UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC
DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dept. of Computer Systems and Computation

Challenges in Humor Recognition: Cross-language Perspective and Hurtfulness Analysis

Master's Thesis

Master's Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging

AUTHOR: Labadie Tamayo, Roberto

Tutor: Rosso, Paolo

External cotutor: CHULVI FERRIOLS, MARIA ALBERTA

ACADEMIC YEAR: 2022/2023

Abstract

In the field of Natural Language Processing (NLP), humor recognition presents distinctive challenges for its complete understanding that stem from the requirement to incorporate not only lexical resources but also encompass phonetic and contextual knowledge. These complexities are worsened when knowledge captured by Machine Learning (ML) systems is employed to make predictions in real-world production scenarios on languages that were not included in their training; in these cases, a high variance in inference is observed. At the same time, there is a multifaceted motivation for delving into the meaning behind humorous expressions. This extends to the application of humor recognition in identifying hate speech, particularly in social media, where messages are often concealed within jokes.

Given the aforementioned considerations, this master's degree thesis tackles some challenging aspects within humor recognition, focusing on cross-language perspectives and examining its potentially hurtful nature.

The first part addresses the limited robustness of transformer models in cross-language and cross-domain humor recognition. It highlights the complexities that arise when dealing with creative wordplay and ambiguous phrases in different languages. The study explores how transformer-based models handle these challenges and proposes incorporating multilingual training to enhance humor recognition, while also considering the potential of translation for monolingual assessment. The second analysis delves into humor's capacity to cause harm. The research introduces a novel dataset designed to investigate humor's role in propagating prejudice against marginalized groups in Spanish tweets. The study evaluates various systems approaches and the characteristics of different dataset instances that impact the on performance of presented models in the "HUrtful HUmour (HUHU): Detection of humor spreading prejudice in Twitter" shared task organized within the research framework. Finally, the third part delves into a novel paradigm within the realm of NLP, known as perspectivism, applied to the analysis of humor intertwined with sexist prejudice. Perspectivism introduces the concept of acknowledging the existence of diverse viewpoints when annotating linguistic elements that pertain to topics subject to societal debate. For the study, a subset of HUHU data conveying sexist messages was re-annotated, relying on a large number of annotators with different attitudinal and ideological profiles.

By integrating these analyses, the thesis offers a comprehensive exploration of humor recognition, addressing challenges posed by cross-lingual contexts, the potential harm embedded in humor, and the intricate relationship between the attitudes of the annotators and their observations in a case of humor and sexism.

Keywords: Humor recognition, Cross-language humor, Hurtful humor, Perspectivism in NLP.

Resumen

En el campo de la Procesamiento del Lenguaje Natural (PLN), el reconocimiento del humor presenta retos distintivos para su completa comprensión que se derivan de la necesidad de incorporar no sólo recursos léxicos sino también abarcar conocimientos fonéticos y contextuales. Estas complejidades se agravan cuando el conocimiento captado por los sistemas de Aprendizaje de Máquina (AM) se emplea para realizar predicciones en escenarios de producción reales sobre lenguas que no fueron incluidas en su entrenamiento, observándose en estos casos una elevada varianza en la inferencia. Al mismo tiempo, existe una motivación polifacética para profundizar en el significado que subyace a las expresiones humorísticas. Esto se extiende a la aplicación del reconocimiento del humor en la identificación del discurso del odio, especialmente en las redes sociales, donde los mensajes se ocultan a menudo dentro de los chistes.

Teniendo en cuenta las consideraciones anteriore, esta tesis de máster aborda algunos aspectos desafiantes dentro del reconocimiento del humor, centrándose en perspectivas interlingüísticas y examinando su naturaleza potencialmente hiriente.

La primera parte aborda la robustez limitada de los modelos *transformers* en el reconocimiento del humor entre lenguas y dominios. Destaca las complejidades que surgen al tratar con juegos de palabras creativos y frases ambiguas en diferentes idiomas. El estudio explora cómo los modelos basados en aquitecturas *transformer* afrontan estos retos y propone incorporar el refinado multilingüe de los modelos para mejorar el reconocimiento del humor, al tiempo que considera el potencial de la traducción para la evaluación monolingüe. El segundo análisis profundiza en la capacidad del humor para causar daño. La investigación presenta un novedoso conjunto de datos diseñado para investigar el papel del humor en la propagación de prejuicios contra grupos marginados en tuits en español. El estudio evalúa varios enfoques de sistemas y las características de diferentes instancias del conjunto de datos que impactan en el rendimiento de los modelos presentados en el tarea *"HUrtful HUmour (HUHU): Detection of humor spreading prejudice in Twitter"* organizada en el marco de la investigación. Finalmente, la tercera parte profundiza en un paradigma novedoso dentro del ámbito de la PLN, conocido como perspectivismo, aplicado al análisis del humor sexista. El perspectivismo introduce el concepto de reconocimiento de la existencia de diversos puntos de vista a la hora de anotar elementos lingüísticos relativos a temas sujetos a debate social. Para el estudio, se volvió a anotar un subconjunto de datos de HUHU que transmitían mensajes sexistas, contando con un número considerable de anotadores con diferentes perfiles actitudinales e ideológicos.

Al integrar estos análisis, la tesis ofrece una exploración exhaustiva del reconocimiento del humor, abordando los retos que plantean los contextos multilingües, el daño potencial incorporado en el humor y la intrincada relación entre las actitudes de los anotadores y sus observaciones en un caso de humor y sexismo.

Palabras clave: Reconocimiento del humor, Humor Interlingüístico, Humor hiriente, Perspectivismo en PLN.

