I2C-UHU at HOMO-MEX 2024: Leveraging Large Language Models and Ensembling Transformers to Identify and Classify Hate Messages Towards the LGBTQ+ Community

Javier Román-Pásaro, Alvaro Carrillo-Casado, Jacinto Mata-Vázquez and Victoria Pachón-Álvarez

I2C Research Group ,University of Huelva, Spain

Abstract

This study presents the strategies advanced by the I2C Group to address the IberLEF-2024 Task HOMO-MEX: Hate speech detection in Online Messages directed towards the Mexican Spanish-speaking LGBTQ+ population. The major contribution has been the integration of Large Language Models (LLMs) for classification through prompting, alongside an ensemble of Transformers. By leveraging the advanced capabilities of LLMs for direct classification tasks, significant improvements in performance were achieved. The ensemble approach, which combines multiple models, further enhanced the results by leveraging the individual strengths of each model.

The experiments highlighted the importance of selecting appropriate hyperparameters during the model training process. Through meticulous experimentation and evaluation of different hyperparameter combinations, the optimal settings for achieving the best performance were identified. In the experiments for Task 1, several models were tested, and multiple ensemblers were created. The first ensembler combined Transformers, and its result was further ensembled with two LLMs, obtaining the best F1-Score for this dataset. The model submitted for Task 1 achieved an F1-Score of 87.64%, ranking in 3rd place in the competition.

Keywords

Hate speech detection, Large Language Models, Transformers, Ensembling methods, Prompting, LGBTQ+ community, IberLEF-2024, Text classification

1. Introduction

In today's digital landscape, the field of Natural Language Processing [1] (NLP) serves as a cornerstone in comprehendi0ng and dissecting the vast amounts of information continually produced by social media platforms. The ability to extract meaningful insights from textual data is crucial across various domains, ranging from social research to political decision-making, and particularly in identifying and addressing social issues. Within this framework, the identification of prejudiced remarks directed towards the LGBTQ+ community has emerged as a pressing concern, emphasizing the need to foster online environments that champion inclusion, respect, and equality.

This paper describes an effort to develop a robust system tailored for pinpointing biased comments targeting the LGBTQ+ community [2], employing advanced natural language processing techniques within the scope of the task defined by the IberLEF 2024 [3] initiative, HOMO-MEX: Hate speech detection towards the Mexican Spanish-speaking LGBTQ+ population [4]. Building upon previous participation and learnings from the last year [5], this task is revisited by integrating the use of Large Language Models [6] (LLMs), a more advanced and contemporary technology, into the methodology. Recognizing the significant advancements and widespread adoption of Transformer models [7], the focus is on harnessing their integration and leveraging the benefits provided by LLMs. These advanced architectures have demonstrated significant potential for understanding and processing natural language in diverse contexts.

The approach capitalizes on these advantages by developing and training models based on LLMs and Transformers to address hate speech detection towards the LGBTQ+ community. Additionally, the

IberLEF 2024, September 2024, Valladolid, Spain

javier.roman780@alu.uhu.es (J. Román-Pásaro); alvaro.carrillo121@alu.uhu.es (A. Carrillo-Casado); mata@uhu.es (J. Mata-Vázquez); vpachon@dti.uhu.es (V. Pachón-Álvarez)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

technique of prompting [8] is incorporated, which guides the model with specific cues or questions to enhance its ability to detect subtle nuances in hate speech. This strategy has significantly enhanced the effectiveness of the classifiers, focusing on the accuracy and sensitivity necessary to identify harmful comments with greater precision and comprehension.

2. Related Works

In recent years, the field of Natural Language Processing (NLP) has witnessed significant advancements, particularly with the advent of Transformer models. Models such as BERT, GPT, and RoBERTa have revolutionized the way text is processed and understood, enabling the tackling of complex linguistic tasks with unprecedented accuracy. One crucial application of NLP technology is the detection of hate messages and discriminatory content, particularly those targeting marginalized communities like the LGBTQ+ community.

