In Unity, There Is Strength: On Weighted Voting Ensembles for Hurtful Humour Detection

Javier Cruz[†], Lucas Elvira[†], Miguel Tabernero[†] and Isabel Segura-Bedmar

Universidad Carlos III de Madrid (UC3M University), Av. de la Universidad, 30, 28911 Leganés, Spain

Abstract

This paper describes our participation in the HUrtful HUmour (HUHU) task at IberLEF 2023, geared towards detecting prejudice-fostering humour on Twitter. A novel weighted voting system of ensembles composed of different popular transformer models is proposed. We empirically demonstrate that ensembles exceed individual transformers in humour and prejudice detection. Our system ranked 12th (beating 46 teams), 1st (beating 48 teams) and 22nd (beating 26 teams) in the binary classification, multilabel classification and regression tasks, respectively. We conclude that combining state-of-the-art transformer models depicts a promising research direction to yield robust systems for detecting humour spreading prejudice in social media. The code is publicly available online: https://github.com/mtabernerop/JUJUNLP.

Keywords

Natural Language Processing, Humour Detection, Ensemble Learning, Transformers

1. Introduction

Hurtful humour refers to a form of humour targeted at a particular individual or group with the objective of causing emotional pain or offense. It often involves making derogatory, insulting, or offensive comments about the physical appearance, beliefs, culture, race, gender, sexual orientation, or other personal attributes of an individual or group. Hurtful humour can contribute to the perpetuation of harmful stereotypes and discrimination, thus leading to feelings of humiliation, shame and marginalization in the target [1].

The detection of offensive comments disguised by the mask of humour and protected by the subjectivity of the latter poses a challenging still worthwhile task [2]. This becomes particularly interesting in social media platforms such as Twitter, where content can be shared widely and quickly, potentially reaching millions of users across the globe. In this context, hurtful humour is often used to reinforce negative stereotypes and discriminatory attitudes towards minorities such as women, the LGBTIQ community or immigrants, among others [3]. Hence, identifying this content in tweets becomes a crucial step towards ensuring a more inclusive and respectful online environment [1].

IberLEF 2023, September 2023, Jaén, Spain

[†] These authors contributed equally.

jacruzde@pa.uc3m.es (J. Cruz); luelvira@pa.uc3m.es (L. Elvira); mtaberne@pa.uc3m.es (M. Tabernero); isegura@inf.uc3m.es (I. Segura-Bedmar)

0009-0001-1172-6419 (J. Cruz); 0009-0007-5423-7627 (L. Elvira); 0009-0000-0800-2252 (M. Tabernero); 0000-0002-7810-2360 (I. Segura-Bedmar)

The HUHU@IberLEF 2023 shared task [4] motivates research aimed at identifying prejudice and stereotyping towards marginalized groups (specifically, women and feminists, the LGBTIQ community, immigrants and individuals who have experienced racial discrimination, or those who are overweight) through the use of humour in Twitter posts, which can be used to disseminate hurtful messages and avoid moral judgment.

This paper describes the participation of our group, JUJUNLP, at the HUHU@IberLEF 2023 competition. Our work proposes the use of ensembles of state-of-the-art transformer models to detect, by joint weighted voting, humorous content, prejudiced groups and degree of prejudice in Spanish-written tweets, which to the best of our knowledge portrays a novel approach to identify hurtful humour in social media content. The empirical results show that the combination of transformer predictions weighted by their individual performance on the task allows achieving competitive results in the aforementioned context.

The rest of the paper is organized as follows. Section 2 reviews the most significant aspects of the HUHU@IberLEF 2023 task. Then, a summary of related work is provided in Section 3. Section 4 thoroughly covers the proposed approach. Sections 5 and 6 describe the empirical evaluation and discuss the results. Finally, Section 7 portrays valuable conclusions and an outline of open avenues for future research.

2. Task Overview

HUHU@IberLEF 2023 [4] is a competition to boost research on the detection of humorous tweets expressing prejudice in social networks towards minorities, including women and feminists, the LGBTIQ community, immigrants and racially discriminated people, and overweight people. For this purpose, the organizers have created a dataset containing a wide spectrum of texts written in Spanish from Twitter. We now describe the subtasks that are defined in this competition and the provided dataset.

2.1. Subtasks

This IberLEF 2023 track allows to participate in three different subtasks. The main specifications of each one are listed below:

HUrtful HUmour Detection (Task 1) Binary classification task aimed at determining whether a prejudicial tweet is meant to be humorous or not. The metric employed will be the F1-score over the positive class.

Prejudice Target Detection (Task 2A) Multilabel classification task where the objective is to identify the aforementioned minority groups on each tweet. The participating systems will be evaluated and ranked using the macro F1-score.

Degree of Prejudice Prediction (Task 2B) Regression task where systems must predict (on a continuous scale from 1 to 5) how prejudicial a tweet is on average among minority groups (5 corresponds to the maximum level of prejudicial). The predictions will be assessed using the Root Mean Squared Error (RMSE).

The core of this article describes the participation of our group in all subtasks.

2.2. Dataset

This section is aimed towards portraying a more thorough insight into the dataset [5] provided by the organizers of HUHU@IberLEF 2023 shared task. The training and test sets contain 2,671 and 778 hand-annotated tweets in Spanish, respectively. Each instance has an ID, a text, a binary value that identifies whether the tweet is humorous or not, a real value from 1 to 5 that determines how prejudicial the message is on average among minority groups, and four binary values to identify the groups that are prejudiced in the text; the target minorities are women and feminists, the LGBTIQ community, immigrants and racially discriminated people, and overweight people. Table 1 illustrates two instances from the dataset with humorous and non-humorous content.

