VEL at HOMO-MEX 2024: Detecting LGBT+phobia in Mexican Spanish Social Media

Devendra Deepak Kayande[1,*], Kishore Kumar Ponnusamy[2], Prasanna Kumar Kumaresan[3], Paul Buitelaar[3] and Bharathi Raja Chakravarthi[3]

[1]*Indian Institute of Information Technology, Allahabad, India*
[2]*Digital University Kerala, Kerala, India*
[3]*Data Science Institute, University of Galway, Ireland*

Abstract

This paper proposes the work of team *VEL* to address IberLEF-2024 shared task Homo-MEX 2024. Hate speech against the LGBT+ community remains a pervasive issue on social media platforms, contributing to a hostile and harmful online environment. Homo-Mex 2024 shared task gives a platform to detect this hatred by introducing datasets and tasks for the LGBT+ community in Mexican Spanish language. Our paper presents an approach to detecting hate speech using natural language processing (NLP) techniques and machine learning algorithms, specifically addressing the Track-1 and Track-3 tasks of the competition, which involve multi-class and binary classification respectively. Our paper presents an approach to detecting hate speech using NLP techniques and machine learning algorithms, addressing the Track-1 and Track-3 tasks of the competition. Track-1 involves predicting the label of each tweet (LGBT+phobic, not LGBT+phobic, or not LGBT+related), while Track-3 involves binary classification of song lyrics phrases as either LGBT+phobic (P) or not LGBT+phobic (NP). We implemented fine-tuning techniques, LSTM-based and XGBoost-based modeling techniques on features extracted using Spanish BERT fed with rigorously preprocessed and augmented data. Out of our submissions, our XGBoost-based method achieved the best macro F1-scores of 74.56 on Track-1 and 47.44 on Track-3 test data. Our research work tries to contribute to the broader goal of creating safer online spaces for the LGBT+ community by providing a robust tool for moderating and mitigating harmful content.

Keywords

Hate Speech, Multi-Class Classification, Binary Classification, Data Augmentation, Social Media Analytics

1. Introduction

LGBT+ people experience a higher prevalence of mental health problems compared to their heterosexual counterparts worldwide [1, 2]. The distinction between LGBT+ individuals and heterosexual individuals can be ascribed to Meyer's (1995) [3] minority stress paradigm. This concept posits that individuals who identify as sexual minorities encounter persistent and unique sources of stress, including prejudice, mistreatment, and the internalization of negative attitudes towards their own sexual orientation [4]. These factors all contribute to the creation of an antagonistic and anxiety-inducing social environment. The various components of stress and the consequent hostility have detrimental impacts on mental well-being [5].

LGBT+ minority individuals exhibit a higher level of enthusiasm in utilizing online platforms. They are twice as inclined to utilize online applications in comparison to heterosexual individuals, with a usage rate of 51% versus 28% [6, 7]. However, the LGBT+ community often faces negativity and hurtful messages [8, 9, 10]. Incidents of abusive content have increased in recent times, primarily due to the surge in popularity of social media [11, 12]. The proliferation of abusive content on social media

IberLEF 2024, September 2024, Valladolid, Spain
*Corresponding author.
devendrakayande427@gmail.com (D. D. Kayande); kishorep161002@gmail.com (K. K. Ponnusamy); kprasannakumar30@gmail.com (P. K. Kumaresan); paul.Buitelaar@universityofgalway.ie (P. Buitelaar); bharathi.raja@universityofgalway.ie (B. R. Chakravarthi)
0009-0007-8607-2468 (D. D. Kayande); 0000-0001-9621-668X (K. K. Ponnusamy); 0000-0003-2244-246X (P. K. Kumaresan); 0000-0001-7238-9842 (P. Buitelaar); 0000-0002-4575-7934 (B. R. Chakravarthi)

platforms, such as Twitter, Facebook, and YouTube, is a significant and escalating problem due to the vast volume of user-generated information available on the internet [13].

Homophobia/transphobia constitutes a form of abuse that can manifest as physical violence, including murder, mutilation, or assault; explicit sexual violence, such as rape, molestation, or penetration; or an invasion of privacy through the revealing of personal information [14]. One example is the statement "Gays should be killed" [15]. Additional instances of homophobia/transphobia remarks encompass statements such as "Homosexual individuals should be subjected to stoning", "An individual of lesbian orientation should be sexually violated to convert her to heterosexuality", "You ought to terminate your existence", "Lesbian individuals, I possess knowledge of your residence and intend to pay you a visit tonight", and "One should forcibly remove the homosexual inclination from him" [16, 17]. These comments have all been specifically targeted at LGBT+ individuals who are socially marginalized [10]. In response to this escalating issue, academics have developed a multitude of traditional machine learning and deep learning algorithms to autonomously identify hate speech on online social platforms [18, 19, 20, 21]. While the majority of automated systems that identify undesirable information rely on natural language processing (NLP) techniques, there is currently a shift towards utilizing advanced machine learning approaches, including deep learning, for this purpose.

The Homo-MEX 24 shared task aims to change that [22, 23]. We participated in the shared task and proposed a comprehensive methodology that integrates advanced natural language processing techniques and machine learning algorithms. Specifically, we employed a hybrid approach combining LSTM networks and XGBoost classifiers, fine-tuned on features extracted using the Spanish BERT model, to adaptively recognize patterns of hate speech within both tweets and song lyrics. This approach was rigorously tested through a series of experiments that benchmarked our models against the competition's diverse datasets.

