Building Robust Models for Detecting Offensive Content and Quantifying Prejudice in Online Platforms

David Borregón Sacristán[1], Antonio Pérez Muñoz[1] and Lucas Sebastián Peris[1]

[1]*Universitat Politècnica de València, Spain*

Abstract

In this paper, we present a comprehensive project focused on developing models that address three critical tasks related to prejudicial humour recognition: Hurtful Humour Detection, Prejudice Target Detection, and Degree of Prejudice Prediction. Offensive and prejudiced language are prevalent in online platforms, posing significant challenges for content moderation and fostering inclusive communities. Therefore, our work aims to contribute to the identification and quantification of such problematic content through the application of state-of-the-art natural language processing (NLP) techniques.

Keywords

natural language processing, offensive content, hurtful humour detection, prejudice target detection, degree of prejudice prediction, hate speech

1. Introduction

The proliferation of online platforms has revolutionized communication and information sharing, enabling individuals from diverse backgrounds to connect and engage in virtual communities. However, this increased accessibility and connectivity has also exposed society to the darker side of online interactions, including offensive content and prejudiced language. The prevalence of hurtful humour, hate speech, and discrimination on digital platforms poses significant challenges in fostering inclusive environments and ensuring the well-being of users.

Addressing these challenges requires the development of effective methods for detecting and quantifying offensive content and prejudice. The field of natural language processing (NLP) offers powerful tools and techniques to analyze and understand text data, making it a promising avenue for combating online harassment and promoting respectful communication. In these working notes on our submission to the *HUHU IberLEF Shared Task 2023*, we present a comprehensive work aimed at building models that address three vital tasks: Hurtful Humour Detection, Prejudice Target Detection, and Degree of Prejudice Prediction.

Hurtful Humour Detection is the first task we tackle, as it serves as a foundational step in identifying offensive content. Distinguishing between humourous content that is innocuous

IberLEF 2023, September 2023, Jaén, Spain

dborsac@etsinf.upv.es (D. Borregón Sacristán); apermuo@etsinf.upv.es (A. Pérez Muñoz); lsebper@etsinf.upv.es (L. Sebastián Peris)

0009-0006-7007-5464 (D. Borregón Sacristán); 0009-0009-6522-608X (A. Pérez Muñoz); 0009-0004-7590-903X (L. Sebastián Peris)

and content that aims to demean or target individuals or groups is crucial in understanding the potential harm caused by certain forms of humour. By developing models capable of recognizing hurtful humour, we can lay the groundwork for subsequent analyses and interventions.

Building upon the foundation of Hurtful Humour Detection, we delve into the task of Prejudice Target Detection. Identifying the specific targets of prejudice within offensive content allows us to gain insights into the various marginalized groups or individuals who are disproportionately affected by online discrimination. By shedding light on these targets, we aim to raise awareness, promote empathy, and encourage targeted interventions to counteract prejudiced discourse.

Lastly, we explore the Degree of Prejudice Prediction task, which involves quantifying the severity of prejudice within offensive language. Recognizing that not all instances of offensive content carry the same degree of harm, our models provide fine-grained analysis to assess the intensity and harmfulness of prejudiced language. This nuanced understanding empowers content moderators and platform administrators to make informed decisions regarding content policies and community guidelines.

The importance of this work lies in its potential to contribute to the field of NLP, offering practical solutions to mitigate the negative impact of offensive content and prejudice in online environments. By developing accurate and efficient models for hurtful humour detection, prejudice target detection, and degree of prejudice prediction, we aim to support online platforms in implementing proactive measures to foster inclusive communities. Through this research, we hope to contribute to the creation of safer online spaces, promoting positive interactions and protecting vulnerable individuals or groups from the detrimental effects of prejudiced content.

