Undefined 1 (2009) 1–5
IOS Press

Detecting Multilingual Hate Speech Targeting Immigrants and Women on Twitter

Olga Kolesnikova †, Mesay Gemeda Yigezu †, Alexander Gelbukh, Selam Abitte and Grigori Sidorov
Centro de Investigación en Computación (CIC), Instituto Politécnico Nacional (IPN), Mexico City 07738, Mexico
Correspondence: kolesnikova@cic.ipn.mx
† These authors contributed equally to this work.

Abstract. Twitter has experienced a tremendous surge in popularity over recent years, establishing itself as a prominent social media platform with a large user base. However, with this increased usage, there has been a concerning rise in the number of individuals resorting to derogatory language and expressing their opinions in a demeaning manner toward others. This surge in hate speech has drawn significant attention to the field of sentiment analysis, which aims to develop algorithms capable of detecting and analyzing emotions expressed in social networks using intuitive approaches.

This paper focuses on addressing the complex task of detecting hate speech and aggressive behavior while performing target classification. We explored various deep-learning approaches, including LSTM, BiLSTM, CNN, and GRU. Each offers unique capabilities for capturing different aspects of the input data. We proposed an ensemble approach that combines the top three performing models. This ensemble approach benefits from the diverse strengths of each individual model showing F1 score of 0.85 for English-HS, 0.94 for English-TR, 0.92 for English-AB, 0.84 for Spanish-HS, 0.86 for Spanish-TR, 0.97 for Spanish-AB, 0.74 for multilingual-HS, 0.94 for multilingual-TR, and 0.88 for multilingual-AB.

Keywords: hate speech, aggressive behavior, target classification, ensemble learning, deep learning, target classification

1. Introduction

Hate speech (HS) is defined as any form of communication, whether spoken, written, or expressed through behavior, that involves the use of derogatory or discriminatory language towards an individual or a group. Such language can specifically target a person or group based on their inherent characteristics, which may include their religion, ethnicity, nationality, race, color, lineage, gender, or any other factor related to their identity [1,2,3].

The rapid progress in mobile technology and internet connectivity has revolutionized the way people communicate, share ideas, interact with others, and exchange information through social media platforms. While social media offers a convenient and effective means of communication, it has also become a platform for the dissemination of HS. The various features available on the internet often contribute to the misuse of social networks, enabling the transmission and widespread propagation of hateful content [4,5].

The dissemination of HS targeting immigrants and women on social media platforms has become a concerning issue. It involves the spread of derogatory and discriminatory content, which perpetuates negative stereotypes, incites violence, and promotes prejudice against these groups [1]. The consequences of HS towards immigrants and women can be significant and detrimental, affecting both individuals and society as a whole [6,7]. The consequences of HS dissemination over immigrants and women on social media include

- Psychological Impact: psychological effects on its targets, leading to increased stress, anxiety, depression, and feelings of isolation. It can also contribute to low self-esteem and reduced well-being among those who are subjected to it [8].

[1] https://www.coe.int/en/web/freedom-expression/hate-speech

- Social Exclusion: HS perpetuates a hostile and unwelcoming environment, leading to social exclusion and marginalization of immigrant communities and women. It can create barriers to their integration into society and hinder their access to equal opportunities and resources [9].
- Threat to safety: HS can escalate into real-world acts of discrimination, harassment, and violence. It creates an environment where individuals are at a higher risk of experiencing physical harm or being targeted based on their ethnicity or gender[10, 11,12].

In order to combat the above-mentioned problems, we introduce the application of deep learning techniques to analyze HS. By leveraging the power of deep learning algorithms [13],[14], we can effectively identify instances of HS targeting migrants and women in tweets. To accomplish this, we employ an ensemble of deep learning models that have been specifically trained to address the unique challenges associated with detecting HS, specifically identifying instances of HS directed towards migrants and women.

These activities encompass several tasks related to addressing HS towards women and immigrants on social media platforms. The first task involves detecting HS, which is a binary classification problem where the goal is to predict whether a given tweet contains HS or not. The second task involves classifying the target range (TR) of the HS. This task aims to categorize tweets into two main categories: those that contain hateful messages specifically directed at an individual or group target. The third task is focused on detecting aggressive behavior (AB) within hateful tweets. The objective here is to classify hateful tweets based on whether they exhibit elements of AB or not.