Index

Acronyms

BERT Bidirectional Encoder Representations from Transformers

CNN Convolutional Neural Network

DL Deep Learning

GRU Gated Recurrent Unit

HAHA Humor Analysis based on Human Annotation

HUHU HUrtful HUmour

IberLEF Iberian Languages Evaluation Forum

IRA Inter-Rater Agreement

LLM Large Language Model

LSTM Long-Short Term Memory

ML Machine Learning

MLM Masked Language Modeling

MT Machine Translation

NLP Natural Language Processing

NSP Next Sentence Prediction

PLM Pretrained Language Model

RF Random Forest

RMSprop Root Mean Squared Propagation

SVM Support Vector Machines

YAKE Yet Another Keyword Extractor

List of Figures

List of Tables

Chapter 1

Introduction

Humans share a set of evolved emotional behaviors; laughter is part of this universal language of basic emotions recognized by all of us (Savage et al., 2017). Precisely because of its ubiquity and relevance in social life, proper comprehension of some humorous expressions goes beyond the semantics involved in messages. It relies on information from the context where jokes are made and the receptor's background knowledge (Tsakona, 2017), which implies a different or even null perception from one individual to another. This situation is more pronounced when humor is nuanced with resources like irony (Reyes et al., 2012), sarcasm, or hateful speech (Frenda et al., 2022). In the same way, language plays a critical role in perceiving the funny meaning when it comes to such creative devices as humor. Particularly, when information flows from one language to another on its way to the receptor, a joke's intended meaning is at risk of vanishing.

Wordplays are examples of language-dependent expressions that can be potentially misunderstood upon literal translation into a different language since they employ the arrangement and phonetics of words to produce humor. For example, in:

A: *Why do male ants float while female ants sink?*

B: *They're buoy-ant*

It is very challenging to translate the phrase to ensure humor understanding by a non-English speaker, regardless of their background knowledge. Whereas, in the case of:

A: *Are you already here?*

B: *No, I'm just a figment of your imagination.*

The literal translation can still provoke laughter.

On the other hand, linguistic diversity on the Internet increases due to its interconnecting nature (Paolillo, 2007). In social media, where people from different cultural backgrounds and ethnicities share information, dealing with this multilingual phenomenon is inherent when identifying and filtering content and behaviors appropriated for specific users.

Numerous NLP tasks have been covered from a multilingual perspective in the social media scenario with ML models (Ghanem et al., 2020; Wang et al., 2019; Al-Hassan and Al-Dossari, 2019). Most works tackle the under-representation of some languages by extending

the knowledge learned from one language to another. In this sense, multilingual transformer-based architectures have become state of the art in almost all of them (Wang et al., 2020; Chauhan et al., 2022). Despite the growing interest in humor in many languages such as English (Ermakova et al., 2022a; Meaney et al., 2021; Hossain et al., 2020), Spanish (Chiruzzo et al., 2021), Portuguese (Clemêncio et al., 2019), and Chinese (Wu et al., 2021), to the date of this research, few efforts had been made to investigate the task of humor recognition from a computational cross-domain and cross-language perspective.

Machine Translation (MT) paves the way for facing the challenge of multilingualism in texts[1]. Nevertheless, although these tools have been adequate for translating literal texts, when dealing with the figurative language their performance drops considerably. In fact, humorous texts that often appeal to cultural knowledge or play on words become a complex problem in Machine Translation (Attardo, 2002; Zabalbeascoa, 2005; Popa, 2005; Low, 2011).

In light of the facts above, this work at first tries to investigate humor recognition in cross-domain and cross-language scenarios. Particularly, approaching the study for state-of-the-art (sota) solutions, whose backbone architectures are based on transformer models. Hence, the first research question to be addressed is: *How robust are transformers models when dealing with translated humorous messages?*

A second challenge in humor recognition is to identify when the intentionality of a joke goes beyond simple amusement and is employed to harm others in a creative manner, often some individuals belonging to a minority or discriminated groups. In this case, humor is used to express prejudice, defined as "the negative pre-judgment of members of a race or religion or any other socially significant group, regardless of the facts contradicting it" (Jones, 1972). The main fact that contradicts this pre-judgment is that social groups are not homogeneous either in their characteristics or in the way they act and when presented as homogeneous, it is made use of stereotypes (Lipmann, 1922).

The challenge of recognizing this toxic humor is especially relevant for NLP research from a social perspective because when minorities strive for equal treatment, humor is used to evade moral judgment and condemn discrimination (Ford and Ferguson, 2004; Ford et al., 2008). This behavior contributes to the persistence of toxic language in easily accessible platforms like social media. Also, despite its seemingly harmless appearance, such humor carries significant repercussions. It serves as a deterrent and a mechanism of social control: individuals actively seek to avoid behaviors or subjects that become objects of ridicule in their respective societies (Freud, 1960; Billig, 2005).

The impact of offensive jokes extends beyond their immediate context, resulting in more severe consequences. For instance, research about sexism has demonstrated that for men exhibiting high levels of hostile sexism (Glick and Fiske, 1996), sexist humor 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 with humor, individuals targeted by prejudice are more likely to endorse and internalize such expressions (Miller et al., 2019). For that, a second research question is addressed: *Does hurtfulness pose challenges in humor recognition when jokes are nuanced with prejudicial messages?*

A third challenge arises in the comprehensive examination of this phenomenon, particularly within the annotation process of "highly subjective tasks," such as identifying hurtful

[1]https://syncedreview.com/2020/05/20/neural-network-ai-is-the-future-of-the-translation-industry/

humor.

When addressing humor and prejudice, it becomes evident that what might be deemed unacceptable and deeply offensive to one individual or a specific group could be perceived as harmless or merely a jest by another group. Unfortunately, this diversity is overlooked when constructing a dataset as a Gold Standard, where the annotation process yields only a singular correct answer. The new paradigm in NLP known as perspectivism, introduces an alternative approach that questions the feasibility of relying solely on a solitary correct answer during the labeling process, especially when dealing with what various authors have identified as a "highly subjective task" (Basile, 2020; Basile et al., 2022).

In addressing this challenge within the scope of this thesis, it is proposed that the disagreement could be considered indicative of the existence of different perspectives among the annotators. This could explain the difficulty of achieving high levels of inter-annotator agreement in some "highly subjective tasks". This study specifically focuses on sexist humor, intending to comprehend the significance of incorporating perspectives from annotators with diverse attitudinal characteristics. From this theoretical framework, sampling a sufficiently representative number of individuals from a population is pivotal. This step is crucial in mitigating the biases in the final annotation that could potentially propagate to ML systems. For that, the third research question to be addressed is: *How could NLP take into account the different perspectives in a given society on phenomena such as humor and prejudice at the annotation level?*

In a nutshell, while examining the challenges associated with humor recognition in the context of Natural Language Processing, this master's degree thesis focuses on addressing the following research questions:

RQ 1. *How robust are transformer models when dealing with translated humorous messages?*

RQ 2. *Does hurtfulness pose challenges in humor recognition when jokes are nuanced with prejudicial messages?*

RQ 3. *How could NLP deal with the different perspectives in a given society on phenomena such as humor and prejudice at the annotation level?*

To answer these three research questions, three empirical research studies are presented. The first one, aligned with RQ 1, aims to investigate both Machine Translation systems and multilingual transformer-based models to identify their feasibility in humor recognition across languages. This involves studying the ability of multilingual transformer models to identify humor in translated messages and exploring the feasibility of utilizing translation or multilingual models to address the cross-lingual challenges in the English-Spanish scenario.

