Dimensionality Reduction Techniques to Detect Hurtful Humour

Pablo Sánchez García [1], Constantino Martínez de la Rosa [1]

[1] *Universitat Politècnica de València, Camino de Vera, Valencia, 46022, Spain*

Abstract

In this paper, we present a concise overview of the approach adopted by the AstonNLP team to address the challenge of Hurtful Humor Detection (HUHU). Our methodology involved utilizing a dataset comprising tweets that encompassed both humorous content and messages containing racial, homophobic, and offensive language targeting various communities. We employed several techniques to derive a meaningful representation of the tweets, enabling the implementation of robust prediction models. These techniques included word embeddings, the novel implementation of dimensionality reduction techniques (PLS-DA), sentiment analysis, and other methodologies. Based on this tweet representation, we developed distinct models to address different aspects of the challenge, namely humor detection (binary classification), prejudice detection (multi-label classification), and quantification of prejudice levels (regression task). Our findings shed light on the complex mechanisms of hurtful humor and try to provide valuable insights for mitigating its impact.

Keywords

Hurtful Humor, Humor Detection, Prejudice Detection, Word Embeddings, Dimensionality Reduction, Sentiment Analysis

1. Introduction

Hurtful humor detection refers to the identification and analysis of jokes, humor, or comedic content that can be potentially harmful, offensive, or derogatory. It involves recognizing humor that targets individuals or groups based on their race, ethnicity, religion, gender, sexual orientation, disability, or other protected characteristics, and determining whether such content crosses the line into harmful territory. The development of hurtful humor detection systems typically involves leveraging natural language processing (NLP) and machine learning techniques to automatically analyze and categorize comedic content. These systems aim to identify and flag potentially harmful jokes or humor based on specific patterns, keywords, or context. Harmful humor detection is particularly relevant in the context of online platforms and social media, where the spread of offensive or harmful content can have significant negative impacts on individuals and communities. By deploying detection systems, platforms can help enforce community guidelines, promote a more inclusive and respectful environment, and mitigate the dissemination of harmful humor.

The Hurtful Humor Detection (HUHU) challenge involves using a set of tweets to apply these techniques. The challenge consists of three tasks:

1. Task 1: Identifying whether the tweet is humorous or not.
2. Task 2a: Identifying if the tweet contains offensive comments towards groups such as LGTBIQ+, overweight individuals, women, or people of different races. In this task, it should be noted that a single tweet may contain offensive comments towards multiple groups.
3. Task 2b: Quantifying, on a scale of 0 to 5, with 0 being the minimum and 5 being the maximum, the level of hatred/prejudice present in the tweet.

IberLEF 2023, September 2023, Jaén, Spain

EMAIL: psangar1@etsinf.upv.es (A. 1); cmardel@etsinf.upv.es (A. 2)

2. Data

As explained in the introduction, the HUHU challenge encompasses multiple tasks. Therefore, the training dataset used to tackle the challenge (provided by the organizing entity) contains specific variables corresponding to each task. For the first task, we have a Boolean variable indicating whether the tweet can be classified as humor or not. For this specific task, there was a complexity element immanent to the nature of this challenge, which was a slightly imbalanced training dataset where approximately the 67% were considered as non-humorous content [1].

In the case of the second task, we have four Boolean categories indicating if the tweet is racist, sexist, homophobic, or fat-shaming. It must be stated out that these categories were not mutually exclusive. This meant that a tweet could denote multiple prejudices. Lastly, for the third task, we have a numeric variable ranging from 0 to 5 (including decimal values) representing the degree of prejudice present in the tweet. Thus, each observation in the dataset consisted of a tweet followed by five One-Hot-Encoding features (used for the first two tasks) and one continuous variable corresponding to the third task. The first dummy variable corresponded to the humor flag, that indicated whether the tweet is humoristic or not and the other four indicated separately whether the different possible prejudice categories were present in the tweet or not. We had a total of 2671 tweets. Some of the characteristics of the dataset are displayed in table 1, where it can be seen positive samples distribution for classification tasks. This is, the percentage of tweets that had humoristic content and, per each of the prejudices, the proportion of samples that targeted that collective.