2. Related Work

Over the past few years, scholars have examined hate speech on social media platforms through several research approaches [24, 25]. The categorization of hate speech in literature is mostly accomplished through the utilization of conventional machine learning and sophisticated deep learning techniques [26]. These methods may generally be classified into two main groups of machine learning: those that rely on feature engineering and those that rely on deep learning [27, 28, 29]. Once the dataset has been obtained and processed, the text must be transformed into numerical vectors to facilitate learning tasks.

Chakravarthi 2023 [10] introduced the task of homophobia and transphobia detection for social media comments and created a dataset for English and Tamil languages. They also [30] worked on the issue of code-mixing in the Tamil-English setting. Kumerasan et al 2023 [31] developed a dataset specifically designed to identify instances of homophobia and transphobia in the Malayalam and Hindi languages. The dataset has 5,193 comments in Malayalam and 3,203 comments in Hindi. Kumerasan et al 2024 [32] created a novel dataset encompassing three languages: Telugu, Kannada, and Gujarati that has been labeled by experts to enable the automatic detection of homophobic and transphobic content [33]. Chakravarthi and his teams created multiple shared tasks on this area to increase the research in homophobia and transphobia detection [21, 19, 20].

Vasquez et al [34] presented Homo-MEX, a corpus for detecting LGBT+Phobia in Mexican Spanish scrapped using Twitter API [35]. Nearly 10,000 tweets were scraped and annotated carefully by 4 annotators. They established a baseline using various machine learning based approaches like feeding TF-IDF vectorizers to SVM, Logistic Regression, Naive Bayes, and Random Forest and also using deep learning approaches like fine-tuning pre-trained large multilingual BERT-based models. Garcia et al [36] addressed this hate speech detection task using features extracted from Spanish-based LLMs integrating knowledge integration strategy using shallow neural networks. Rosauro et al [37] used TF-IDF vectorizing strategy for the raw Mexican Spanish text alongside Multinomial Naive Bayes and SVC. They also used transformer-based approaches for using BETO and multilingual DeBERTa. Shahiki et al [38] used BERT as a sole model for fine-tuning the Mexican Spanish hate speech data. Erika et al

[39] addressed the multi-label classification task of the hate speech using TF-IDF weighted features and used classical machine learning approaches like Gaussian Mixture Models, SVMs, and Random Forests. They also implemented a second approach using a Bag of Words text representation coupled with dimensionality reduction techniques and these features fed to Logistic Regression with OneVSRest strategy. Morina et al [40] proposed an end-to-end approach for hate speech classification. They used a back-translation data augmentation technique to address the data scarcity and used an ensemble of properly fine-tuned BETO [41], XLM-RoBERTa [42] and Spanish corpus pre-trained RoBERTa.

3. Task Description

Homo-MEX 24[1] represents a pivotal effort within the NLP field to enhance detection systems capable of identifying discriminatory language against the Mexican Spanish-speaking LGBT+ community [23, 22, 43]. This competition[2], crucial due to the ongoing prevalence of discrimination based on sexual orientation and gender identity, organizes its challenges into three focused tracks.

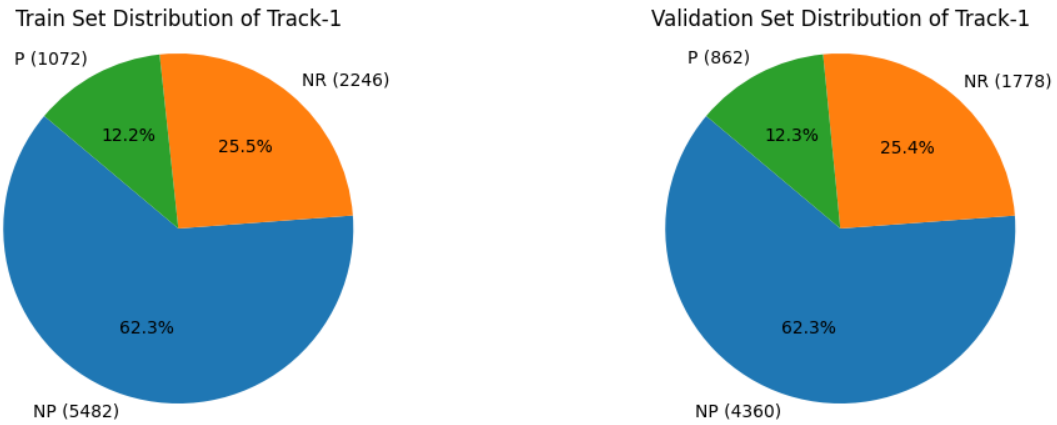
- **Track 1: Hate speech detection track (Multi-class):** Participants classify each tweet under one of three potential categories: explicit LGBT+phobic (P), non-discriminatory but LGBT+ mentioning (NP), and irrelevant to LGBT+ issues (NR). This track aims to refine the accuracy of digital tools in recognizing and categorizing various forms of communication about the LGBT+ community.
- **Track 2: Fine-grained hate speech detection track (Multi-labeled):** This more granular task requires identifying specific types of discriminatory remarks. Tweets might display one or several forms of hate speech, such as lesbophobia (L), gayphobia (G), biphobia (B), transphobia (T), or other defined LGBT+phobias (O). The objective is to apply labels that reflect the specific biases present, thereby deepening the understanding of how hate speech manifests.
- **Track 3: Homophobic lyrics detection track (Binary):** This track challenges participants to evaluate whether phrases within songs perpetuate hate speech against the LGBT+ community, classifying them as either LGBT+phobic (P) or not (NP). It tests the capability of models to interpret and judge content within a cultural and artistic medium, differing significantly from typical social media text.

Engagement in the Competition, Our research group contributed to **Track 1** and **Track 3**, focusing on direct and indirect forms of hate speech in both social and artistic mediums. These efforts are not just academic; they address real-world needs for better content moderation that can support a safer, more inclusive online discourse. The Homo-MEX 24 competition provides a valuable venue for researchers to advance understanding and technology against LGBT+phobia, reflecting broader social advancements through the application of NLP technologies.