The second empirical research is conducted in the organization frame of an NLP shared task to study how targeting social minorities with prejudice impacts humor recognition regarding difficulty for ML and Deep Learning (DL) models. The analysis involves the creation of a dataset designed to examine the role of humor in causing harm and prejudice towards minorities. The latter concerns specifically Spanish tweets that are prejudicial against: (i) women and feminists; (ii) the LGBTIQ community; (iii) immigrants and racially discriminated people; and (iv) overweight people.

In the third research, the focus narrows down to a subset of tweets that encompass sexist and humorous content from the dataset mentioned earlier. This investigation involves a

substantial group of 76 annotators, aiming to investigate how varied attitudes impact the annotation process. The goal is to develop a methodology to ensure that the annotators reproduce a representative mix of perspectives in a given annotation task. Specifically, the inter-annotator agreement is explored as a metric to gauge the level of perspectivism inherent in the subjective task of recognizing sexist humor. By delving into these aspects, it is possible to gain valuable insights into the complexities of this task and the subjective nature of hurtful humor recognition when it comes to diverse perspectives. Moreover, in this research, the hypothesis that the disagreement is not individual but social is tested.

The reminder of this master's degree thesis is structured as follows:

- In **Chapter 2** is described the current state of the art, focusing on tasks that relate to the topics examined in this thesis. Specifically, we provide an overview of the most advanced approaches presented at conferences and shared tasks that aim at addressing various facets of humor recognition. These shared tasks encompass a spectrum, ranging from conventional and straightforward humor recognition to the intricate analysis of humor intertwined with offensive language. Also, it is provided a brief review on the landscape of cross-language humor recognition, highlighting the advancements in this domain. Finally, this chapter delves into the efforts directed towards the evaluation of the hurtful aspects present in humorous messages.

- **Chapter 3** presents a study on the cross-language scenario for humor recognition and the robustness of deep learning models to handle it, particularly of transformer models. The experimental frame focuses on English and Spanish languages, and to comprehensively examine the cross-language dynamics, two strategies are adopted: the first is based on multilingual transformer models for exploiting the cross-language knowledge distilled by them, and the second introduces machine translation to learn, and make predictions in a single language.

- In **Chapter 4**, an analysis of the phenomenon wherein humor is employed to convey hurtful messages is conducted, especially its computational implications for ML architectures. Aligned with this, there are described the methodology, dataset, and results of the HUHU shared task organized in the framework of this work on hurtful humor.

- In **Chapter 5** concepts underpinning the construction of the dataset proposed in Chapter 3 are analyzed. Especially a further analysis is performed on the perspectivism of annotators in highly subjective tasks such as sexist humor recognition when they exhibit certain attitudes. This study explores how varying the number of annotators impacts on measures of inter-annotator agreement and whether it adequately encompasses the full spectrum of perspectives.

- Finally, in **Chapter 6** overall conclusions and the interesting findings on the issues analyzed in the manuscript are summarized. Moreover, future lines of work are presented to study more in-depth how to construct more robust systems to address efficiently all these problems.

Chapter 2

State of the Art

2.1 Humor Recognition

Computational humor recognition is a widely explored issue since the 2000s. Some of the first empirical pieces of evidence in this task's feasibility were given by Mihalcea and Strapparava (2005). From there on, several works have been conducted to integrate contextual, visual, and acoustic information in multimodal approaches (Yang et al., 2019b; Vásquez and Aslan, 2021; Song et al., 2021b; Chauhan et al., 2022; Tomás et al., 2022).

This issue has been studied in diverse forums and shared tasks held in conferences, recognizing humor either in its isolated form or nuanced with the presence of other communicative devices such as irony, sarcasm, or hateful speech. These platforms have provided researchers with benchmark datasets and evaluation metrics, enabling the comparison and validation of different approaches. As a result, both traditional machine learning techniques (ML), deep learning (DL), and their combination have been proposed and explored to tackle the complexity of humor recognition. For instance, at the SemEval evaluation forum, during the last years, humor in English was addressed from a computational perspective. In 2017 for Task 6 #HashtagWars: Learning a Sense of Humor (Potash et al., 2017), where participants were asked to detect the top funniest tweets from a given set.

Here, Donahue et al. (2017), combined the use of handcrafted features with deep learning-based approaches. Their system utilized Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), which processed phonetic feature representations combined with GloVe incontextual embeddings (Pennington et al., 2014). The hidden representation of this model was at the time combined with a character-level Convolutional Neural Network (CNN) (Kiranyaz et al., 2021), and an XGBoost (Chen and Guestrin, 2016) component in an ensemble. Baziotis et al. (2017b) presented a fully DL system; specifically they proposed a Siamese architecture (Koch et al., 2015) with a bidirectional Long Short-Term Memory (Bi-LSTM) (Schuster and Paliwal, 1997), augmented with an attention mechanism (Rocktäschel et al., 2015) and based on word embedding representation. On the other hand, Cattle and Ma (2017) explored the use of handcrafted features based on word association features to feed a Random Forest (RF), a classic ML algorithm.

In SemEval 2020, Task 7: Assessing Humour in Edited News Headlines (Hossain et al., 2020), was assessed with the aim of investigating how machine learning systems deal with the recognition of humor caused by local modification on headlines, i.e., short edits applied to

a text which can turn it from non-funny to funny. This time, and after the introduction of the transformer architecture (Vaswani et al., 2017), the best results were obtained by the use of Pretrained Language Models (PLMs), namely Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), RoBERTa (Liu et al., 2019), GPT-2 (Radford et al., 2019) and XLNet (Yang et al., 2019a), and incontextual word embeddings, such as Word2Vec (Mikolov et al., 2013), FastText (Joulin et al., 2017) and GloVe word vectors to exploit other neural networks architectures. Specifically, the two first ranked systems (Morishita et al., 2020; Rozental et al., 2020) were based on ensembling different transformer-based models predictions. Another interesting approach was the one employed by Tomasulo et al. (2020) (ranked 3rd), which combined the predictions made by Gated Recurrent Unit (GRU) (Cho et al., 2014) with a XGBoost regressor. Their proposal fed GRUs with headline representations produced by contextual (transformer-based) and incontextual embeddings to finally ensemble these predictions.