The novelty of this paper revolves around several key contributions.

- It involves conducting a comprehensive analysis through the implementation of data balancing techniques.
- It delves into a comparative study of deep learning models and our proposed approach. By conducting this comparison, we aim to highlight the strengths and weaknesses of both approaches, ultimately shedding light on the advantages offered by our proposed approach.
- The paper emphasizes the significance of constructing and evaluating monolingual and multilingual HS detection models.

This entails developing models that can effectively identify and classify HS in both single-language and multilingual contexts. By evaluating the performance of these models, we contribute to the field's understanding of HS detection in various linguistic settings.

The paper is structured in the following manner: To comprehend the latest advancements in the field, related works are presented in Section 2. Our dataset is described in Section 3. In Section 4, the proposed architecture for addressing the problem is showcased. The experiments and their results are presented in Section 6. Lastly, an overall summary and future work of the research activities are provided in Section 7.

2. Related Work

In [15], the authors focused on improving the analysis of HS using a deep learning (DL) approach. They conducted experiments to classify HS against women and detect whether the HS was performed by individuals or groups of people using DL models. To implement their approach, the authors employed CNNs as the core DL models and trained them on Spanish and English data. The models were configured with various parameters. Another thing the authors did was combine word embedding models, such as inverse Glove (global vector) [16], term and inverse document frequency (TF-IDF), and transformer-based embeddings [17].

The implemented models were combinations of CNNs, BiLSTM, and multilayer perceptron (MLP). In conclusion, the researchers found that the implemented model, which combined the transformer model with the DL models, achieved higher accuracy, with an F1 score of 0.93 based on the English dataset for hate speech, and 0.93 on the Spanish dataset for hate speech.

In [18], the researchers directed their attention to the detection of HS and sexism in multilingual contexts, specifically targeting two groups: immigrants and women. Their investigation encompassed English and French tweets, highlighting the importance of a multilingual perspective. To address such challenge, the researchers employed feature-based models and conducted a series of experiments. For the identification of HS, the Random Forest algorithm showed itself as the most effective approach which achieved an F1 score of 0.78. In contrast, the detection of sexism relied on the Support Vector Machine algorithm, which yielded promising results that is an F1 score of 0.68.

In [19] the authors presented HaterNet, an intelligent system currently in use by the Spanish National Office Against Hate Crimes of the Spanish State Secretariat for Security. HaterNet serves as an innovative solution for identifying and monitoring the evolution of HS on Twitter [2]. Additionally, the authors provide a new publicly available dataset specifically focused on HS in the Spanish language. This dataset comprises 6,000 expert-labeled tweets, facilitating further research in this domain. Furthermore, the authors conducted a comprehensive comparison of various classification approaches based on different document representation strategies and text classification models. Among the approaches tested, the best-performing method involved a combination of an LSTM model and an MLP neural network, which achieved an AUC of 0.828.

The authors of [20] focused on HS detection and explored a straightforward ensemble approach utilizing transformers. The study involved two ensembles that leveraged a pretrained RoBERTa model [21], initially trained on one million tweets from the OffensEval competition [22], and subsequently fine-tuned on different folds of the data. The evaluation of these ensembles was conducted on the HASOC test [23] set. The research findings revealed interesting observations. In the case of the HASOC *only* model, the ensemble demonstrated a superior F1 score. However, the situation differed for the HASOC *offensEval* model, where the ensemble's best performance was achieved by utilizing the model trained on the fifth fold. Notably, regardless of these variations, both ensemble methods outperformed the winner of the HASOC competition by achieving an F1 score of 0.85

[24] presented deep ensembles as a method for estimating predictive uncertainty in deep learning models. The researchers build upon some previous approaches and propose deep ensembles as a simple and scalable method for predictive uncertainty estimation. Their approach involves training an ensemble of deep neural networks with a shared architecture but different random initializations. By combining the predictions of these models, they obtained reliable uncertainty estimates that outperformed other methods in terms of accuracy and scalability.

