Detecting HUrtful HUmour on Twitter using Fine-Tuned Transformers and 1D Convolutional Neural Networks

Iván Árcos[1,†], Jaime Pérez[1,†]

[1]Polytechnic University of Valencia, Buildings 1G - 1E - 1H, ETS of Computer Engineering, Camí de Vera, s/n, 46022 Valencia, Spain

Abstract

This paper presents a comprehensive approach to detect humor that spreads prejudice on Twitter. Our methodology utilizes embedding extraction and fine-tuning techniques, employing 1D Convolutional Neural Networks (CNN) to capture relationships among embeddings and enhance model performance. Additionally, we leverage sentiment analysis, along with other extracted variables, to further improve the effectiveness of the models. We address three distinct tasks in our evaluation. The first task focuses on distinguishing tweets that express prejudice through humor from those that express prejudice without humor. In the second task, we perform multilabel classification to identify the targeted minority groups in prejudiced tweets, including women and feminists, the LGBTIQ community, immigrants and racially discriminated people, and over- weight individuals. The third task involves predicting the degree of prejudice on a continuous scale ranging from 1 to 5 for tweets targeting minority groups. Experimental results demonstrate the efficacy of our approach, highlighting the significance of capturing relationships among embeddings using 1D Convolutional Neural Networks. Additionally, the incorporation of sentiment analysis and other extracted variables further enhances model performance. Our findings contribute to advancing sentiment analysis and prejudice detection in social media, fostering a more inclusive online environment. The proposed methodology opens up avenues for future research and development in this domain

Keywords

Humor detection, Prejudice detection, Sentiment analysis, 1D Convolutional Neural Networks, Embedding extraction, Fine-tuning, Minority groups, Social media, Twitter. CEUR-WS

1. Introduction

The expression of prejudice is a common strategy used to harm individuals from minority groups [1]. Prejudice is defined as the ”negative pre-judgment of members of a race, religion, or any other socially significant group, regardless of contradicting facts” [1]. Prejudice expression is closely related to stereotypes, which are beliefs about the characteristics of a social group originating from preconceived judgments or prejudices that perceive a certain group as ”different.” These beliefs can emphasize negative or positive aspects, as the core of discriminatory strategies is to present the other group as distinct from ourselves [2]. The study of this phenomenon has

IberLEF 2023, September 2023, Jaén, Spain

[†]These authors contributed equally.

iarcgab@etsinf.upv.es (I. Árcos); jpernav@etsinf.upv.es (J. Pérez)

been a focus in social sciences since the early 20th century, but it remains an ongoing challenge, especially in the era of social media platforms that provide new avenues for the dissemination of prejudice. Interestingly, humor is often employed in these messages to evade moral judgment and condemnation of discrimination. In fact, as a society starts to overcome its prejudices towards certain social groups, humor can become a space where these prejudiced attitudes persist [1]. Previous research has explored the use of offensive language in humor, such as in the HAHA task at IberEval 2018 [3] and IberLEF 2019 and 2021 [4, 5], as well as the dissemination of stereotypes through irony [6]. There have also been efforts to study the hurtfulness of other forms of figurative language, such as sarcasm [7]. In the HUHU task [2], the focus is specifically on examining the use of humor to express prejudice towards minority groups in Spanish tweets. Other studies [8][9], uses computational linguistics to identify characteristics that distinguish high and low levels of offense in humorous texts. The study focuses on how hate speech is disguised as humor, particularly in Spanish tweets targeting minority groups.

2. Task Descriptions

2.1. Subtask 1: HUrtful HUmour Detection

The first subtask aims to determine whether a prejudicial tweet is intended to cause humor. Participants are required to distinguish between tweets that use humor to express prejudice and tweets that express prejudice without using humor. This task involves binary classification, where systems are evaluated and ranked based on the F1-measure over the positive class. The F1-measure provides a balanced evaluation of both precision and recall, capturing the system's ability to correctly identify hurtful humor instances.

2.2. Subtask 2A: Prejudice Target Detection

In the second subtask, participants are asked with identifying the targeted minority groups in the tweets. The specified minority groups include women and feminists, the LGBTIQ community, immigrants and racially discriminated people, and overweight people. This task is formulated as a multilabel classification problem, where systems need to assign relevant labels to each tweet. The evaluation metric employed for this task is the macro-F1 score, which considers the average performance across all labels.