At SemEval 2021, a HaHackathon was organized for detecting and rating humor and Offense (Meaney et al., 2021). This was the first shared task with the aim of detecting offensive language in humorous messages; one of the subtasks asked participants to predict the rate of offense in texts, although from a general perspective. It is worth mentioning that in the proposed dataset, besides humor, instances were annotated along other dimensions, including the presence of offensive language, how much funny texts were, and whether the humor perception resulted in controversy according to the standard deviation from the annotation process. The top-ranked participants made extensive use of PLMs such as BERT, ERNIE 2.0 (Sun et al., 2020), ALBERT (Lan et al., 2020), DeBERTa (He et al., 2020) or RoBERTa, by ensembling them (Song et al., 2021a; Gupta et al., 2021). On the other hand, considering the annotation dimensions, some participants also included multi-task learning strategies, sharing knowledge across tasks. These approaches tended to be more successful across sub-tasks (Pang et al., 2021).

In 2018, at the IberLEF evaluation forum, humor was addressed in Spanish language in the well-known Humor Analysis based on Human Annotation (HAHA) task (Castro et al., 2018) aiming to detect humor and measure it on a continuous scale from 0 to 5, defining its mean degree of funnies. Here, the best-ranked approaches proposed by Ortiz-Bejar et al. (2018) were based mainly on traditional ML approaches exploring Support Vector Machines (SVM), Nearest Centroid, Kernel Ridge, or Ridge. On the other hand, Ortega-Bueno et al. (2018) presented systems based on Bi-LSTM neural networks with attention mechanisms. This network was fed using word2vec pre-trained embeddings, and its output was combined with a representation based on a set of linguistic features (stylistic, structural and content, and effective ones).

On its 2019 edition (Chiruzzo et al., 2019), systems were more focused on the use of deep learning approaches, especially recurrent and transformers architectures; for instance, Farzin et al. (2019) and Ortega-Bueno et al. (2019) employed recurrent-based strategies. The former introduced the use of the Universal Language Model Fine-tuning (ULMFIT) approach (Howard and Ruder, 2018a), whereas the winner system (Ismailov, 2019) as well as most of the top-ranked participants (Mao and Liu, 2019) proposed strategies based on transformer architectures, specifically BERT.

In 2021 it was presented a third edition of HAHA, focused on a more fine-grained analysis of humor where the organizers aimed at detecting the linguistic device employed to convey humor: e.g., irony, wordplay, hyperbole, etc., as well as the content of which jokes were based on, distinguishing among racist jokes, sexist jokes, dark humor, dirty jokes, and others categories (fifteen in total). Almost all the presented systems relied again on the use of neural networks, in most cases PLMs such as BERT, GPT-2, or BETO (Cañete et al., 2020), a BERT-based model trained entirely with Spanish texts. The task organizers reported some that teams also trained other types of models (for example, SVM or Decision Trees) to make comparisons, but none of the participants that sent their system descriptions submitted any of these models. The winner systems (Grover and Goel, 2021), were based on ensembles of multiple transformer architectures fine-tuned on the provided dataset. In the same way Annamoradnejad (2021); Wang et al. (2021); García-Díaz and Valencia-García (2021) proposed BERT-based models fine-tuned to the explored tasks. It is worth mentioning the strategy used by Rodriguez et al. (2021), who starting from transformers-based representations of instances, trained a Siamese Neural Network to classify them according to prototypes determined by a Deep Reinforcement Learning algorithm.

Despite significant progress, humor recognition remains an intriguing and challenging problem. The subjective nature of humor, the cultural context, and individual variations in comedic taste present ongoing obstacles for developing accurate and robust humor recognition systems. However, continued research and interdisciplinary efforts are expected to lead to further breakthroughs in this field, ultimately enhancing the development of more sophisticated and context-aware humor recognition models.

2.2 Humor Across Languages

In the formerly mentioned forums, the issue of cross-linguality in humor recognition has not been explored. In fact, just a few works examine the phenomenon of humor from a cross-language and multilingual view (Chauhan et al., 2022).

Systems based on large PLMs have outperformed state of the art in many NLP tasks, including humor recognition (Grover and Goel, 2021; Subies et al., 2021) and machine translation (Vaswani et al., 2017).
However, humor translation remains a field with a huge room for improvement due to its subjectivity and linguistic complexity. Some of the most recent works (Miller, 2019) provide an interactive method for the computer-assisted translation of puns.

In this line, the task JOKER@CLEF 2022: Automatic Wordplay and Humor Translation Workshop (Ermakova et al., 2022b), where participants were asked to perform translation of humorous texts and identify its nature, was the first attempt to construct a parallel and multilingual humor corpus, specifically with 5K parallel one-liner puns and 1.5K parallel instances of wordplay in named entities. Here, most participants' approaches relied again on transformers-based models, this time reinforced with templates featuring (Arroubat, 2022; Anne-Gwenn et al., 2022).

Recent developments in machine learning and artificial intelligence have greatly improved the quality of MT, but puns are often held to be untranslatable, particularly by statistical or neural MT (Ardi et al., 2022), which cannot robustly deal with texts that deliberately

disregard or subvert linguistic conventions (Miller, 2019). In JOKER 2023: Automatic Wordplay Analysis (Ermakova et al., 2023), the first corpus for wordplay detection in French in the frame of its third task: Pun translation from English to French and to Spanish. Also, it presented a parallel corpus of jokes across these three languages.

Besides the poor existence of parallel corpora, the above-referred issues in humor translation and recognition have worsened the scarcity of works transferring humor knowledge from one language to another.