The remainder of this paper is structured as follows: Section 2 delves into the state of the art, describing the most used tools for the tasks at hand and naming related papers on hurtful humour detection and prejudice analysis. Section 3 describes the methodology and dataset used in our work. In Section 4, we present detailed results and analyses for each of the three tasks. Finally, Section 5 discusses the implications of our findings, limitations of the study, and avenues for future research, concluding with the significance of our project in addressing the challenges of offensive content and prejudice detection in online platforms.

2. State of the art

The field of natural language processing has witnessed remarkable advancements in recent years, particularly with the emergence of transformer-based models [1] such as BERT, GPT, and RoBERTa. These models have revolutionized various NLP tasks, including offensive content detection and prejudice analysis. Transformers excel in capturing contextual dependencies and semantic nuances in text, enabling them to comprehend and interpret complex language patterns. Researchers have leveraged pre-trained transformer models to develop fine-tuned architectures that exhibit superior performance in tasks related to hurtful humour detection, prejudice target detection, and degree of prejudice prediction. Additionally, specialized datasets annotated with offensive language and prejudice markers have been curated to train and evaluate these models effectively. The combination of transformer architectures and domain-specific datasets has greatly enhanced the state of the art in NLP tasks related to prejudice in humour, empowering the development of more accurate and robust models for content moderation and fostering

inclusive online communities.

In fact, previous tasks have investigated the use of offensive language in humour, in particular for Spanish HAHA at IberEval 2018 (Castro et al., 2018) [2] and IberLEF 2019 y 2021 (Chiruzzo et al., 2019; Chiruzzo et al., 2021) [3, 4] or the dissemination of stereotypes using irony (Ortega-Bueno et al., 2022) [5], and previous work was done to study the hurtfulness of other types of figurative language such as sarcasm (Frenda et al., 2022) [6]. More related work followed with the use of linguistic features to foster explainability (Merlo et al., 2022) [7] and HUHU at IberLEF 2023 (Labadie-Tamayo et al., 2023) [8].

Nonetheless, in HUHU, our focus is on examining the use of humour to express prejudice towards minorities, specifically analyzing Spanish tweets that are prejudicial towards:

- Women and feminists
- LGBTIQ community
- Immigrants and racially discriminated people
- Overweight people

3. Methodology

Our dataset comprises a diverse collection of tweets obtained from the shared task website, providing a rich and varied corpus for analysis. The dataset consists of a structured format with several key features that capture important information about each tweet. These features include the following:

- **index**: A unique identifier for each tweet.
- **tweet**: The actual content of the tweet, which may include text, hashtags, mentions, and URLs.
- **humour**: A binary label indicating whether the tweet contains hurtful humour or not.
- **fatphobia**: A binary label indicating whether the prejudice expressed in the tweet is targeting overweight people.
- **prejudice_woman**: A binary label indicating whether the prejudice expressed in the tweet is targeting women.
- **prejudice_lgbtiq**: A binary label indicating whether the prejudice expressed in the tweet is targeting the LGTBIQ+ collective.
- **prejudice_inmigrant**: A binary label indicating whether the prejudice expressed in the tweet is targeting inmigrants.
- **mean_prejudice**: A real number indicating the degree of prejudice present in the tweet.

Table 1 provides a glimpse of the dataset, showcasing a sample tweet along with its associated features.

Feature	Sample
index	65226
tweet	No todo lo qe brilla es oro, como por ejemplo... tu cara grasosa XD *Not all that glitters is gold, like for example... your greasy face XD*
humour	1
prejudice_woman	0
prejudice_lgbtiq	0
prejudice_inmigrant_race	0
fatphobia	1
mean_prejudice	2.8

Table 1: Sample tweet and its features.

3.1. Preprocessing

Before delving into the specific tasks, it is important to discuss the preprocessing steps we performed on the "tweet" column. We carried out two distinct preprocessing approaches.

The first preprocessing step focused on basic text cleaning techniques, including converting the text to lowercase, removing punctuation marks, and eliminating words such as hashtags, mentions, URLs, and stopwords. This initial preprocessing was performed to eliminate potential sources of noise that could hinder the effectiveness of subsequent techniques, such as models based on the transformers architecture.