[2] **Twitter was renamed to X. In this paper, we use the previous name Twitter and refer to messages on Twitter as tweets.**

3. Datasets

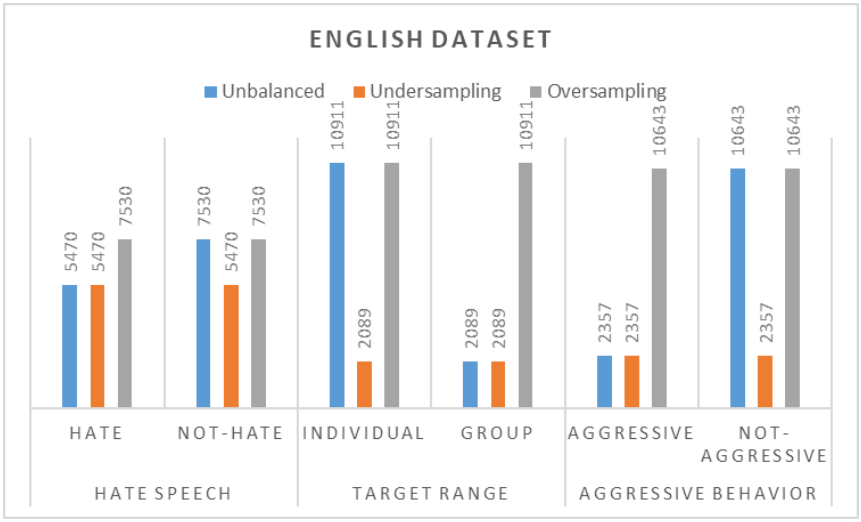
The data utilized in our study was obtained from the publicly available HatEval dataset, more specifically from the SemEval-2019 Task 5 sharing task [25]. The entire HatEval dataset consists of 19,600 tweets, with 13,000 tweets in English and 6,600 tweets in Spanish. These tweets are distributed across two targets: 9,091 tweets about immigrants and 10,509 tweets about women. This distribution holds true for both English and Spanish data that are publicly accessible [3]. Fig. 1 depicts the statistics of the dataset in detail.

The data is not balanced in all aspects and can bias the detection of a desired class. In order to reduce such possible bias, we applied two data balancing techniques: random under-sampling and random over-sampling. Random under-sampling involves reducing the number of instances in the majority class to match the number of instances in the minority class [26]. This technique helps to create a more balanced dataset but may discard valuable information present in the majority class. Random over-sampling involves increasing the number of instances in the minority class to match the number of instances in the majority class. This can be achieved by duplicating existing instances or generating synthetic instances using such techniques as random under-sampling and random over-sampling [27]. It is important to choose the appropriate technique(s) based on the problem at hand and carefully evaluate their impact on model performance [28],[29].

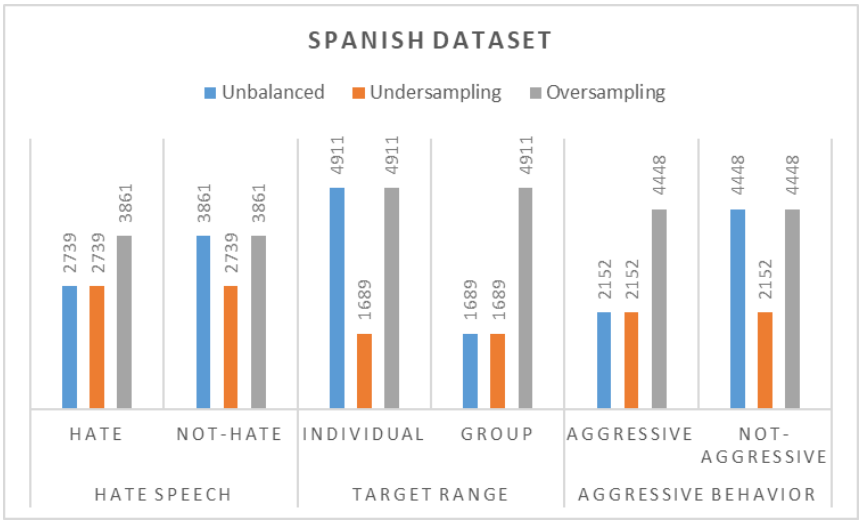
4. Proposed Architecture

In our research, we aim to address the challenging task of detecting HS and AB while also performing target classification. To tackle this task, we explore and propose several deep learning approaches, namely, LSTM, BiLSTM, CNN, and GRU. Each of these models offers unique capabilities in understanding and capturing different aspects of the input data. To ensure a thorough evaluation, we conduct experiments using each of these models individually for the aforementioned tasks. We analyze their performance and compare the results obtained from each model. While each model shows promising results individually, we observe that no single model consistently outperforms the others across all tasks. Based on these findings, we propose an ensemble approach by combining the top