2.3 Hurtfulness in Humor

In this work, the hurtful character of humor is closely tied to hate speech, defined as offensive expressions directed towards specific groups or individuals (Mondal et al., 2017). This definition is consistent with the meaning that the UN Strategy and Plan of Action on Hate Speech provides: *"any form of communication, be it spoken, written, or behavioral, that attacks or uses pejorative and discriminatory language against individuals or groups based on characteristics such as religion, ethnicity, nationality, race, color, descent, gender, or other identity factors."* This notion aligns with the concept of negative prejudice, as described by Jones (1972), that is to say, a particular instance of offensive speech where individuals are targeted based on preconceived psycho-demographic characteristics that are often stereotyped.

Among the research in NLP, some works have shed light on using humorous messages to convey toxic and stereotyped language in social media. HaHackathon from SemEval 2021, as described in Section 2.1, was the first shared task with the aim of detecting offensive language in humorous messages. Nevertheless, this analysis just included a general perspective without focusing on the expression of prejudice. Also, in HAHA@IberLEF 2021, where the target of the humor was analyzed, some categories, such as `racist joke` and `sexist jokes`, involved stereotyped groups of persons.

Finally, the IROSTEREO shared task in PAN Lab at CLEF 2022 was organized on Profiling Irony and Stereotype Spreaders on Twitter (Ortega-Bueno et al., 2022). Here, participants were asked to determine whether an author of a Twitter feed in English was 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 humor 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), as of the completion of this research, no prior work has specifically assessed the utilization of humor as a means to propagate prejudice against minority groups.

Aside from shared tasks context, recent research in NLP has shown that offensive jokes can be distinguished from non-offensive ones based on the presence of negative stereotypes and ethnic slurs (Merlo et al., 2023; Merlo, 2022). These works revealed that features such as content-related, syntactic, morphological, and affective ones contribute significantly to the differentiation between the two classes of humor. Notably, negative stereotypes, moral and behavioral defects, and the use of swear words emerged as markers of offensive humor.

2.4 Perspectivism and Prejudice in Humor

Previous research addressing the recognition of humor or its offensive character typically relied on training the models with annotated corpora according to the most common procedure to date in NLP: creating a Gold Standard, on the basis of sufficient agreement among annotators.

However, it is well known that both prejudice and humor are controversial issues in our societies. Precisely for this reason, it has been considered relevant to delve into the process of annotating humor that contains prejudice from the perspective of an emerging paradigm which is making its way into NLP: perspectivism or Learning from Disagreements paradigm (for a recent review, see (Uma et al., 2021)). This new approach aims at avoiding the bias of considering a unique and correct vision of one phenomenon captured by a gold standard corpus, even when the problem addressed is the object of a strong social debate such as hate speech or sexist language.

Aligned with the concerns raised by this new paradigm, some research in NLP has addressed the potential biases in the labeling process. For instance, Sap et al. (2022) emphasize the importance of considering demographic, ideological, and attitudinal variations among annotators. To the date of this thesis, there is not previous research that explores the perspectivism paradigm in the annotation of prejudice in humor, despite the fact that both prejudice and humor are controversial topics.

Chapter 3

Humor Across Languages

The portability of humor from one language to another remains challenging for computer machines and even humans. Nevertheless, from the introductory part of this manuscript, it has remained clear the need to extend the capabilities of recognizing it into a cross-language scenario. In this Chapter , an empirical analysis on the capabilities of transformer models is conducted. These models, having demonstrated state-of-the-art performance across numerous NLP tasks, including humor recognition, are evaluated within the context of cross-language scenarios. Also, some alternatives are proposed to smooth the impact of using knowledge between domains and languages. To this aim, this study relies on two strategies: the first is based on multilingual transformer models for exploiting the cross-language knowledge distilled by them, and the second introduces machine translation to learn and make predictions in a single language.

3.1 Methodology

Appraising the robustness of transformer models within the context of cross-lingual analysis mandates the utilization of a parallel corpus enriched with annotated instances of humor. However, due to the scarcity of such a corpus, the approach was adapted to involve the compilation of a dataset considering different sources and languages. This compilation process was particularly focused on the English and Spanish languages, primarily due to the extensive amount of work related to humor recognition on them.

Taking into account the expanding utilization of pre-trained transformer models across various NLP tasks, this study incorporates three multilingual variants of them to assess their performance. However, as this work is not intended to outperform the sota in humor recognition, for classification, it was simply stacked a ReLU-activated layer between the encoder module and a softmax output layer for each model.

Following this architecture, the models were fine-tuned and evaluated with an end-to-end fashion by feeding the ReLU-activated layer with the [CLS] vector from the last encoder block of the transformer. For comprehensive insights, further details pertaining to this process can be found in Section 3.3. The adoption of multilingual models stems from a hypothesis that they possess the capacity to assimilate and share foundational knowledge, regardless of the language employed during the fine-tuning phase or the subsequent evaluation. This

choice reflects the intention to assess the models' capability to capture cross-lingual contextual nuances.

The first model, BERT-multilingual-base (*ml-base*), was pre-trained on the top 104 languages with the largest Wikipedia using a Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives. The second, *ml-sentiment*, is a fine-tuned version of the latter in a sentiment analysis task on texts from six languages[1], among them English and Spanish, which are the ones addressed in this work. This model was employed because although its pre-training knowledge comes from *ml-base*, the information introduced by the sentiment-tuning could provide us with criteria diversity to empirically characterize the general behavior of humor. Finally, it is studied another variation of the BERT-base model (*ml-distil*) into a smaller and distilled architecture proposed by Sanh et al. (2019), trained on the top 104 languages with the largest Wikipedia.

3.2 Datasets

Monolingual datasets from 4 shared tasks were gathered:

(i) SemEval-2020 Task 7: Assessing Humor in Edited News Headlines.

(ii) SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense.

(iii) HAHA@IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish.

(iv) JOKER@CLEF 2022 Task 1: Classify and Explain Instances of Wordplay.

These datasets are annotated with several aspects related to humor, as described in Chapter 2. However, they assess it with texts of different genres and writing styles, including tweets, headlines, or isolated wordplays, and representing different knowledge domains. This enables us to see two perspectives of their aggregation: the *language-level*, where datasets in the same language are grouped into a single corpus, and the *domain-level*, where each dataset is analyzed separately regardless of its language.