Table 1

Positive Samples distribution per Classification Target

	Humor	Anti-Feminist	LGTB-Phobia	Racism	Fatphobia
Percentage of Tweets (%)	32.5	48.4	22.7	24.8	8.0

We see that the problem of imbalance was generalizable for multilabel classification task, which was certainly an issue to take into consideration when developing our solution. The same way, it was considered to proceed with oversampling techniques by using LLM such as GPT-3 or LLaMA to generate synthetic tweets [2] with the appropriate characteristics to balance our dataset, but the idea was abandoned due to that the lack of time to experiment and infer if using this technique could induce biased classification models or conversely, to improve the performance. This is a pendent subject which remains to be studied in further steps of the project.

On the other hand, addressing the regression task, it also had to be considered that the variable to infer had a subtle different distribution depending on whether the tweet had humor or not. We were noticed about this issue by the organization of the challenge, as it can be seen in Figure 1.

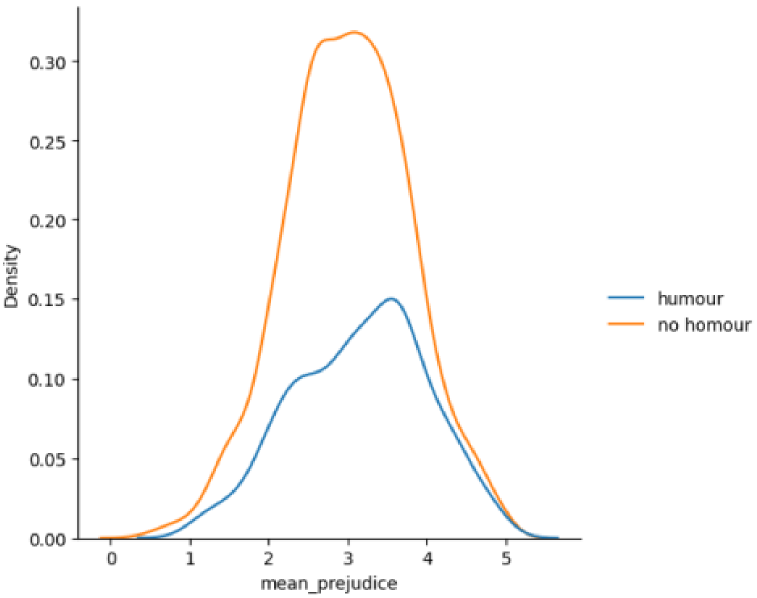

Figure 1: Degree of Prejudice Distribution per Humoristic Content class [1]

2.1 Tweet Preprocessing and Embedding Creation

The data is raw tweet data from Twitter and hence data cleaning and pre-processing is required. When tackling preprocessing, we refer to tokenizing tweets, inferring exogenous variables that could lead to a more representative way to use the samples to address the different challenge tasks. The same way, it involves generating flexible state-of-the-art embedding representations of tweets which are capable of capturing relationships between different words including their syntactic & semantic relationships [3].

We developed our own tokenizer; this involved using special packages from ‘spacy’ (Spanish corpus-based transformer: ‘es_dep_news_trf’) and ‘nltk’ Python libraries which helped us to delete useless textual information by lemmatizing and normalizing the text, deleting stop words, punctuation, and special characters such as emojis. Tweets were lowercased as well and the ‘numlet’ library was incorporated to traduce integers and decimal numbers into typed numbers (e.g. 500 to ‘quinientos’).

After tokenizing the tweet, we created a 400-dimensional embedding based on skip-gram Word2Vec model with a window size of 6. We chose it in place of other embedding representations because of its capability for contextual understanding, to capture semantic similarity, and because it offers computational efficiency. These strengths could enable an effective detection of hurtful humor by considering the tweet's surrounding words, encoding semantic information, and handling diverse linguistic variations. Nevertheless, we will counter its weaknesses by introducing additional variables that will be described in the following sections.