Several recent investigations have focused on leveraging Transformer models for detecting hate messages against the LGBTQ+ community. Fortuna and Nunes (2018) provide a comprehensive survey on automatic detection of hate speech in text, highlighting the challenges and approaches in the field [9]. Their work underscores the importance of developing robust models to address the complexity of hate speech, including its various forms and targets.

Chhaya et al. (2023) specifically address the need for annotated datasets in hate speech detection by introducing SamPar, a Marathi hate speech dataset focused on homophobia and transphobia [10]. They demonstrate that fine-tuning pre-trained Transformer models on such annotated datasets significantly improves performance in detecting hate messages against the LGBTQ+ community. By leveraging the contextualized representations learned by Transformer models, it is possible to capture the subtle nuances and linguistic patterns indicative of hate speech.

Zaken et al. (2021) explore the effectiveness of parameter-efficient fine-tuning techniques for Transformer-based models [11]. Their findings indicate that targeted fine-tuning, even with simple methods like Bitfit, can enhance the models' ability to accurately identify and categorize discriminatory content. This research showcases the potential of Transformer-based models in capturing the intricate linguistic characteristics of hate speech, allowing for more effective moderation of online platforms, the protection of vulnerable communities, and the promotion of a safer and more inclusive digital environment.

Furthermore, the rise of Large Language Models (LLMs) has expanded the capabilities of NLP in detecting hate speech and discriminatory content. Shi et al. (2023) conduct a comparative analysis of BERT and LLM-based approaches for multivariate hate speech detection on Twitter, demonstrating the superior performance of these models [12]. They highlight the impressive capabilities of LLMs, such as GPT-3 and Falcon, in understanding and generating natural language, making them valuable tools in identifying subtle nuances and linguistic patterns indicative of hate speech. Incorporating LLMs into hate speech detection systems has the potential to enhance accuracy and sensitivity, thereby contributing to the creation of safer online spaces for marginalized communities like the LGBTQ+ community.

These investigations collectively underscore the advancements in Transformer and LLM-based models for detecting hate speech, emphasizing the importance of annotated datasets, fine-tuning techniques, and the powerful capabilities of modern NLP models in addressing complex linguistic tasks.

3. Task and Dataset Description

The task undertaken in this research involves predicting the label of each individual tweet concerning the LGBT+ community. Specifically, the task requires categorizing tweets into one of three possible labels: LGBT+phobic, not LGBT+phobic, and not LGBT+related. The definitions of these labels are as follows:

- **LGBT+phobic (P):** Tweets that contain hate speech directed at individuals whose sexual orientation and/or gender identity differs from cis-heterosexuality.
- **Not LGBT+phobic (NP):** Tweets that do not contain hate speech against the LGBT+ community but do mention this community.
- **Not LGBT+related (NR):** Tweets that do not relate to the LGBT+ community in any way.

The Corpus provided by the organizers is described at Codabench (https://www.codabench.org/competitions/2229/). The training dataset provided by the organizers consists of 7000 tweets. To effectively train and evaluate our models, the dataset has been divided into three subsets in a stratified manner: 4480 tweets (64%) for training, 1120 tweets (16%) for validation, and 1400 tweets (20%) for testing. This stratification ensured that the class distribution was consistent across the training, validation, and test sets. Each entry in the dataset includes the tweet content and the corresponding label.

The percentages were achieved by initially performing an 80/20 division of the dataset into training and testing subsets. Subsequently, the training subset was further divided into training and validation subsets using another 80/20 split. This stratified division ensured that the class distribution remained consistent across the training, validation, and test sets. Each entry in the dataset comprises the tweet content along with its corresponding label.

The class distribution for each subset is shown in Table 1.

Table 1
Class distribution

Class	Train Dataset	Valid Dataset	Test Dataset
P	552	138	172
NP	2790	698	872
NR	1138	284	356
Total	**4480**	**1120**	**1400**

Table 2 provides examples of tweets from each class along with their labels. Each tweet is presented in its original language, followed by a translated version in parentheses.