SemEval 2020 Task 7 Dataset

In the dataset of this task (*Headlines*), for each headline, there was annotated: the replacement, the humor rating given by six annotators on a 0 to 3 scale, and the mean value of humor rating. From there, negative examples were considered the original headline, and as positive examples of humor those micro-editions whose mean humor rating was above 2 (*i.e.,* moderately funny and funny).

SemEval 2021 Task 7 Dataset

The dataset from this task (*Hahackathon*) contains English texts from Twitter mixed with instances from the Kaggle Short Jokes dataset, described with the presence of humor as well as humor rating, controversy, and offensiveness rating of the messages by 20 different annotators. Being of interest for the present study just the binary annotation regarding whether a text can be considered as funny.

[1] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

IberLEF 2021 HAHA Dataset

In the shared task HAHA 2021, it was proposed a dataset (*HAHA*) composed of tweets written in Spanish, annotated regarding the presence of humor, funniness score, the humor mechanism employed (*e.g.,* parody, stereotype, etc.), and the humor target, *i.e.,* for a humorous tweet, the target of the joke from a set of classes such as racist jokes, sexist jokes, etc.

JOKER@CLEF 2022 Task 1 Dataset

For this task, given a wordplay in the English language, participants were asked to classify it, attending different criteria. They also must identify and disambiguate the target words as an explanation of the wordplay. In the dataset (*JOKER*), the criteria annotated for each example include whether the source and the target of the wordplay co-occur in the text (horizontal/vertical), the manipulation type, *viz.* identity, similarity, permutation, abbreviation, if cultural reference is needed in order to understand the instance of wordplay, whether it is offensive or not, and whether the wordplay is in conventional form. Also, the target words and the disambiguation of the wordplay are annotated.

From these data examples, just were taken those whose manipulation is by permutation (the textual material is given a new order, as in anagrams or spoonerism. e.g. "Dormitory = dirty room"), similarity (source and target are not perfectly identical, but the resemblance is obvious, e.g. "They're called lessons because they lessen from day to day") or Identity (source and target are formally identical, e.g. "How do you make a cat drink? Easy: put it in a liquidizer").

Table 3.1 shows the distribution of the examples in each dataset. The balance between positive (humor) and negative (non-humor) classes follows the one proposed by their authors. For training and testing the models from a cross-language perspective, two partitions are assumed, one composed of examples originally in English from *SemEval 2021 Task 7 Dataset*, *JOKER@CLEF 2022 Task 1 Dataset* and *SemEval-2020 Task 7 Dataset*, with a representation of the positive class at training of 43%. The second is represented just by *IberLEF 2021 HAHA Dataset* dataset with a 39% of humorous examples for training. The instances in both partitions come from different knowledge domains and humor styles; then, besides the cross-language difficulty, a cross-domain phenomenon is faced during evaluation.

Language	Dataset	Train ($\mathcal{T}$)		Test ($\mathcal{D}$)	
		Humor	Non-Humor	Humor	Non-Humor
English	*SemEval 2021 Task 7*	3436	5564	385	615
	JOKER@CLEF 2022 Task 1	531	0	4516	0
	SemEval-2020 Task 7 Dataset	890	890	88	88
Spanish	*IberLEF 2021 HAHA*	11595	18405	3000	3000

Table 3.1: Statistics of the datasets.

3.3 Experimental Framework

When conducting experiments, in the fine-tuning process of every model, parameters were optimized with the Root Mean Squared Propagation (RMSprop) algorithm (Hinton et al., 2012) by employing an increasing learning rate from the shallower layers to the deeper ones (Howard and Ruder, 2018b), starting from 1e-5 and increasing it on each layer with a factor of 0.1 units. Regarding translation steps involved in this work, they were accomplished with googletrans library[2], using Spanish as the complementary language for English, and vice versa. In the same way, when approaches based on back-translation were conducted, this complementarity relation was applied to select the pivot language. For evaluating each strategy, Micro-F1 over the positive class is employed taking into account that at the domain-level, there are cases of slight unbalance or extreme scenarios where there are no examples for the non-humor class, as in *JOKER*.

Given the characteristics of the constructed dataset, it is essential to initially assess the extent of noise present when synthesizing parallel instances using machine translation. Consequently, its impact on the semantics of humorous messages has been examined. In pursuit of this objective, it is evaluated how the perception of humor by these models is altered when applied to a back-translated instance. In doing so, it is ascertained whether the underlying semantic content of the texts (Attardo, 2017) remains intact, even when the inherent humorous incongruity is diminished in the pivot language during the process of back-translation (a result of humor characteristics described in the Introduction).

In Table 3.2 are shown the results in terms of F1 for every model (detailed at *domain-level*), where $\mathcal{D}$ stands for the original version of the test sets described in Table 3.1, $\mathcal{D}^*$ is the version of $\mathcal{D}$ where each instance is translated into its complementary language and $\mathcal{D}^{**}$ corresponds to the back-translated version of $\mathcal{D}$. An estimation of the F1 95%-confidence interval (*ci*) by Percentile bootstrapping according to DiCiccio and Efron (1996) is included.

Model	Dataset	$\mathcal{D}$	*ci*	$\mathcal{D}^{**}$
ml-base	*Hahack.*	0.921	0.015	0.923
	JOKER	0.941	0.005	0.939
	Headlines	0.778	0.062	0.772
	HAHA	0.870	0.008	0.869
ml-sent	*Hahack.*	0.914	0.017	0.916
	JOKER	0.934	0.005	0.933
	Headlines	0.814	0.050	0.802
	HAHA	0.871	0.008	0.870
ml-distil	*Hahack.*	0.905	0.018	0.903
	JOKER	0.945	0.005	0.944
	Headlines	0.709	0.070	0.716
	HAHA	0.863	0.009	0.861

Table 3.2: Variation in humor perception by multilingual transformer models after back-translation.

[2] https://pypi.org/project/googletrans/

Here it can be observed that the error in $\mathcal{D}^{**}$ (a perturbed instance of $\mathcal{D}$) is not significant concerning the results on $\mathcal{D}$ if the learned parameters of the models are assumed.

Since this examination only seeks an empirical probe of the model's capability to find a similar interpretation of the back-translated data w.r.t. the original, for this experiment, every multilingual model was trained by employing all the domains and languages at the same time, allowing the knowledge-sharing among all the *domain-level* datasets.