The second preprocessing step, which was more in-depth, aimed to prepare the data for techniques that do not have their own embeddings, unlike transformers. In this step, we repeated the aforementioned basic cleaning processes and additionally performed tasks such as stopword removal, lemmatization, and obtaining the keyed vector representation for each tweet. In our case, the chosen approach to vectorize the words was based on word embedding, a technique by which each word is associated with an n-dimensional vector, which means that closely related words have very close vector representations, and allows the models to understand the meaning of the words in a certain way. Specifically, the fastText algorithm [9] was employed to compute the word embeddings. This algorithm was preferred due to its optimized nature for obtaining word embeddings quickly. Additionally, its open-source nature, availability, and cost-free usage made it accessible to all, facilitating replication of our experiments.

These additional preprocessing steps allowed us to apply techniques that rely on traditional feature engineering approaches. By understanding the different preprocessing methods employed throughout our research, we can now delve into the specific tasks at hand.

3.2. Addressing Task 1: Hurtful Humour Detection

The first task consists in determining whether a prejudicial tweet is intended to cause humour, using the "humour" feature from our dataset as the target feature. Prior to undertaking the task, it was of particular interest to study the distribution of classes in the "humour" feature, revealing an extreme class imbalance. Despite the potential information loss, undersampling was applied to the data since models seemed to classify predominantly into the majority class when this

step was omitted. Moreover, employing oversampling techniques like SMOTE considerably worsened the results, as generating meaningful synthetic text samples proved to be highly delicate [10]. Consequently, after lightly preprocessing the data and applying undersampling, we utilized the tokenizer provided by the transformer model chosen, specifically BETO [11], which exhibited slightly superior performance compared to ROBERTA [12] after multiple trials. We then computed the embeddings for the data and performed fine-tuning of the transformer [13] through a 5-fold cross-validation process, incorporating early stopping techniques to prevent overfitting and obtain a genuine assessment of the model's performance.

To further enhance the predictions provided by BETO, we compared its errors with those of other techniques that also yielded satisfactory results, such as SVM and XGBoost. Given the discernible differences in error patterns between these models, and considering their relatively similar precision levels, we decided to combine them to achieve improved results. After fine-tuning the hyperparameters for SVM and XGBoost, the best outcomes were obtained using a voting ensemble method based on soft voting. In this approach, given an input, we extracted the probabilities assigned to each class by the models and ultimately determined the class of the observation based on the highest percentage returned by the three models.

The approach described in detail above can be simplified in a diagram, as shown in Figure 1.

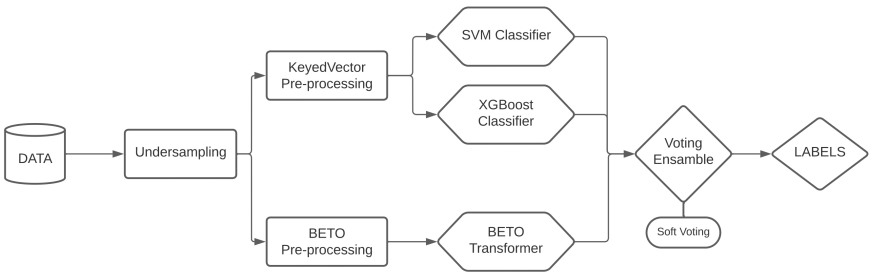

Figure 1: Diagram showing the structure of the approach chosen for subtask 1 (Hurtful Humour Detection).

3.3. Addressing Task 2a: Prejudice Target Detection

For the second task of Prejudice Target Detection, a multilabel classification problem, we followed a systematic methodology to build an effective model. The steps involved in this process are as follows. First, we performed basic cleaning on the 'tweet' column to preprocess the text data. This step ensured that the input data was in a standardized format and ready for further processing. Next, we computed sentence embeddings using BETO (a BERT-based model specifically trained for Spanish) for the preprocessed tweets. Sentence embeddings transform the textual data into numerical representations that capture the semantic meaning of the text. These embeddings serve as input to the neural network model for training and prediction.