Table 2
Examples of instances

Content	Label
Como que últimamente ando muy intolerante a la mirada lasciva de los pinches jotos #ARomperHocicos (Lately I've been very intolerant to the leering eyes of the fucking jotos #ToBreakSnouts.)	P
Si eres Bi vas a los dos baños. Si eres pan vas mitad a uno mitad a otro #LoveWins url (If you're Bi you go to both bathrooms. If you're bread you go half to one half to the other #LoveWins url)	NP
La mejor vestida de los #Emmysna0na1 sin lugar a dudas es #AnyaTaylorJoy url (The best dressed of the #Emmysna0na1 without a doubt is #AnyaTaylorJoy url.)	NR

4. Methodology

This section details the methodology employed in this study, consisting of several key steps. First, due to the diversity in users' writing styles and the multitude of writing possibilities, text normalization and preprocessing [13] were performed. Next, given the data imbalance regarding the phobic class, a data augmentation approach based on backtranslation [14] was utilized. Subsequently, an exhaustive search for hyperparameters [15] was conducted to identify the optimal training parameters for the Transformer models specific to this task. Finally, a classification model was developed using an ensemble approach [16], combining the top three Transformer models and ensembling that result with two LLMs, employing a hard voting approach to enhance performance.

Figure 1 illustrates the comprehensive methodology pipeline, encompassing initial text preprocessing, data augmentation via backtranslation, exhaustive hyperparameter search, and the final ensemble text classification approach.

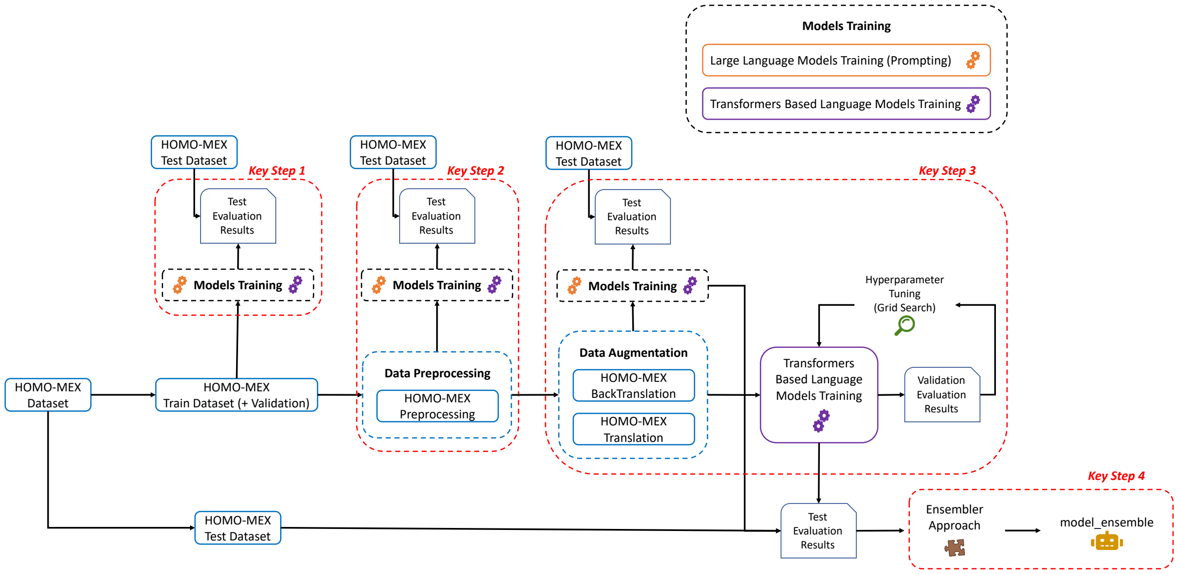

Figure 1: Methodology Pipeline

Since the datasets are in the Spanish language, pretrained models in this language were primarily used. However, considering that Mexican Latin American Spanish includes a significant amount of Anglo-Saxon vocabulary, some multilingual models were also selected to explore alternative options. The selected pretrained models, obtained from the Hugging Face library (https://huggingface.co/), were as follows (the model alias to be used throughout the paper is indicated in parentheses):