2.3. Subtask 2B: Degree of Prejudice Prediction

The third subtask focuses on predicting the degree of prejudice expressed in the tweets on a continuous scale ranging from 1 to 5. Participants are required to assign a numerical value to indicate the level of prejudice exhibited towards the minority groups. The evaluation metric used for this task is the Root Mean Squared Error (RMSE), which measures the average difference between the predicted values and the ground truth values. A lower RMSE indicates better performance in accurately predicting the degree of prejudice.

3. Data

The dataset used in this study consists of 2,671 tweets collected exclusively from the Twitter platform. These tweets were manually labeled and annotated regarding attributes such as humor and prejudice. The aim of this dataset is to provide information and analysis on the presence and manifestation of humor and prejudice in the context of Twitter.

Regarding the dataset characteristics, there is an uneven distribution in the classes of humor and prejudice targets. In terms of humor, approximately 32.5% of the tweets are classified as humorous, while the remaining 67.5% do not contain humorous elements.

In relation to prejudice, the following proportions are observed: 48.4% of the tweets show some form of prejudice towards women, 22.7% towards the LGBTQIA+ community, 24.9% towards immigrants or racial groups, and 8.0% specifically exhibit prejudice towards overweight individuals (fatphobia).

Lastly, the average value of "mean_prejudice" in the dataset is approximately 3.05, with a standard deviation of 0.81. This indicates that, on average, the tweets exhibit a moderate level of prejudice in their content.

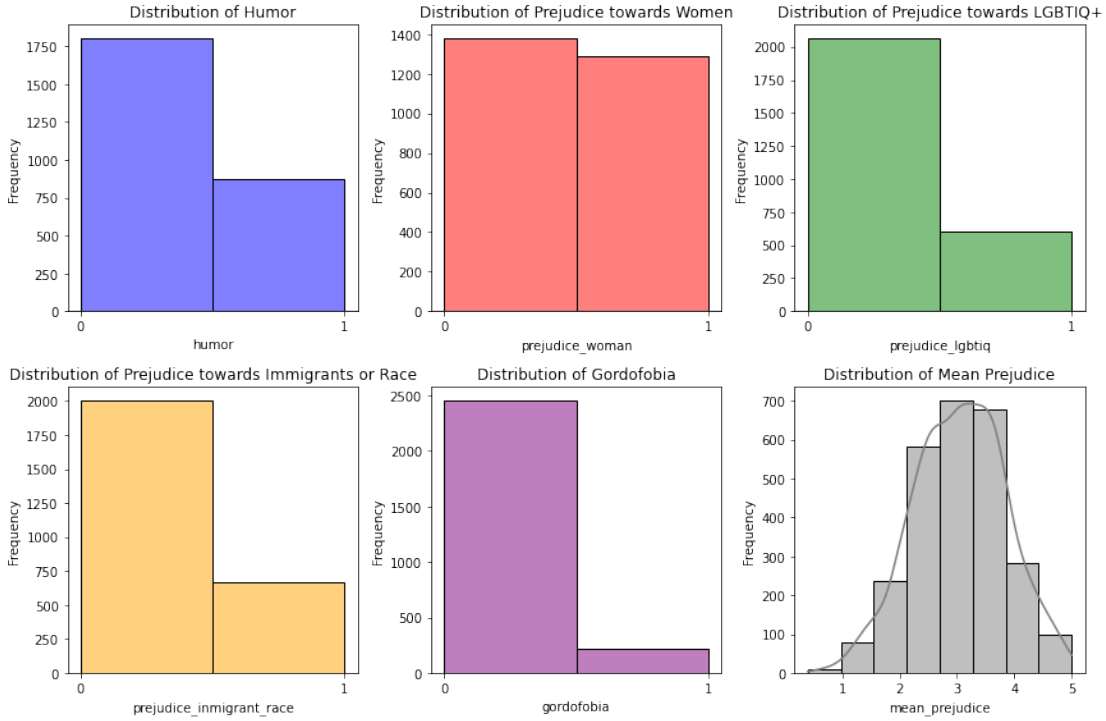

Figure 1: Distribution of Data

4. Features Extraction

In this section, we explain the process of features extraction from the Twitter data. We employed the following resources and methods:

4.1. HurtLex

HurtLex [10] is a lexicon that consists of offensive, aggressive, and hateful words in 50+ languages. It is categorized into 17 categories, which include negative stereotypes, professions, disabilities, moral defects, and more. The lexicon indicates the presence of stereotypes and follows a two-level structure: conservative (offensive senses) and inclusive (all relevant senses). To extract variables from the tweets associated with HurtLex, we counted the occurrences of words in each category and normalized them by the total word count.