4. Dataset Statistics

The Homo-MEX 24 competition provided a structured approach to dataset availability across various phases—Development, Training, and Testing—to facilitate progress with the task of hate speech detection towards the Mexican Spanish-speaking LGBT+ population. During the initial Development Phase, organizers were given a preliminary view of the data to the participants, which allowed for early model experimentation and strategy adjustments. This was crucial for understanding the distribution of data points as illustrated in Figure 1 and Figure 2, which depict the dataset composition for Tracks 1 and 3, respectively. For Track 1, the data consisted of tweets categorized into three classes: LGBT+phobic (P), not LGBT+phobic (NP), and not LGBT+related (NR). The distribution in the training set was predominantly NP (62.3%), followed by NR (25.5%) and P (12.2%), with a similar distribution observed in the validation set.

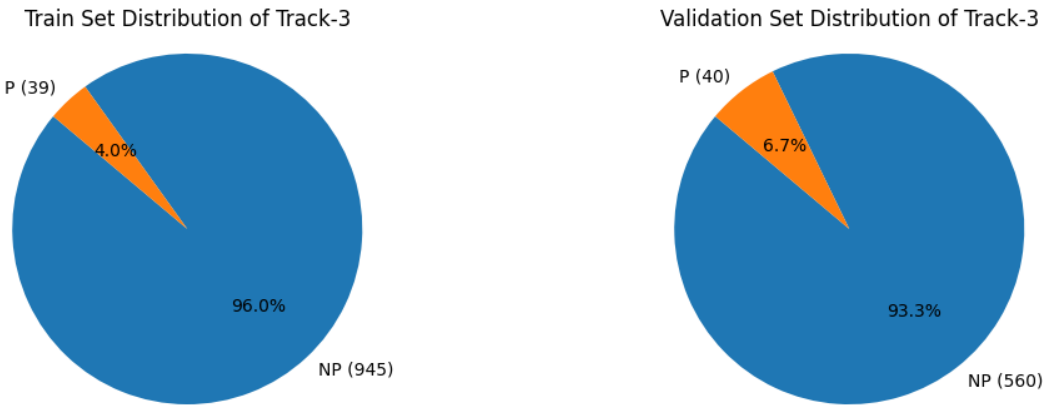
[1] https://sites.google.com/view/homomex/home
[2] https://www.codabench.org/competitions/2229/



(a) Training (b) Validation

Figure 1: Pie Chart of the number of data points of Track-1.

In the subsequent Training Phase, the full training dataset was provided, enabling in-depth tuning and refinement of the models. The comprehensive dataset supported the development of robust models capable of accurately classifying and predicting the tweets categories of hate speech and its absence, as detailed in the phased descriptions of the tasks. The final Testing Phase challenged participants to apply their models to an unlabeled testing dataset, the results of which were evaluated using predefined metrics. This phased approach not only structured the modeling challenges effectively but also offered multiple touch points for model improvement and validation against a progressively disclosed dataset, reflecting a realistic and rigorous testing environment.



(a) Training (b) Validation

Figure 2: Pie Chart of the number of data points of Track-3.

5. Methodology

We used the Spanish BERT, BETO [44] 'dccuchile/bert-base-spanish-wwm-uncased' as our backbone for feature extraction from the raw text Mexican Spanish Data for Track-1 and Track-3 of the competition. BETO outputs two kinds of features, Sequential features, i.e. a 768-dimensional embedding vector for each token in the input data, and also average pooled tensor of these Sequential Features, also called

Sentence Embedding of each input sentence to the mode. The data was imbalance hence we also applied augmentation strategies, raw oversampling of raw text sentences of the minority classes to feed to BETO, and for our XGBoost[45] model, we fed Sentence Embeddings with SMOTE[46] augmenting technique. The following section details our methodology implemented and the overall method is represented pictorially in Figure 3.

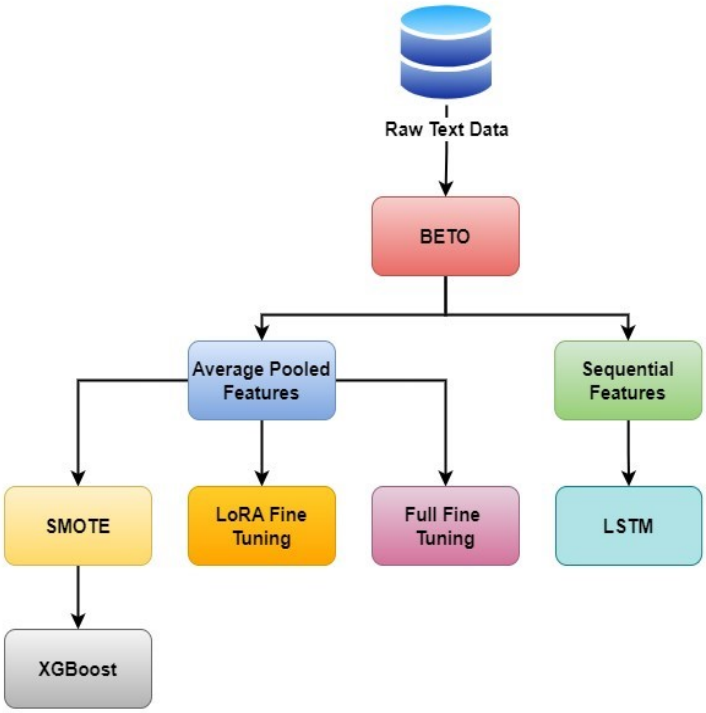


Figure 3: Overall methodology we implemented.