- **Transformers:**
  - `dccuchile/bert-base-spanish-wwm-uncased` [17] (**BETO**): This model is a Spanish version of BERT, adapted to understand and process Spanish text through pretraining on a large Spanish corpus.
  - `microsoft/mdeberta-v3-base` [18] [19] (**mDeBERTa**): A version of DeBERTa that supports multiple languages, including Spanish, and is designed to enhance context understanding and semantic relationships in text.
  - `PlanTL-GOB-ES/roberta-base-bne` [20] (**RoBERTa**): Based on the RoBERTa base model, this model has been pretrained using the largest known Spanish corpus to date, optimizing its ability to handle Spanish text.
  - `xlm-roberta-large` [21] (**XLM**): A multilingual model based on RoBERTa, capable of processing and understanding text in multiple languages, including Spanish, making it ideal for handling linguistic variations and Anglo-Saxon vocabulary present in Latin American Spanish.
- **LLMs:**

- `tiiuae/falcon-7b` [22] [23] (**Falcon**): A large language model renowned for its ability to handle complex NLP tasks. Due to its capacity, two variants of this model were created: one trained with data in the original language (Falcon) and another with data translated into English (FalconENG).

All models were trained on the original language of the tweets, except for the LLM Falcon model, for which the two aforementioned variants were generated to evaluate its performance with data in different languages [24].

To utilize the large language models (LLMs) for classification tasks, a technique known as "prompting" was employed. This method involves providing a structured prompt to the LLMs to guide their responses. For this study, distinct prompts were designed in both Spanish and English to match the language of the input data (see Table 3).

Table 3
Prompting Examples

Language	Content
Spanish	[INST]Analiza la relación del tweet encerrado entre corchetes con la comunidad LGBTQ+, determina si es fóbico (P), no fóbico (NP) o no relacionado (NR), y devuelve la respuesta como únicamente la etiqueta de sentimiento correspondiente "fóbico" o "no fóbico" o "no relacionado". [/INST] [Barak Obama es gay y su esposa Michelle es transgénero dijo Joan Rivers la famosa conductora de TV en Estados Unidos. Que tal?] =
English	[INST] Analyze the relation of the tweet enclosed in square brackets with the LGBTQ+ community, determine if it is phobic, non-phobic, or non-related, and return the answer just as the corresponding sentiment label "phobic" or "non-phobic" or "non-related" [/INST] [Barak Obama is gay and his wife Michelle is transgender said Joan Rivers the famous TV driver in the United States. How about that?] =

These prompts were designed to instruct the model to classify the sentiment of the tweet in relation to the LGBTQ+ community. The models were prompted to return one of three possible labels: "phobic," "non-phobic," or "non-related," thereby providing a clear and structured approach for sentiment analysis.

Regarding performance, it is worth mentioning that all models are trained using an NVIDIA GeForce RTX 4090 graphics card with 24 GB of RAM. This powerful hardware setup allows for efficient training of the models, ensuring that the large datasets and complex computations required for fine-tuning the pre-trained language models are handled effectively.

To compare the results obtained by the different models and developed strategies, a baseline using the selected pre-trained models was established. Given the impossibility of knowing the optimal hyperparameter values beforehand, commonly used values were employed for fine-tuning the pre-trained language models: a batch size of 32, a learning rate of 5e-5, a maximum length of 128, a weight decay of 0.001, and the adamw_torch optimizer [25].

The results, as summarized in Table 4, provide insights into the performance of each model under the baseline conditions.

4.1. Data Preprocessing

The data preprocessing involved several steps. Firstly, all text was converted to lowercase to ensure consistency. Secondly, hyperlinks and emoticons were removed from the text. Thirdly, usernames were replaced with a fixed word "@user". Additionally, a synonym dictionary was created by analyzing the frequency of the most common insults (https://es.wiktionary.org/wiki/Wikcionario:homosexual/Tesauro) in which specific Mexican Spanish insults were substituted with more common alternatives that retained the same meaning but fit better within the vocabulary of the pretrained models.

