VerbaNexAI Lab at HOMO-MEX 2024: Multiclass and Multilabel Detection of LGBTQ+ Phobic Content Using Transformers

Roger David Gonzalez-Henao[1,*,†], Duvan Andres Marrugo-Tobon[1,†], Juan Carlos Martinez-Santos[1,†] and Edwin Puertas[1,*,†]

[1]*Universidad Tecnológica de Bolívar, Faculty of Engineering, Cartagena de Indias 17013001, Colombia*

Abstract

In natural language processing, accurate categorization of tweets, including detecting hate speech, is crucial for efficient information organization and analysis. This paper presents a Natural Language Contents Evaluation System designed explicitly for multi-class tweet categorization, focusing on hate speech detection. Our system leverages the power of Transformers, namely BERT, to enhance classification accuracy and efficiency. We capture essential tweet information using feature extraction techniques, enabling practical analysis and categorization. We employ techniques ensuring fair representation of different tweet categories to address imbalanced corpora. Our system demonstrated impressive performance in the HOMO-Mex 2024 competition: first place in Track 1 with a 91% F1 score, third place in Track 2 with a 93% F1 score, and second place in Track 3 with a 56% F1 score. These results highlight the robustness and generalizability of our trained models for hate speech detection. This system contributes to advancing automated tweet categorization, providing a reliable and efficient solution for organizing and analyzing diverse tweet datasets.

Keywords

Natural Language Processing, Hate Speech Detection, Tweet Categorization, BERT, Imbalanced Data,

1. Introduction

Social media platforms like Twitter have revolutionized communication in the digital age, allowing rapid information sharing and global real-time interaction. The evolution of the internet led to social media usage, reviewing sites, and many more platforms. The online communication of social media platforms has increased exponentially in overall languages globally. This stage permits users to post and share content and express their views in consideration of anything at any time [1].

However, this progress comes with challenges, notably the rise of hate speech. Hate speech on the internet is on the rise around the world, with approximately 60% of the global population (4.54 billion) using social media to communicate. According to studies, approximately 53% of Americans have encountered online harassment and hatred. This score is 12 points higher than the findings of a similar survey performed in 2017 [2]. Hate speech often uses abbreviations and familiar words to mask malicious intent, making detection harder [2]. Moreover, the detection of online hate speech has some challenges also. One issue is that the definition of hate speech can be arbitrary and different depending on the culture. Another issue is that hate speech can be covered in various ways, such as using code words or symbols [3].

IberLEF 2024, September 2024, Valladolid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

rhenao@utb.edu.co (R. D. Gonzalez-Henao); marrugod@utb.edu.co (D. A. Marrugo-Tobon); jcmartinezs@utb.edu.co (J. C. Martinez-Santos); epuerta@utb.edu.co (E. Puertas)

https://github.com/Rogerd97 (R. D. Gonzalez-Henao); https://github.com/DuvanMarrugo14 (D. A. Marrugo-Tobon); https://www.utb.edu.co/profesores/juan-carlos-martinez-santos-2/ (J. C. Martinez-Santos); https://github.com/EdwinPuertas (E. Puertas)

0000-0002-0877-7063 (R. D. Gonzalez-Henao); 0009-0002-9537-0209 (D. A. Marrugo-Tobon); 0000-0003-2755-0718 (J. C. Martinez-Santos); 0000-0002-0758-1851 (E. Puertas)

Natural language processing (NLP) techniques are crucial in addressing this issue, thanks to the availability of large text datasets and the need for sophisticated human-computer interactions [4]. Advances in NLP have led to models like BERT, RoBERTa, and DistilBERT, which excel at capturing contextual and semantic nuances. RoBERTa is an enhancement of BERT, which is trained on a bigger dataset to improve performance [5]. LSTM networks are also valuable for understanding complex language patterns in hate speech [6].

