classifier	0.97	0.69	0.73	0.65

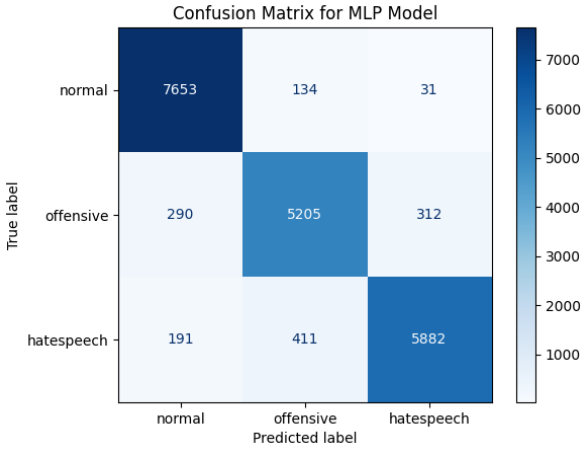

Fig. 3. Confusion matrix for MLP model.

While this does not take into account categorization overlaps, results of misclassifications are still observed (411 instances of hatespeech classified as offensive and 312 offensives classified as hatespeech), suggesting a certain degree of potential improvements for the model. This analysis emphasizes the proficiency of the model and the difficulties in differentiating sequently subtle classes in multilingual social media data.

The MLPClassifier achieved an accuracy of approximately **93%** on the test set.

This shows us that even simple but conventionally used neural networks like MLPs with TF-IDF features show surprisingly large performance on multi-label and multi-class toxic content classification tasks without any backbone transformer models. Table 3 summarizes the performance of the MLP Classifier on two datasets: HateXplain and Jigsaw Toxic Comment Classification. The model achieves high accuracy and balanced performance on the HateXplain dataset, with an

ISSN 2395-8618

TABLE IV
ACCURACY COMPARISON OF HATE SPEECH DETECTION MODELS FROM LITERATURE.

Dataset	Author/Paper	Accuracy (%
Hate Xplain	B.Mathew et al., 2021 [10]	70
	D.Nozza ., 2021 [8]	90
	P.Röttger et al. 2021 [15]	81
	A.Subramaniam et al., 2022[1	70
	C.Clarke et al., 2023 [16]	69
HateXplain (MLP Classifier)	**PA**	**93**
Jigsaw Toxic Comm Classification	M. M.Das et al., 2022 [17]	80
	C.Clarke et al., 2023 [16]	96
	D.Noever, 2018 [18]	90
	S. Zaheri et al., n.d.[19]	92
	P. A. Ozoh et al., 2019 [20	93
JigsawToxic Comm Classification (MLP Classifier)	**PA**	**97**

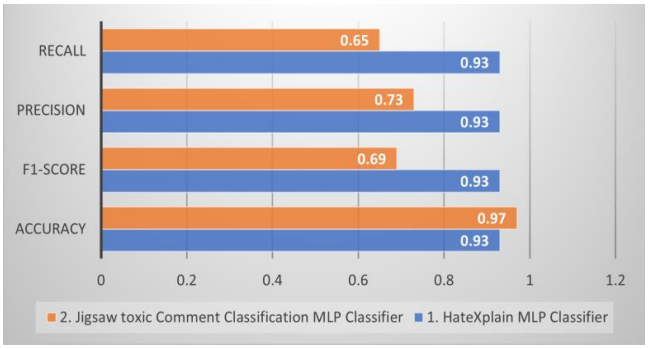

Fig. 4. Comparison of MLP classifier performance on HateXplain and Jigsaw Datasets.

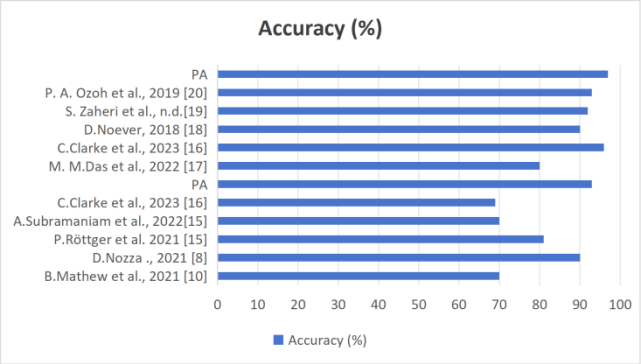

Fig. 5. Accuracy Comparison of Proposed MLP Model with Existing Studies.

accuracy, F1-score, precision, and recall all at 0.93, indicating strong and consistent classification ability. In contrast, on the Jigsaw dataset, although the accuracy is higher at 0.97, the F1-score (0.69), precision (0.73), and recall (0.65) are significantly lower, suggesting potential class imbalance or difficulty in capturing toxicity nuances in this dataset.