2.2 Exogenous Variables

Our approach on this challenge was not mainly focused on reaching the highest model performance. This has been commonly achieved in similar challenges by fine-tuning state-of-the-art transformers [4]. However, in the name of eXplainable Artificial Intelligence (XAI) we consider that behind a good performing methodology there should be some interpretability strategies that could lead us to understand the underlying linguistic, syntactic or stylistic mechanisms of hurtful humor, while having a decent performance [5][6]. To provide a procedure able to ease interpretation, we opted for extracting additional features which could feature the before mentioned aspects.

2.2.1 Sentiment and Emotions Analysis Features

Introducing this kind of variables might seem conflicting when looking for interpretability, at least because of the models that are used to infer this information. Nonetheless, the information they provide is quite interpretable and feasible to understand, and it has been proved that introducing sentiment analysis information could lead to a better understanding of language classification model decisions in other contexts [7]. To create emotion features for each tweet, we used the XLM-RoBERTa fine-tuned model, which was presented in the EmoEvalEs competition, part of IberLEF 2021 Conference. This model returns the probability for a given tweets to express surprise, sadness, anger, disgust, or fear. In Figure 2, it can be seen the comparison between the mean probability for each of the emotions depending on whether the tweet had humoristic content or not.

We saw that including these features could lead to better explanations since for some emotions, the difference in the mean probability for humoristic and non-humoristic tweets was significant. The same way, as it can be seen in Figure 3, we noticed that this observation was applicable when comparing tweets depending on the prejudice strength variable. In fact, the sentiments which difference between groups is higher tend to be the same than when using humor as discriminatory variable. These were joy, anger and disgust which are common sentiments which presence could lead to differentiate between categories of tweet in terms of offensiveness or discrimination.

To complete the meaningful impact that sentiment analysis features could have, we introduced a polarity feature that determines the degree of positiveness of a tweet in a range of 0 to 1. The combination of both types, emotion and polarity is displayed in Figure 4, where it can be inferred that tweets with positive polarity imply having higher probabilities of emotions like joy or surprise, conversely to negative content polarity, which is usually related to higher probabilities of anger or sadness. The polarity feature was inferred using the Python library ‘sentiment_analysis_spanish’.

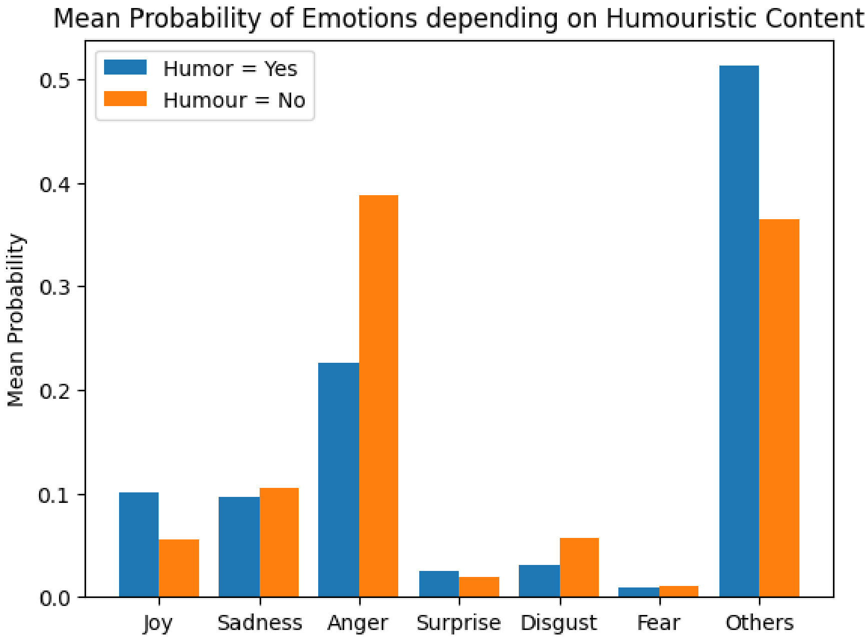

Figure 2: Mean Probability of Emotions depending on Humoristic Content.