Transformers have significantly improved tweet classification accuracy and efficiency. It's essential to use feature extraction techniques to enhance tweet categorization systems further that capture relevant tweet information [7]. Various methods, such as character n-grams [8], Word2Vec embeddings [9], and CNNs [10], help models grasp tweet nuances and improve classification.

Addressing imbalanced tweet datasets is crucial for fair category representation during training [11]. Strategies like oversampling and undersampling create balanced category distributions, ensuring the model learns from all classes equally [12]. Advanced methods like SMOTE [13] and ADASYN [14] generate synthetic samples or adjust sampling rates for better balance.

By integrating Transformers, effective feature extraction, and corpus balancing techniques [15], we aim to improve tweet classification accuracy and robustness, enabling precise information retrieval and analysis in social media [16]. This comprehensive framework addresses the challenges of imbalanced datasets and enhances tweet classification systems' performance.

This paper examines the effectiveness of BERT in detecting various types of hate speech in tweets, explicitly focusing on LGBT+phobic content. The contributions of this work include the use of advanced Transformer models for tweet classification, the application of feature extraction techniques to improve classification accuracy, and a thorough evaluation using real-world datasets. We organized the structure of this paper as follows: Initially, it explains the methodology used for the classification models, starting with the description of the dataset and concluding with the representation of the Transformer. Subsequently, it addresses the experimental validation and discusses the results. Finally, the document ends with a summary of the findings and potential directions for future research.

2. Methodology

2.1. Dataset Description

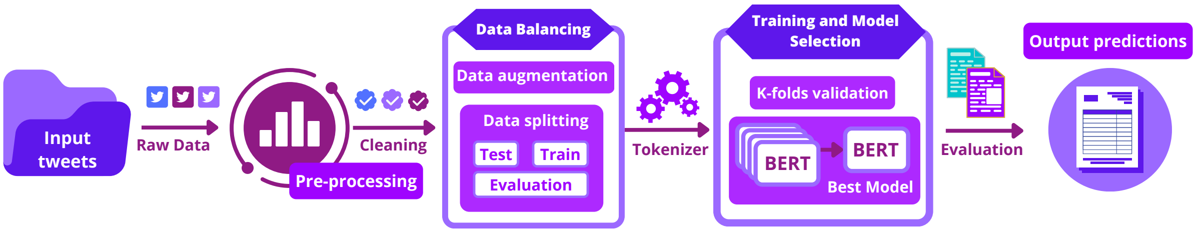


Figure 1: Transformative framework for tweet classification.

HOMO-Mex 2024, a competition belonging to IberLEF 2024 [17], organized by the Grupo de Ingeniería Lingüística at the Universidad Nacional Autónoma de México, leverages meticulously curated datasets aimed at detecting and categorizing LGBT+phobic content across diverse platforms, including social media and music. The Hate Speech Detection track specifically focuses on classifying tweets in Mexican Spanish for LGBT+phobic content, utilizing a dataset comprising approximately 12,000 tweets sourced from 2012 to 2022. It provides a decade-long temporal span for robust analysis [18][19].

The competition consists of three tracks. Track 1 is a multi-class classification challenge in which the models must classify each tweet as LGBT+phobic (P), non-LGBT+phobic (NP), or non-LGBT+ related (NR). In addition, the challenge extends the shared task to detailed hate speech detection in Track 2, which includes about 5,000 instances labeled with specific phobia types such as Gayphobia (G), Lesbophobia (L), Transphobia (T), Biphobia (B), Aphobia (A), and unrelated content (NR), also split 80/20 for training and testing. Track 3 dives into the music domain, using a dataset of approximately 1,000 song lyrics classified as LGBT+phobic (P) or non-LGBT+phobic (NP), collected using web scraping techniques. A detailed description of the data set in terms of label and amount of data per label can be seen in Table 1. All tracks have an unbalanced data distribution per class or label, highlighting the need for meticulous model training to address these imbalances.