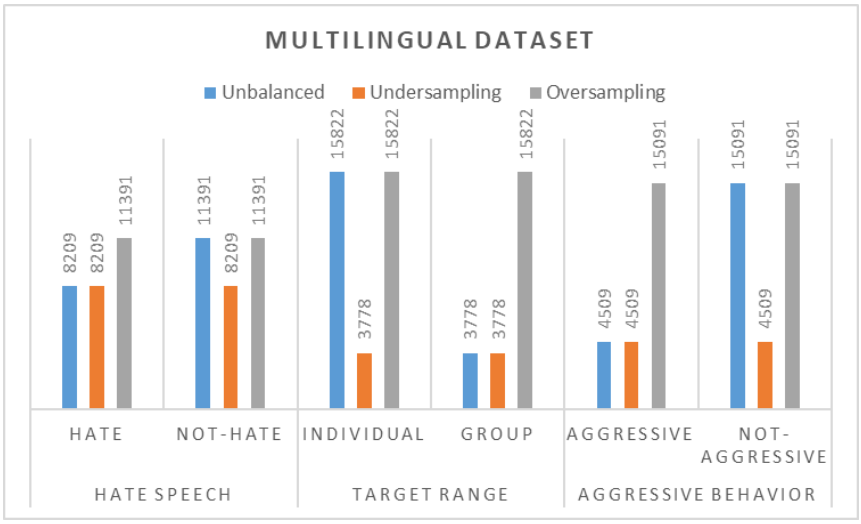
[3] http://hatespeech.di.unito.it/hateval.html



(a) English dataset distribution



(b) Spanish dataset distribution



(c) Multilingual dataset distribution

Fig. 1. Dataset distribution

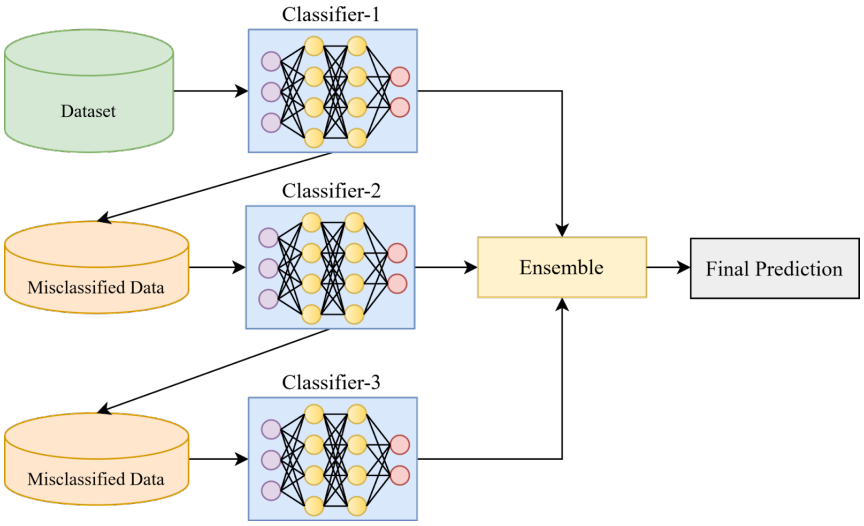


Fig. 2. The proposed ensemble architecture (*Dataset can be Spanish, English, and both or multilingual and classifier 1, 2, and 3 indicate the top three deep learning approaches based on their performance*)

three performing models, that will benefit from the diverse strengths of each individual model.

We employed boosting in an ensemble model [30] which works by iterative training weak learners (base models) and combining their predictions to create a strong predictive model. The key idea behind boosting is to sequentially train each weak learner to focus on the instances that were misclassified or had high errors by the previous learners. This way, the ensemble gradually learns to correct the mistakes made by earlier models and improve overall predictive accuracy. This approach captures a wider range of linguistic and contextual features, leading to enhanced performance in detecting HS, AB, and performing TR classification. Fig. 2 shows an overview of the ensemble approach.