4.2. Emotions, Irony, and Cyberbullying

We utilized pretrained transformer models to extract variables related to emotions, irony, and cyberbullying from the tweets. Specifically, we used the *twitter-xlm-roberta-emotion-es* [11] model to obtain scores for different types of emotions, including sadness, joy, anger, surprise, disgust, fear, and others. For irony detection, we employed the *roberta-base-bne-irony* [12] model, and for cyberbullying detection, we utilized the *roberta-base-bne-finetuned-cyberbullying-spanish* [13] model.

The extraction of these features enables us to capture the presence of offensive language, stereotypes, emotions, irony, and cyberbullying in the Twitter data. This information contributes to enhance the performance of our models and provides valuable insights into the nature of the tweets.

5. Pre-trained embeddings

Embeddings are obtained from the pretrained transformer model *bertin-roberta-base-spanish* [14] with 768 dimensions. These embeddings capture the contextual representation of the input text and provide rich semantic information.

To optimize the performance of our models, a grid search with 10-fold cross-validation is performed. This allows us to explore different hyperparameter combinations and select the best configuration for each model.

Four models are compared in our experiments: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), XGBoost, and Random Forest. Each model is trained on the embedded tweet data to learn the patterns and relationships between the input features and the target labels.

In addition to the tweet embeddings, we also analyze the effect of using sentiment features. These features capture the sentiment expressed in the tweets, which can provide valuable insights for prejudice detection. To feed the machine learning models, we utilize the embedding of the special token [CLS], which represents the overall meaning of the entire tweet. Additionally, to incorporate sentiments, we concatenate seven emotion-associated components (sadness, joy, anger, surprise, disgust, fear, and others) to this embedding.

This process is applied to all three tasks: HUrtful HUmour Detection, Prejudice Target Detection, and Degree of Prejudice Prediction. The same set of models and hyperparameter optimization techniques are employed for each task, allowing us to evaluate their performance consistently across the different aspects of prejudice detection.

6. Results on pre-trained embeddings

The results of the cross-validation for the proposed method are presented below.

6.1. Task 1: HUrtful HUmour Detection

For Task 1, the performance of SVM, MLP, and XGBoost models was similar, with F1 scores approaching 0.7. The inclusion of emotions and HurtLex variables had a noticeable effect, particularly on XGBoost and Random Forest models. Figure 1 shows the cross-validation results for Task 1.

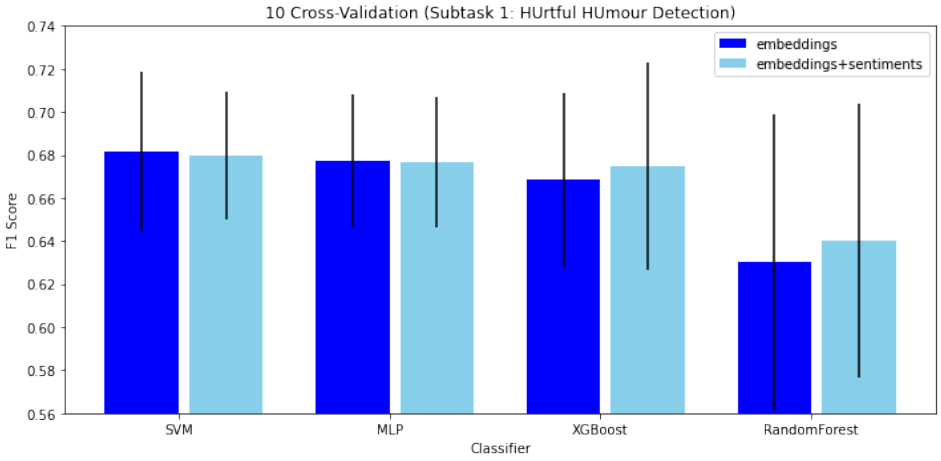

Figure 2: Cross-validation results for Task 1

6.2. Task 2a: Prejudice Target Detection

For Task 2, which involved binary classifications for each target, the models achieved macro F1-scores of over 0.85. Once again, the inclusion of emotions and HurtLex variables had a noticeable effect, especially on MLP and Random Forest models. Figure 2 presents the cross-validation results for Task 2a.