Table 1
Class Distribution in HOMO-Mex 2024 Shared Task

Track	Label	Quantity
Track 1	LGBT+phobic (P)	1089
	Not LGBT+phobic (NP)	5765
	Not LGBT+related (NR)	2328
Track 2	Lesbophobia (L)	72
	Gayphobia (G)	714
	Biphobia (B)	10
	Transphobia (T)	79
	Other LGBT+phobia (O)	64
	Not LGBT+related (NR)	0
Track 3	LGBT+phobic (P)	39
	Not LGBT+phobic (NP)	904

2.2. Classification Process

Figure 1 illustrates the process for obtaining the optimal model for detecting phobic hate in tweets. The approach begins with the design of a pipeline that addresses the three tasks outlined in Section 2.1. The primary variation lies in the model's inputs and the number of outputs, dependent on the defined task. Note that although we utilized the same structural framework, the internal code configuration is modified to fulfill these specific requirements. The dataset is read and stored in a DataFrame, with labels for n columns varying according to the task. We conducted data preprocessing and cleaning alongside correcting class imbalances relevant to the task. Furthermore, we performed an exploratory analysis and a split of data for training, validation, and testing, followed by training through cross-validation to estimate the most effective classification model.

Feature extraction, a critical step in the model training process, represents phrases or documents by assigning a probability of occurrence to words. This step is instrumental in determining the model's ability to understand the data. We trained the system using 80% of the data, with the remaining 20% allocated for verification—10% for validation and 10% for post-training testing.

Data Cleaning

Data cleaning of tweets involves several steps to preprocess the text and remove unnecessary or irrelevant information, as shown in Figure 2. The process includes removing special characters, punctuation marks, URLs, and mentions and converting the text into individual words. We removed empty words to reduce noise, and hashtags and emoticons were treated depending on the objectives of the analysis. In addition, text normalization is performed by converting the text to lowercase letters, processing abbreviations and contractions, and removing or substituting numbers.

URL and Metadata Removal

The initial phase of preprocessing involves removing URLs and retweet tags from the tweets. At the same time, user mentions (@) were retained as they were useful in identifying instances of hate speech.

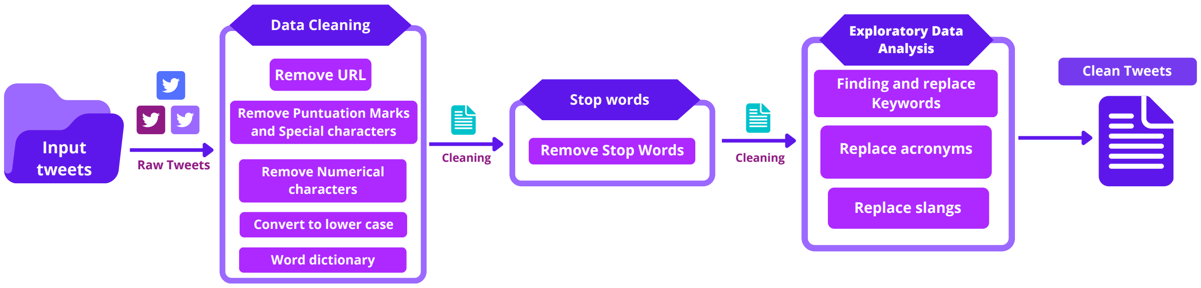


Figure 2: Data cleaning pipeline.

This selective removal is essential to eliminating irrelevant external references and metadata that could negatively influence the analysis outcomes.

Handling Informal Language

Social media platforms often feature informal language use, including leetspeak and slang. The preprocessing script addresses this by translating leetspeak to standard text and replacing Spanish slang terms with their standard equivalents, using an extensive predefined dictionary. This step is critical for maintaining the semantic integrity of the dataset.

- **Example:** "q tal? estoy bn, tqm!" transforms to "que tal? estoy bien, te quiero mucho!"