5. What was the Rationale behind Choosing this Approach?

An ensemble of deep learning models is proposed to enhance the overall effectiveness and resilience of the classification or prediction task. An ensemble of deep learning models is often selected for several reasons:

Increased accuracy: Ensemble models have the potential to achieve higher accuracy compared to individual models. By combining the predictions of multiple models, the ensemble can capture different aspects of the data and leverage the collective knowledge of the individual models.

Reduced overfitting: Deep learning models are prone to overfitting, especially when the dataset is small or noisy. By combining multiple models with different architectures, initializations, or training strategies, the ensemble can reduce overfitting by averaging out individual model biases and errors.

Improved generalization: Ensemble models have better generalization capabilities, meaning that they can perform well on unseen or test data. The ensemble can capture diverse patterns, features, and relationships present in the data, leading to more robust predictions.

Increased model stability: Ensemble models tend to be more stable than individual models. They are less sensitive to small variations in the training data or model configurations, which can lead to more consistent and reliable predictions.

Complementary learning: Each deep learning model in the ensemble may specialize in capturing different aspects of the data or learning different representations. By combining their predictions, the ensemble can benefit from the complementary strengths of the individual models.

Model interpretability: Ensemble models can provide more insights into the decision-making process. By analyzing the contributions of the individual models, it becomes possible to understand the importance of different features or patterns in the data.

Overall, an ensemble of deep learning models offers a powerful approach to enhance the performance, stability, and generalization of the classification or pre-

diction task. It leverages the diversity and collective intelligence of multiple models to produce more accurate and robust results. Data balancing techniques are used to address the issue of class imbalance in machine learning tasks, where the number of instances in different classes is significantly uneven. These techniques aim to mitigate the impact of class imbalance and ensure that the model does not become biased towards the majority class.

6. Experiments and Results

In our study, we conducted a series of experiments using various datasets and deep learning approaches. The models were trained on the training subset and subsequently evaluated on the test subset for each of three datasets in Fig. 1. The experiments were conducted on a single GPU-NVIDIA V100 and A100 equipped with 52GB of RAM, ensuring computational efficiency.

To obtain a comprehensive evaluation, we considered both monolingual and multilingual datasets. We explored different data balancing techniques, such as random under-sampling and random over-sampling, to address class imbalances in the datasets. For each dataset, we performed three different experiments, examining the effectiveness of these balancing techniques based on unbalanced data.

In order to implement our deep learning approaches, we employed various functions. One notable technique involved training a Word2Vec model [31] using the distributed training capabilities provided by TensorFlow [4]. This allowed us to leverage the power of a single GPU device for training the model effectively.

To prepare the text data for modeling, we employed tokenization, which involved converting the text into sequences of integer indices. By doing so, we represented each word as a numeric value. Additionally, we determined the vocabulary size based on the training data, which influenced the dimensions of the input representations.

Furthermore, we applied the pad-sequence function to both the training and test sequences. This step involved adjusting the sequences to a fixed length of MAX-LENGTH, which ensured consistent input shape for deep learning models. By padding or truncating the sequences, we facilitated compatibility with models that require inputs of the same shape.

[4]https://www.tensorflow.org/

Following the evaluation of deep learning models, we opted to employ the gbtree booster type from XGBoost [32] to construct an ensemble approach.

Our results are presented in Tables 1-9, they clearly demonstrate the superiority of our proposed model across all experiments conducted. Notably, BiLSTM consistently achieved the best performance in terms of deep learning models. In terms of data balancing techniques, random over-sampling showed better results in most experiments, except for the Spanish-HS and Multi-HS datasets, where unbalanced data performed better.

Surprisingly, the random under-sampling technique yielded superior results specifically for the Spanish-HS dataset. These findings emphasize the effectiveness of the ensemble approach in achieving enhanced classification outcomes. Furthermore, when the dataset exhibits moderate variability in terms of size, employing random over-sampling techniques typically leads to enhanced performance. However, it is important to note that if the dataset size becomes excessively large, there is a risk of overfitting.