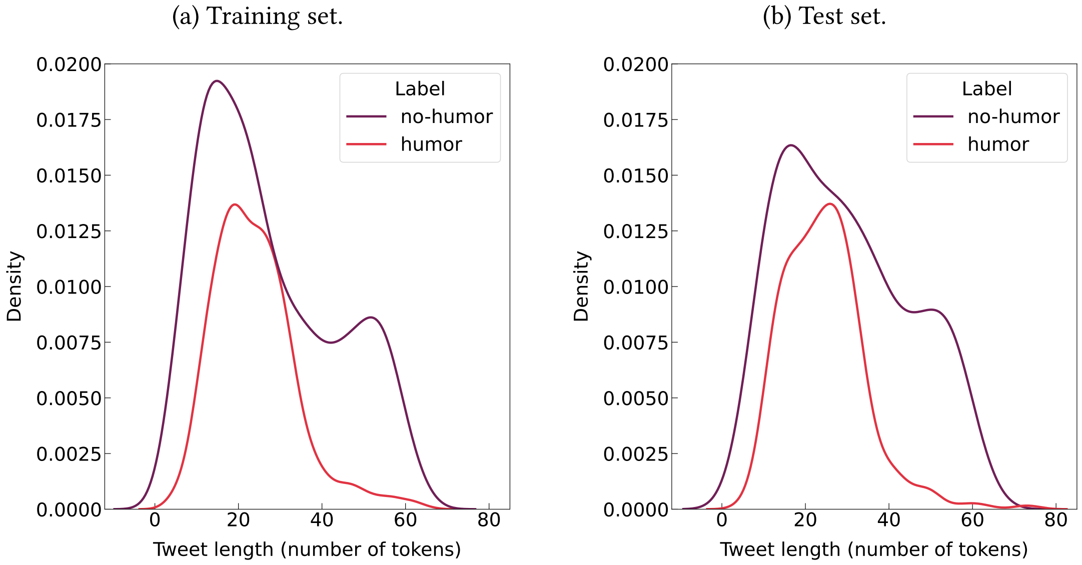
Table 1
Sample of humorous and non-humorous texts from the dataset.

tweet	humor	prejudice woman	prejudice lgbtiq	prejudice inmigrant race	gordofobia	mean prejudice
No hay nada mas indefenso que una mujer con las uñas recién pintadas.	1	1	0	0	0	2.0
“Putos negros los considero la raza inferior, ojala vuelvan los nazis”	0	0	0	1	0	4.6

In the training set, 67% (1,802) of the instances were identified as humour tweets, while the remaining 33% (869) were considered non-humorous. This distribution is maintained in the test division (522 tweets marked as “no-humor” and 256 tagged as “humor”), thus evidencing a strong unbalanced distribution of the classes in favour of non-humorous texts. Figure 1 shows the distribution of tweet length for humour and non-humour instances in both dataset divisions. Funny tweets tend to be longer, with 27.5 ± 16.0 and 29.9 ± 15.4 tokens on average in the training and test datasets, while non-humorous texts are slightly shorter, with 23.8 ± 9.9 and 24.4 ± 9.6 tokens, respectively.

Attending to the class distribution in the multilabel classification task, a particularly interesting aspect was identified. In the provided training dataset, all tweets were labeled to target at least one minority group and at most two, i.e., either one or two of the four labels take the value 1, while the others take the value 0. However, this event does not occur in the test dataset, where many instances are marked as prejudicial towards three or even all four groups. Considering this feature, Figure 2 plots the number of instances labeled with each class in the training and test datasets. For simplicity, each label is referred to by a representative capital letter, namely “W” for women and feminists, “L” for the LGBTIQ community, “I” for immigrants and racially discriminated people, and “G” for fatphobic prejudices (“gordofobia” in the original datasets); this label encoding is maintained throughout the rest of this paper. Note that in the analysis of the training set the main diagonal contains the instances that are only tagged with a single class (see Figure 2a); recall that since subtask 2B is posed as a multilabel classification problem, the matrix trace does not necessarily have to be equal to the total number of instances in the dataset (and in fact it is not), provided that several instances are labeled as prejudicial towards more than one minority group. Here, a clear correlation between labels can be seen; for instance, tweets that are offensive to women and feminists have a high probability of expressing prejudice against overweight people as well. In the test set, texts offensive towards women are the most common. Tweets that at the same time target this group and the LGBTIQ community are frequent as well.

Figure 1: Density graph of tweet length in the training and test datasets for humorous and non-humorous texts.



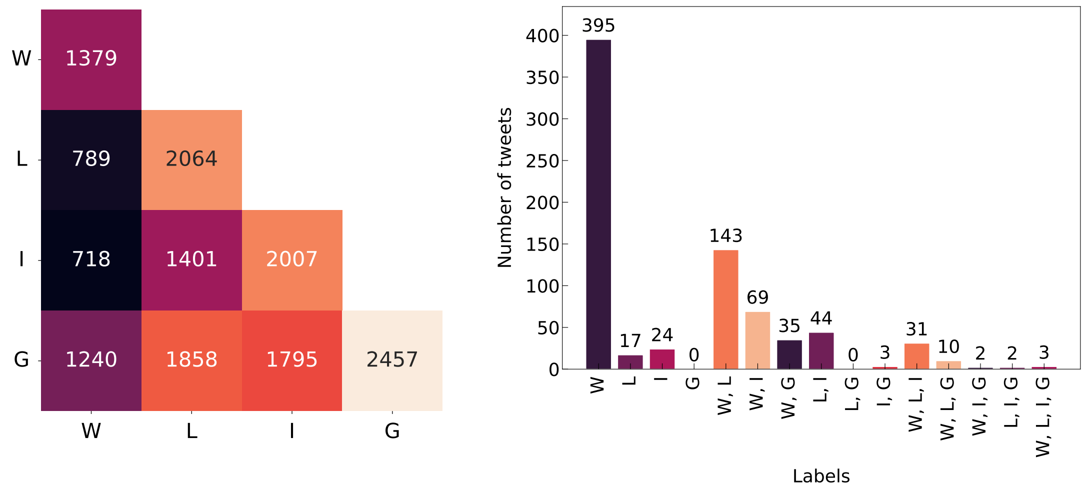
(a) Training set.

(b) Test set.

Lastly, Figure 3 plots a density graph of the prejudice scores in the training and test datasets, separated into two curves to differentiate between humorous tweets and those with only hurtful content (i.e., “no-humor” class). It is straightforward to determine that humorous texts tend to register a higher prejudice score. This again emphasizes the relevance of the HUHU@IberLEF 2023 shared task: identifying offensive comments in social media content that may be hidden behind humorous undertones is essential to ensure a safe, respectful, and diverse online environment.

