Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 383-395 recibido 25-06-2023 revisado 10-07-2023 aceptado 19-07-2023

Everybody Hurts, Sometimes Overview of HUrtful HUmour at IberLEF 2023: Detection of Humour Spreading Prejudice in Twitter

Todos herimos, a veces Resumen de HUHU en IberLEF 2023: Detección de Humor que Difunde Prejuicios en Twitter

Roberto Labadie Tamayo,[1] **Berta Chulvi,**[1,2] **Paolo Rosso**[1]
[1]PRHLT Research Center, Universitat Politècnica de València
[2]Social Psychology Department, Universitat de València
rlabtam@posgrado.upv.es, berta.chulvi@upv.es, prosso@dsic.upv.es

Abstract: Humour is an efficient strategy to spread prejudice because, most of the time, it evades moral judgement. However, it perpetuates stereotypes and doing so justifies discriminatory acts. At HUHU we propose a frame to study how humour is used to discriminate minorities and to analyse their interplay with the degree of prejudice expressed against specific groups. To this end, we provide a corpus of prejudiced tweets in Spanish annotated with the presence of humour, its prejudice degree and the targeted groups: women and feminists, the LGBTI+ community, immigrants and racially discriminated people, and over-weighted people. This paper analyses the results achieved by the 46 teams that participated in HUHU.

Keywords: hurtful humour, prejudice, minority groups, Twitter.

Resumen: El humor es una estrategia eficiente para propagar prejuicios porque, la mayoría de las veces, elude el juicio moral. Sin embargo, perpetúa estereotipos y, al hacerlo, justifica actos discriminatorios. En HUHU roponemos un marco para estudiar cómo el humor se utiliza para discriminar a las minorías y analizar su interacción con el grado de prejuicio expresado contra grupos específicos. Con este fin, proporcionamos un corpus de tweets prejuiciosos en español anotados en cuanto a la presencia de humor, su grado de prejuicio y los grupos de: mujeres y feministas, comunidad LGBTI+, inmigrantes y personas discriminadas racialmente, así como personas con sobrepeso. Este artículo analiza los resultados obtenidos por los 46 equipo que participaron en HUHU.

Palabras clave: humor hiriente, prejuicio, grupos minoritarios, Twitter.

1 Introduction

Sometimes people hurt other people in a creative way: they use humour. More frequently than not, the target of the joke is part of a minority or discriminated group. In this case, humour is used for the expression of prejudice, defined as "the negative pre-judgement of members of a race or religion or any other socially significant group, regardless of the facts that contradict it" (Jones, 1972). The main fact that contradicts this pre-judgement is that social groups, whatever they are, are not homogeneous either in their characteristics or in the way they act. When we present them as homogeneous, we make use of a stereotype (Lipmann, 1922).

Moreover, where minorities fight to raise egalitarian treatment, we can observe that humour becomes a space in which prejudiced attitudes and stereotypes are maintained. In fact, recent research in NLP shows that one of the features that distinguishes offensive jokes from non-offensive ones is the presence of negative stereotypes and ethnic slurs (Merlo et al., 2023). Some authors consider that prejudicial messages make use of humour to avoid the moral judgement that penalises discrimination (Ford and Ferguson, 2004; Ford

ISSN 1135-5948 DOI 10.26342/2023-71-30

et al., 2008). But, despite its inoffensive appearance, this humour is not harmless; it has a deterrent effect and serves as a social control mechanism: people don't want to be ridiculed, so they try to avoid what is laughed at in a given society (Freud, 1960; Billig, 2005). In addition, the effects of these offensive jokes spill over into other spaces with far more serious consequences. For instance, research about sexism has demonstrated that for men exhibiting high levels of hostile sexism (Glick and Fiske, 1996), sexist humour can have important social consequences, such as rape proclivity (Romero-Sánchez et al., 2017). Furthermore, it has been observed that when prejudiced content is nuanced in humour, individuals targeted by prejudice are more likely to endorse and internalise such expressions (Miller et al., 2019).

Due to the relevant implications of all praxes that discriminate against minorities, this task aims to shed light on the complex interplay between prejudiced language, minority groups and humour when it serves to convey hurtful content.

2 Related Works

Dealing with figurative language has placed many challenges for NLP. Most of the complexity comes with the contextual and sociocultural concepts involved in comprehending this kind of communicative device. Humour, as an instance of this form of communication, has been investigated in several shared tasks from different perspectives, ranging from classical humour recognition to more fine-grained analysis focusing on the underlying mechanism that provokes humour, the target of the jokes, and its offensive and hurting aspects.

At the SemEval evaluation forum, during the last years, humour in English was addressed from a computational perspective in: (i) 2017 in Task 6 #HashtagWars: Learning a Sense of Humour (Potash, Romanov, and Rumshisky, 2017), where participants were asked to detect the top funniest tweets from a given set; (ii) 2020 in Task 7 on Assessing Humour in Edited News Headlines (Hossain et al., 2020), where the aim was investigating the presence of humour after local modification on headlines; and (iii) 2021 in Task 7 where a HaHackathon was organised for Detecting and Rating Humour and Offence (Meaney et al., 2021). This was the first shared task with the aim of detecting offensive language in humorous messages; one of the subtasks aimed at predicting the rate of offence in texts, although from a general perspective without focusing on the expression of prejudice.