Nonetheless, a comprehensive exploration was conducted to assess the implications of this knowledge-sharing on outcomes within the cross-domain scenario specifically present in the English *language-level* dataset. The diverse domain combinations employed for fine-tuning the *ml-base* model are outlined in Table 3.3. In these combinations, *K*, *H*, and *J* correspond to *Hahackathon*, *Headlines*, and *JOKER* respectively[3].

Setting	Test Set		
	JOKER	Headlines	*Hahack.*
H	-	0.737	-
K	-	-	0.913
K+H	0.713	-	-
K+J	-	0.667	-
H+J	-	-	0.764
K+H+J	0.906	0.749	0.920

Table 3.3: Cross-domain settings for English datasets.

Looking at the results, it becomes evident that employing a purely cross-domain scenario (rows 3-5) detrimentally impacts the model's performance. Nevertheless, leveraging external knowledge as a method of data augmentation (last row) proves notably effective in enhancing the achieved results.

Also in all cases, the results are inferior with respect to those obtained in Table 3.2, even when the fine-tuning is carried out across all the domains in the English language (last row). The latter suggests that the model employed knowledge from *HAHA* (Spanish corpus) to make inferences in English-written texts. This phenomenon prompted an exploration into the potential effectiveness of employing a multilingual system within a cross-language scenario through a zero-shot approach.

To investigate this, each model underwent further fine-tuning using data from the English *language-level* dataset to assess its performance on the Spanish *language-level* dataset, and vice versa. The outcomes of each case are presented in Table 3.4.

If the results from Table 3.4 obtained using *ml-bert* are compared with those from Table 3.3, it can be observed that the model performance diminishes in each dataset, which supports the critical situation a cross-language scenario represents in comparison with making inferences on different domains and writing styles of humor but in the same language.

After studying the impact of cross-language and cross-domain factors on humor recognition, the feasibility of extending knowledge through translation is examined during the evaluation phase. Additionally, considering the results presented in Table 3.2, it becomes

[3]It is worth mentioning that the model trained solely on the JOKER dataset, could not be included in the evaluation, since this data only consisted of positive examples of humor

Fine-tuning Language	Dataset	ml bert	ml sentiment	ml distil
Spanish	*Hahackathon*	0.760	0.753	0.754
	JOKER	0.666	0.661	0.650
	Headlines	0.534	0.528	0.500
English	*HAHA*	0.754	0.713	0.729

Table 3.4: Cross-language scenario results.

apparent that it is possible to delve into the challenges faced by humor recognition systems, irrespective of any potential *meaning changes* introduced by machine translation

3.3.1 Humor Recognition in Translated Instances

In this analysis, it is explored the strategy of extending knowledge within the context of a cross-language scenario, employing automated instance translation. This exploration pertains to multilingual transformers that have been fine-tuned to recognize humor using English-written messages and subsequently evaluated on Spanish samples. The core addressed question is whether it is more effective to employ automated translation of these Spanish samples into English, thereby mitigating the inherent cross-lingual intricacies. The latter, also for the opposite direction, when using multilingual transformers fine-tuned with Spanish-written messages and evaluated on English samples. Table 3.5 shows the results of the evaluation in these translated $\mathcal{D}^*$ datasets.

Dataset ($\mathcal{D}^*$)	ml bert	ml sentiment	ml distil
Hahackathon	0.808	0.825	0.787
JOKER	0.736	0.719	0.743
Headlines	0.553	0.554	0.512
HAHA	0.767	0.734	0.731

Table 3.5: Language inversion to reduce cross-language effect.

Here, there emerges a discernible improvement in comparison to the earlier findings from Table 3.4, which means a certain degree of humorous perception is preserved after the process of translation and makes more useful the information learned during the model fine-tuning process.

The latter studies do not allow to isolate the vanishing of humor recognition for transformers introduced when instances are translated. To this end, for evaluating the $\mathcal{D}^*$ dataset, it is employed the same model parameters from the experiments referred to in Table 3.2, where cross-domain knowledge sharing was allowed.

Looking over the results presented in Table 3.6 alongside those pertaining to $\mathcal{D}$ and $\mathcal{D}^{**}$ in Table 3.2, a notable lack of robustness of the transformer models becomes apparent within the humor translation context. During the prediction phase, these models encounter shared

Model	Dataset	$\mathcal{D}^*$
ml-base	*Hahack.*	0.880
	JOKER	0.875
	Headlines	0.659
	HAHA	0.811
ml-sent	*Hahack.*	0.856
	JOKER	0.861
	Headlines	0.641
	HAHA	0.803
ml-distil	*Hahack.*	0.833
	JOKER	0.885
	Headlines	0.616
	HAHA	0.789

Table 3.6: Results for evaluation in translated instances.

challenges revolving around polysemous words, ambiguities in phrases originating in the source language and their subsequent translation into the target language, and word rearrangements, particularly when dealing with wordplays. For a more illustrative understanding, specific instances highlighting this issue from both the *Hahackathon* and *HAHA* datasets are provided in Table 3.7.

India is a very peaceful country because nobody has any beef over there. India es un país muy pacífico porque nadie tiene problemas allí.
Two dyslexics walk into a bra Dos disléxicos entran en un sostén
—¿Follamos? —No, que yo recuerde. —"Shall we fuck?" -Not that I remember.

Table 3.7: Translation ambiguities examples.

In the case of the *Headlines* dataset, which exhibits the greatest drop in performance, it can be noticed that besides the translation degeneration, examples are culturally dependent and related to knowledge and vocabulary distant from the one employed in the pre-training and fine-tuning phase of the evaluated models [4]. That is, HAHA vocabulary represents informal Twitter texts, and *Headlines* involves in some way "journalistic" and more formal vocabulary.

[4]In these cases models were fine-tuned with data originally in Spanish (*HAHA*).

Table 3.8 shows some examples related to the *Headlines* phenomenon.

Gov. Kasich slams President Trump's move on haircut care subsidies
White House spokesman does not rule out Trump-Putin July cuddling in Germany

Table 3.8: Contextual dependency of HAHA translated examples.

The experiments developed in this Section showed that humor translation helps the model to extend the knowledge learned in one language for inference in examples written in another one, i.e., it helps to mitigate the cross-language effect in some cases. Nevertheless, these models still struggle in front of the humor complexity as a communicative device when it is translated, effectively tracking a degeneration in the humor perception when messages flow from one language to another.