6.3. Task 2b: Degree of Prejudice Prediction

For Task 2b, the SVM model achieved the best performance with an RMSE of 0.7. However, there was no observable effect when including emotions and HurtLex variables in this task. Figure 3 presents the cross-validation results for Task 2b.

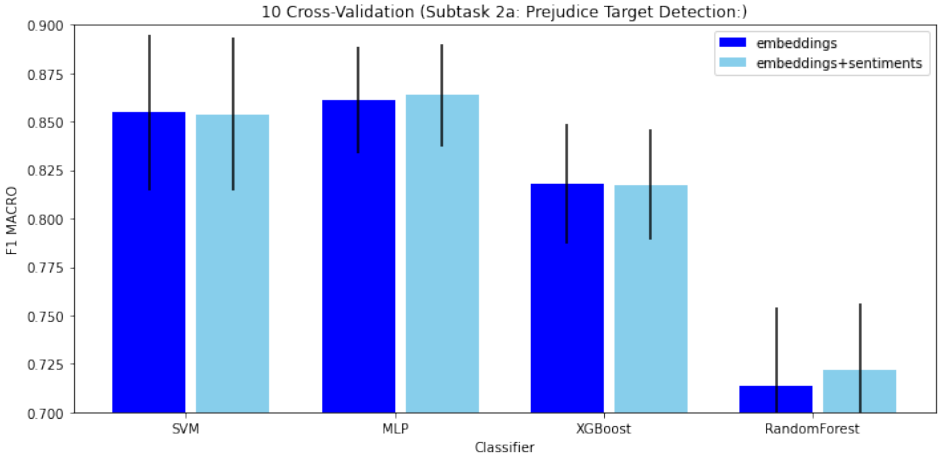

Figure 3: Cross-validation results for Task 2a

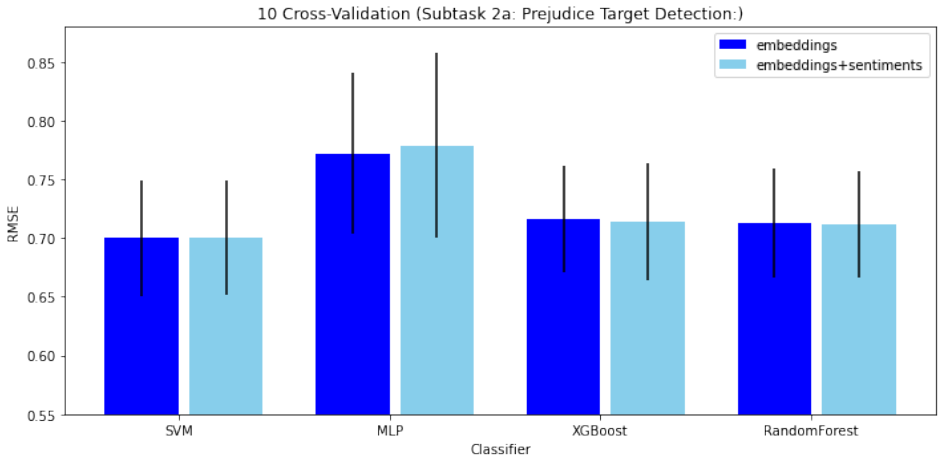

Figure 4: Cross-validation results for Task 2b

6.4. Fine-tuning Transformers for Task-specific Adaptation

Fine-tuning Transformers enables task-specific adaptation and potentially better performance compared to using precalculated embeddings with classical classifiers. The idea is to feed the Transformer model with 80 tokens representing a tweet. Instead of using the embedding associated with the [CLS] token or taking the average, we propose the use of 1-dimensional convolutional neural networks (CNN) with varying numbers of filters followed by 1-dimensional Max Pooling layers to obtain a final representation, which is then flattened into a vector.