Offensive Language Neutralization

To make the data suitable for broader audience analysis and to mitigate bias, the script replaces offensive terms within the tweets with neutral equivalents. This neutralization process employs a predefined dictionary that systematically substitutes sensitive words with more general terms.

- **Example:** "Eres un joto" changes to "Eres un gay"

Stopword Removal and Lexical Filtering

As part of refining the text, we removed stopwords-words that are prevalent yet carry little meaningful weight. The script also filters out short words, focusing the dataset on terms that are more likely to contribute significantly to understanding the text's context or sentiment.

Data Balancing and Augmentation

The final step in preprocessing involves balancing and augmenting the dataset to correct class imbalances, a common issue in machine learning dataset preparation. Class imbalances can significantly bias the performance of machine learning models, leading to poorer generalization, especially in underrepresented classes. This balancing process is crucial for training models that can accurately interpret and respond to diverse data inputs without skew.

Streamlined Training Methodology

This method uses a K-Fold Cross-Validation approach to train the BERT model. As described in Algorithm 1, the training process initializes the BERT model with predefined settings. It divides the dataset into folds, ensuring comprehensive exposure to various data subsets. Each fold rotationally acts as a validation set, allowing every data segment to contribute to model validation.

2.3. Bidirectional Encoder Representations from Transformers (BERT)

Architecture

This model utilizes a pre-trained BERT (Bidirectional Encoder Representations from Transformers) as its core for feature extraction and a neural network classifier for class predictions. The architecture is detailed as follows:

Algorithm 1 Streamlined Training Process for BERT Model

```
Initialize BERT model with pre-trained weights.
Set training parameters: learning rate, epochs, epsilon.
Prepare optimizer and learning rate scheduler.
procedure K-Fold Cross-Validation Training
    Split dataset into $k$ folds.
    for each fold do
        Train model on $k - 1$ folds and validate on the remaining fold.
        Track validation metrics and save best model parameters.
    end for
    Select and retrain the best model on the entire dataset.
end procedure
procedure Evaluation
    Set model to evaluation mode.
    for each validation batch do
        Compute predictions and measure performance metrics.
    end for
    Output average validation loss and F1 score.
end procedure
```

- **Transformer Model (BERT)**: The BERT model processes input data and generates feature-rich embeddings. It consists of multiple layers of transformer blocks that perform bidirectional text encoding. The parameters of BERT can be optionally frozen, which allows for customization between using BERT solely for feature extraction or further fine-tuning on specific tasks. The BERT parameters are in the Table 2.
- **Neural Network Classifier**: Figure 3 illustrates a one-layer feed-forward neural network attached to the BERT model, serving as a classifier. It includes a linear layer mapping BERT's output to an intermediate representation, a ReLU activation function for non-linearity, dropout for regularization, and a final linear layer outputting class logits.

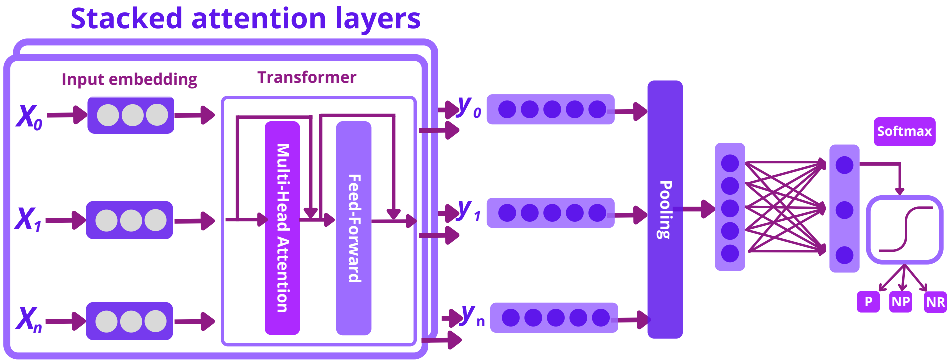


Figure 3: The multi-head attention layer used in the Transformer architecture shows, step by step, how the Transformer model (BERT) processes each stage.