We assessed the performance of monolingual and multilingual datasets with varying degrees of balance, including **unbalanced**, **random under-sampling**, and **random over-sampling** datasets. Each experiment was evaluated using accuracy **(ACC)**, F1 score **(F1)**, and Area Under the Curve **(AUC)** as metrics. In Tables 1-9, the models with **bold** values indicate superior performance compared to other models in terms of both model performance and data balancing techniques.

7. Conclusion and Future Work

We considered both monolingual and multilingual datasets and addressed class imbalance through random under-sampling and random over-sampling techniques. Three experiments were performed for each dataset, evaluating the effectiveness of these balancing techniques using unbalanced data.

Our ensemble deep learning approach provided improved accuracy and robustness in addressing the complex challenges of detecting and classifying HS, TR, and AB.

Overall, utilizing random over-sampling techniques generally leads to improved performance when the dataset exhibits moderate variability in size.

In future work, we will focus on further improving the results of our research. One avenue we will explore is the evaluation of pretrained models to as-

Table 1
Results for English-HS

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	0.77	0.73	0.87	0.79	0.79	0.88	**0.84**	**0.85**	**0.93**
BiLSTM	0.69	0.65	0.75	0.68	0.71	0.76	0.7	0.73	0.78
CNN	0.67	0.63	0.73	0.67	0.69	0.73	0.7	0.72	0.77
GRU	0.56	0.49	0.5	0.49	0.66	0.5	0.5	0.66	0.5
LSTM	0.56	0.47	0.49	0.48	0.51	0.5	0.49	0.51	0.48

Table 2
Results for English-TR

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	0.9	0.65	0.94	0.88	0.88	0.96	**0.94**	**0.94**	**0.98**
BiLSTM	0.86	0.51	0.88	0.77	0.78	0.82	0.92	0.92	0.96
CNN	0.85	0.44	0.86	0.76	0.77	0.83	0.92	0.92	0.96
GRU	0.83	0.41	0.5	0.5	0.66	0.5	0.52	0.55	0.51
LSTM	0.83	0.39	0.49	0.5	0.66	0.5	0.5	0.52	0.5

Table 3
Results for English-AB

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	0.85	0.41	0.89	0.82	0.83	0.91	**0.92**	**0.92**	**0.97**
BiLSTM	0.81	0.25	0.75	0.53	0.67	0.58	0.81	0.83	0.9
CNN	0.80	0.16	0.74	0.61	0.69	0.69	0.86	0.87	0.94
GRU	0.81	0.15	0.5	0.48	0.52	0.5	0.5	0.66	0.5
LSTM	0.81	0.15	0.49	0.51	0.68	0.5	0.52	0.58	0.52

Table 4
Results for Spanish-HS

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	**0.87**	**0.84**	**0.94**	0.84	0.84	0.93	0.79	0.8	0.88
BiLSTM	0.94	0.64	0.76	0.49	0.65	0.52	0.5	0.25	0.49
CNN	0.67	0.62	0.73	0.56	0.56	0.58	0.55	0.54	0.58
GRU	0.56	0.53	0.52	0.5	0.66	0.5	0.5	0.67	0.5
LSTM	0.56	0.51	0.5	0.5	0.66	0.52	0.5	0.67	0.5

Table 5
Results for Spanish-TR

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	0.88	0.63	0.94	0.84	0.83	0.94	**0.86**	**0.86**	**0.95**
BiLSTM	0.8	0.25	0.79	0.56	0.67	0.65	0.75	0.77	0.84
CNN	0.8	0.22	0.75	0.56	0.66	0.63	0.8	0.81	0.88
GRU	0.79	0.21	0.5	0.52	0.23	0.5	0.47	0.64	0.5
LSTM	0.79	0.19	0.49	0.52	0.22	0.5	0.47	0.64	0.5

Table 6
Results for Spanish-AB

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	0.84	0.51	0.91	**0.97**	**0.97**	**0.99**	0.8	0.79	0.89
BiLSTM	0.76	0.21	0.57	0.55	0.29	0.59	0.61	0.57	0.64
CNN	0.77	0.14	0.6	0.58	0.56	0.59	0.71	0.72	0.79
GRU	0.77	0.12	0.5	0.49	0.66	0.5	0.5	0.5	0.5
LSTM	0.77	0.12	0.5	0.5	0.5	0.5	0.5	0.49	0.49