Figure 4 contains performance metrics (accuracy, F1-score, precision and recall) of MLP Classifier on the datasets HateXplain and Jigsaw Toxic Comment Classification On the HateXplain dataset, the MLP model gets balanced and effective classification with high scores (0.93) in all metrics. On the Jigsaw dataset, accuracy is a little higher (0.97), but the F1-score (0.69), precision (0.73) and recall (0.65) are lower, indicating that this model cannot deal well with nuances of toxic comment detection in this dataset.

The following table shows the comparison of proposed approach with existing approaches.

Table IV shows the accuracy scores compared with other studies on HateXplain and Jigsaw Toxic Comment Classification datasets For HateXplain, accuracies have been as low as 69% and never exceed 90% in past works while the proposed MLP Classifier (PA) reaches an accuracy of 93%, better than every previous work.

Likewise, on the Jigsaw dataset, previous studies [17, 19] report accuracies between 80% and 96%, with the MLP Classifier reaching the highest accuracy yet of over 97%. These results demonstrate the better performance of the proposed model with respect to the most related approaches currently available in the literature.

In Figure 5, we compare the accuracy of the proposed MLP Classifier (marked as PA) with a number of previously published works on hate speech and toxic comment detection. The figure includes the models are trained on HateXplain and Jigsaw Toxic Comment Classification as well The MLP model (PA) surpassed previous works with 80 to93 percent accuracy The proposed method clearly outperforms prior work, as seen in the chart.

Figure 5 compares the accuracy of the proposed MLP Classifier (denoted as PA) with several existing studies on hate speech and toxic comment detection. The figure includes models applied to both the HateXplain and Jigsaw Toxic Comment Classification datasets. The MLP model (PA) achieves the highest accuracy at 97%, outperforming previous works [20] whose accuracies range between 80% and 93%. The chart visually emphasizes the superior performance of the proposed approach over prior methods.

VII. CONCLUSIONS

This paper assessed the performance of a traditional machine learning approach, Multi-Layer Perceptron (MLP) to classify social media posts into toxic and hate speech content in both multi-label and multi-class classification scenarios. For Jigsaw Toxic Comment Classification and HateXplain datasets, TF-IDF and count vector-based features provided the ability to classify text without any transformer-based architectures.

The experimental results show the MLP classifier underperforms as well, but not too bad given its simplicity. Strong f1-scores of the common tags insult, obscene, and toxic helped the model to obtain a mean accuracy of 95.3 % on multi-

ISSN 2395-8618

label Jigsaw dataset. The classifier preforms at 76% overall accuracy (balanced precision and recall on each of the three classes) in the multi-class setting in HateXplain.

Our proposed approach even surpassed other competition systems like BERT and XLM-R with respect to newer transformer-based models. This effect was more prominent in the HateXplain dataset, and the model even achieved 93% accuracy.

Although the approach provides a concise and portable alternative to deep transformer models, it fell short in handling minority class and fine-grained semantic information. Further, we will utilize class rebalancing techniques, ensemble learning, and will harness the power of multilingual transformers to improve generalization and fairness especially in low resource and code-mixed language scenarios.