5.1. Data Augmentation and Preprocessing

Random Oversampling for raw text data and SMOTE for Sentence Embeddings were used as main augmentation techniques. The dataset of Track-1 had the distribution shown in Figure 1. The Not LGBT+phobic (NP) class had the most number of data points so we randomly over-sampled the data points of the remaining 2 classes Not LGBT+related (NR) and LGBT+phobic (P) to 5,400. For the preprocessing part, we removed '@Username' patterns, numbers, and URLs, removed all the characters except punctuation marks, and newline patterns, and removed all the emojis from all the sentences. After the preprocessing, we analyzed the distribution of sequence lengths and decided to keep the max sequence length as 100 tokens, hence sentences having a token count greater than 100 were truncated and sentences having a token count lower than 100 were padded. We extracted the sentence embeddings from BETO for input to the XGBoost model. To augment these sentence embeddings we used the SMOTE augmentation technique. Similar preprocessing and augmentation steps were done for data of Track-3 which had Not LGBT+phobic (NP) as the majority class and LGBT+phobic (P) as the minority class.

SMOTE works by creating synthetic examples along the line segments between existing minority class examples. The step-by-step process in mathematical form is as follows:

1. Select a minority class sample x_i.
2. Find its k-nearest minority class neighbors. Let's denote one of these neighbors as $x_{i,k}$.
3. Generate a synthetic sample x_{new} as follows:

$$x_{\text{new}} = x_i + \delta \cdot (x_{i,k} - x_i) \tag{1}$$

where δ is a random number between 0 and 1.

5.2. Full Fine Tuning of Spanish BERT

First, we unfreezed the whole BETO model and added a one-layer dense neural net acting as a classifier with a softmax at the end. The whole model received the gradient updates for 10 epochs, with a batch size of 16 using AdamW optimizer with a learning rate of 2e-5 with a linear schedule and rest default parameters of the optimizer also with Early Stopping callback. Cross Entropy loss was used to train the model to reduce misclassifications.

5.3. LoRA Fine Tuning of Spanish BERT

Full fine-tuning resulted in performance degradation on the new span classification adaption task, the model after training was an overfit model with very bad generalizing on validation data. Hence we used LoRA [41] to fine-tune a new paradigm for parameter-efficient fine-tuning of the model. LoRA helps LLMs to learn new tasks without any catastrophic forgetting from the previous knowledge without a very large number of parameter updates hence with less computation power.

In LoRA, we have a weight matrix $W \in \mathbb{R}^{d \times k}$ in the pre-trained model. We introduce two smaller matrices $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$, where r is the rank, typically much smaller than both d and k. The adapted weight matrix W_{adapted} is then computed as:

$$W_{\text{adapted}} = W + BA \tag{2}$$

Here, A and B are the trainable parameters during fine-tuning. The rank r is chosen to be much smaller than d and k to reduce the computational cost and the number of parameters that need to be trained.

The forward pass of the model with LoRA fine-tuning involves computing the output Y using the adapted weight matrix:

$$Y = XW_{\text{adapted}} = X(W + BA) \tag{3}$$

where $X \in \mathbb{R}^{n \times d}$ is the input to the layer (with n being the batch size).

During back-propagation, the gradients with respect to A and B are computed to update these matrices:

$$\frac{\partial L}{\partial A} = \frac{\partial L}{\partial Y} X B^T \tag{4}$$

$$\frac{\partial L}{\partial B} = A^T X^T \frac{\partial L}{\partial Y} \tag{5}$$

where L is the loss function.

The weights W are typically frozen during this process, and only A and B are updated. For the LoRA fine-tuning the rank 'r' was set to 16 and the LoRA adapters were applied to all 'linear' layers of the BETO model. The model was trained for 10 epochs with Cross Entropy loss with a batch size of 8 using an AdamW optimizer with a learning rate of 3e-4 with a linear rate schedule on raw text.

5.4. XGBoost on Sentence Embeddings

The third approach, our best submission on Track-1 and Track-3, was based on eXtreme Gradient Boosting or XGBoost trained on the sentence embeddings taken from BETO of the raw text. XGBoost working can be summarized as follows:

1. *Objective Function:* XGBoost aims to minimize an objective function $\mathcal{L}$ that combines a loss function L, measuring prediction errors, and a regularization term Ω to control model complexity. Mathematically, it can be expressed as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{6}$$

where θ represents the parameters of the model, y_i is the true label, and $\hat{y}_i$ is the predicted value.

2. *Boosting Process:* XGBoost uses boosting, where each new tree in the ensemble is trained to correct the errors (residuals) of the previous trees. Let's denote the prediction of the t-th iteration as $\hat{y}_i^{(t)}$. The prediction of the next tree $t+1$ is:

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + f_{t+1}(\mathbf{x}_i) \tag{7}$$

where f_{t+1} is the new tree added at iteration $t+1$.

3. *Gradient Boosting:* In gradient boosting, each new tree is trained to minimize the gradient of the loss function concerning the predicted values. This can be mathematically represented as:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} - \eta \cdot \nabla_{\hat{y}_i^{(t-1)}} L(y_i, \hat{y}_i^{(t-1)}) \tag{8}$$

where η is the learning rate and $\nabla_{\hat{y}_i^{(t-1)}} L(y_i, \hat{y}_i^{(t-1)})$ is the gradient of the loss function.

4. *Regularization:* XGBoost includes regularization terms to control the complexity of the model. One common form of regularization is L2 regularization, which penalizes the complexity of the individual trees. Mathematically, it can be represented as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|\mathbf{w}\|^2 \tag{9}$$

where T is the number of leaves in the tree, $\mathbf{w}$ are the leaf weights, and γ and λ are regularization parameters.