Table 4
Baseline results

Model	F1-Score (Macro Average)
BETO	0.800386
mDeBERTa	0.802570
RoBERTa	0.799538
XLM	0.792636
Falcon	0.779674
FalconENG	0.765180

4.2. Data Augmentation and Hyperparameter Search

To address the class imbalance [26] within the dataset, a data augmentation approach based on back-translation was employed. This method aimed to increase the instances of the P (Phobic) class by duplicating the number of phobic instances. The augmentation process involved translating the Spanish text to English, followed by translation from English to German, and then back to Spanish. This multi-step translation process was facilitated using pretrained models such as "Helsinki-NLP/opus-mt-es-en", "Helsinki-NLP/opus-mt-en-de", and "Helsinki-NLP/opus-mt-de-es" [27].

The hyperparameter search is a crucial step for fine-tuning Transformer models. Therefore, multiple iterations of training and evaluation were conducted using different combinations of key hyperparameters. To mitigate training time costs, the datasets were proportionally reduced before experimentation. Specifically, this experimentation was carried out using 80% of the original dataset size. The search for optimal hyperparameters was facilitated using the Optuna [28] library for Python. Optuna is an open-source hyperparameter optimization framework that automates the process of finding the best hyperparameters for machine learning models.

By combining hyperparameter search with data augmentation [29], significant improvements in model performance were observed. This synergy arises from the fact that data augmentation expands the diversity of the training data, enabling the model to learn more robust and generalizable representations. Moreover, the optimized hyperparameters fine-tune the model's architecture and training process, further enhancing its ability to capture intricate patterns and nuances within the data.

In the hyperparameter search process, the search space explored is shown in Table 5. The number of epochs was set to 10 due to the implementation of early stopping to prevent overfitting. In addition, looking at the number of tokens in each tweet, it was decided to set the maximum length value to 128.

Table 5
Hyperparameter Space

Hyperparameter	Values
Batch Size	[8, 16, 32]
Learning Rate	[2e-5, 3e-5, 5e-5]
Weight Decay	[0.001, 0.01, 0.1]
Optimizer	[adamw_hf, adamw_torch, adafactor]

In addition, the best hyperparameters for each model were identified, as shown in Table 6. It is notable that the optimal hyperparameter values are similar among the different models, suggesting that certain configurations are effective overall in improving model performance on this task.

The improved results from these steps are shown in Table 7.

4.3. Ensemble Approach

Two ensemblers were developed to address the task of hate speech detection. The first ensembler combines the three best Transformer models on the test set: mDeBERTa, RoBERTa, and XLM. In this

Table 6
Best Hyperparameters per Model

Hyperparameter	BETO	RoBERTa	mDeBERTa	XLM
Batch Size	16	16	16	16
Learning Rate	5e-5	3e-5	3e-5	3e-5
Weight Decay	0.01	0.01	0.01	0.01
Optimizer	adamw_hf	adamw_hf	adamw_hf	adamw_hf

Table 7
Results after Data Augmentation and Hyperparameter Search

Model	F1-Score (Macro Average)
BETO	0.802733
mDeBERTa	0.817852
RoBERTa	0.817342
XLM	0.816338
Falcon	0.784406
FalconENG	0.770541

ensemble, mDeBERTa acts as the tiebreaker due to its highest individual F1-Score [30]. Given that this is a hard-voting [31] multiclass classification task, a tie can occur when each model selects a different class. In such cases, the final decision is based on the prediction from mDeBERTa because of its superior performance.

The result from this first ensembler is then subjected to a second round of voting against two additional models, Falcon. These Falcon models provide additional robustness to the overall solution. In this second voting stage, the LLM model acts as the tiebreaker, adding an extra layer of accuracy and reliability to the classification process.

This two-stage ensemble approach [32], first combining the best Transformers and then integrating LLM models, leverages the individual strengths of each model and enhances the overall system's capability to accurately identify hate speech directed towards the LGBTQ+ community.