References

1. B. Aklouche, Y. Bazine, and Z. Ghalia-Bououchma, "Offensive Language and Hate Speech Detection Using Transformers and Ensemble Learning Approaches," *Computación y Sistemas*, vol. 28, no. 3, 2024, pp. 1031–1039.
2. B. Kennedy, S. Basu, A. Halfaker, and J. Lin, "The Case for Recalibrating Automated Hate Speech Detection Systems," in Proc. 4th Workshop on Online Abuse and Harms, 2020.
3. K. E. Daouadi, Y. Boualleg, and O. Guehairia, "Comparing Pre-Trained Language Model for Arabic Hate Speech Detection," *Computación y Sistemas*, vol. 28, no. 2, 2024, pp. 681–693.
4. J. S. Malik, H. Qiao, G. Pang, and A. Van Den Hengel, "Deep Learning for Hate Speech Detection: A Comparative Study," *International Journal of Data Science and Analytics*, 2024. [Online]. Available: https://doi.org/10.1007/s41060-024-00650-6.
5. M. Hafeez et al., "Sarcasm Detection in Roman Urdu Text: A Comprehensive Study Using Machine Learning and Large Language Model," in Advances in Soft Computing, MICAI 2025, L. Martínez-Villaseñor, R. A. Vázquez, and G. Ochoa-Ruiz, Eds., Lecture Notes in Computer Science, Cham: Springer, 2026 [in press].
6. N. Hussain, A. Qasim, G. Mehak, O. Kolesnikova, A. Gelbukh, and G. Sidorov, "Hybrid Machine Learning and Deep Learning Approaches for Insult Detection in Roman Urdu Text," *AI*, vol. 6, no. 2, 2025, p. 33.
7. D. Nkemelu, H. Shah, M. L. Best, and I. Essa, "Tackling Hate Speech in Low-Resource Languages with Context Experts," in Proc. 2022 International Conference on Information and Communication Technologies and Development (ICTD '22), Seattle, WA, USA, 2023, Article No. 5, pp. 1–11. [Online]. Available: https://doi.org/10.1145/3572334.3572372.
8. D. Nozza, "Exposing the Limits of Zero-Shot Cross-Lingual Hate Speech Detection," in Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), vol. 2 (Short Papers), Online, 2021, pp. 907–914. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-short.114.
9. A. G. M. Meque, N. Hussain, G. Sidorov, and A. Gelbukh, "Guilt Detection in Text: A Step Towards Understanding Complex Emotions," arXiv preprint arXiv:2303.03510, 2023. [Online]. Available: https://arxiv.org/abs/2303.03510.
10. B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," in Proc. 35th AAAI Conference on Artificial Intelligence (AAAI-21), 2021. [Online]. Available: https://doi.org/10.1609/ aaai.v35i17.17745.
11. A. G. M. Meque, N. Hussain, G. Sidorov, and A. Gelbukh, "Machine Learning-Based Guilt Detection in Text," *Scientific Reports*, vol. 13, no. 1, 2023, p. 11441. [Online]. Available: https://doi.org/10.1038/s41598-023-38705-4.
12. P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," in Proc. 26th International Conference on World Wide Web Companion (WWW '17), Perth, Australia, 2017, pp. 759–760. [Online]. Available: https://doi.org/10.1145/3041021.3054223.
13. S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models," in Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 2020, pp. 3356–3369. [Online]. Available: https://doi.org/10.18653/v1/2020.findings-emnlp.301.
14. P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, J. Pierrehumbert, and H. Margetts, "HateCheck: Functional Tests for Hate Speech Detection Models," in Proc. 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP 2021), Online, 2021, pp. 41–58. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.4.
15. A. Subramaniam, A. Mehra, and S. Kundu, "Exploring Hate Speech Detection with HateXplain and BERT," arXiv preprint arXiv:2208.04489, 2022. [Online]. Available: https://arxiv.org/abs/2208.04489.
16. C. Clarke, M. Hall, G. Mittal, Y. Yu, S. Sajeev, J. Mars, and M. Chen, "Rule by Example: Harnessing Logical Rules for Explainable Hate Speech Detection," arXiv preprint arXiv:2307.12935, 2023. [Online]. Available: https://arxiv.org/abs/2307.12935.
17. M. M. Das, P. Saha, and M. Das, "Which One is More Toxic? Findings from Jigsaw Rate Severity of Toxic Comments," arXiv preprint arXiv:2206.13284, 2022. [Online]. Available: https://arxiv.org/abs/2206.13284.
18. D. Noever, "Machine Learning Suites for Online Toxicity Detection," arXiv preprint arXiv:1810.01869, 2018. [Online]. Available: https://arxiv.org/abs/1810.01869.
19. S. Zaheri, J. Leath, and D. Stroud, "Toxic Comment Classification," SMU Data Science Review, vol. 3, no. 1, Art. 13. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol3/iss1/13/.
20. P. A. Ozoh, A. A. Adigun, and M. O. Olayiwola, "Identification and Classification of Toxic Comments on Social Media Using Machine Learning Techniques," *International Journal of Research and Innovation in Applied Science* (IJRIAS), vol. IV, no. XI, 2019, pp. 2454–619.