3. Experimental Results

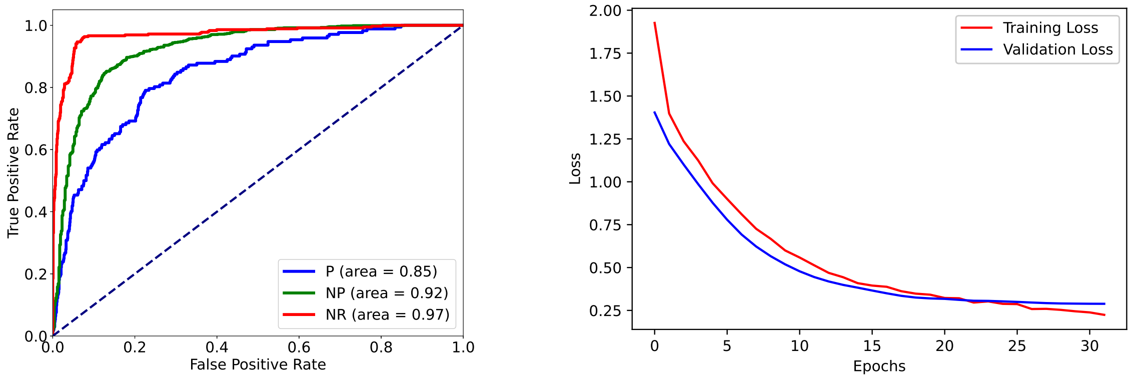
Table 2 details the configuration parameters for the BERT model used in our experiments.

Table 2
BERT Model Configuration

Parameter	Value
Model Version	bert-base-uncased
Number of Layers	12
Hidden Size	768
Attention Heads	12
Epochs	32
Optimizer	AdamW
Learning Rate	1e-5
Epsilon	1e-8

Training and Validation Loss

Figure 4b illustrates the loss curve for the BERT model over 32 epochs. It is evident that both the training and validation losses decrease steadily as the number of epochs increases, demonstrating the model's learning process. The training loss (in red) consistently trends lower than the validation loss (in blue), indicating that the model fits the training data well but generalizes effectively to unseen data.



(a) ROC Curve per class - Track 1.

(b) Loss BERT model with 32 epochs.

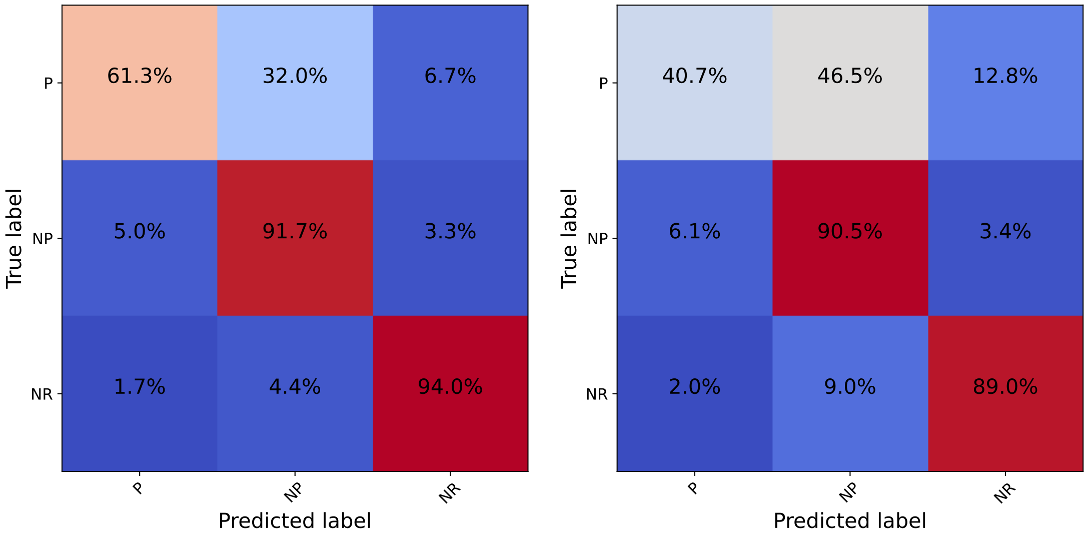
Figure 4: ROC Curve and Loss for BERT model.