The results of the ensemble methods are summarized in Table 8.

Table 8
Results after Ensemble Approach

Model	F1-Score (Macro Average)
Transformer Ensembler	0.817852
Final Ensembler	0.818974

5. Results

The competition results are summarized in Table 9. The table displays the username, F1-score, precision, recall, and place for each participant.

The ensemble approach used in this study demonstrated robust performance, resulting in a third-place finish with an F1-score of 0.876497. The precision achieved was 0.909812, while the recall was 0.853129.

This performance highlights the effectiveness of the implemented techniques, despite the highly competitive nature of the task. The ensemble method's ability to combine multiple models and leverage their strengths proved beneficial in accurately identifying hate speech directed towards the LGBTQ+ community.

Table 9
Competition Results

Username	F1-Score	Precision	Recall	Place
verbanex	0.914326	0.936409	0.896260	1
atoro491	0.914326	0.936409	0.896260	1
rogerd97	0.914326	0.936409	0.896260	1
quanle709	0.877537	0.929059	0.847655	2
i2chuelva	**0.876497**	**0.909812**	**0.853129**	**3**
sdamians	0.871312	0.919452	0.840588	4
metztli	0.856299	0.086978	0.845787	5

6. Error Analysis

Figure 2 illustrates the classifier's performance when predicting classes NP (Not Phobic) and NR (Not Related) in Task 1. However, the classifier is less reliable at predicting class P (Phobic). This discrepancy may be attributed to the significant imbalance in the training dataset, where the phobic class has the lowest representation.

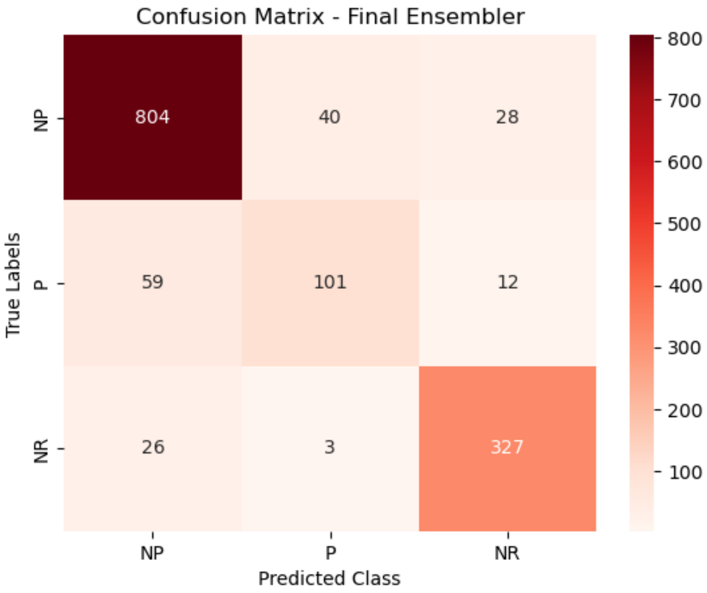

Figure 2: Confusion Matrix for Task 1

Despite achieving good results in Task 1, occasional errors in prediction were observed. Table 10 presents some instances where the model made incorrect predictions. These errors can be attributed to the limited context and similar vocabulary in the content, which can lead to confusion.

The analysis reveals that while the classifier performs well for certain classes, improvements are needed for better handling of the phobic class. Addressing the class imbalance in the training data could be a potential solution to enhance the model's performance in future iterations.

In the first tweet, the model likely flagged it as LGBT+phobic due to the presence of potentially offensive terms like "marica" and hashtags such as "#ridicula" and "#noaguntas," although the context might not be hateful. This suggests a lack of nuanced understanding of colloquial usage. In the second tweet, the model failed to recognize the derogatory connotation of "jotita," leading to a misclassification as Not LGBT+phobic, highlighting the model's difficulty with culturally specific slang and subtle negative comparisons. The third tweet was misclassified as Not LGBT+related possibly due to its ambiguous language and indirect reference to transgender individuals, indicating the model's struggle with subtle or euphemistic expressions. Enhancing contextual understanding, addressing class imbalance, incorpo-