XGBoost builds an ensemble of decision trees that collectively minimize the objective function, producing accurate predictions while controlling overfitting. We did hyper-parameter tuning on the validation data to control the tree growing parameters like 'max-depth', 'learning-rate', 'n-estimators', and 'min-child-weight' and some regularization parameters like 'reg-alpha' and 'reg-lambda' to control the overfitting. XGBoost was trained on *multiclass logloss* for Track-1 and *binary logloss* for the Track-3 task.

5.5. LSTM on Sequential Spanish BERT Features

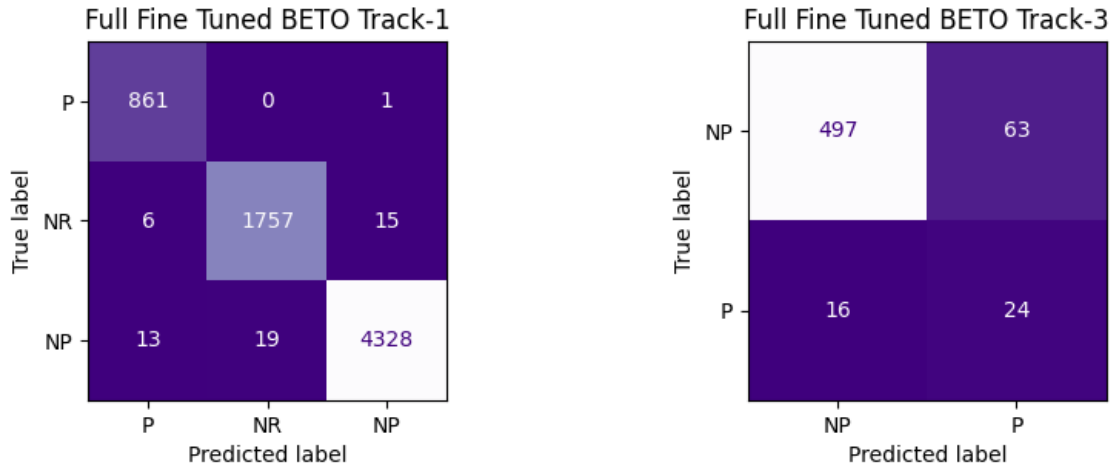
Training a model on sequential features allows it to capture temporal dependencies within the data, enabling a better understanding of the underlying task dynamics. LSTM (Long Short-Term Memory) networks excel in this regard by effectively modeling long-range dependencies and mitigating the vanishing gradient problem encountered in traditional recurrent neural networks (RNNs). Their gated architecture enables them to retain and selectively update information over extended sequences. LSTM enhances the model's ability to learn intricate patterns and relationships within sequential data, leading to improved performance on various tasks. BERT's ability to capture global context through its attention mechanisms and LSTM's capability to model sequential patterns in data. BERT provides contextualized embeddings that encapsulate rich semantic information for each token. Feeding these embeddings into an LSTM allows the model to further learn temporal dependencies and sequential patterns in the data. We trained 2 layers of LSTM with a hidden dimension of 192 and a single softmax layer at the end with a dropout of 0.3 between the final and LSTM layers on the frozen BETO sequential features for 30 epochs with Cross Entropy loss using AdamW optimizer with a learning rate of 3e-4 with a linear learning rate schedule and a batch size of 32.

Table 1
Track-1 Results on Validation Data

Model	Precision (macro)	Recall (macro)	F1 (macro)	Accuracy
Full Fined BETO	99.60	99.60	99.60	99.22
LoRA Fine Tuned BETO	85.94	91.49	88.63	97.81
LSTM	95.64	98.15	96.83	97.55
XGBoost	97.00	99.03	98.05	98.20