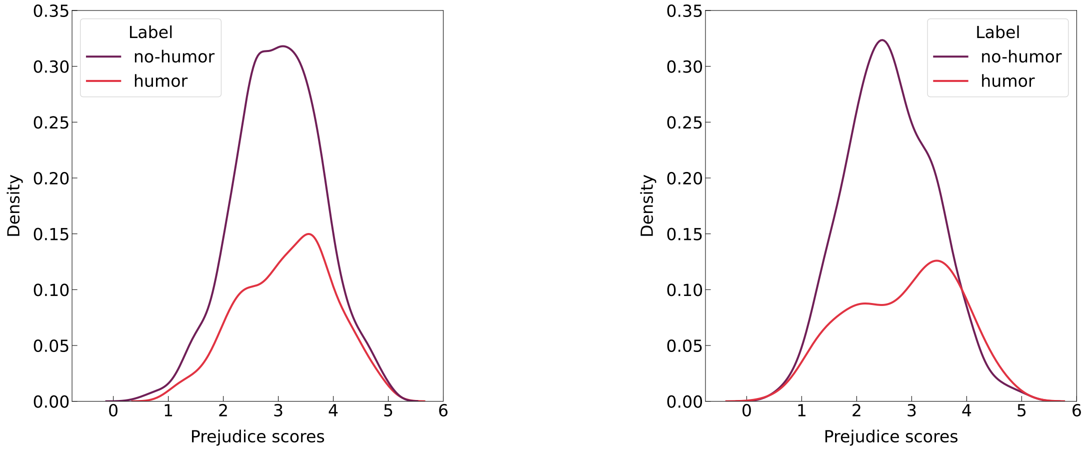
During the development phase, the training dataset [5] provided by the organizers was divided into three splits with a ratio of 70:20:10, i.e., 1,870 tweets for training, 534 for validation, and 267 for testing. Data stratification was performed for the binary and multilabel classification tasks in order to preserve the class distribution of the original dataset in each split. For regression, the data subsets obtained through random sampling resulted representative enough.

Figure 2: Class distribution in the training and test datasets for texts prejudicial to minority groups.


(a) Training set.

(b) Test set.

Figure 3: Density graph of prejudice score in the training and test datasets for humorous and non-humorous texts.


(a) Training set.

(b) Test set.

3. Background

The computational detection of humour is a well-established and actively researched topic within the field of Natural Language Processing (NLP) [6, 7, 2]. In 2017, Zhang et al. introduced the concept of Contextual Knowledge and diverse features to capture the emotionality and subjectivity behind humorous content [8]. Recent efforts have been also directed towards detecting humour in social media content, such as the work from Zhang and Liu (2014) [9], which resorts to the use of Machine Learning (ML) techniques to handle sentiment analysis and opinion mining to distinguish humorous texts from non-humorous posts.

However, based on the dynamic nature of language, cultural context and the subtleties involved in detecting sarcasm, irony and other forms of non-benign humour, its automatic recognition is far from triviality [2]. Actually, the fun may sometimes be indistinguishable even for the human reader. For the sake of involving scientists and fostering research in this field, various annual events on humour recognition are held. SemEval-2015 Task 11 [10] was oriented to the study of three broad classes of figurative language: irony, sarcasm and metaphor. Further, Task 6 of the 11$^{\text{th}}$ edition of this workshop [11] was aimed towards capturing the specific sense of humour in tweets submitted to a comedy show. HAHA@IberLEF 2018 [12, 13] was the first Spanish-language humour detection challenge, followed by the celebration of the same competition in 2019 [6]. Here, humour detection was posed as a binary classification task and the funniness of crow-annotated tweets had to be scored as a regression problem. Additionally, SemEval-2021 Task 7 [2] later extended the participating tracks to recognize offensive content in controversial humorous posts. In the same line, SemEval-2017 Task 7 [14] channeled studies towards the analysis of puns.

There is little doubt that the introduction of Transformers [15] marked the beginning of a new thrilling chapter in the NLP domain. After the proposal of BERT [16], the "blue-eyed boy" of this emerging era, multiple alternative architectures have been designed to handle complex language processing tasks [17, 18, 19]. It is no wonder that their application has ranged through various practical scenarios, including the recognition of humorous content. For instance, Weller and Seppi (2019) proposed a transformer-based method that acquires the ability to recognize jokes by analyzing ratings obtained from Reddit pages [20]. They empirically demonstrated that this contribution outperformed previous approaches in this domain, obtaining an F1-score of 93.1% and 98.6% for two datasets of puns (32,003 instances) and jokes (231,657 instances), respectively.

Reasonably, the individual success of transformer models raises the question of whether their combination could potentially ease humour detection. In this context, ensembles that use multiple ML techniques jointly have shown robust performance in humour detection tasks. In particular, HITACHI [21], the winning team at the SemEval-2020 Task 7 [22], uses stacking in ensembles of pre-trained language models (PLMs) to compute the final predictions. In this workshop, two tasks were defined: scoring of funniness in the range [0, 3] and prediction of the funnier headline between pairs of these. HITACHI ranked first in both substasks, achieving an RMSE of 0.449 and an accuracy of 67.4%, respectively. The dataset originally contained a total of 5,000 news headlines.

Furthermore, the winner of the HAHA@IberLEF 2019 shared task [23] introduced an ensembling system of a fine-tuned multilingual version of BERT and a Naïve Bayes classifier, yielding an F1-score of 82.1% and a 0.736 RMSE for humour detection (binary classification) and funniness score prediction (regression) tasks, respectively. The dataset consisted of 30,000 hand-annotated Spanish tweets, out of which 38.7% were labeled as humorous.

The top system of the 2021 edition of the HAHA competition was JOCOSO [24], an ensemble of diverse transformer architectures (plus a Naive Bayes classifier) fine-tuned on the dataset provided by the organizers. The training and development splits of the latter matched the training and test sets of the 2019 edition; in addition, a new test split of 6,000 tweets was provided. JOCOSO ranked first in the (binary) humour classification task (F1-score = 88.5%) while performing competitively in the rest: it obtained the third place in the (regression) humour rating task (0.6296 RMSE) and was runner-up in the (multiclass) humour logic mechanism classification and (multilabel) humour target classification tasks, with F1-scores of 29.1% and 35.8%, respectively. These last works have heavily inspired the notions presented in this paper.

4. System Overview

4.1. Models

We now provide a brief description of the state-of-the-art transformers that were used during the development phase.

Presented in 2018, BERT [16] is the most popular transformer model due to its outstanding performance in many NLP tasks. Since its release, many state-of-the-art transformers have been developed based on it, including RoBERTa [17], ALBERT [25] and DistilBERT [18], among many others. BERT was trained under two tasks: masked language modeling (MLM) and next sentence prediction (NSP). In particular, the multilingual version used in this work was pre-trained in a self-supervised fashion on the top 104 languages with the most extensive Wikipedia.