The goal of this approach is to capture relationships between the tokens and the embeddings, allowing us to improve the performance of the models in all three tasks. To the resulting vector from the convolutional operations, we add the variables explained earlier: HurtLex, emotions, irony, and cyberbullying extracted using pretrained transformers. We then apply dense layers,

and the final layer depends on the specific task. For the first task, we use a neuron with sigmoid activation, and for the second task, we use four neurons with sigmoid activation. It's worth noting that now, unlike before with independent binary classifications, we are modeling the possible relationships between the different targets. For example, if a tweet expresses prejudice against women, it might also be more likely to exhibit fatphobia.

In these two tasks, binary cross-entropy is used as the optimization function. For the third task, a neuron without activation is used. The batch size used was 32, and for the second task, oversampling was performed to balance the batches.

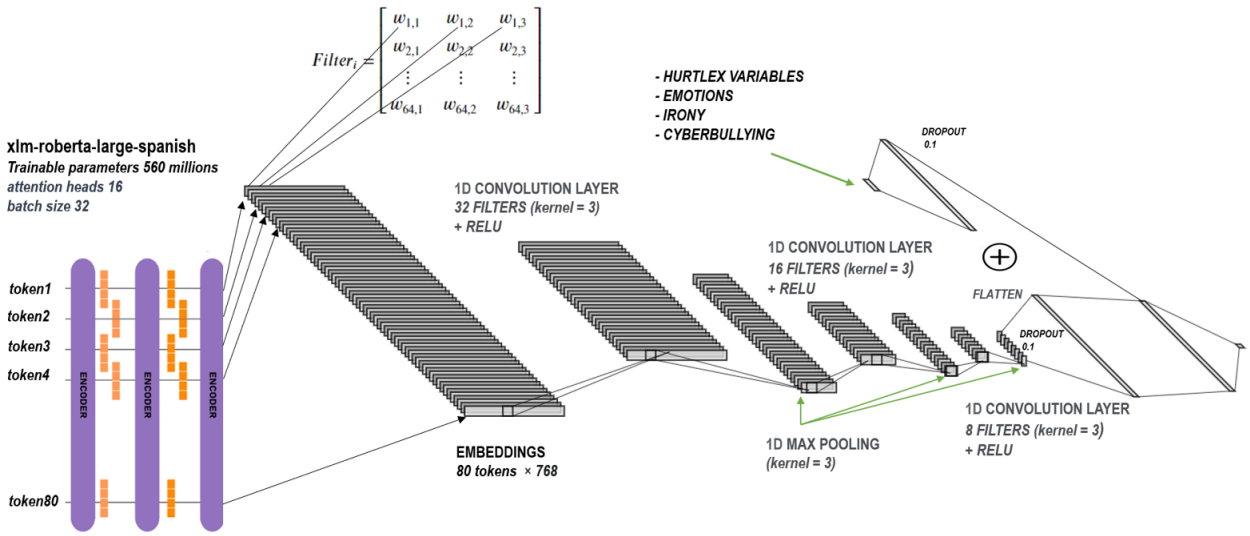

Figure 5: Architecture diagram illustrating the proposed approach

7. Attention Matrix

Attention matrices provide insights into the relative importance of each word in relation to others within a sentence. They offer a representation of how tokens interact with each other during the model's processing.

In our approach, we calculate attention matrices for each input sentence. Each matrix consists of values that indicate the level of attention or importance assigned to different word pairs within the sentence. By analyzing these matrices, we can identify the strongest interactions between tokens.

To highlight the most significant interactions, we extract the maximum value from all attention matrices. This allows us to identify the pairs of words that have the highest attention or influence on each other. By focusing on these key interactions, we gain valuable insights into the relationships between words and their impact on the overall meaning of the sentence.