ROC Curve Analysis

The ROC curve per class, shown in Figure 4a, highlights the model's performance in distinguishing between classes. The AUC (Area Under the Curve) values for classes P, NP, and NR are 0.85, 0.92, and 0.97, respectively. These high AUC values indicate predictive solid performance, with class NR achieving near-perfect discrimination.

Confusion Matrix

Figure 5a and Figure 5b present the confusion matrices for the training and evaluation phases. The training confusion matrix shows the model's prediction performance on the training set, where it achieved high precision and recall for the NR class (94.0%) but slightly lower performance for the P



(a) Confusion matrix for training.

(b) Confusion matrix for evaluation.

Figure 5: Confusion matrices.

and NP classes. The evaluation confusion matrix indicates similar trends, with the model maintaining robust performance on the NR class (89.0%) but showing some misclassifications in the P and NP classes.

Summary of Model Performance

Our initial training phase involved 80% of the provided data. The BERT model achieved an accuracy of 0.945. The final evaluation on a provided unlabeled dataset yielded accuracies of 82.4% and 81.3%, respectively, confirming the model's effectiveness.

Table 3 compares the general performance metrics across different tracks, showcasing F1-Score and recall values for each track. Tracks 1 and 2 exhibit high F1 scores and recall, emphasizing the model's reliability and robustness. In Track 1, we achieved first place; in Track 2, we secured third place; and in Track 3, we attained second place.

Table 3
General Performance Metrics for Each Track

Track 1		Track 2		Track 3	
F1-Score	Recall	F1-Score	H. Loss	F1-Score	Recall
0.92	0.90	0.93	0.03	0.57	0.69

4. Conclusions

The study demonstrates the effectiveness of the BERT model in detecting phobic hate in tweets across three different tasks by only varying the inputs and the number of outputs without altering the underlying model structure. The BERT model and a neural network classifier showed robust feature extraction and class prediction capabilities by employing effective data-cleaning techniques and a well-structured pipeline. K-Fold Cross-Validation ensured reliable and generalized model performance, as evidenced by high accuracy, F1-Score, and Hamming Loss values. The model successfully handled data imbalance issues for Tracks 1 and 2. However, for Track 3, which focuses on classifying song lyrics as LGBT+phobic or non-LGBT+phobic, the model's performance could have been more effective, indicating the need for further improvements in handling imbalanced datasets in this specific task.

In summary, future work should address the remaining challenges in accurately predicting the NR category and refine the model's performance through advanced techniques. Expanding the model's language capabilities to encompass a more diverse range of languages would be beneficial. By iterating and improving upon the existing models, we can advance the field of hate speech detection and contribute to developing more robust and inclusive natural language processing solutions.

References

[1] P. K. Kumaresan, R. Ponnusamy, R. Priyadharshini, P. Buitelaar, B. R. Chakravarthi, Homophobia and transphobia detection for low-resourced languages in social media comments, Natural Language Processing Journal 5 (2023) 100041. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000389. doi:https://doi.org/10.1016/j.nlp.2023.100041.

[2] A. A. Hind Saleh, K. Moria, Detection of hate speech using bert and hate speech word embedding with deep model, Applied Artificial Intelligence 37 (2023) 2166719. doi:10.1080/08839514.2023.2166719.

[3] A. Das, S. Nandy, R. Saha, S. Das, D. Saha, Analysis and detection of multilingual hate speech using transformer based deep learning, 2024. arXiv:2401.11021.