To optimize the model's hyperparameters, we created an Optuna study [14]. Optuna is a hyperparameter optimization framework that explores different combinations of hyperparameters to find the best configuration. We defined an objective function to maximize the evaluation

metric's performance (e.g., F1 score) and utilized Optuna to search for the optimal hyperparameters. The best trial's value and corresponding parameters were printed for reference. It is important to note that in developing this function we took two additional precautionary measures to avoid overfitting: adjusting the learning rate during training and triggering an early stop if the validation loss does not improve for a certain number of epochs.

Based on the optimal hyperparameters, we constructed a Keras Sequential model [15]. The model architecture consisted of dense layers, dropout layers, and an output layer. Dense layers learn complex patterns and representations from the input data, while dropout layers help prevent overfitting by randomly dropping out a fraction of neurons during training. The output layer, with a sigmoid activation function, produced a probability distribution over the different prejudice groups being predicted. All this information can be compressed in a diagram like the one shown in Figure 2.

To assess the model's performance and ensure its generalizability, we employed 10-fold cross-validation. This technique involved splitting the dataset into 10 subsets, training the model on 9 subsets, and evaluating its performance on the remaining subset. By performing k-fold ("k" being 10, in this case) cross-validation, we obtained more reliable estimates of the model's performance and identified any potential overfitting or underfitting issues.

Figure 2: Diagram showing the structure of the approach chosen for subtask 2 (Prejudice Target Detection).

3.4. Addressing Task 2b: Degree of Prejudice Prediction

One of the labels available in the provided dataset was "prejudice_degree", representing the degree of prejudice reflected in each tweet. This information is valuable as not all tweets have the same impact on readers. Identifying the degree of prejudice allows us to distinguish between mildly offensive and highly offensive messages. The "prejudice_degree" label is a numerical feature ranging from 0 to 5, where 0 signifies a mildly offensive tweet and 5 represents a highly offensive one.

The task at hand involves predicting the degree of prejudice in text using regression models, with evaluation based on the root mean squared error (RMSE) on the test set. Our initial approach was relatively simple: finding individual models that could provide acceptably small RMSE values and combining them using an ensemble technique. It was preferable for these models to be as diverse as possible so that any errors in one model could be compensated by the others, preventing overfitting. Some of the models that yielded the best results were SVR (Support Vector Machine for Regression) [16], XGBoost for regression, and Random Forest. The next step was to tune the hyperparameters of each model and select an ensemble method. Since it is a regression task, we considered stacking as one of the best options for the ensemble method [17, 18, 19]. We experimented with L1 and L2 penalties and found that L2 with Ridge regression

produced the best results. We also considered using ElasticNet, but the combination of penalties worsened the results with higher percentages of L1 penalties. In the end, we presented the results of the stacking model with L2 penalty using Ridge regression, employing SVR, XGBoost, and RandomForest models.

Given two submissions were admitted to the shared task, we aimed to be more innovative in the next iteration and opted for a more complex method that made conceptual sense to us. During our tests with regression models, we noticed that due to the normal distribution of the data, the models tended to make predictions around the mean for the majority of instances. Consequently, extreme value tweets (between 1 and 2, and between 4 and 5) had a higher classification error. To address this, we trained three distinct models [20]: one for central values (2-4), one for low prejudice tweets (1-2), and one for high prejudice messages (4-5). Predictions from the model were then obtained through stacking, considering the decisions made by the three models. The main challenge of this model was that, when training with different subsets of data, there were instances where two models had to predict a degree of prejudice that was not present in their training samples.