DistilBERT [18] is a smaller version of BERT that can be trained faster. This is achieved through distillation, i.e., the number of layers in the initial version of BERT is reduced by a factor of 2, and token embeddings and poolers are removed to yield a cheaper and lighter transformer model. In this work, a multilingual version of DistilBERT is assessed.

RoBERTa [17] seeks to provide a highly optimized version of BERT by tweaking various methodological parameters. It was originally trained using texts from five English language datasets: the BookCorpus dataset [26], the English Wikipedia, the CC-News data, the Stories dataset [27], and the Open Web data.

Lastly, BETO [19] is a variation of BERT trained on a big Spanish corpus [28] with the Whole Word Masking technique, which masks, in addition to the original token, all tokens of the same word.

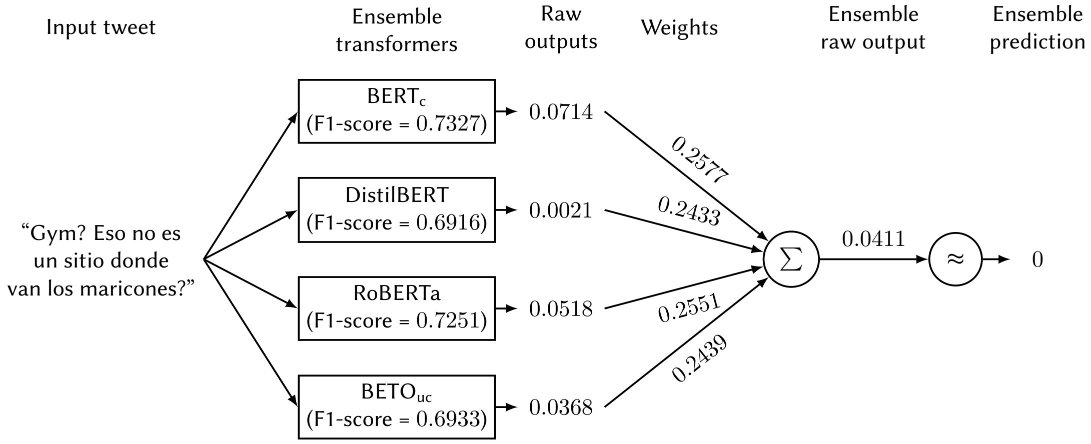
For the sake of completeness, this study uses both the cased and uncased versions of BERT and BETO. In the remainder of the document, this aspect will be specified with subscript “c” or “uc” for cased and uncased, respectively.

4.2. Ensemble voting system

We present a novel system for detecting hurtful humour in tweets by using ensembles of the transformer models described in subsection 4.1. Our system combines the raw predictions of transformer models (which were fine-tuned separately) using a weighted voting system, resulting in joint predictions that leverage the strengths and address the weaknesses of individual transformers. The weights assigned to each transformer correspond to the normalized value of the task score, which is computed over the validation test. Thus, greater importance is given to the predictions of the transformers that offer better performance in the task at hand, without disregarding the contributions of other transformers that would allow a stronger prediction consensus to be reached. Accordingly, we employ the F1-score over the positive class for the binary classification task (subtask 1) and the macro F1-score for the multilabel classification problem (subtask 2A), while for the regression task (subtask 2B) we use the inverse of the RMSE, all calculated over the validation set. By summing the weighted predictions, a raw output is yielded. Remark that subtasks 1 and 2A are posed as binary and multilabel classification tasks, respectively; hence, this raw value is approximated to the nearest binary value in these scenarios.

Figure 4 illustrates how the label of a given instance is predicted in the binary classification task. The ensemble used in this case is composed of four models. The weights refer to the normalized F1-scores, so that the sum of all the resulting values equals 100%; thus, these weights represent the percentage of importance assigned to each transformer. Further, the ensemble output has to be rounded off to produce the final prediction.

Figure 4: Example of an ensemble of four transformer models using a F1-score-based weighted voting system in task 1. In this example, the ensemble was composed only of four transformers.


Input tweet
Ensemble transformers
Raw outputs
Weights
Ensemble raw output
Ensemble prediction
"Gym? Eso no es un sitio donde van los maricones?"
BERT$_c$ (F1-score = 0.7327)
DistilBERT (F1-score = 0.6916)
RoBERTa (F1-score = 0.7251)
BETO$_{uc}$ (F1-score = 0.6933)
0.0714
0.0021
0.0518
0.0368
0.2577
0.2433
0.2551
0.2439
Σ
0.0411
≈
0

For the sake of completeness, all possible ensembles are defined as a variation with repetition of the 6 transformer models studied (BERT$_{c,uc}$, RoBERTa, DistilBERT, and BETO$_{c,uc}$). Each ensemble can be represented by a binary string of 6 bits (since we are evaluating 6 transformers), where each bit i determines whether the model i is present in the ensemble at hand (1) or not (0). This results in a total of 2^6-1 = 63 ensembles (one is subtracted since the empty ensemble, encoded by the string containing six 0's, is neglected). It is important to note that this only involves calculating the predictions by varying the weight used in the voting phase. For each subtask, the ensemble model that performed best on the test division represented the architecture used to estimate the predictions that were submitted to the competition.

5. Experimental Setup

Unlike rule-based models or recurrent neural networks (RNNs), transformer models can learn complex language patterns without extensive preprocessing or human intervention [15]. Their architecture, including multi-head attention and encoding-decoding layers, allows them to handle various NLP tasks without modifying the input data. Transformers excel in unstructured or loosely structured language scenarios, thus eliminating the need for extensive data preparation. For the participation in the HUHU@IberLEF 2023 competition, preprocessing tasks such as tokenization, stemming, lemmatization or stop-word removal did not give better results than when using the raw data; consequently, these were kept in their original form. A noteworthy issue is that the treatment of hashtags, URLs, and mentions to other Twitter users were already addressed in the dataset provided by the organizers, represented in a unified way in the training and test sets by the words "HASHTAG", "URL" and "MENTION", respectively.

All transformer models were individually trained for 10 epochs and a batch size of 8 on the training split. To avoid overfitting, early stopping was applied with 3-epochs patience. For the sake of achieving better performance of the transformers in the tasks, hyperparameter tuning was done via grid-search using different learning rates ({2*e*-5, 4*e*-5, 8*e*-5}) and optimizers ({*AdamW*, *Adafactor*}). The best hyperparameter values were chosen evaluating the models on the validation split. All transformer models were trained on NVIDIA T4 Tensor Core GPUs on Google Colab.