Table 7
Results for Multi-HS

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	**0.80**	**0.74**	**0.88**	0.70	0.69	0.77	0.74	0.74	0.82
BiLSTM	0.72	0.64	0.79	0.67	0.68	0.75	0.69	0.73	0.78
CNN	0.7	0.6	0.76	0.49	0.66	0.49	0.50	0.67	0.50
GRU	0.59	0.41	0.56	0.49	0.66	0.50	0.68	0.71	0.76
LSTM	0.59	0.37	0.52	0.65	0.65	0.72	0.67	0.67	0.48

Table 8
Results for Multi-TR

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	0.9	0.65	0.94	0.8	0.8	0.89	**0.94**	**0.94**	**0.98**
BiLSTM	0.86	0.55	0.88	0.78	0.79	0.85	0.92	0.92	0.97
CNN	0.85	0.51	0.86	0.50	0.60	0.50	0.51	0.43	0.49
GRU	0.83	0.46	0.5	0.77	0.78	0.84	0.89	0.89	0.95
LSTM	0.83	0.4	0.49	0.50	0.58	0.48	0.51	0.41	0.51

Table 9
Results for Multi-AB

Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
Proposed model	0.86	0.48	0.89	0.79	0.71	0.81	**0.87**	**0.88**	**0.94**
BiLSTM	0.82	0.34	0.77	0.67	0.66	0.74	0.83	0.84	0.91
CNN	0.82	0.32	0.76	0.66	0.63	0.72	0.50	0.61	0.49
GRU	0.81	0.29	0.5	0.82	0.34	0.77	0.82	0.34	0.77
LSTM	0.81	0.26	0.49	0.51	0.61	0.50	0.50	0.60	0.49

sess their effectiveness in addressing the task of detecting hate speech and aggressive behavior. Pretrained models have shown promising results in various natural language processing tasks, and we believe they can potentially enhance the performance of our models as well.

Additionally, we plan to employ more extensive data augmentation techniques. Data augmentation involves generating additional training samples by applying various transformations to the existing data. This may help in mitigating overfitting and improving the generalization capabilities of our models. By augmenting the dataset with diverse variations, we aim to capture a wider range of linguistic patterns and nuances present in hate speech and aggressive behavior.

Furthermore, we will explore and experiment with different data-balancing techniques. While our research already considered random under-sampling and random over-sampling methods, there are other approaches that can be explored. Techniques such as SMOTE (Synthetic Minority Random Oversampling Technique) [33] and ADASYN (Adaptive Synthetic Sampling) [34] are worth investigating, as they generate synthetic samples to address class imbalance. By employing a combination of different data balancing techniques, we aim to achieve better model performance and robustness across different datasets.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

[1] Nockleby, JT, Levy, LW, Karst, KL, Mahoney, DJ (2000). Encyclopedia of the American constitution. Detroit, MI: Macmillan Reference , 3 (2).

[2] Yigezu, M. G., Kolesnikova, O., Sidorov, G., Gelbukh, A. (2023). Transformer-Based Hate Speech Detection for Multi-Class and Multi-Label Classification.

[3] Balouchzahi, F., Shashirekha, H.L., Sidorov, G. (2021). HSSD: Hate Speech Spreader Detection using N-grams and Voting Classifier. In CLEF (Working Notes) (pp. 1829-1836).

[4] Z. Mossie, J.-H. Wang, Social network HS detection for amharic language, Com- puter Science Information Technology (2018) 41–55

[5] Yigezu, MG, Kanta, S., Kolesnikova, O., Sidorov, G., Gelbukh, A. (2023, September). Habesha@ DravidianLangTech: Abusive Comment Detection using Deep Learning Approach. In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (pp. 244-249).

[6] Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., López, H. M. H. (2021). Internet, social media and online HS. Systematic review. Aggression and Violent Behavior, 58, 101608.

[7] Yigezu, M. G., Mehamed, M. A., Kolesnikova, O., Guge, T. K., Gelbukh, A., Sidorov, G. (2023, October). Evaluating the Effectiveness of Hybrid Features in Fake News Detection on Social Media. In 2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) (pp. 171-175). IEEE.