At the IberLEF evaluation forum, humour in Spanish was addressed in the HAHA shared task on: Humor Analysis based on Human Annotation in (i) 2018 (Castro, Chiruzzo, and Rosá, 2018) and (ii) 2019 (Chiruzzo et al., 2019) to detect and rate humorous messages in a scale from 0 to 5; and (iii) 2021 HAHA focused more on a fine-grained analysis of humour where the organizers aimed at detecting the linguistic device employed to convey humour: e.g., irony, wordplay, hyperbole, etc., as well as the content of which the joke is based on distinguishing among racist jokes, sexist jokes, dark humour, dirty jokes, and others categories (fifteen in total).

Finally, in the PAN Lab at CLEF in 2022, the IROSTEREO shared task was organized on Profiling Irony and Stereotype Spreaders on Twitter (Ortega-Bueno et al., 2022). Participants were asked to determine whether or not an author of a Twitter feed in English is keen to spread stereotypes via the usage of irony. In this task, stereotypes were approached as a set of widespread beliefs associated with a group category presented in a homogeneous way.

Although some of the previous shared tasks investigated the use of offensive language in humour or the dissemination of stereotypes using irony, and previous work was done to study the hurtfulness of other types of figurative language such as sarcasm (Frenda et al., 2022). To the best of our knowledge no previous work assessed the use of humour to spread prejudice against minorities. Therefore, we propose HUHU as the first shared task in Spanish focusing on studying humour in prejudicial messages against: (i) *women and feminist*, (ii) *the LGBTI+ community*, (iii) *immigrants and racially discriminated people*, and (iv) *overweighted people.*

3 Dataset

We constructed the dataset together with 80 students of psychology that manually tracked down Twitter accounts to study the characteristics of users who spread hate speech using humour. This characterisation comprised identifying hurtful texts targeting various societal groups, including women, feminists, the LGBTI+ community, immigrants, racially discriminated people, politically aligned population segments, vegans, and other stereotypically perceived groups. Based on the obtained results, we proceed with our information retrieval strategy using 898 user accounts. Among all the targets identified in the preliminary corpus, in the context of this shared task, our interest lies in studying the four aforementioned groups (section 2). Hence, we considered this initial set of tweets as a corpus of toxic language containing instances belonging to a positive or negative macro-classes from 898 distinct Twitter users. The positive macro-class comprised the four minority groups under investigation, while the negative class comprised the remaining groups.

From the Twitter accounts posting tweets belonging to the positive macro-class, we retrieve the last 1000 tweets posted after January 1st, 2020, taking them as potential prejudicial speech spreaders. The latter process yields us a set of roughly 80 thousand instances; we proceed to filter them by employing discriminative keywords representing the topic modelled by the interest macro-class. To this end, we fuse a set of keywords obtained by: (i) KeyBERT (Grootendorst, 2020), (ii) YAKE (Campos et al., 2020), and the top 100 terms[1] according to the information gain in the distribution of the two classes.

3.1 Annotation Process

The filtering yielded a reduction of nearly 30 thousand tweets. For each account, we observed duplicated instances. Inspired by (Chiruzzo et al., 2021) we constructed graphs interconnecting tweets for each user and grouped together those pairs with a Jaccard similarity above 0.7 by a cut-off. Later, we employed the Grivan Newman algorithm (Girvan and Newman, 2002) to find communities of similar texts and provided annotators with an ordination according to this to speed up the detection and removal of duplicated instances.

Annotation was carried out in two main steps. The first annotation stage consisted in taking the majority vote from 3 annotators who decided whether the tweets actually conveyed prejudicial content and whether they perceived any humorous intention by answering yes or no to the two following questions:

1. *Does this tweet express prejudice towards one of the following minorities: women or feminists, immigrants or racialized groups, LGBTI+ or other sexual minorities, overweight people?*
2. *Does the tweet's author intend to be humorous?*

Two teams consisting of one male and two female university students were hired by the Universitat Politècnica de València (UPV) to accomplish this annotation task. From this step, just prejudicial tweets were kept, and several rounds were done considering all the Twitter accounts, giving a larger representation of those which seemed to use humour to convey prejudice in the initial set of manually annotated tweets. The latter was due to the poor balance detected in the preliminary corpus exploration. Once we had our potential dataset, a second annotation step was approached by a team of five annotators (three female and two male students) hired by the UPV. They were asked to identify the minority group being the target of prejudice and for each of them the prejudice degree in *discrete* scale ranging from 1 to 5 where 1 means a lower prejudice degree and 5 the opposite.

First, the overall degree of prejudice towards each minority in a given instance was determined by the average scoring provided by the five annotators (Equation 1). Subsequently, the prejudice score of the tweet was defined as the mean prejudice value towards *targeted* minorities as in Equation 2:

$$A_k^{(i)} = \frac{1}{5}\sum_j T_{jk}^{(i)} \qquad \forall k, i \quad (1)$$

$$S^{(i)} = \frac{1}{\sum_k \mathbb{1}(A_k^{(i)} > 0)} \sum_k A_k^{(i)} \qquad \forall i \quad (2)$$

Here, $T_{jk}^{(i)}$ represents the scoring provided by the j^{th} annotator to the k^{th} target in the i^{th} tweet and $A_k^{(i)}$ is the average score of prejudice for the i^{th} tweet under k^{th} target.

Examination of the mean prejudice distribution among annotators with the

[1] Here we exclude punctuation marks, stop words, and other semantically meaningless structures

Kolmogorov-Smirnov test yields a non-normal distribution of the degree of prejudice in all targeted groups ($p < 0.001$). This skewed data distribution leads to a low level of agreement among different raters when using traditional Inter-Rater Agreement (IRA) measures (Eugenio and Glass, 2004). To address this issue, we employ Gwet's AC1 measure of IRA (Gwet, 2008). Table 1 shows the IRA for each prejudiced target individually for the whole set of instances and for prejudicial texts nuanced with humour.