We performed an extensive evaluation of all possible ensembles with different hyperparameters on our test split. As the organizers allowed competitors to send two submissions of predictions for each subtask, we chose the two approaches that reported the best scores on our test split. Table 2 summarizes these approaches.

The work presented in this paper is implemented in Python 3.10. The development of the ensembles of transformers is primarily based in the `simpletransformers` library (version 0.63.9) [29], which allows to quickly train, evaluate and make predictions with fine-tuned state-of-the-art transformer models within few lines of code. To date, it offers support to various NLP tasks including text and token classification, question answering, language modeling and generation, multi-modal classification, conversational AI, and text representation generation.

Many other libraries have been used to plot, visualize and evaluate the dataset and the empirical results. Some of these include but are not limited to (in alphabetical order) `matplotlib`, `numpy`, `pandas`, `seaborn` and `scikit-learn`.

6. Results

The HUHU@IberLEF 2023 shared task allowed for the submission of two different runs of predictions per task. Hence, our team, JUJUNLP, participated in the three subtasks by using the two best-performing ensembles in each. These alongside the hyperparameter values that yielded the best results on the test split (10% of the training dataset divided for experimentation) are reported in Table 2.

After the the organizers published the labeled test dataset, we have been able to evaluate all transformers on this set (see Table 3). As it was explained above, our experimentation during the development phase shows that individual transformers perform worse than some ensembles when they were evaluated on our test split, which was created by randomly stratified sampling. However, the assessment of all transformers and ensembles on the final test dataset does not show the same behavior. In fact, some transformers, such as BETO_{c} or BETO_{uc} (both fine-tuned

Table 2
Best approaches on our test split. Score represents the value (calculated over the test split of the training set) of the metric used in the corresponding task, i.e., F1-score, macro F1-score, and RMSE in subtasks 1, 2A, and 2B, respectively.

Subtask	System (run) ID	Ensemble	Learning rate	Optimizer	Score
1	1	BERT$_{c}$+BETO$_{c}$	4e-5	Adafactor	77.5%
	2	All transformer models	2e-5	Adafactor	77.4%
2A	1	BERT$_{c}$+RoBERTa+BETO$_{c}$+BETO$_{uc}$	4e-5	AdamW	94.8%
	2	BERT$_{c}$+BERT$_{uc}$+RoBERTa+BETO$_{c}$+BETO$_{uc}$	2e-5	AdamW	94.6%
2B	1	BERT$_{c}$+BETO$_{c}$+BETO$_{uc}$	4e-5	AdamW	0.640
	2	DistilBERT+BETO$_{c}$+BETO$_{uc}$	2e-5	AdamW	0.644

using AdamW optimizer and 4*e*-05 as learning rate), overcome the best ensembles that we found during the development phase (see Table 2). We have pointed out in subsection 2.2 that the class distribution for subtask 2A (multilabel text classification) is different in the training and test dataset provided by the organizers. As previously stated, all tweets in the training dataset were labeled with at most two out of the four labels. However, many tweets in the test dataset are annotated with three or even four labels.

Table 3
Transformer results for each subtask on the final test dataset. Score is the metric used in the corresponding task, i.e., F1-score, macro F1-score, and RMSE in subtasks 1, 2A, and 2B, respectively. The best results per task are highlighted in bold.

Transformer	Optimizer	Learning rate	Score		
			Subtask 1	Subtask 2A	Subtask 2B
BERT$_{c}$	Adafactor	4*e*-05	74.5%	73.5%	0.980
		2*e*-05	70.5%	74.4%	0.980
BERT$_{uc}$	Adafactor	4*e*-05	72.1%	72.8%	0.998
		2*e*-05	67.4%	72.6%	1.035
RoBERTa	AdamW	4*e*-05	75.9%	74.5%	1.042
		2*e*-05	69.7%	72.0%	1.037
DistilBERT	AdamW	4*e*-05	73.6%	75.5%	0.956
		2*e*-05	68.2%	73.9%	0.996
BETO$_{c}$	**AdamW**	**4*e*-05**	**77.0%**	**80.1%**	0.951
		2*e*-05	75.9%	79.5%	0.957
BETO$_{uc}$	**AdamW**	**4*e*-05**	73.4%	76.5%	**0.928**
		2*e*-05	70.3%	76.1%	0.952

Additional valuable conclusions can be drawn from Table 2 regarding the use of ensembles of transformers: although transformer models were also evaluated individually in each subtask, none exhibited better performance than when they were combined through the voting system described above. This proves that ensembles achieve better results in the hurtful humour detection tasks of this competition than any transformer model independently. Attending to the transformer models, it can be directly observed that BETO is present in all best ensembles of each task, while BERT does not appear only in the second-best system for the regression track.

On the other hand, DistilBERT exhibited the worst performance among the 6 state-of-the-art models, being present only in the second-best ensemble for subtask 2B.

Based on the official results reported by the organizers of the HUHU@IberLEF 2023 shared task, 58, 49 and 48 teams participated in subtasks 1 (binary classification), 2A (multilabel classification) and 2B (regression), respectively. Table 4 summarizes the participation of our team (JUJUNLP) in the competition, which managed to rank 12th, 1st and 22nd in the aforementioned tracks with the first run of predictions, while the second run finished 27th, 4th and 25th. In the following analysis, JUJUNLP$_i$ stands for the system that produced the results for run i (see Table 2). Notice that JUJUNLP$_1$ for subtask 1 (BERT$_c$+BETO$_c$) differs, for instance, from JUJUNLP$_1$ for subtask 2B (BERT$_c$+BETO$_c$+BETO$_{uc}$); the task being analyzed will be explicitly mentioned as appropriate.

Table 4
Official results of the HUHU@IberLEF 2023 shared task. The metrics recorded by the best (winning) approach in each task and the best performing baseline are indicated alongside the name of the system that registered them. For the two runs submitted of our system (JUJUNLP$_1$ and JUJUNLP$_2$, respectively), the position achieved in the final ranking is shown in parentheses. The metrics are the F1-score, macro F1-score and RMSE in subtasks 1, 2A and 2B, respectively.