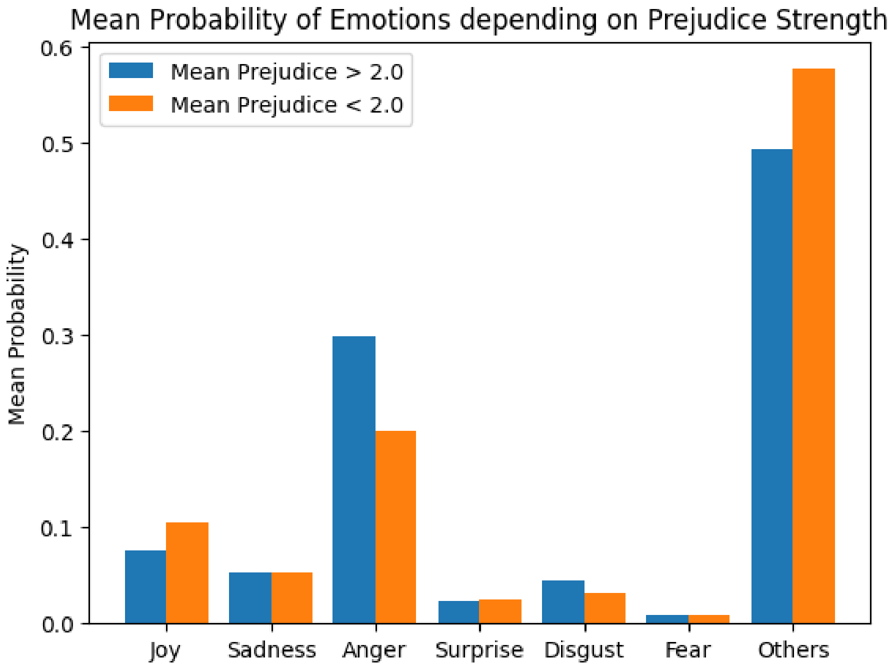

Figure 3: Mean Probability of Emotions depending on Prejudice Strength.

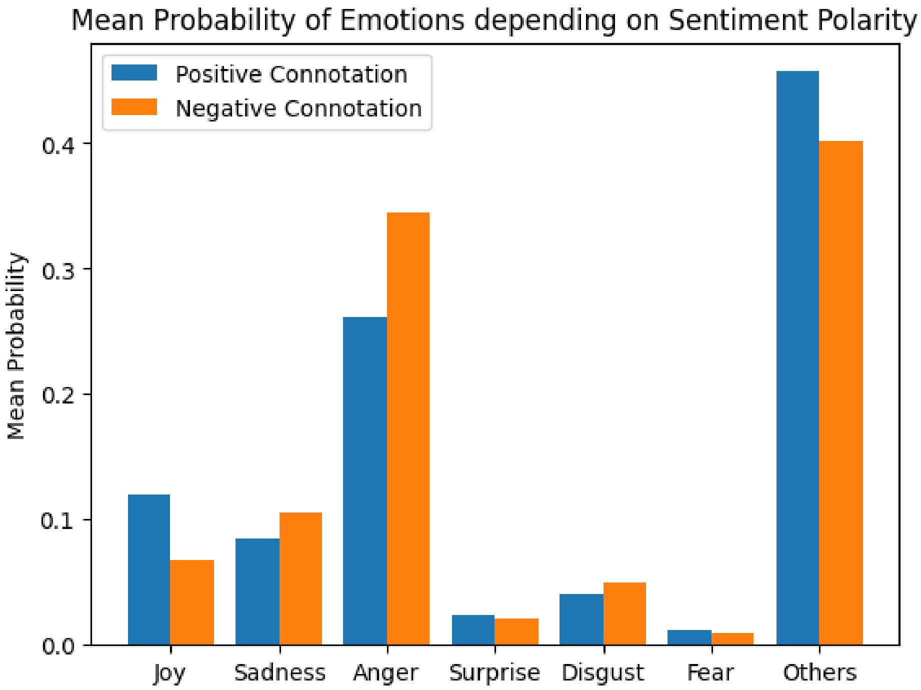

Figure 4: Mean Probability of Emotions depending on Sentiment Polarity.