[8] Saha, K., Chandrasekharan, E., De Choudhury, M. (2019, June). Prevalence and psychological effects of hateful speech in online college communities. In Proceedings of the 10th ACM conference on web science (pp. 255-264).

[9] Essed, P. (2004). Naming the unnameable: sense and sensibilities in researching racism. Researching race and racism, 119-133.

[10] Simon, P., Commission européenne contre le racisme et l'intolérance. (2007). " Ethnic" statistics and data protection in the Council of Europe countries: study report. Strasbourg: Council of Europe.

[11] Yigezu, M. G., Kebede, T., Kolesnikova, O., Sidorov, G., Gelbukh, A. (2023, September). Habesha@ DravidianLangTech: Utilizing Deep and Transfer Learning Approaches for Sentiment Analysis. In Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages (pp. 239-243).

[12] Balouchzahi, F., Gowda, A., Shashirekha, H., Sidorov, G. (2022, May). MUCIC@ TamilNLP-ACL2022: Abusive Comment Detection in Tamil Language using 1D Conv-LSTM. In Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages (pp. 64-69).

[13] Yigezu, M. G., Bade, G. Y., Kolesnikova, O., Sidorov, G., Gelbukh, A. (2023). Multilingual Hope Speech Detection using Machine Learning.

[14] Yigezu, M. G., Tonja, A. L., Kolesnikova, O., Tash, M. S., Sidorov, G., Gelbukh, A. (2022, December). Word Level Language Identification in Code-mixed Kannada-English Texts using Deep Learning Approach. In Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts (pp. 29-33).

[15] Hasan, A., Sharma, T., Khan, A., Hasan Ali Al-Abyadh, M. (2022). Analysing HS against migrants and women through tweets using ensembled deep learning model. Computational Intelligence and Neuroscience, 2022.

[16] Pennington, J., Socher, R., Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[17] González, J. Á., Hurtado, L. F., Pla, F. (2020). Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter. Information Processing Management, 57(4), 102262.

[18] Chiril, P., Benamara, F., Moriceau, V., Coulomb-Gully, M., Kumar, A. (2019, July). Multilingual and multitarget HS detection in tweets. In Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019) (pp. 351-360). ATALA.

[19] Pereira-Kohatsu, J. C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M. (2019). Detecting and monitoring HS in Twitter. Sensors, 19(21), 4654.

[20] Alonso, P., Saini, R., Kovács, G. (2020, September). HS detection using transformer ensembles on the hasoc dataset. In International conference on speech and computer (pp. 13-21). Cham: Springer International Publishing.

[21] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[22] Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., ... Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). arXiv preprint arXiv:2006.07235.

[23] Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., Patel, A. (2019, December). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (pp. 14-17).

[24] Lakshminarayanan, B., Pritzel, A., Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30.

[25] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... Sanguinetti, M. (2019, June). Semeval-2019 task 5: Multilingual detection of HS against immigrants and women in twitter. In Proceedings of the 13th international workshop on semantic evaluation (pp. 54-63).

[26] Fernández, A., Garcia, S., Herrera, F., Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.

[27] Mohammed, R., Rawashdeh, J., Abdullah, M. (2020, April). Machine learning with random over-sampling and random random under-sampling techniques: overview study and experimental results. In 2020 11th international conference on information and communication systems (ICICS) (pp. 243-248). IEEE.

[28] Junsomboon, N., Phienthrakul, T. (2017, February). Combining random over-sampling and random under-sampling techniques for imbalance dataset. In Proceedings of the 9th international conference on machine learning and computing (pp. 243-247).

[29] Arisholm, E., Briand, L. C., Johannessen, E. B. (2010). A systematic and comprehensive investigation of methods to build and evaluate fault prediction models. Journal of Systems and Software, 83(1), 2-17.

[30] Liu, X. Y., Zhou, Z. H. (2013). Ensemble methods for class imbalance learning. Imbalanced learning: Foundations, algorithms, and applications, 61-82.

[31] Rong, X. (2014). word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

[32] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... Zhou, T. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

[33] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: synthetic minority random over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

[34] He, H., Bai, Y., Garcia, E. A., Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). Ieee.