System	Subtask 1	Subtask 2A	Subtask 2B
Top system	82% (RETUYT-INCO$_1$)	79.6% (JUJUNLP$_1$)	0.855 (M&C$_2$)
JUJUNLP$_1$	77.2% (12)	79.6% (1)	0.934 (22)
JUJUNLP$_2$	72.2% (27)	77.4% (4)	0.939 (25)
Best baseline	78.9% (BLOOM-1B1)	76% (BETO)	0.874 (BETO)

Regarding subtask 1, our team did not make the top 10. The performance of our system is clearly improvable, since the difference of F1-scores between RETUYT-INCO$_1$ and JUJUNLP$_1$ is almost 5 points. Further, the best performing baseline ranked atop of both of our submitted entries in this subtask, recording a value of the aforementioned metric almost 2 points better.

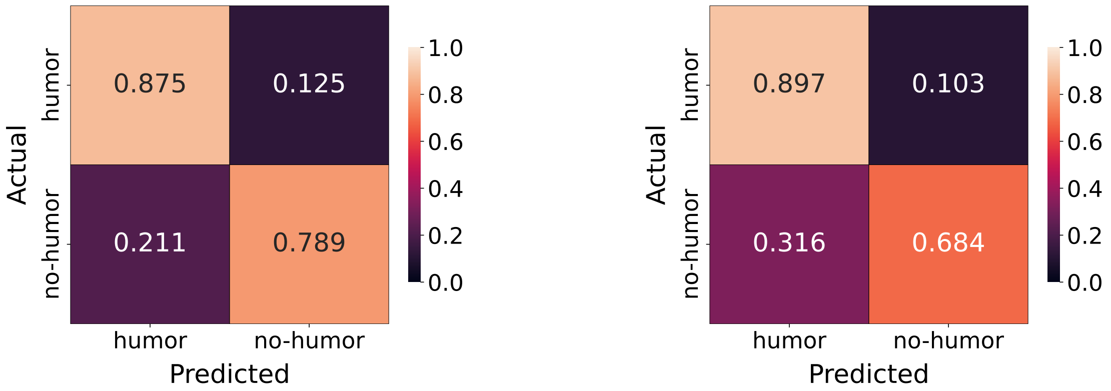
In subtask 2A, we managed to win the rest of participants. Here, the ensembles of transformers used appear to suit the detection of prejudiced groups in Spanish tweets, since JUJUNLP$_2$ also achieved a valuable position in the ranking (4th). Further, our team exceeds all baseline approaches in this track. A plausible justification for this fact is that ensembles allow to incorporate diverse perspectives and to take into account the independence of labels in the context of multilabel classification. This diversity ensures that ensembles can handle cases where labels are interdependent or co-occur in complex ways. In addition, ensemble systems can handle noisy labels (due to the inherent subjectivity of those who label the data) more effectively by leveraging predictions from multiple models.

Lastly, in subtask 2B JUJUNLP obtained an RMSE 0.079 units worse than that of the winner of this task (0.084 for run 2). In fact, all baseline approaches outperform our system, including BETO which ranked 2nd. Although BETO is also present in JUJUNLP$_1$ and JUJUNLP$_2$, a different dataset division or model fine-tuning process (among other options) may have been followed by the organizers, which would explain the notable differences in the results achieved by our approach.

For the sake of performing a thorough error analysis and further experiments with our systems, Figures 5 to 8 evaluate the results of JUJUNLP$_1$ and JUJUNLP$_2$ on the test dataset, whose labels were publicly released after the deadline for submission of prediction runs was reached.

Figure 5 shows a similar performance by JUJUNLP$_1$ and JUJUNLP$_2$. Both display a lower precision (0.806 and 0.739) in comparison to their recall (0.875 and 0.897), implying that multiple non-humorous tweets are incorrectly marked as humour. However, funny tweets are correctly identified in almost 90% of the cases. These can be considered as competitive results, since several errors are attributed to instances that are on the borderline of humour (and, in fact, could give rise to discussion). As an example, the humorous tweet "Lo géneros son como las torres gemelas, antes eran dos pero ahora es un tema sensible." is classified by our systems as "no-humor".

Figure 5: Confusion matrices on the test dataset for subtask 1.



(a) Predictions computed by JUJUNLP$_1$.

(b) Predictions computed by JUJUNLP$_2$.

The confusion matrices corresponding to the four binary classes that comprise subtask 2A based on the results achieved by JUJUNLP$_1$ and JUJUNLP$_2$ are portrayed in Figures 6 and 7. Again, both systems exhibit a similar performance. In every class, the precision achieved by the ensembles is higher than the recall (calculated over the positive class). In other words, the number of False Negatives (FN) is higher than the amount of False Positives (FP). JUJUNLP$_1$ and JUJUNLP$_2$ excel in determining whether a tweet is offensive towards overweight people, i.e., "Gordofobia" (fatphobia) is the label where JUJUNLP$_1$ and JUJUNLP$_2$ achieve a higher F1-score: 0.930 and 0.898, respectively. The second class in which they show decent results is "prejudice_woman". The opposite scenario is presented by the "prejudice_lgbtiq" class (0.683 and 0.669 F1-scores), where these ensembles are practically unable to distinguish tweets that express prejudice towards the LGBTIQ community. JUJUNLP$_1$ and JUJUNLP$_2$ also behave poorly in recognizing texts with offense towards immigrants or expressing racial discrimination. A remarkable aspect that was identified in this subtask is that both ensembles tend to set the "prejudice_woman" label to true (1) whenever the text at hand mentions women, even when it does not apply; this explains why the proportion of FPs in the first class is higher than in any other label. In addition, if the word "negro" appears in a tweet, the ensemble systems directly mark it as racist. Definitely, this kind of scenarios that give rise to incorrect predictions could be solved if a superior and more varied training dataset was available.

Figure 6: Confusion matrices on the test dataset with the predictions by JUJUNLP$_1$ for subtask 2A.