2.2.2 Dimensionality Reduction Features over BoW

The principal ‘novelty’ of our approach is based on dimensionality reduction methodologies, which are commonly used in the context of process monitorization. Its presence in other contexts such as NLP is promising, since are models capable of condensing high dimensional data information into a few features. Our conception has a different point of view of textual latent information, differing from other approaches that have been taken [8][9].The methodology consists in, firstly, obtaining the Bag of Words of the training part of the dataset, in our case, the 67% of the tweets. After it, we applied PLS-DA (Partial Least Squares Discriminant Analysis) to that matrix and having as target variables the One-Hot-Encoding features which are subject to be inferred in the classification tasks (both binary and multilabel) and obtaining 200 principal components. To comprehend why we chose this methodology, we will briefly state out some theoretical basis about PLS.

It is a multivariate statistical model, which shares its fundamentals with PCA. Nonetheless, it is supervised and has two objective functions simultaneously. To extract as much pertinent information from both sets of variables as feasible, the first objective function concentrates on maximizing the covariance between the predictor variables and the response variables. Similar to the variance-maximizing objective in PCA is this function. The within-set covariance of the predictor variables, which aids in reducing predictor redundancy, is related to the second objective function in PLS. This objective function makes sure that multicollinearity problems are avoided while the created latent variables accurately reflect the predictor variables. To wrap up, PLS finds a set of latent variables that optimize the link between the predictor variables and the response variables while minimizing redundancy within the predictor set by optimizing these two objective functions.

By this way, we could create latent variables over BoW that sum up the relevance of some words in order to predict prejudice categories and hurtful humoristic content. In this sense, one of the main advantages and reasons for opting for this method resides in its interpretability. This comes because the resulting values (technically speaking, scores) of the tweets in the latent features can be used to discriminate between the groups of the target variables. Just to give a quick insight into this in Figure 5, we display the scores of the tweets in the first two most relevant components and colored by if they were targeted to have prejudice against feminist or women or not. We can see that the first latent variable is able to discriminate between tweets targeting women and the ones which not. This plot shows the potential of using dimensionality reduction techniques in NLP and it is the reason why we opted to implement it. The same way, if we combined this interpretation with weight plots, we could find out which words have more relevance to discriminate between groups.