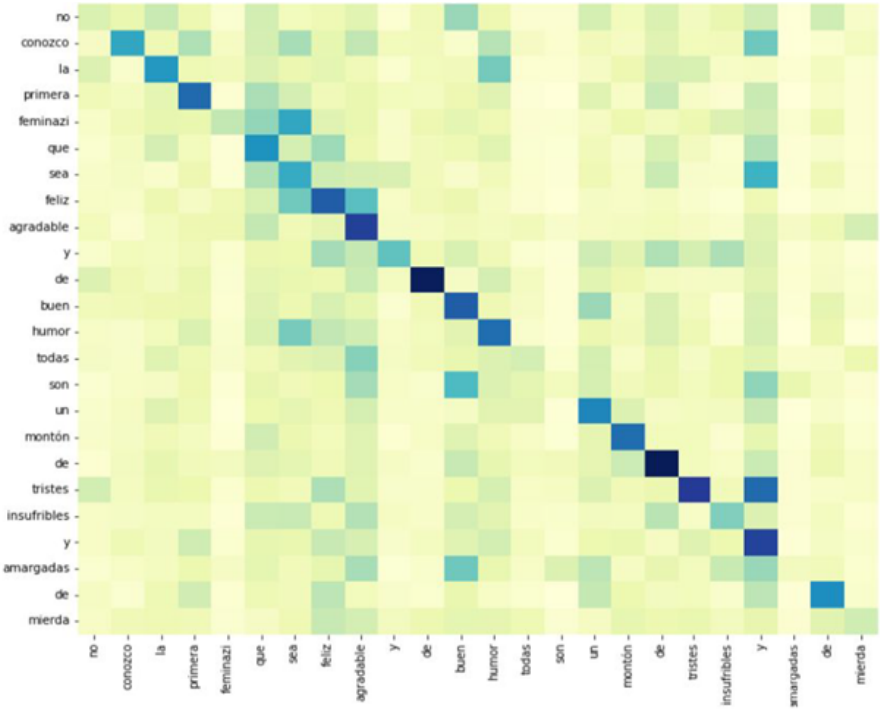

Figure 6: Example of the attention matrix of a tweet

Table 1
Performance Comparison of Different Models

Task	Model	Baseline	Emotions and Hurtlex
Subtask 1 (F1 SCORE)	SVM	0.7449	0.7500
	MLP	0.7236	0.7327
	XGBoost	0.7609	0.7624
	Random Forest	0.6982	0.7514
	xlm-roberta-large-spanish	-	**0.8473**
Subtask 2a (MACRO F1)	SVM	0.8608	0.8608
	MLP	0.8545	0.8622
	XGBoost	0.8244	0.8463
	Random Forest	0.7629	0.7725
	bertin-roberta-base-spanish	-	**0.9218**
Subtask 2b (RMSE)	SVM	0.7063	0.7035
	MLP	0.7767	0.7579
	XGBoost	0.7223	0.7186
	Random Forest	0.7299	0.7268
	bertin-roberta-base-spanish	-	0.7050
	Stacking (SVM + XGBoost)	-	**0.6904**

8. Analysis of results

Our experimentation on the test set reveals the significant impact of incorporating emotions and HurtLex features in improving metrics compared to using solely pre-trained transformer embeddings. This demonstrates the value of leveraging additional linguistic features to enhance model performance. Furthermore, the power of fine-tuning a transformer for a specific task is evident in our results. We conducted fine-tuning using *xlm-roberta-large-spanish* [15] for Subtask 1 and *bertin-roberta-base-spanish* for Subtask 2a. In both cases, we observed substantial improvements in metrics compared to using precalculated embeddings with classical classifiers. However, fine-tuning alone did not achieve the desired RMSE in Subtask 2b. Therefore, we opted for an ensemble approach by combining SVM and XGBoost, which outperformed individual models.

The models developed using the proposed architecture have allowed us to achieve a commendable position in the ranking. The incorporation of emotions and HurtLex variables, along with fine-tuning transformers, has proven to be effective in improving model performance across all three tasks. These results reflect the dedication and effort invested in building robust models.

Subtask 1: Hurtful Humour Detection

- Rank: 5th
- F1 score: 0.784

Subtask 2A: Prejudice Target Detection

- Rank: 10th
- Macro F1 score: 0.746

Subtask 2B: Degree of Prejudice Prediction

- Rank: 3rd
- RMSE (Root Mean Squared Error): 0.881

9. Conclusions

In conclusion, our current models have demonstrated their effectiveness and enabled us to secure a strong position in the ranking. We are optimistic about the future optimization and experimentation possibilities, which we believe can yield even better results. With each iteration, we strive to push the boundaries of model performance and make meaningful contributions to the field of prejudicial language detection.