Another version of this model involved creating a classifier model first. In this approach, a sample was classified into one of the three groups, and the corresponding model within that group was responsible for regression. However, the main issue with this model was its strong dependence on having a good classifier since even if the regression models were perfect, a poor prior classification would result in a significant difference in RMSE. The proposed approaches can be summarized in a scheme, as shown in Figure 3.

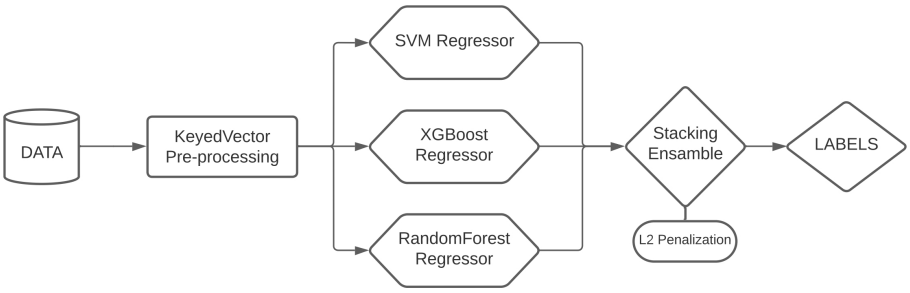

Figure 3: Diagram showing the structure of the approach chosen for subtask 3 (Degree of Prejudice Prediction).

4. Results

This section provides an overview of the results obtained from the trained models for each task. For the first task (Hurtful Humour Detection), after fine-tuning the hyperparameters for SVM and XGBoost, the best results were achieved using a soft voting ensemble method. This method involves extracting the class probabilities provided by each model when given an input and making the final decision for the class of an observation based on the highest percentage returned by the three models. By employing this approach, we obtained an average F1 score

of 0.83 using cross-validation on the test data. However, our model performed worse than expected when used on the unseen data, obtaining an F1 score of 0.744.

Our approach for the second task (Prejudice Target Detection) showed promising results when tested on the data provided by the shared task organizers, achieving a macro F1 score of 0.79. That being said, our model did not prove to be as reliable as we had first thought, as the published results revealed it had only achieved a 0.607 macro F1 score, leaving us to think if we could have done more to avoid overfitting.

As for the third task, Degree of Prejudice Prediction, we experimented with three main models. Regarding the results, our best-performing model, based on our validation process, was the original stacking model. The subsequent models involved stacking of subtasks with a prior classifier. The models submitted to the competition were the first two mentioned, achieving RMSEs of 0.99 and 1.07, respectively. Nonetheless, we believe that the model utilizing the classifier has the potential to deliver promising results with some improvements in its implementation.

5. Conclusions

The implications of our work are twofold: practical and societal. On a practical level, our models for hurtful humour detection, prejudice target detection, and degree of prejudice prediction have significant implications for content moderation on online platforms. By accurately identifying offensive content, platforms can take proactive measures to remove or flag such content, promoting a safer and more inclusive environment for users. Additionally, the ability to detect the specific targets of prejudice enables a targeted approach in combating discrimination and promoting empathy and understanding.

Furthermore, our models provide a nuanced understanding of the degree of prejudice within offensive language. This insight empowers platform administrators and policymakers to make informed decisions regarding content moderation policies and community guidelines. By quantifying the severity of prejudice, platforms can prioritize the enforcement of stricter measures for highly offensive content, thereby reducing the potential harm inflicted upon marginalized individuals or groups.

Societally, our work contributes to raising awareness about the prevalence and impact of offensive content and prejudice in online discourse. By shedding light on these issues, we foster conversations surrounding the responsible use of language, respect for diverse perspectives, and the importance of fostering inclusive communities. The availability of accurate models for detecting and quantifying prejudice can lead to a more informed public discourse, challenging harmful narratives and promoting social harmony.

Moreover, our research opens avenues for further exploration and development in the field of NLP and content moderation. As online platforms continue to grapple with the challenges posed by offensive content, our work provides a foundation for future research in refining and expanding the capabilities of models in detecting and combating different forms of harmful language.