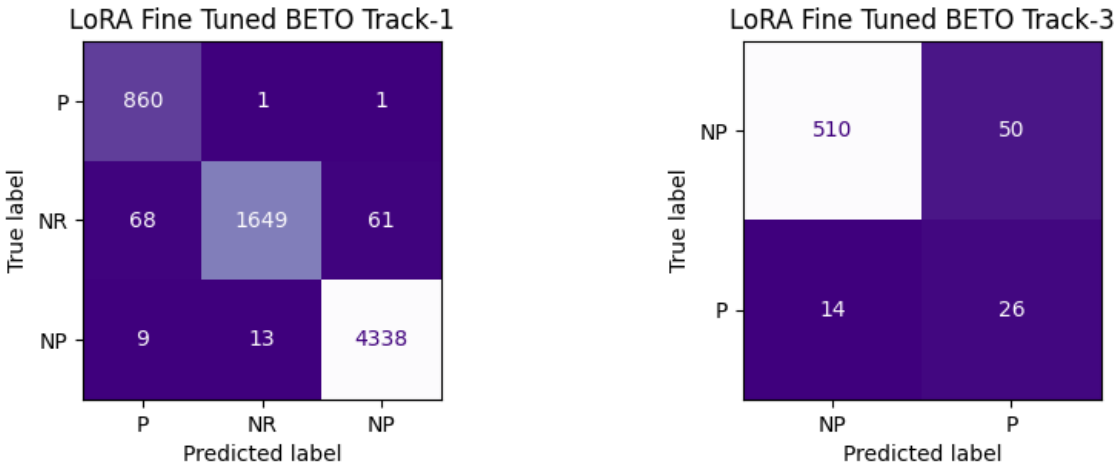
6. Results

Trained models were evaluated on macro-average Precision, Recall, and F1-Score for both Track-1 and Track-3. For Track-1 our best model XGBoost trained on SMOTE augmented sentence embeddings achieved an F1-score of 74.56 on test data as depicted by the leaderboard while 98.058 on validation data. Full Fine tuned BETO achieved very high F1-scores of 99.604 and 99.085 on train and validation data respectively but a very low F1-score of 32.39 on test data as shown on the leaderboard after our second submission. LoRA fine-tuned BETO achieved F1-scores of 98.712 and 88.632 on training data and validation data respectively while LSTM trained on temporal features resulted in scores of 98.311 and 96.830 on training and validation data. For Track-3 LoRA Fine tuned BETO on raw text data resulted in an F1-score of 69.461 on validation data and 87.195 on the training data whereas full fine-tuning exhibited worse performance due to less generalizing capability by achieving an F1-score of 94.134 on training data and 57.897 on validation data. LSTM trained on sequential features of BETO of Track-3 resulted in an F1-score of 95.443 on training data and 61.760 on validation data while our best submission of the XGBoost model resulted in an F1-score of 47.440 on test data as shown on the leaderboard, 68.391 on validation data. Overall the models showed less generalizing power to the unseen distribution of test data with very prone to overfitting on the training dataset even after applying regularizing techniques like dropout, early stopping, and L-2 regularizing which resulted in a significant gap between the scores on validation and test data. The macro scores and accuracy on the validation dataset of Track-1 are shown in Table 1 and for Track-3 are provided in Table 2 while the confusion matrices of each model on both the tasks are provided in the following figures.



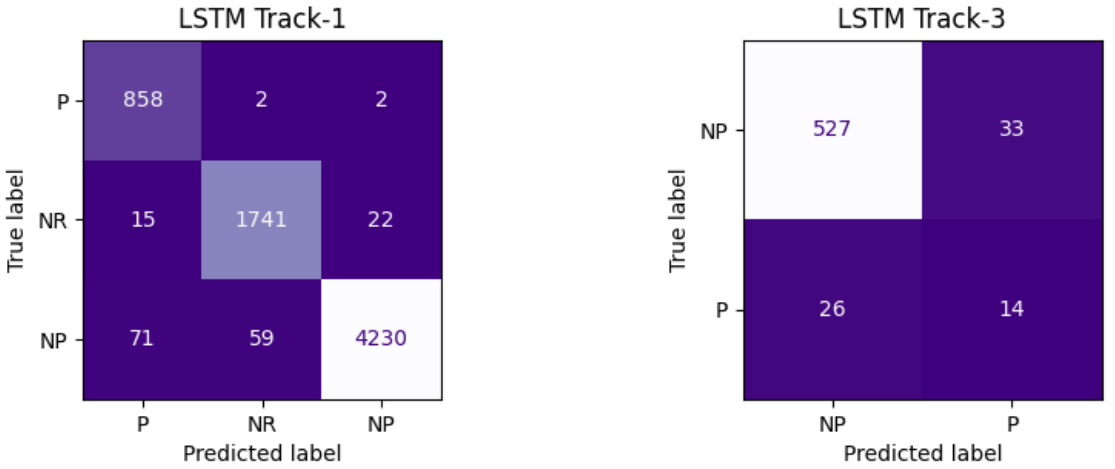
(a) Track-1 (b) Track-3

Figure 4: Confusion Matrix of validation data for Full Fine Tuned BETO.



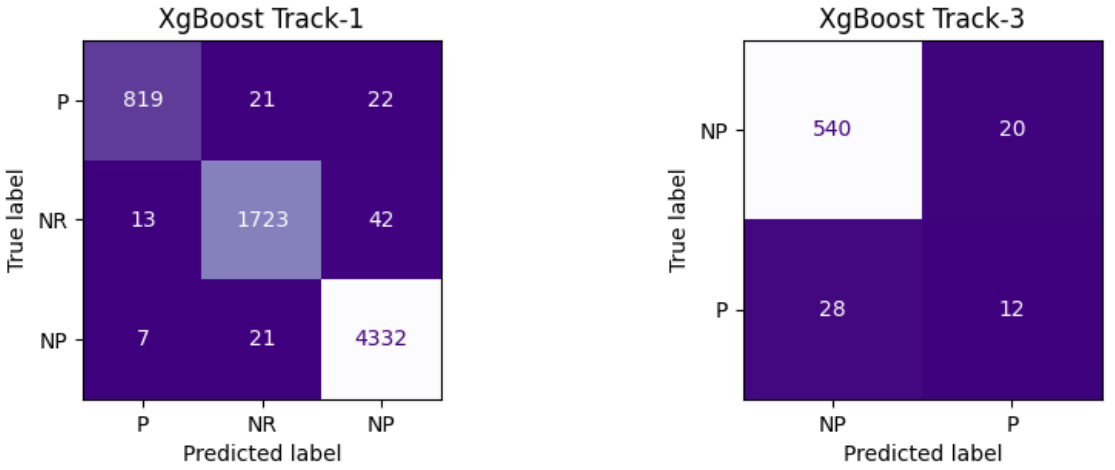
(a) Track-1 (b) Track-3

Figure 5: Confusion Matrix of validation data for LoRA Fine Tuned BETO.



(a) Track-1 (b) Track-3

Figure 6: Confusion Matrix of validation data for LSTM on Sequential BETO Features.



(a) Track-1 (b) Track-3

Figure 7: Confusion Matrix of validation data for XgBoost.