	G_1	G_2	G_3	G_4
All	$0.49_{0.02}$	$0.79_{0.01}$	$0.81_{0.01}$	$0.94_{0.01}$
Humor	$0.51_{0.03}$	$0.85_{0.02}$	$0.80_{0.02}$	$0.90_{0.02}$

Table 1: Gwet's AC1 measure of IRA across annotators from the second phase for each prejudiced minority.

From here, we can observe particularly low IRA values for the target related to the women and feminist movement. This difference is intriguing because all tasks are subjective tasks regarding the definition given by (Wong, Paritosh, and Aroyo, 2021) of subjective tasks with genuine ambiguity judging toxicity of online discussions (Aroyo et al., 2019), which typically reach values of IRA ranging between 0.2 and 0.4 in their annotation process.

3.1.1 HAHA 2021

In the first annotation step, we included tweets from the dataset proposed in (Chiruzzo et al., 2021). In this corpus of tweets in the Spanish language, the authors included annotation of the jokes' target, i.e., if somebody is being laughed at (the butt of the joke) and who that entity is. We filter out positive examples of humour comprising entities related to the studied minority groups, incorporating 1402 instances to the annotation flow, which were finally reduced to 503 in the final dataset.

3.2 Dataset Statistics

After both annotation stages, the final distribution of tweets remained as shown in Table 2.

From the columns denoted with emojis 😆 and 😐, representing humorous and not humorous instances respectively, we can notice an important imbalance in the final dataset, nevertheless this relation of quantities does not suppose a critical scenario for most machine learning systems.

Source	😆	😐	G1	G2	G3	G4
Crawled	607	2323	1652	791	753	169
HAHA	518	1	328	66	89	100
Total	1125	2324	1980	857	842	269

Table 2: Final dataset statistics.

On the other hand, it is important to note that a single tweet may contain prejudice towards multiple minorities. Therefore, the values in columns G_1 to G_4 represent the sizes of sets that are not mutually exclusive.

Our analysis revealed that when targets of prejudice are combined, the most common pattern was an overlap of at most two classes. However, it is crucial to highlight that this overlapping was not observed in the majority of instances. For a more comprehensive understanding of this phenomenon, refer to Figure 1, where we specifically focused on pairwise relations.

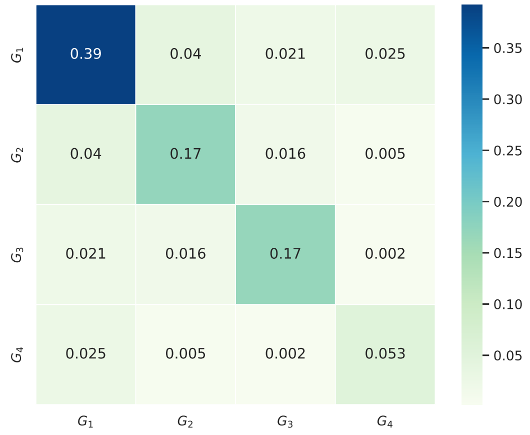

Figure 1: Pairwise co-occurrence of minorities being the target of prejudice.

Based on the graph, it also becomes evident that instances of prejudice against women (G_1), which is the dimension with the highest controversy in terms of IRA, are over-represented in our dataset compared to the other minorities. This observation highlights a converse situation where the minority represented by (G_4) is disproportionately weighted in our dataset.

Regarding the proportion of humorous and non-humorous messages targeting each minority, we noticed a consistent pattern in abusive tweets across most groups, except for the one targeting overweight individuals (G_4). Specifically, the quantity of humorous and non-humorous messages in this particular group was nearly equal.

Finally, we investigated how levels of prejudice, as measured by Equation 2, are distributed in both positive and negative cases of humour. The distribution of these prejudice levels, as depicted in Figure 2, reveals that there is a shift towards more hurtful messages among tweets that convey jokes. This observation holds some implications for the dataset, giving empirical evidence that humour when used to make people laugh at certain aspects of a minority group, can amplify the hurtful connotations of prejudiced messages. This phenomenon is aligned with the research that points out the potential impact of humour in reinforcing and perpetuating prejudice.

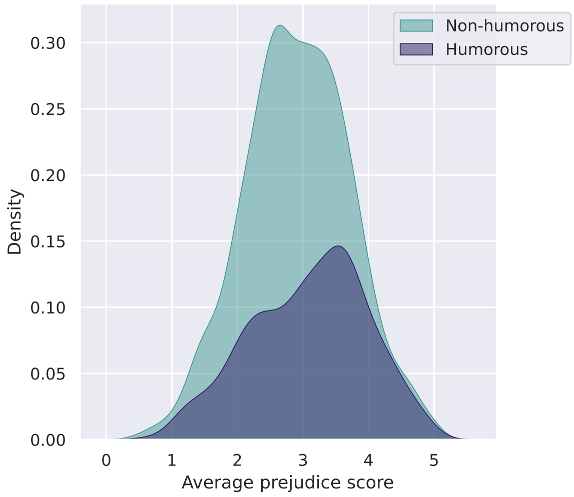

Figure 2: Degree of prejudice in humorous/non-humorous texts.