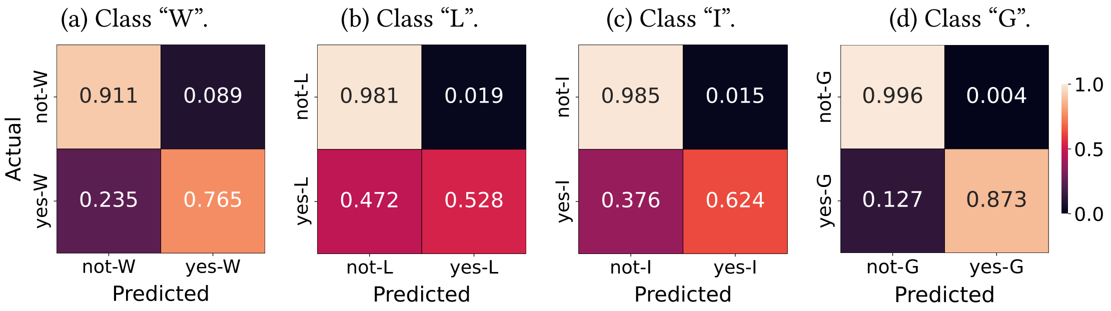
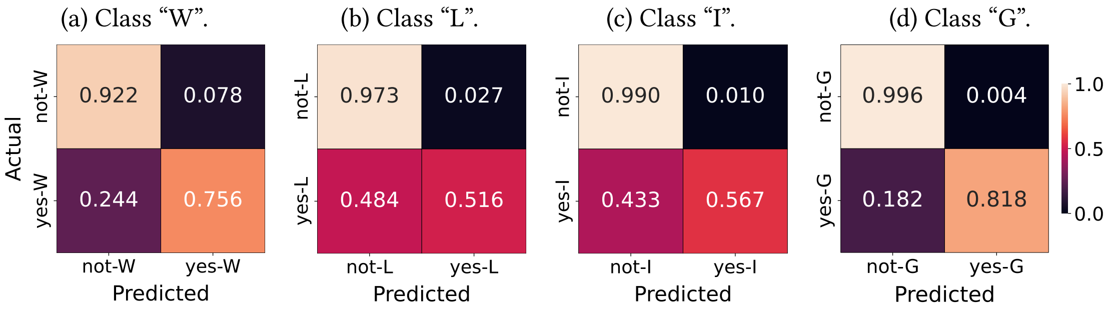
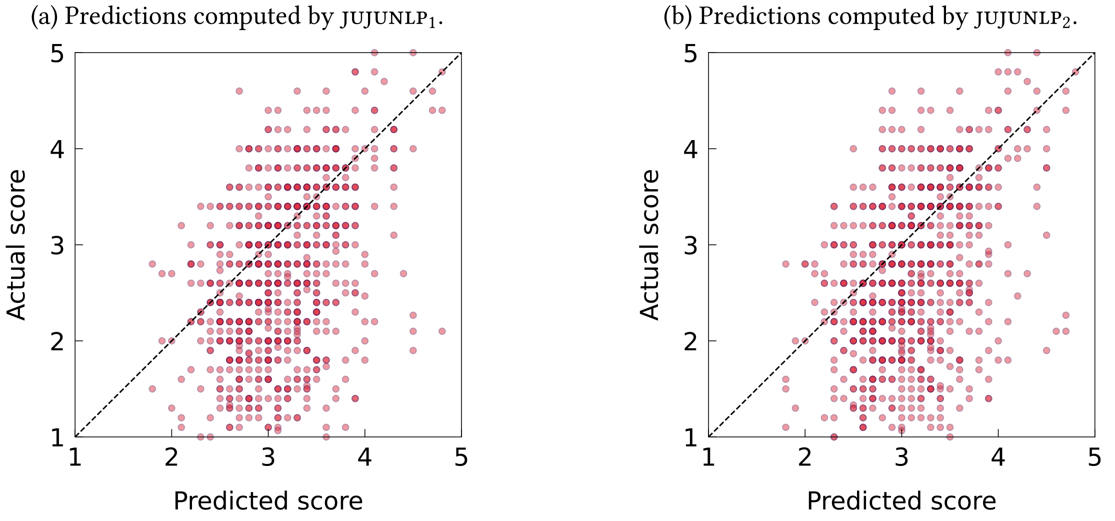

(a) Class "W". (b) Class "L". (c) Class "I". (d) Class "G".

Figure 7: Confusion matrices on the test dataset with the predictions by JUJUNLP$_2$ for subtask 2A.


(a) Class "W". (b) Class "L". (c) Class "I". (d) Class "G".

Finally, the regression task (subtask 2B) posed arguably the most complex prediction scenario. Figure 8 plots the predicted versus actual scores of prejudice degree for the tweets in the test dataset. Remark that correct predictions lie on the main diagonal. The number of points portrayed under this line evidence that the systems tend to find the content of the tweets in the dataset more prejudicial than what their annotators have deemed. Handling this task by ensembling transformer models does not seem to offer many benefits. In fact, BETO (used as a baseline approach in the competition) yielded a lower RMSE on the test dataset. A possible explanation for this is that in this subtask ensembles find it rather difficult to distinguish between humorous and non-humorous texts. For this reason, when their content is offensive, they tend to be rated as highly prejudicial, while for human understanding they may not be so hurtful.

Figure 8: True versus predicted scores on the test dataset for subtask 2B.



(a) Predictions computed by JUJUNLP$_1$.

(b) Predictions computed by JUJUNLP$_2$.

7. Conclusion

In this paper, we introduced a novel approach of transformer ensembles that use a weighted voting system to make predictions on Spanish tweets. In particular, it has been applied to detect hurtful humour, prejudiced groups, and degree of prejudice in the form of binary classification, multilabel classification, and regression tasks, respectively. We empirically assessed the performance of several state-of-the-art transformer models, including RoBERTa, DistilBERT, BERT and BETO (the last two were evaluated on both cased and uncased versions). The predictions of the ensemble systems were calculated as the sum of the individual transformer predictions, weighted by the (normalized) value of the metric they achieved in each task. The experimentation carried out on our test split (a random and stratified sample from the training dataset) reveals that ensembles consistently exceed individual transformers in all studied tasks.

The participation of our approach at the HUHU@IberLEF 2023 competition obtained competitive results. In particular, our ensembles achieved an F1-score of 77.2% in hurtful humour detection, an F1-score of 79.6% in prejudice target detection, and an RMSE of 0.934 in degree of prejudice prediction, thus ranking 12th (27th), 1st (4th) and 22nd (25th), respectively. Specifically, they exhibited strong performance on the multilabel classification task (subtask 2A), outperforming the rest of competitors, by leveraging model specialization, balancing the accuracy and recall of the individual models, managing label imbalance, mitigating biases, and improving the generalization capability of the system to unseen cases during training. We must emphasize that the class distribution in the training dataset for subtask 2A (multilabel) seems to be quite different from the class distribution of the test set. Further, certain instances of the test dataset register a prejudice score smaller than 1 even though this is not contemplated in the description of subtask 2B. Overall, we value these results as encouraging outputs and we firmly believe that with a larger corpus formed of training and test datasets with similar class distributions, the learning process could be considerably improved.