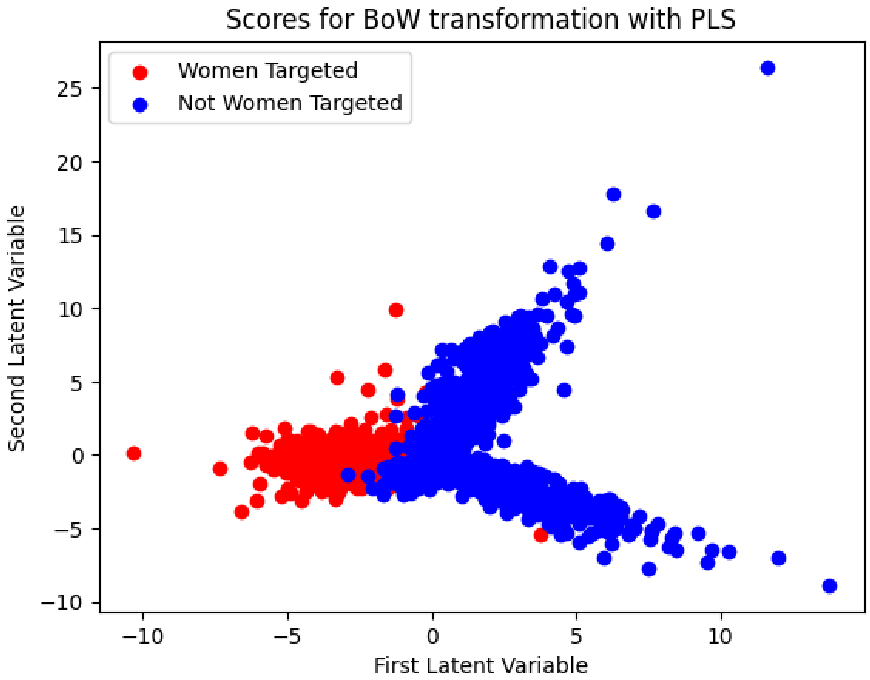

Figure 5: BoW scores after PLS-DA transformation

2.2.3 Capturing Humoristic and Emphatic Language Style

Spanish is a complex language, and its native speakers are usually very emotional when using it, even when writing it. This is usually reflected at their syntaxis and style, which people use to express their sentiments. Due to this, we though that it could be interesting to introduce some variables that can reproduce some relevant information about the tweet emphasis, polarity or style in relation to humor or prejudices. These features were introduced as One-Hot-Encoding features and are summed up in Table 2. The same way, Twitter social media allows users to communicate using some special features which are mentioning, retweeting, using emojis, using hashtags to talk about specific topics etc. For some contexts, this information might be relevant and can denote emotional emphasis put into the tweet, so we introduced the use of them with One-Hot-Encoding features as well.

Table 2

Humoristic and Emphatic Language Style Features and Description

Feature Name	Feature Description
Exclamation	If abusive use of exclamation is committed, this feature is turned to 1. We determined that exclamation is abusive when exceeds in more than two exclamation signs in just a tweet.
Capital Letters	If abusive use of uppercase letters is applicable, this feature is turned to 1. Due to that tweets are not longer than 250 characters, when there are more than 10 capital letters, we consider that the use of them is abusive.
Retweet	If the tweet contains the RETWEET interaction word, this feature is turned into 1.
Emojis	If the tweet contains emojis, this feature is turned into 1.
Hashtag	If the tweet contains the HASTAG interaction word, this variable is turned into 1.
URL	If the tweet contains an URL, this feature is turned into 1.
Mention	If the tweet contains the MENTION interaction word, this feature is turned into 1.

2.2.4 Common Vocabulary to Describe Prejudices and Humor

In order to summarize and introduce common vocabulary used for humoristic purposes and for expressing prejudices, we created dummy features, each one of them associated with the one of the 25 most common words used in the training part of the dataset for tweets classified as humoristic or that revealed some prejudice of the ones that are matter of interest in this challenge. If any of these words appeared in a tweet, the value of the variable was switched to 1. This vocabulary revealed to be promising to discriminate between prejudices as it included offensive words such as “feminazi”, “puto”, “mierda” or “maricón” among others.

3. Models and results

In problems of this nature, it is challenging to choose a model without conducting any form of prior experimentation. Therefore, the approach we have adopted to select the models has been to test those that we believed could best predict and choose the ones that yielded the best results.

3.1 Task 1: Hurtful Humour Detection

For the first task, we tested that the two best models were Support Vector Machines (SVM) and Neural Networks. We attempted to employ various ensemble methods such as bagging, voting, boosting, or stacking to achieve improved model performance. The results obtained revealed that the stacking approach yielded the best performance, using 3 SVM models as base classifiers and a neural network as the meta-classifier. The achieved result for the metric used to evaluate this task (F1 Score) is 0.8519, indicating that the model achieves highly accurate predictions. In the following table we provide a quick summary of the other models that were tested together with the performance metric.

Table 3
Task 1 Training Models’ Performance

	Voting	Bagging	Random Forest	AdaBoost	XGBoost	Stacking
Binary F1-Score	0.8497217	0.84	0.8145	0.83	0.817	0.852

3.2 Task 2a: Prejudice Target Detection

For the second task, as it involves multi-label classification, we considered using K-Nearest Neighbors (KNN) and Random Forest, in addition to models identical to those explained in the previous section. The results obtained revealed that the model that performed the best is the Multi-Layer Perceptron, utilizing 6 hidden layers with 150 neurons each, employing the hyperbolic tangent (tanh) activation function and a learning rate of 0.0002. The achieved result for the metric used to evaluate this task (Macro F1 Score) is 0.9241. The same way we did with task 1, we can see the training performance of our tested models in Table 4.