3.3 Provided Partitions

Before partitioning, we carried out some prepossessing steps to provide the dataset to the participants; essentially, we masked URLs and mentioned users, thereby protecting sensitive information. Regarding the hashtags, we designed a reduced set of specific terms expressing laugh, e.g., *haha*, *jeje*; or the explicit intention of humour, e.g., *rie*, *humor*. Then, we masked hashtags containing those terms, while the remaining hashtags were segmented using the ekphrasis library proposed by (Baziotis, Pelekis, and Doulkeridis, 2017).

When constructing the training and test datasets, we ensure they closely reflect the distributions observed in the corpus. This is achieved by maintaining a proportional split of approximately 75% for the training set and 25% for the test set. The specific distribution for each category can be seen in Table 3. We keep in the training set instances at most containing prejudice toward two minorities simultaneously, given the reduced number of tweets with three or more targets.

Source	😆	😐	G1	G2	G3	G4
Train	869	1802	1292	607	664	214
Test	256	522	688	250	178	55

Table 3: Distribution into training and test set.

We also ensure that the distributions depicted in Figure 2 were preserved in training-test partitioning, including the skewness towards more hurtful content for jokes.

4 *Tasks Description*

Three subtasks were proposed to assess the hurtful humour observed in the dataset and the dimensions of the prejudice.

HUrtful HUmour Detection

Subtask 1 consisted in determining whether a prejudicial tweet is intended to cause humour. Participants had to distinguish between tweets that use humour to express prejudice and tweets that express prejudice without humour. Systems were evaluated and ranked employing the F1-score over the positive class.

Prejudice Target Detection

In **Subtask 2a**, considering the minority groups analysed, participants were asked to identify the targeted groups on each tweet as a multi-label classification task. To this end, systems were evaluated using the macro-F1 measure taking into account the unbalance observed in section 3.2.

Degree of Prejudice Prediction

Subtask 2b consisted of predicting on a continuous scale from 1 to 5 to evaluate how prejudicial the messages are on average among minority groups. It was evaluated employing the Root Mean Squared Error.

4.1 Baselines

We present three baseline models to establish a comparative framework. These models encompass different approaches, with one utilising a classic machine learning approach and the other two leveraging state-of-the-art transformer architectures.[2]

The first baseline model utilises a straightforward linear classification technique, employing a support vector machine based on bags of 3-grams of characters. The second involves fine-tuning a pre-trained BETO model (Cañete et al., 2020), which is based on the BERT architecture and trained on Spanish texts. Finally, we incorporated a fine-tuning version of the BLOOM model (Scao et al., 2022), a multilingual model, on its bloom-1b1 variant.

During the transformer-based models fine-tuning process, we employed the RMSprop algorithm (Hinton, Srivastava, and Swersky, 2012) for parameter optimisation. We gradually increase the learning rate from shallower layers to deeper ones (Howard and Ruder, 2018), starting from 1e-5 incrementally adjusted by a factor of 0.1 for each subsequent layer. We use a batch size of 32 examples for BETO and 16 for BLOOM.

In addition to the aforementioned models, we explored a more naive approach for the classification task. This approach involved predicting the positive class for subtask 1 and assigning ”true” labels to all four classes in the multi-label subtask 2a. This serves as a baseline to compare the performance of the more sophisticated models.

5 *Participating Systems*

Participants were allowed to send up to two submissions for each subtask. The usage of external data was restricted due to the presence of some instances from the HAHA dataset.

For the final evaluation, 46 teams of the 77 registered in HUHU, made at least one submission. Table 4, Table 5 and Table 6 show the top-ranked systems along with the results of the proposed baselines for each subtask respectively.[3]

The majority of the participating teams preprocessed the tweets of the dataset and employed traditional Machine Learning (ML) as well as Deep Learning (DL) models, specifically transformer-based architectures.

Team	run	F_1-score ↑
RETUYT-INCO	1	0.820
BERT 4EVER	2	0.799
CISHUHUC	1	0.796
BLOOM-1b1		*0.789*
MosquitosBiased	1	0.784
HUHU-RMA-2023	1	0.782
amateur37	1	0.781
MJR	1	0.779
JPK	2	0.778
INGEOTEC	1	0.775
CAVIROS	2	0.774
JUJUNLP	1	0.772
mesichiquito	1	0.766
LaVellaPremium	2	0.764
BETO		*0.759*
SVM-3gram-char		*0.679*
allTrue		*0.492*

Table 4: Top-ranked systems for subtask 1.

Team	run	Macro-F_1 ↑
JUJUNLP	1	0.796
Joe	1	0.783
Ratolins	1	0.778
RETUYT-INCO	1	0.773
BETO		*0.760*
BERT 4EVER	2	0.758
LaVellaPremium	1	0.753
MosquitosBiased	1	0.746
FENRIRFENIX	1	0.741
amateur37	1	0.739
Patata	2	0.732
mesichiquito	1	0.729
CAVIROS	2	0.727
Chincheta	1	0.722
SVM-3gram-char		*0.603*
allTrue		*0.482*

Table 5: Top-ranked systems for subtask 2a.