Table 10
Examples of model errors for Task 1

Content	Label	Prediction
Y estaba de nuevo en él cumbres y me puse a llorar del dolor tan grande neta #ridicula #marica #noaguntas (And I was back on the summit and I started to cry from the pain so big #ridicula #marica #noaguntas)	NP	P
para qué ser una jotita discreta si puedes ser hetero-moderna. (why be a discreet little girl if you can be hetero-modern.)	P	NP
Ya parecen transformers esas madres, y ciertamente soy pobre url. (They already look like transformers, and I'm certainly poor url.)	P	NR

rating cultural and linguistic nuances, and implementing human-in-the-loop review could improve the model's performance.

7. Conclusion

This paper presents an approach for the IberLEF-2024 Task HOMO-MEX, focusing on the detection of hate speech in online messages directed towards the Mexican Spanish-speaking LGBTQ+ community. The primary contributions include leveraging Large Language Models (LLMs) for classification through prompting and employing an ensemble of Transformer models to enhance performance.

The integration of LLMs provided a substantial boost in performance, as these models capture complex linguistic patterns and nuances essential for accurate hate speech detection. Furthermore, the ensemble approach, combining the strengths of multiple models, proved to be highly effective. By using Transformers such as BETO, RoBERTa, mDeBERTa, and XLM, and incorporating LLMs like Falcon, a significant improvement in classification accuracy was achieved.

The experiments underscored the importance of hyperparameter tuning and data preprocessing. Through meticulous experimentation and evaluation of different hyperparameter combinations, the optimal settings that led to the best performance were identified. The use of data augmentation techniques, particularly backtranslation, addressed the class imbalance issue and contributed to the robustness of the models.

The results of this approach are promising. In Task 1, the model achieved an F1-Score of 87.64%, securing 3rd place in the competition. This outcome demonstrates the efficacy of the methods and their potential for real-world applications in detecting hate speech.

Future work could explore further enhancements, such as incorporating more sophisticated data augmentation techniques, experimenting with additional LLMs, and exploring different ensembling strategies. Additionally, extending the approach to other languages and cultural contexts could further validate the generalizability and effectiveness of the methodology. It may also be interesting to study the engineering of prompting and how it affects the model's behavior, as well as to investigate new techniques as they emerge.

In conclusion, this study highlights the power of combining advanced NLP techniques, such as LLMs and Transformer-based ensembling, in tackling the critical issue of hate speech detection. The insights gained from this research contribute to the broader field of NLP and offer practical solutions for fostering safer and more inclusive online environments.

Acknowledgments

This paper is part of the I+D+i Project titled *"Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people*

[NONCONSPIRA-HATE!]", PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF A way of making Europe".

This paper is part of the I+D+i Project titled *"Conspiracy Theories and hate speech online: Comparison of patterns in narratives and social networks about COVID-19, immigrants, refugees and LGBTI people [NON-CONSPIRA-HATE!]", PID2021-123983OB-I00, funded by MCIN/AEI/10.13039/501100011033/ and by "ERDF/EU".*

References

[1] P. M. Nadkarni, L. Ohno-Machado, W. W. Chapman, Natural language processing: an introduction, Journal of the American Medical Informatics Association 18 (2011) 544–551.

[2] V. G. Aravindh, V. Kirubanand, M. F. Mirza, Triangulation study on lgbtq inclusion with sustainable development goal 10 using twitter data and topic modelling, in: Interdisciplinary Perspectives on Sustainable Development, CRC Press, 2023, pp. 131–135.

[3] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.

[4] H. Gómez-Adorno, G. Bel-Enguix, G. Sierra, S.-T. Andersen, S. Ojeda-Trueba, T. Alcántara, M. Soto, C. Macias, H. Calvo, Overview of homo-mex at iberlef 2024: Homo-mex: Hate speech detection in online messages directed towards the mexican spanish speaking lgbtq+ population, Procesamiento del Lenguaje Natural 73 (2024).