As future work, we plan to include more transformer models pre-trained on Spanish text as part of the ensemble mechanism. We seek to empirically compare the weighted voting system described in this work with alternative ensemble methods, including classical (soft) voting, stacking, bagging and boosting.

In addition, a motivating line is depicted by the translation of Spanish tweets to English, thus opening the possibility to use state-of-the-art transformers pre-trained on large English corpora. Back translation also emerges as a promising approach for data augmentation.

References

[1] L. I. Merlo, B. Chulvi, R. Ortega, P. Rosso, When humour hurts: linguistic features to foster explainability, Procesamiento del Lenguaje Natural 70 (2023).

[2] J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, W. Magdy, 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, Online, 2021, pp. 105–119. URL: https://aclanthology.org/2021.semeval-1.9. doi:10.18653/v1/2021.semeval-1.9.

[3] R. Ortega-Bueno, B. Chulvi, F. Rangel, P. Rosso, E. Fersini, Profiling irony and stereotype spreaders on twitter (irostereo)., in: PAN 2022, 2021.

[4] 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.

[5] R. Labadie, B. Chulvi, P. Rosso, HUrtful HUmour (HUHU): Detection of humour spreading prejudice in Twitter, 2023. URL: https://doi.org/10.5281/zenodo.7967255. doi:10.5281/zenodo.7967255.

[6] L. Chiruzzo, S. Castro, M. Etcheverry, D. Garat, J. J. Prada, A. Rosá, Overview of haha at iberlef 2019: Humor analysis based on human annotation., in: IberLEF@ SEPLN, 2019, pp. 132–144.

[7] L. Chiruzzo, S. Castro, S. Góngora, A. Rosá, J. Meaney, R. Mihalcea, Overview of haha at iberlef 2021: Detecting, rating and analyzing humor in spanish., Procesamiento de Lenguaje Natural 67 (2021) 132–144.

[8] D. Zhang, W. Song, L. Liu, C. Du, X. Zhao, Investigations in automatic humor recognition, in: 2017 10th International Symposium on Computational Intelligence and Design (ISCID), volume 1, IEEE, 2017, pp. 272–275.

[9] R. Zhang, N. Liu, Recognizing humor on twitter, in: Proceedings of the 23rd ACM international conference on conference on information and knowledge management, 2014, pp. 889–898.

[10] A. Ghosh, G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, A. Reyes, SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter, in: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado, 2015, pp. 470–478. URL: https://aclanthology.org/S15-2080. doi:10.18653/v1/S15-2080.

[11] P. Potash, A. Romanov, A. Rumshisky, 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, Vancouver, Canada, 2017, pp. 49–57. URL: https://aclanthology.org/S17-2004. doi:10.18653/v1/S17-2004.

[12] S. Castro, L. Chiruzzo, A. Rosa, Overview of the haha task: Humor analysis based on human annotation at ibereval 2018, in: CEUR workshop proceedings, volume 2150, 2018, pp. 187–194.

[13] S. Castro, L. Chiruzzo, A. Rosá, D. Garat, G. Moncecchi, A crowd-annotated Spanish corpus for humor analysis, in: Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 7–11. URL: https://aclanthology.org/W18-3502. doi:10.18653/v1/W18-3502.

[14] T. Miller, C. Hempelmann, I. Gurevych, SemEval-2017 task 7: Detection and interpretation of English puns, in: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 58–68. URL: https://aclanthology.org/S17-2005. doi:10.18653/v1/S17-2005.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[16] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018). URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019). URL: http://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[18] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108 (2019). URL: http://arxiv.org/abs/1910.01108. arXiv:1910.01108.

[19] 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.

[20] O. Weller, K. Seppi, Humor detection: A transformer gets the last laugh, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3621–3625. URL: https://aclanthology.org/D19-1372. doi:10.18653/v1/D19-1372.

[21] T. Morishita, G. Morio, H. Ozaki, T. Miyoshi, Hitachi at semeval-2020 task 7: Stacking at scale with heterogeneous language models for humor recognition, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 791–803.

[22] N. Hossain, J. Krumm, M. Gamon, H. Kautz, SemEval-2020 task 7: Assessing humor in edited news headlines, in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, International Committee for Computational Linguistics, Barcelona (online), 2020, pp. 746–758. URL: https://aclanthology.org/2020.semeval-1.98. doi:10.18653/v1/2020.semeval-1.98.

[23] A. Ismailov, Humor analysis based on human annotation challenge at iberlef 2019: First-

place solution., in: IberLEF@ SEPLN, 2019, pp. 160–164.

[24] K. Grover, T. Goel, Haha@ iberlef2021: Humor analysis using ensembles of simple transformers., in: IberLEF@ SEPLN, 2021, pp. 883–890.

[25] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, CoRR abs/1909.11942 (2019). URL: http://arxiv.org/abs/1909.11942. arXiv:1909.11942.

[26] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: arXiv preprint arXiv:1506.06724, 2015.

[27] T. H. Trinh, Q. V. Le, A simple method for commonsense reasoning, CoRR abs/1806.02847 (2018). URL: http://arxiv.org/abs/1806.02847. arXiv:1806.02847.

[28] J. Cañete, Compilation of large spanish unannotated corpora, 2019. URL: https://doi.org/10.5281/zenodo.3247731. doi:10.5281/zenodo.3247731.

[29] T. C. Rajapakse, Simple Transformers, https://github.com/ThilinaRajapakse/simpletransformers, 2019.

A. Evaluation Metrics

The fundamental metrics considered in this work during the evaluation of the performance of the transformer ensembles are described below:

- F1-score ranges from 0 to 1 and represents the harmonic mean of precision and recall.

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

- Macro F1-score provides the arithmetic mean of the F1-score for the different classes.

$$\text{Macro F1-Score} = \frac{\sum_{i=1}^{N} \text{F1-score}_i}{N},$$

  where F1-score$_i$ is the F1-score of class i and N is the data split size.

- Root Mean Squared Error (RMSE) is the root of the squared distance between actual and predicted values.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}},$$

  where $\hat{y}_i$ and y_i are the predicted and actual values for instance i, respectively, and N is the data split size.