5.1 Preprocessing

As part of their preprocessing strategy, several teams employed various techniques such as converting all tweets to lowercase, lemmatising or stemming words, removing stopwords, eliminating punctuation marks and special characters (Aguirre and Cadena, 2023; Árcos and Pérez, 2023). Some teams even experimented with removing emojis, which could potentially aid in detecting humorous intentions (García and de la Rosa, 2023). In addition, other teams eliminated URL, MENTION, and HASHTAG tokens introduced during the data partitioning pro-

[2] Source code can be found here https://github.com/labadier/HUHU

[3] Full Ranking can be found in the shared task web page at https://sites.google.com/view/huhuatiberlef23/results

Team	run	RMSE ↓
M&C	1	0.855
Huhuligans	1	0.874
BETO		*0.874*
MosquitosBiased	1	0.881
Zeroimagination	1	0.881
CIC-NLP	1	0.881
ByteMeIfYouCan	1	0.887
cocalao	1	0.890
mesichiquito	1	0.891
MJR	1	0.893
MJR	2	0.893
FENRIRFENIX	1	0.895
LaVellaPremium	2	0.898
Climent	1	0.899
SVM-3gram-char		*0.907*
BLOOM-1b1		*0.915*

Table 6: Top-ranked systems for subtask 2b.

cess, discarding any remaining unmasked instances.

Moreover, specific teams introduced word correction by replacing words not found in the embedding dictionaries with the nearest element based on the Levenshtein distance criterion. Another noteworthy preprocessing step undertaken by the team (García-Díaz and Valencia-García, 2023) involved the removal of jargon proper from social networks.

Conversely, another group of participants, primarily those proposing systems based on transformer-based models, opted to tokenise the tweets directly (Kaoshik and Kather, 2023; Peng and Lin, 2023; Inácio and Oliveira, 2023).

5.2 Text Representation and Models

Most of the proposed systems relying on machine-learning methods employed representations based on Bag of Words or n-grams tokens weighted with their respective tf-idf value to feed Support Vector Machines (SVM), Random Forest (RF), Gradient Boosting regressors, etc. For instance, in their work, (Aguirre and Cadena, 2023) combined these representations with linguistic features from the HurtLex lexicon (Bassignana, Basile, and Patti, 2018) to address all three subtasks. Similarly, (Árcos and Pérez, 2023) employed these representations to train a system consisting of stacked SVM and Gradient Boosting regressors for prejudice degree estimation.

Another emerging trend was the integration of traditional approaches with representations obtained from pre-trained word embeddings based on deep learning techniques, both contextual and non-contextual. For instance, (Sastre et al., 2023) experimented with the application Principal Components Analysis to reduce the embeddings obtained from RoBERTuito (Pérez et al., 2022) and employed them as input for a Multilayer Perceptron in subtask 1, and Gradient Boosting regressors and SVMs for subtasks 2a and 2b respectively. In a similar way, (García and de la Rosa, 2023) utilised word embeddings from the Word2Vec matrix and a pre-trained XLM-RoBERTa model to predict emotion probabilities, polarity features, and stylistic features. These features were fed into an ensemble of SVMs and a shallow Neural Network model for subtask 1, and a Multilayer Perceptron for subtasks 2a and 2b. (Bonet, Rincón, and López, 2023) adopted a similar strategy but using Decision Trees Regressors and SVMs instead. Finally, (Inácio and Oliveira, 2023) and (Sacristán, Muñoz, and Peris, 2023) employed contextual embeddings coming from Large Language Models (LLM) to feed SVMs in subtask 1.

On the other hand, some systems solely relied on pre-trained and fine-tuned LLMs based on transformer architectures, like the best-performing system for regression subtask proposed by the team M&C, which consisted in a simple fine-tuning of RoBERTa model (Liu et al., 2019). Relying also on transformer architectures, (Kaoshik and Kather, 2023) proposed an ensemble approach for subtask 1, using predictions from DistilBERT Cased (Sanh et al., 2019), XLM-RoBERTa Spanish (Lange, Adel, and Strötgen, 2021), RoBERTuito Cased, BERT Cased (Devlin et al., 2018) and mBERT Cased which is a multilingual version of the former. They adopted a similar strategy for subtasks 2a and 2b, excluding the RoBERTuito model. In the same way, (Inácio and Oliveira, 2023)and (García-Díaz and Valencia-García, 2023) combined different of these state-of-the-art pre-trained models. The latter, employing a Knowledge Integration technique that combines linguistic features with representations learned from the LLMs into a multi-input neural network. Whereas (Cruz et al., 2023), who achieved the best performance in competition for subtask 2a, combined the predictions of the fine-tuned LLMs by weighting them with respect to their individual performance.

It is worth mentioning the multitask learning system proposed by (Peng and Lin, 2023), which incorporated a cross-task interaction mechanism to share knowledge across tasks and used BERT-based model as a backbone. This, in addition to the attempt to balance the dataset by introducing a back-translation technique on the work of (Sastre et al., 2023) to fine-tune a RoBERTuito, which resulted in the best-performing system for subtask 1.

6 *Analysis and Discussion*

After the evaluation phase, we analyzed the predictions made by participants. In this section, we provide some findings related to the systems' performance and the difficulties placed by instances from the dataset.

6.1 Systems Performance

As stated in section 5, we observed a near-balanced number of submissions using both DL and traditional ML architectures. We study how this particular distinction in the approaches defined any difference in performance. Figure 3 shows the distribution of F1-scores for both subtask 1 (left) and subtask 2a (right) on DL and ML approaches.

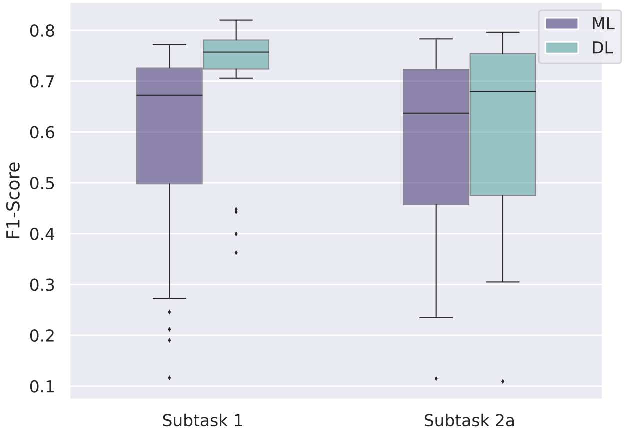

Figure 3: ML and DL systems performance for subtask 1 and subtask 2a.