[5] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S.-T. Andersen, S. Ojeda-Trueba, Overview of homo-mex at iberlef 2023: Homo-mex: Hate speech detection in online messages directed towards the mexican spanish speaking lgbtq+ population, Procesamiento del Lenguaje Natural 71 (2023).

[6] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, A. Mian, A comprehensive overview of large language models, arXiv preprint arXiv:2307.06435 (2023).

[7] A. Gillioz, J. Casas, E. Mugellini, O. Abou Khaled, Overview of the transformer-based models for nlp tasks, in: 2020 15th Conference on Computer Science and Information Systems (FedCSIS), IEEE, 2020, pp. 179–183.

[8] S. Arora, A. Narayan, M. F. Chen, L. Orr, N. Guha, K. Bhatia, I. Chami, C. Re, Ask me anything: A simple strategy for prompting language models, in: The Eleventh International Conference on Learning Representations, 2022.

[9] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.

[10] B. Chhaya, P. K. Kumaresan, R. Ponnusamy, B. R. Chakravarthi, Sampar: A marathi hate speech dataset for homophobia, transphobia, in: International Conference on Speech and Language Technologies for Low-resource Languages, Springer, 2023, pp. 34–51.

[11] E. B. Zaken, S. Ravfogel, Y. Goldberg, Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, arXiv preprint arXiv:2106.10199 (2021).

[12] X. Shi, J. Liu, Y. Song, Bert and llm-based multivariate hate speech detection on twitter: Comparative analysis and superior performance, in: International Artificial Intelligence Conference, Springer, 2023, pp. 85–97.

[13] C. P. Chai, Comparison of text preprocessing methods, Natural Language Engineering 29 (2023) 509–553.

[14] D. R. Beddiar, M. S. Jahan, M. Oussalah, Data expansion using back translation and paraphrasing for hate speech detection, Online Social Networks and Media 24 (2021) 100153.

[15] M. Feurer, F. Hutter, Hyperparameter optimization, Automated machine learning: Methods, systems, challenges (2019) 3–33.

[16] G. Brown, Ensemble learning., Encyclopedia of machine learning 312 (2010) 15–19.

[17] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[18] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[19] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=XPZIaotutsD.

[20] A. G. Fandiño, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[21] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, CoRR abs/1911.02116 (2019). URL: http://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[22] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic, B. Noune, B. Pannier, G. Penedo, Falcon-40B: an open large language model with state-of-the-art performance (2023).

[23] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, J. Launay, The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only, arXiv preprint arXiv:2306.01116 (2023). URL: https://arxiv.org/abs/2306.01116. arXiv:2306.01116.

[24] Z. Li, Y. Shi, Z. Liu, F. Yang, N. Liu, M. Du, Quantifying multilingual performance of large language models across languages, arXiv preprint arXiv:2404.11553 (2024).

[25] R. Llugsi, S. El Yacoubi, A. Fontaine, P. Lupera, Comparison between adam, adamax and adam w optimizers to implement a weather forecast based on neural networks for the andean city of quito, in: 2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM), IEEE, 2021, pp. 1–6.

[26] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent data analysis 6 (2002) 429–449.

[27] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the World, in: Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT), Lisbon, Portugal, 2020.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2623–2631.

[29] D. Wagner, F. Ferreira, D. Stoll, R. T. Schirrmeister, S. Müller, F. Hutter, On the importance of hyperparameters and data augmentation for self-supervised learning, arXiv preprint arXiv:2207.07875 (2022).

[30] G. Abramowitz, Model independence in multi-model ensemble prediction, Australian Meteorological and Oceanographic Journal 59 (2010) 3–6.

[31] M. A. Fauzi, A. Yuniarti, et al., Ensemble method for indonesian twitter hate speech detection, Indonesian Journal of Electrical Engineering and Computer Science 11 (2018) 294–299.

[32] M. S. U. Miah, M. M. Kabir, T. B. Sarwar, M. Safran, S. Alfarhood, M. Mridha, A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and llm, Scientific Reports 14 (2024) 9603.