In summary, the implications of our work encompass practical advancements in content moderation, societal awareness and discourse, and the potential for continued research and

development. By addressing the challenges of offensive content and prejudice detection, we contribute to creating more inclusive online spaces that foster respectful interactions and protect individuals from the adverse effects of prejudiced language.

Acknowledgments

We would like to express our sincere gratitude to Professor Paolo Rosso and Professor Reynier Ortega for their invaluable guidance, support, and expertise throughout the duration of this project. Their insights and knowledge in the field of natural language processing have been instrumental in shaping our research and methodologies.

We would also like to extend our appreciation to IberLEF for organizing the shared task that provided us with the opportunity to enhance our NLP skills and contribute to the advancement of the field. The platform's dedication to fostering collaboration and innovation in NLP research is commendable, and we are grateful for the opportunity to participate in this meaningful project.

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. arXiv:1810.04805.

[2] S. Castro, L. Chiruzzo, A. Rosá, Overview of the HAHA Task: Humor Analysis Based on Human Annotation at IberEval 2018, 2018.

[3] L. Chiruzzo, S. Castro, M. Etcheverry, D. Garat, J. Prada, A. Rosá, Overview of HAHA at IberLEF 2019: Humor Analysis Based on Human Annotation, 2019.

[4] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. A. Meaney, R. Mihalcea, Overview of HAHA at Iberlef 2021: Detecting, Rating and Analyzing Humor in Spanish, in: Procesamiento del Lenguaje Natural (SEPLN), volume 67, 2021, pp. 257–268.

[5] R. Ortega-Bueno, B. Chulvi, F. Rangel, P. Rosso, E. Fersini, Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO) at PAN 2022, 2022.

[6] S. Frenda, C. A., V. Basile, C. Bosco, V. Patti, P. Rosso, The Unbearable Hurtfulness of Sarcasm. Expert Systems with Applications (ESWA), volume 193, 2022.

[7] L. Merlo, B. Chulvi, R. Ortega-Bueno, P. Rosso, When Humour Hurts: Linguistic Features to Foster Explainability, in: Procesamiento del Lenguaje Natural (SEPLN), volume 70, 2023, pp. 85–98.

[8] 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.

[9] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information, arXiv preprint arXiv:1607.04606 (2016).

[10] T. Wongvorachan, S. He, O. Bulut, A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining, Information 14 (2023) 54. URL: https://www.mdpi.com/2078-2489/14/1/54. doi:10.3390/info14010054, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[11] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, ALBETO and DistilBETO: Lightweight Spanish Language Models, 2023. arXiv:2204.09145.
[12] 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.
[13] Fine-tuning a BERT model | Text | TensorFlow, 2023. URL: https://www.tensorflow.org/tfmodels/nlp/fine_tune_bert.
[14] Optuna - A hyperparameter optimization framework, 2023. URL: https://optuna.org/.
[15] El modelo secuencial | TensorFlow Core, 2023. URL: https://www.tensorflow.org/guide/keras/sequential_model?hl=es-419.
[16] T. Sharp, An Introduction to Support Vector Regression (SVR), 2023. URL: https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2.
[17] V. Margot, An original method to combine regression estimators in Python, 2022. URL: https://towardsdatascience.com/an-original-method-to-combine-regression-estimators-in-python-b9247141263.
[18] Y. Verma, A beginner's guide to stacking ensemble deep learning models, 2022. URL: https://analyticsindiamag.com/a-beginners-guide-to-stacking-ensemble-deep-learning-models/.
[19] 1.11. Ensemble methods, 2023. URL: https://scikit-learn/stable/modules/ensemble.html.
[20] Building an Ensemble Learning Based Regression Model using Python, 2023. URL: https://www.section.io/engineering-education/ensemble-learning-based-regression-model-using-python/.