From here we can see that most DL-based systems, as well as the average within this category, exhibit superior performance for humour recognition task. However, the situation differs when considering subtask 2a, which involves identifying the underlying topic of prejudicial tweets. In this case, there is considerable reliance on specific terms related to the ground truth minority. Consequently, even straightforward techniques like Bag of Words (BoW) can yield predictions that are nearly as precise as the more intricate modelling approaches employed by transformer-based models.

Regarding subtask 2b in Figure 4, where the Kernel Density Estimation of the systems' performance is depicted, we can observe a greater representation of ML submissions, specifically just 17 of the 55 documented submissions were based on DL. Nevertheless, systems under this paradigm again presented a more regular and shifted distribution towards lower RMSE values, i.e., a better performance. In fact, we must point out that the top three ranked systems were based on transformer-based architectures.

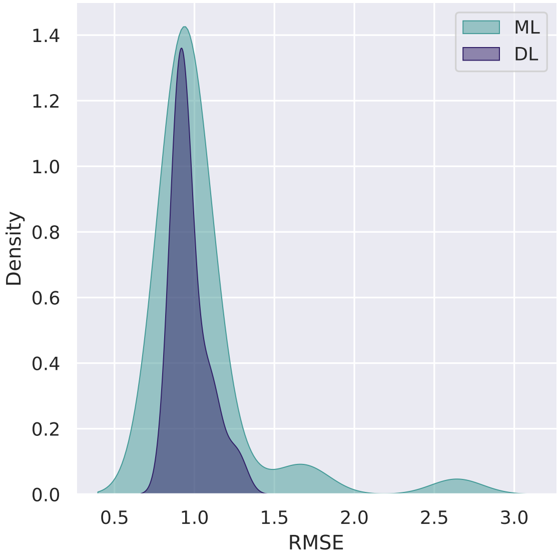

Figure 4: ML and DL systems performance for subtask 2b.

6.2 Error Analysis

To explore the errors made by the systems in subtask 1, we categorised the texts based on their difficulty level in each submission. Four difficulty categories were established: very difficult (more than 75% of submissions failed), difficult (between 75% and 50% failed), easy (less than 50% but more than 25% failed), and very easy (less than 25% failed).

We notice a significant relationship (Pearson-χ^2 = 238.545, df = 3, p < 0.001) between the difficulty level and whether the text was a joke or not. Among the non-

Difficulty	Not humour	Humour	Humour G1	Humour G2	Humour G3	Humour G4
very easy	28.93	0	0	0	0	0
easy	36.97	7.81	6.8	5.4	8.6	14.3
difficult	19.16	45.31	57.1	13.5	41.4	39.3
very difficult	14.94	46.88	36.1	81.1	50	46.4
Total	100.00	100.00	100	100	100	100

Table 7: Percentage of correctly classified according to its difficulty category.

humorous texts, approximately 65% of instances were classified as *easy* or *very easy*. However, for the humorous class, this percentage dropped significantly to only 7.8% of texts. These findings are summarised in Table 7.

In addition, we analysed if there was a particular target group that introduced some specific difficulty in the recognition of humour. Among the non-humours texts we did not find any significant relation between the level of difficulty and the target group mentioned in the text. However, among the jokes we found a significant relation (Pearson-χ^2 = 34.071, df = 16, $p < 0.005$) between the degree of difficulty for humour recognition and the groups mentioned in the text. As we can see in Table 7, the most serious difficulty for humour recognition is related to the fact of mentioning the LGBTI+ group (Humor G2).

For subtask 2a, we examined the number of teams that successfully identified the mention of each minority group in the instances. As we can see in Table 8, systems had more difficulty recognising all targeted groups when more than one was mentioned in the same tweet. Since the data do not follow a normal distribution, we applied the Mann-Whitney Test and all means differed significantly ($p < 0.001$). We explored also if humour introduced significant differences in the task of recognising which group is mentioned. The Mann-Whitney Test indicates that only two groups show significant differences regarding humour: the LGBTI+ group (G2) is better recognised in humorous texts than in non-humorous and the over-weighted people (G4) are better recognised in non-humorous texts than in humorous texts. Moreover, we observed a significant Spearman correlation between the number of teams that correctly recognised the target group and the level of prejudice (mean) that the five annotators gave to the tweet. This correlation indicates that systems had an increased ability to recognise all target groups in tweets have been judged more prejudicial by annotators (see Table 8).

Regarding subtask 2b, we analysed the predictions of the best-ranked submission and we computed the differences between the prejudice scores given by the annotators and the scores predicted. Positive values represented cases of overestimation of the degree of prejudice and negative values the opposite. We explored the potential effects of the presence of humour and the number of targeted minority groups on the performance of the best system. The Kolmogorov-Smirnov test revealed a violation of the normality assumption for these measures. However, considering the robustness of ANOVA to violations of normality in previous research (Blanca et al., 2017; Schmider et al., 2010), we proceeded with a parametric analysis to explore the interaction effect of these two variables. By performing an ANOVA, we observed a significant interaction between the two independent variables (F(1) = 15.008, $p < 0.001$) and a significant main effect of mentioning only one minority group or more than one (F(1) = 299.953, $p < 0.001$). When tweets mentioned more than one group, the system tended to overestimate the degree of prejudice, and this overestimation was significantly more pronounced in humorous texts. Conversely, when tweets targeted only one group, the system underestimated the degree of prejudice, and there was no significant difference between humour and non-humorous texts in terms of this underestimation (see Figure 5).