Table 4
Task 2 Training Models' Performance

	Random Forrest	Support Vector Machine	KNN	Decision Trees	Multilayer Perceptron
Macro F1-Score	0.86	0.88	0.79	0.9	0.92

3.3 Task 2b: Mean Prejudice Degree Inference

For the final task (2b), as it involves a regression problem, we considered using models such as linear regression, MLP (Multi-Layer Perceptron), and some ensemble methods, particularly boosting and Random Forest. The results obtained indicate that the model that performs the best is Random Forest, utilizing 50 estimators and a maximum depth of 15. The achieved result for the metric used to evaluate this task (RMSE - Root Mean Squared Error) is 0.75. The summary of training results is displayed in Table 5

Table 5
Task 3 Training Models' Performance

	Voting	AdaBoost	XGBoost	Random Forest	Multilayer Perceptron
RMSE	0.78	0.76	0.79	0.75	0.92

4. Conclusions and Future Work

Despite the results of the challenge did not materialize the effort put and the performance that our models had in the training phase. However, we believe that are methodology, with certain improvement and corrections can have a meaningful impact on hurtful humor detection. After presenting the results, we noticed that when preparing the final model for predicting the test tweets, the BoW of the training data was not prepared properly, what lead us to a very likely overfitted models. Nonetheless, we are eager to keep improving our methodology introducing PoS tagging applications, using transformers embedding representations and adding subjectivity and syntactical features. As future work, as it was stated out in Section 2, the problem of unbalanced classes is very common in tasks like hurtful humor detection. To counter this issue, we will work to deliver a methodology based on state-of-the art LLM (such as GPT, Bard or LLaMa). This procedure will mainly consist in training the models with some samples from the shared task, allowing them to learn the structure of the samples, the style and slang of the tweets and the patterns that make a tweet humoristic, offensive or both at the same time. After it, we will try to generate positive samples for those classes, in this case, prejudices that were in a clear imbalanced situation. If this worked properly, we would dispose from synthetic tweets that could make our models' performance more reliable. The official results of the submitted shared tasks were 0.2 and 0.3 of F1-Score for the first two challenges, while the outcome for the third sub-task was 1.1 of RMSE.

5. Acknowledgements

We want to express out gratitude to Paolo Rosso and Reynier Ortega Bueno, our NLP subject instructors, for their invaluable advice and assistance in dealing with the Hurtful Humor Challenge.

Their expertise and support have been essential to our development as students, and we really appreciate the opportunity to learn from them.

References

[1] Labadie-Tamayo, R., Chulvi, B., & Rosso, P. (2023, September). Everybody Hurts, Sometimes. Overview of HUrtful HUmour at IberLEF 2023: Detection of Humour Spreading Prejudice in Twitter. *Procesamiento Del Lenguaje Natural (SEPLN)*, *71*.

[2] L. Faulbruck, A. Muminovic, T. Peschenz, "Generating Synthetic Comments to Balance Data for Text Classification", Seminar Information Systems, Humboldt-Universitat zu Berlin (2020),

[3] S. Chandran, Text Representations for Language Processing, Volume 2, Towards Data Science, 2020

[4] A. Cherif, "BERT-based ensemble learning for multi-aspect hate speech detection", Springer Link, 2023

[5] Frenda S., Cignarella A., Basile V., Bosco C., Patti V., Rosso P. The Unbearable Hurtfulness of Sarcasm. In: Expert Systems with Applications (ESWA), vol. 193 https://www.sciencedirect.com/science/article/pii/S0957417421016870?via%3DihubS,

[6] L. Kerem Senel, "Semantic Structure and Interpretability of Word Embeddings" (2018) arXiv:711.00331v3

[7] G.A. Marchellim, Y. Ruldeviyani, "Sentiment analysis of hate speech as an information tool to prevent riots and environmental damage" (2021), IOPScience, doi:10.1088/1755-1315/700/1/012024

[8] B. Fayyazuddin Ljungberg, "Dimensionality reduction for bag-of-words models: PCA vs LSA", Standford University

[9] Y. Kim, S. Wiseman, A.M. Rush, "Deep Latent Variable Models of Natural Language" (2019), Harvard University

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