Our final approach involved fine-tuning the *xlm-roberta-large-spanish* transformer model for subtask 1. Similarly, for subtask 2a, we fine-tuned the *bertin-roberta-base-spanish* transformer model to effectively detect the targeted minority groups in the tweets.

Furthermore, we incorporated emotions and HurtLex features to enhance the models' understanding of the linguistic context and improve their performance. This additional information proved valuable in capturing nuanced patterns and further improving F1 score.

Finally, we employed ensemble techniques, combining Support Vector Machine (SVM) and XGBoost models, to address the challenges of Subtask 2b and achieve superior results. This ensemble approach allowed us to leverage the strengths of each model and achieve better predictive performance.

10. Future work

However, we believe that there is still room for further optimization and experimentation with the proposed architecture. By exploring different hyperparameter settings, conducting more extensive grid searches, and fine-tuning the model, we anticipate that even better results can be achieved.

With a deeper understanding of the task requirements and the potential of the proposed architecture, we are confident that future iterations of our models can surpass the current results. By continuously refining and iterating on our approach, we aim to contribute to the development of state-of-the-art models for prejudicial language detection.

References

[1] D. B. Jones, A study on prejudice, Journal of Social Psychology 45 (1972) 123–145.

[2] R. Labadie-Tamayo, B. Chulvi, P. Rosso, Everybody hurts, sometimes. overview of hurtful humour at iberlef 2023: Detection of humour spreading prejudice in twitter, in: Procesamiento del Lenguaje Natural (SEPLN), volume 71, 2023.

[3] S. Castro, L. Chiruzzo, A. Rosá, Overview of the haha task: Humor analysis based on human annotation at ibereval 2018, IberEval@SEPLN (2018).

[4] L. Chiruzzo, S. Castro, M. Etcheverry, D. Garat, J. Prada, A. Rosá, Overview of haha at iberlef 2019: Humor analysis based on human annotation, IberLEF@SEPLN (2019).

[5] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. Meaney, R. Mihalcea, Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish, Procesamiento del Lenguaje Natural 67 (2021) 257–268.

[6] R. Ortega-Bueno, B. Chulvi, F. Rangel, P. Rosso, E. Fersini, Profiling irony and stereotype spreaders on twitter (irostereo) at pan 2022, CEUR-WS.org (2022).

[7] S. Frenda, C. A., V. Basile, C. Bosco, V. Patti, P. Rosso, The unbearable hurtfulness of sarcasm, Expert Systems with Applications (ESWA) 193 (2022).

[8] L. I. Merlo, When Humour Hurts: A Computational Linguistic Approach, Bachelor's thesis, Universitat Politècnica de València, 2022. URL: http://hdl.handle.net/10251/188166.

[9] L. Merlo, B. Chulvi, R. Ortega-Bueno, P. Rosso, When humour hurts: Linguistic features to foster explainability, Procesamiento del Lenguaje Natural (SEPLN) (2023).

[10] E. Bassignana, V. Basile, V. Patti, Hurtlex: A multilingual lexicon of words to hurt, in: Proceedings of the Conference on Language Resources and Evaluation (LREC), 2018, pp. 5–6. URL: https://ceur-ws.org/Vol-2253/paper49.pdf.

[11] Daveni, Twitter xlm-roberta emotion model, https://huggingface.co/daveni/twitter-xlm-roberta-emotion-es, 2021.

[12] Dtomas, Roberta base bne irony model, https://huggingface.co/dtomas/roberta-base-bne-irony, 2022.

[13] JonatanGk, Roberta base bne finetuned cyberbullying (spanish) model, https://huggingface.co/JonatanGk/roberta-base-bne-finetuned-cyberbullying-spanish, 2021.

[14] Bertin-roberta base spanish model, https://huggingface.co/bertin-roberta-base-spanish, 2021.

[15] Xlm-roberta large spanish model, https://huggingface.co/llange/xlm-roberta-large-spanish, 2022.