We further examined the estimation of the degree of prejudice in tweets that mentioned a singular group to determine if different groups introduced significant differences. While the interaction with humour was not significant, there was a significant main effect of the group mentioned in the tweet (F(2) = 24.446, $p < 0.001$). The results showed that the model tended to overestimate the degree of prejudice against the LGBTI+ group

Target	Non-humour	Humour	Single group	Multiple groups	Correlation with prejudice degree
G1	38.61	38.09	45.01	30.08	0.310**
G2	41.39	46.44	52.06	31.56	0.499**
G3	45.88	46.67	52.21	38.40	0.561**
G4	53.84	51.72	55.00	50.77	0.354**

**Correlation is significant at the 0.01 level.

Table 8: Number of teams that identify correctly the group target.

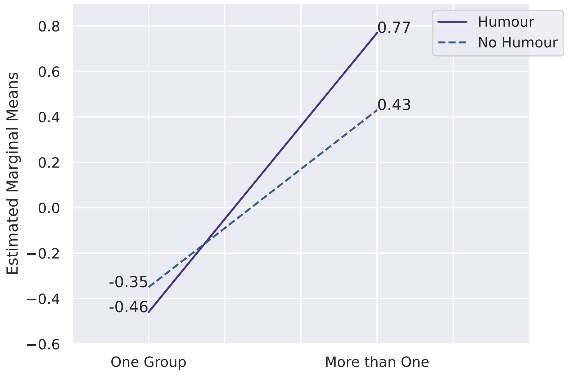

Figure 5: Prejudice degree according to humour presence and number of targeted groups on instances.

(Mean = 0.58) and underestimate the degree of prejudice against immigrants (Mean = -0.73) and women (Mean = -0.41). It is worth noting that the prejudice against overweighted people always appeared alongside another target group, and thus, it was not included in this particular analysis.

7 Conclusions

In this overview paper we described the HUHU shared task that we organized IberLEF 2023 and presented the results obtained by the 46 teams that participated. We provided a novel dataset related to humour recognition consisting of prejudiced tweets in Spanish annotated along the dimensions of the degree of prejudice perceived by individuals and the minority targets involved in text. We observed in the construction of the corpus a tendency to effectively increase the perception of prejudice when the minority being targeted also is an object of mocking. For subtask 1 in competition, 66 systems were submitted, where transformer-based models obtained the higher performance with respect to traditional machine learning algorithms, specifically the best performing system tried to balance the classes representation using back-translation to fine-tune a RoBERTuito model. In subtask 2b we observed a similar phenomenon across the 58 received submissions, resulting in the best estimators of prejudice degree RoBERTa-based models. In contrast for subtask 2a, where we had 56 submitted runs, in spite of the top-ranked system being based on ensembling multiple transformers models, the subsequent two teams in ranking, approached traditional machine learning techniques. In this case, we noticed that the performance of both paradigms was quite balanced due to the highly vocabulary-dependent characteristic of this task.

Based on the error analysis, it is evident that humour recognition is still a challenge for most systems. A higher number of errors were observed in the positive class, indicating difficulties in identifying instances of humour. Interestingly, we found that the degree of prejudice played a crucial role in aiding systems to recognise the target group: when humans label a tweet as more prejudicial the systems recognise better the victim of this prejudice. We also noticed that certain groups - especially the LGBTI+ community- introduced special difficulties for systems to recognise humour. Regarding the estimation of prejudice degree, we observed that the presence of multiple target groups and the use of humour leads the systems to overestimate the prejudice.

Acknowledgements

This work has been partially developed with the support of valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and co-funded by the European Union. The work of Berta Chulvi and Paolo Rosso is supported by FairTransNLP-Stereotypes PID2021–124361OB-C31 funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe.

References

Aguirre, M. C. and A. Cadena. 2023. Using vector embeddings and feature vectors to humor identification. In *IberLEF@SEPLN*. CEUR-WS.org.

Aroyo, L., L. Dixon, N. Thain, O. Redfield, and R. Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19. Association for Computing Machinery.

Bassignana, E., V. Basile, and V. Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Italian Conference on Computational Linguistics*.

Baziotis, C., N. Pelekis, and C. Doulkeridis. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August. Association for Computational Linguistics.

Billig, M. 2005. *Laughter and Ridicule: toward a social critique of humour*. Sage, London.

Blanca, M. J., R. Alarcón, J. Arnau, R. Bono, and R. Bendayan. 2017. Non-normal data: Is anova still a valid option? *Psicothema*, 29(4):552–557.

Bonet, H. A., A. M. Rincón, and A. M. López. 2023. Detection, classification and quantification of hurtful humor (huhu) on twitter using classical models, ensemble models, and transformers. In *IberLEF@SEPLN*. CEUR-WS.org.

Campos, R., V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 01.

Castro, S., L. Chiruzzo, and A. Rosá. 2018. Overview of the haha task: Humor analysis based on human annotation at ibereval 2018. In *IberEval SEPLN*.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Chiruzzo, L., S. Castro, M. Etcheverry, D. Garat, J. J. Prada, and A. Rosá. 2019. Overview of haha at iberlef 2019: Humor analysis based on human annotation. In *IberLEF SEPLN*.