[4] N. Patwardhan, S. Marrone, C. Sansone, Transformers in the real world: A survey on nlp applications, Information 14 (2023). URL: https://www.mdpi.com/2078-2489/14/4/242. doi:10.3390/info14040242.

[5] R. T. Mutanga, N. Naicker, O. O. Olugbara, Hate speech detection in twitter using transformer methods, International Journal of Advanced Computer Science and Applications 11 (2020). URL: https://api.semanticscholar.org/CorpusID:222450148.

[6] B. Wei, J. Li, A. Gupta, H. Umair, A. Vovor, N. Durzynski, Offensive language and hate speech detection with deep learning and transfer learning, 2021. arXiv:2108.03305.

[7] H. A. Madni, M. Umer, N. Abuzinadah, Y.-C. Hu, O. Saidani, S. Alsubai, M. Hamdi, I. Ashraf, Improving sentiment prediction of textual tweets using feature fusion and deep machine ensemble model, Electronics 12 (2023). URL: https://www.mdpi.com/2079-9292/12/6/1302. doi:10.3390/electronics12061302.

[8] K. L. Tan, C. P. Lee, K. M. Lim, A survey of sentiment analysis: Approaches, datasets, and future research, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/7/4550. doi:10.3390/app13074550.

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, 2013. arXiv:1310.4546.

[10] A. Severyn, A. Moschitti, Twitter sentiment analysis with deep convolutional neural networks, in: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, 2015, pp. 959–962. doi:10.1145/2766462.2767830.

[11] R. K. Behera, M. Jena, S. K. Rath, S. Misra, Co-lstm: Convolutional lstm model for sentiment analysis in social big data, Information Processing & Management 58 (2021) 102435. URL: https://www.sciencedirect.com/science/article/pii/S0306457320309286. doi:https://doi.org/10.1016/j.ipm.2020.102435.

[12] G.-D. Pilar, S.-B. Isabel, P.-M. Diego, G. Ávila José Luis, A novel flexible feature extraction algorithm for spanish tweet sentiment analysis based on the context of words, Expert Systems with Applications 212 (2023) 118817. URL: https://www.sciencedirect.com/science/article/pii/S0957417422018358. doi:https://doi.org/10.1016/j.eswa.2022.118817.

[13] D. Elreedy, A. F. Atiya, A comprehensive analysis of synthetic minority oversampling technique (smote) for handling class imbalance, Information Sciences 505 (2019) 32–64. URL: https://www.sciencedirect.com/science/article/pii/S0020025519306838. doi:https://doi.org/10.1016/j.ins.2019.07.070.

[14] A. S. Hussein, T. Li, D. M. Abd Ali, K. Bashir, C. W. Yohannese, A modified adaptive synthetic sampling method for learning imbalanced datasets, in: Developments of Artificial Intelligence Tech-

nologies in Computation and Robotics: Proceedings of the 14th International FLINS Conference (FLINS 2020), World Scientific, 2020, pp. 76–83.

[15] M. Aloraini, A. Khan, S. Aladhadh, S. Habib, M. F. Alsharekh, M. Islam, Combining the transformer and convolution for effective brain tumor classification using mri images, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/6/3680. doi:10.3390/app13063680.

[16] B. Jang, M. Kim, G. Harerimana, S.-u. Kang, J. W. Kim, Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism, Applied Sciences 10 (2020). URL: https://www.mdpi.com/2076-3417/10/17/5841. doi:10.3390/app10175841.

[17] 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.

[18] G. Bel-Enguix, H. G'omez-Adorno, G. Sierra, J. V'asquez, S. T. Andersen, S. Ojeda-Trueba, Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed toowards the mexican spanish speaking lgbtq+ population, Natural Language Processing 71 (2023).

[19] H. G'omez-Adorno, G. Bel-Enguix, H. Calvo, J. V'asquez, S. T. Andersen, S. Ojeda-Trueba, T. Alc'antara, M. Soto, C. Macias, Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population, Natural Language Processing 73 (2024).