Table 2
Track-3 Results on Validation Data

Model	Precision (macro)	Recall (macro)	F1 (macro)	Accuracy
Full Fined BETO	54.70	61.43	57.89	86.83
LoRA Fine Tuned BETO	65.76	78.03	69.46	89.33
LSTM	59.86	63.78	61.76	90.16
XGBoost	66.28	63.21	64.53	92.00

7. Conclusion

In this paper we presented our proposal for hate speech detection for the LGBT+ community speaking Mexican Spanish Language in Twitter comments through IberLEF-2024 Homo-MEX 24 shared tasks. Our approach leveraged the Spanish BERT model - BETO as the main feature extractor. We used full fine-tuning and LoRA fine-tuning approaches. LSTM-based sequential modeling was also implemented. XGBoost trained on SMOTE augmented sentence embeddings emerged as our best model achieving macro F1-scores - 74.56 on Track-1 and 47.44 on Track-3. Further, we plan to explore extensive hyper-parameter optimization, better data augmentation methods, ensemble approaches, and more robust regularizing techniques to avoid over-fitting.

8. Acknowledgments

Author Bharathi Raja Chakravarthi was supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2). Prasanna Kumar Kumaresan was supported in part by a research grant from the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References

[1] D. Burgess, R. Lee, A. Tran, M. Van Ryn, Effects of perceived discrimination on mental health and mental health services utilization among gay, lesbian, bisexual and transgender persons, Journal of LGBT health research 3 (2007) 1–14.

[2] A. Marciano, Y. David, N. Antebi-Gruszka, The interplay of internalized homophobia, compulsive use of dating apps, and mental distress among sexual minority individuals: Two moderated mediation models, Computers in Human Behavior 156 (2024) 108241. URL: https://www.sciencedirect.com/science/article/pii/S0747563224001092. doi:https://doi.org/10.1016/j.chb.2024.108241.

[3] I. H. Meyer, Minority stress and mental health in gay men, Journal of health and social behavior (1995) 38–56.

[4] B. Rodríguez-Expósito, J. A. Rieker, S. Uceda, A. I. Beltrán-Velasco, V. Echeverry-Alzate, M. Gómez-Ortega, A. Positivo, M. Reiriz, Psychological characteristics associated with chemsex among men who have sex with men: Internalized homophobia, conscientiousness and serostatus as predictive factors, International Journal of Clinical and Health Psychology 24 (2024) 100465. URL: https://www.sciencedirect.com/science/article/pii/S1697260024000309. doi:https://doi.org/10.1016/j.ijchp.2024.100465.

[5] I. H. Meyer, Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence., Psychological bulletin 129 (2003) 674.

[6] C. McClain, R. Gelles-Watnick, From looking for love to swiping the field: Online dating in the us (2023).

[7] L. L. Sharabi, C. V. Ryder, L. C. Niess, A space of our own: exploring the relationship initiation

experiences of lesbian, gay, bisexual, transgender, queer, intersex, and asexual dating app users, Journal of Social and Personal Relationships 40 (2023) 2277–2297.
[8] B. R. Chakravarthi, Hope speech detection in youtube comments, Social Network Analysis and Mining 12 (2022) 75.
[9] B. R. Chakravarthi, Multilingual hope speech detection in english and dravidian languages, International Journal of Data Science and Analytics 14 (2022) 389–406.
[10] B. R. Chakravarthi, Detection of homophobia and transphobia in youtube comments, International Journal of Data Science and Analytics (2023) 1–20.
[11] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, B. R. Chakravarthi, Multimodal hate speech detection from bengali memes and texts, in: International Conference on Speech and Language Technologies for Low-resource Languages, Springer, 2022, pp. 293–308.
[12] M. Vegupatti, P. K. Kumaresan, S. Valli, K. K. Ponnusamy, R. Priyadharshini, S. Thavaresan, Abusive social media comments detection for tamil and telugu, in: International Conference on Speech and Language Technologies for Low-resource Languages, Springer, 2023, pp. 174–187.
[13] S. Rajiakodi, B. R. Chakravarthi, R. Ponnusamy, P. Kumaresan, S. Thangasamy, B. Sivagnanam, C. Rajkumar, Overview of shared task on caste and migration hate speech detection, in: B. R. Chakravarthi, B. B, P. Buitelaar, T. Durairaj, G. Kovács, M. Á. García Cumbreras (Eds.), Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 145–151. URL: https://aclanthology.org/2024.ltedi-1.14.
[14] D. A. Haaga, " homophobia"?, Journal of Social Behavior and Personality 6 (1991) 171.
[15] R. Ponnusamy, P. K. Kumaresan, K. K. Ponnusamy, C. Rajkumar, R. Priyadharshini, B. R. Chakravarthi, Team_tamil at hodi: Few-shot learning for detecting homotransphobia in italian language (2023).
[16] 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.
[17] K. Lande, R. Ponnusamy, P. K. Kumaresan, B. R. Chakravarthi, Kaustubhsharedtask@ lt-edi 2023: Homophobia-transphobia detection in social media comments with nlpaug-driven data augmentation, in: Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, 2023, pp. 71–77.
[18] H. Kibriya, A. Siddiqa, W. Z. Khan, M. K. Khan, Towards safer online communities: Deep learning and explainable ai for hate speech detection and classification, Computers and Electrical Engineering 116 (2024) 109153. URL: https://www.sciencedirect.com/science/article/pii/S0045790624000818. doi:https://doi.org/10.1016/j.compeleceng.2024.109153.
[19] B. R. Chakravarthi, R. Ponnusamy, M. S, P. Buitelaar, M. Á. García-Cumbreras, S. M. Jimenez-Zafra, J. A. Garcia-Diaz, R. Valencia-Garcia, N. Jindal, Overview of second shared task on homophobia and transphobia detection in social media comments, in: B. R. Chakravarthi, B. Bharathi, J. Griffith, K. Bali, P. Buitelaar (Eds.), Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 38–46. URL: https://aclanthology.org/2023.ltedi-1.6.
[20] B. R. Chakravarthi, P. Kumaresan, R. Priyadharshini, P. Buitelaar, A. Hegde, H. Shashirekha, S. Rajiakodi, M. Á. García, S. M. Jiménez-Zafra, J. García-Díaz, R. Valencia-García, K. Ponnusamy, P. Shetty, D. García-Baena, Overview of third shared task on homophobia and transphobia detection in social media comments, in: B. R. Chakravarthi, B. B, P. Buitelaar, T. Durairaj, G. Kovács, M. Á. García Cumbreras (Eds.), Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 124–132. URL: https://aclanthology.org/2024.ltedi-1.11.
[21] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. McCrae, P. Buitelaar, P. Kumaresan, R. Ponnusamy, Overview of the shared task on homophobia and transphobia detection in social media comments, in: B. R. Chakravarthi, B. Bharathi, J. P. McCrae, M. Zarrouk, K. Bali, P. Buitelaar (Eds.), Proceedings of the Second Workshop on Language Technology for Equality, Diversity and

Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 369–377. URL: https://aclanthology.org/2022.ltedi-1.57. doi:10.18653/v1/2022.ltedi-1.57.
[22] 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.
[23] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, Procesamiento del Lenguaje Natural 73 (2024).
[24] R. Priyadharshini, B. R. Chakravarthi, M. S, S. Cn, K. S V, P. B, A. Murugappan, P. K. Kumaresan, Overview of shared-task on abusive comment detection in Tamil and Telugu, in: B. R. Chakravarthi, R. Priyadharshini, A. K. M, S. Thavareesan, E. Sherly (Eds.), Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages, INCOMA Ltd., Shoumen, Bulgaria, Varna, Bulgaria, 2023, pp. 80–87. URL: https://aclanthology.org/2023.dravidianlangtech-1.11.
[25] B. R. Chakravarthi, R. Priyadharshini, S. Banerjee, M. B. Jagadeeshan, P. K. Kumaresan, R. Ponnusamy, S. Benhur, J. P. McCrae, Detecting abusive comments at a fine-grained level in a low-resource language, Natural Language Processing Journal 3 (2023) 100006. URL: https://www.sciencedirect.com/science/article/pii/S2949719123000031. doi:https://doi.org/10.1016/j.nlp.2023.100006.
[26] A. Founta, L. Specia, A survey of online hate speech through the causal lens, in: A. Feder, K. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. Roberts, U. Shalit, B. Stewart, V. Veitch, D. Yang (Eds.), Proceedings of the First Workshop on Causal Inference and NLP, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 74–82. URL: https://aclanthology.org/2021.cinlp-1.6. doi:10.18653/v1/2021.cinlp-1.6.
[27] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: L.-W. Ku, C.-T. Li (Eds.), Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101. doi:10.18653/v1/W17-1101.
[28] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys (CSUR) 51 (2018) 1–30.
[29] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources and Evaluation 55 (2021) 477–523.
[30] B. R. Chakravarthi, A. Hande, R. Ponnusamy, P. K. Kumaresan, R. Priyadharshini, How can we detect homophobia and transphobia? experiments in a multilingual code-mixed setting for social media governance, International Journal of Information Management Data Insights 2 (2022) 100119. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000623. doi:https://doi.org/10.1016/j.jjimei.2022.100119.
[31] 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.
[32] P. K. Kumaresan, R. Ponnusamy, D. Sharma, P. Buitelaar, B. R. Chakravarthi, Dataset for identification of homophobia and transphobia for Telugu, Kannada, and Gujarati, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 4404–4411. URL: https://aclanthology.org/2024.lrec-main.393.
[33] B. R. Chakravarthi, P. Kumaresan, R. Priyadharshini, P. Buitelaar, A. Hegde, H. Shashirekha, S. Rajiakodi, M. Á. García, S. M. Jiménez-Zafra, J. García-Díaz, et al., Overview of third shared task on homophobia and transphobia detection in social media comments, in: Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, 2024, pp. 124–132.
[34] J. Vásquez, S. Andersen, G. Bel-Enguix, H. Gómez-Adorno, S.-L. Ojeda-Trueba, Homo-mex: A

mexican spanish annotated corpus for lgbt+ phobia detection on twitter, in: The 7th Workshop on Online Abuse and Harms (WOAH), 2023, pp. 202–214.
[35] P. K. Kumaresan, K. K. Ponnusamy, S. Kogilavani, S. Cn, R. Priyadharshini, B. R. Chakravarthi, Vel@ lt-edi: Detecting homophobia and transphobia in code-mixed spanish social media comments, in: Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, 2023, pp. 233–238.
[36] J. A. García-Díaz, S. M. Jiménez-Zafra, R. Valencia-García, Umuteam at homo-mex 2023: Fine-tuning large language models integration for solving hate-speech detection in mexican spanish (2023).
[37] C. F. Rosauro, M. Cuadros, Hate speech detection against the mexican spanish lgbtq+ community using bert-based transformers (2023).
[38] M. Shahiki-Tash, J. Armenta-Segura, Z. Ahani, O. Kolesnikova, G. Sidorov, A. Gelbukh, Lidoma at homomex2023@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), 2023.
[39] E. Rivadeneira-Pérez, M. de Jesús García-Santiago, C. Callejas-Hernández, Cimat-nlp at homo-mex2023@ iberlef: Machine learning techniques for fine-grained speech detection task (2023).
[40] A. J. M. Moriña, J. R. Pásaro, J. M. Vázquez, V. P. Álvarez, I2c-uhu at iberlef-2023 homo-mex task: Ensembling transformers models to identify and classify hate messages towards the community lgbtq (2023).
[41] 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.
[42] 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.
[43] 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).
[44] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, arXiv preprint arXiv:2308.02976 (2023).
[45] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.