Chiruzzo, L., S. Castro, S. Góngora, A. Rosá, J. Meaney, and R. Mihalcea. 2021. Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish. *Procesamiento del Lenguaje Natural*, 67.

Cruz, J., L. Elvira, M. Tabernero, and I. Segura-Bedmar. 2023. In unity, there is strength: On weighted voting ensembles for hurtful humour detection. In *IberLEF@SEPLN*. CEUR-WS.org.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Eugenio, B. D. and M. Glass. 2004. The Kappa Statistic: A Second Look. *Computational Linguistics*, 30, 03.

Ford, T. and M. Ferguson. 2004. Social Consequences of Disparagement Humor: A Prejudiced Norm Theory. *Personality and Social Psychology Review*, 8(1):79–94.

Ford, T. E., C. F. Boxer, J. Armstrong, M. Moya, and J. R. Edel. 2008. More than "just a joke": the prejudice-releasing function of sexist humor. *Personality & social psychology bulletin*, 34(2):159–70.

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

Freud, S. 1960. *Jokes and their Relation to the Unconscious*. Norton, Harmondsworth, England.

García, P. S. and C. M. de la Rosa. 2023. Dimensionality reduction techniques to detect hurtful humour. In *IberLEF@SEPLN*. CEUR-WS.org.

García-Díaz, J. A. and R. Valencia-García. 2023. Umuteam at huhu 2023: Detecting prejudices in humour using ensemble learning and knowledge integration. In *IberLEF@SEPLN*. CEUR-WS.org.

Girvan, M. and M. E. Newman. 2002. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99.

Glick, P. and S. T. Fiske. 1996. Ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, 70:491–512.

Grootendorst, M. 2020. Keybert: Minimal keyword extraction with bert.

Gwet, K. L. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61.

Hinton, G., N. Srivastava, and K. Swersky. 2012. Lecture 6a overview of mini–batch gradient descent. *Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online]*.

Hossain, N., J. Krumm, M. Gamon, and H. Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, December.

Howard, J. and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, July.

Inácio, M. L. and H. G. Oliveira. 2023. Attempting to recognize humor via one-class classification. In *IberLEF@SEPLN*. CEUR-WS.org.

Jones, E. 1972. *Prejudice and racism*. Addinson-Wesley.

Kaoshik, J. and S. B. Kather. 2023. Leveraging ensemble voting and fine-tune strategies in pre-trained transformers to detect prejudicial tweets and hurtful humour. In *IberLEF@SEPLN*. CEUR-WS.org.

Lange, L., H. Adel, and J. Strötgen. 2021. Boosting transformers for job expression extraction and classification in a low-resource setting. In *Proceedings of The Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings.

Lipmann, W. 1922. *Public Opinion*. New York:Harcourt Brace.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Meaney, J. A., S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy. 2021. SemEval 2021 task 7: HaHackathon, detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics, August.

Merlo, L., B. Chulvi, R. Ortega-Bueno, and P. Rosso. 2023. When humour hurts: linguistic features to foster explainability. *Procesamiento del Lenguaje Natural*, 70(0):85–98.

Miller, S. S., C. J. O'Dea, T. J. Lawless, and D. A. Saucier. 2019. Savage or satire: Individual differences in perceptions of disparaging and subversive racial humor. *Personality and Individual Differences*, 142.

Ortega-Bueno, R., B. Chulvi, F. Rangel, P. Rosso, and E. Fersini. 2022. Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO). Overview for PAN at CLEF 2022. In G. Faggioli, N. Ferro, A. Hanbury, and M. Potthast, editors, *CLEF 2022 Labs and Workshops, Notebook Papers*. CEUR-WS.org, September.

Peng, M. and N. Lin. 2023. Cross-task interaction mechanism for humour prejudice detection. In *IberLEF@SEPLN*. CEUR-WS.org.

Potash, P., A. Romanov, and A. Rumshisky. 2017. SemEval-2017 task 6: HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, August.

Pérez, J. M., D. A. Furman, L. Alonso Alemany, and F. M. Luque. 2022. Robertuito: a pre-trained language model for social media text in spanish. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France, June. European Language Resources Association.

Romero-Sánchez, M., H. Carreterio-Dios, J. L. Megías, M. Moya, and T. Ford. 2017. Sexist Humor and Rape Proclivity: The Moderating Role of Joke Teller Gender and Severity of Sexual Assault. *Violence against women*, 23(8):951–972.

Sacristán, D. B., A. P. Muñoz, and L. S. Peris. 2023. Building robust models for detecting offensive content and quantifying prejudice in online platforms. In *IberLEF@SEPLN*. CEUR-WS.org.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Sastre, I., A. Baladón, M. Berois, F. Cánepa, A. Lucasa, S. Castro, S. Góngora, and L. Chiruzzo. 2023. Retuyt-inco submission at huhu 2023: Detecting humor and prejudice through supervised methods. In *IberLEF@SEPLN*. CEUR-WS.org.

Scao, T. L., A. Fan, C. Akiki, E. Pavlick, et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Schmider, E., M. Ziegler, E. Danay, L. Beyer, and M. Buehner. 2010. Is it really robust? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6:147–151, 01.

Wong, K., P. Paritosh, and L. Aroyo. 2021. Cross-replication reliability - an empirical approach to interpreting inter-rater reliability. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, August.

Árcos, I. and J. Pérez. 2023. Detecting hurtful humour on twitter using fine-tuned transformers and 1d convolutional neural networks. In *IberLEF@SEPLN*. CEUR-WS.org.