3.4 Conclusions

Humor relies on the incongruences of two semantic planes that, when contrasted by the receptor, produce it in a natural way. Its translation comes with different implications that make pre-trained transformer-based models not robust to recognize it in a cross-language scenario. The main concerns are related to contextual information, background knowledge dependency, and lexical characteristics of the language. This vanishing becomes more severe in creative ways of humor, such as wordplays involving phonetics, word polysemy, and phrasal ambiguity. Nevertheless, neural machine translation is capable of individually preserving the humorous semantics, as we examined in this Chapter. Also, despite the referred humor recognition vanishing, when the samples are translated and assessed directly in the language of the models' fine-tuning process, they achieve better performance for recognizing humor in a cross-language scenario.

Chapter 4

Hurtful Humor

In Section 2.3, a brief review of the state of the art on research involving relations in humor and prejudiced language was conducted. From there, it was found that no prior work analyzing specifically the confluence of humor and prejudice was done in NLP domain. Therefore, in the framework of this research, we organized HUHU: Detection of Humour Spreading Prejudice in Twitter, as the first shared task in Spanish language focusing on studying humor in prejudicial messages against: (i) *women and feminists*, (ii) *the LGBTI+ community*, (iii) *immigrants and racially discriminated people*, and (iv) *over-weighted people*. The fundamental objective of this study was to address and answer the query posed by RQ2: *Does the presence of hurtfulness pose challenges in humor recognition when jokes are nuanced with prejudicial messages?*

In this Chapter the methodology, dataset, and results from HUHU are described; furthermore, some insights into the interplay between prejudiced language, minority groups, and humor when it serves to convey hurtful content are provided.

4.1 HUHU Dataset Construction

As a first step for this research, an initial corpus was constructed together with 80 students of psychology that manually tracked down Twitter accounts to study the characteristics of users who spread prejudice against minorities using humor. This characterization 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, it was conducted an information retrieval strategy using 898 user accounts. Among all the targets identified in the preliminary corpus, in the context of this research, the interest lay in studying the four aforementioned groups, i.e.: women and feminists, the LGBTI+ community, immigrants and racially discriminated people, and over-weighted people. Hence, this initial set of tweets was considered as a corpus of toxic language containing instances belonging to 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. It must be noticed that these macro-classes were assumed regardless the existence of humor in any example.

From the Twitter accounts posting tweets belonging to the positive macro-class, the last 1000 tweets posted after January 1st, 2020 were retrieved, taking these accounts as potential

prejudicial speech spreaders. The latter process yielded a set of roughly 80 thousand instances. To filter these instances and focus on the topic of interest represented by the macro-class, a set of discriminative keywords was employed (see Table A.1 Appendix A.1). These keywords were obtained from various sources: (i) KeyBERT (Grootendorst, 2020), (ii) Yet Another Keyword Extractor (YAKE) (Campos et al., 2020), and the top 100 terms[1] according to the information gain in the distribution of the two classes.

4.1.1 Annotation Process

The filtering yielded a reduction of nearly 30 thousand tweets. For each account, duplicated instances were observed. Inspired by (Chiruzzo et al., 2021), graphs interconnecting tweets for each user were constructed, and those with a Jaccard similarity above 0.7 were grouped together by a cut-off. Later, it was employed the Grivan Newman algorithm (Girvan and Newman, 2002) to find communities of similar texts, and annotators were provided 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 step 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 respectively:

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 that seemed to use humor 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 the potential dataset was obtained, a second annotation step was done by a team of five annotators (three female and two male students) hired by UPV. They were asked to identify the minority group being the target of prejudice and for each of them the prejudice degree in a *discrete* scale ranging from 1 to 5 where 1 means a lower prejudice degree and 5 the opposite (see annotators guidelines in Appendix A.2).

The overall degree of prejudice towards each minority in a given instance was determined by the average scoring provided by the five annotators (Equation 4.1). Subsequently, the prejudice score of the tweet was defined as the mean prejudice value towards *targeted* minorities as in Equation 4.2:

$$A_k^{(i)} = \frac{1}{5}\sum_j T_{jk}^{(i)} \qquad \forall k, i \tag{4.1}$$

$$S^{(i)} = \frac{1}{\sum_k \mathbb{1}(A_k^{(i)} > 0)}\sum_k A_k^{(i)} \qquad \forall i \tag{4.2}$$

[1]Here punctuation marks, stop words, and other semantically meaningless aspects were excluded

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 the k^{th} target.

Examination of the mean prejudice distribution among annotators with the Kolmogorov-Smirnov test yielded a non-normal distribution of the degree of prejudice in all targeted groups ($p < 0.001$). This skewed data distribution leads to low agreement among different raters when using conventional Inter-Rater Agreement (IRA) measures (Di Eugenio and Glass, 2004). To address this issue, it was employed the Gwet's AC_1 measure of IRA (Gwet, 2008), which utilizes a probabilistic model of agreement (Paun et al., 2022). This approach estimates the difficulty levels of the items within the corpus through probabilistic inference and then separately estimates the probability of chance agreement for easy and hard items. This probabilistic modeling approach helps mitigate the impact of the skewed data distribution on the agreement assessment process. Table 4.1 shows the IRA for each prejudiced target individually, i.e., women and feminists (G_1), LGBTI+ community (G_2), immigrants and racially discriminated people (G_3) and over-weighted people (G_4) for the whole set of instances and for prejudicial texts nuanced with humor.

	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 4.1: Gwet's AC_1 measure of IRA across annotators from the second phase for each prejudiced minority. Sub-indices on each entry represent the size of the confidence interval for $\alpha = 0.05$

From here, it can be observed particularly low IRA values for the target related to women and the feminist movement. This difference is intriguing because all tasks are subjective tasks regarding the definition given by Wong et al. (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.

HAHA 2021

Due to some difficulties in the retrieval of humorous texts, in the first annotation step, some tweets from the dataset proposed in (Chiruzzo et al., 2021) were included. 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. Positive examples of humor comprising entities related to the studied minority groups were filtered out (see Table A.3 in Appendix A), incorporating 1402 instances to the annotation flow, which were reduced to 503 in the final dataset.

4.1.2 Dataset Statistics

After both annotation steps, the final distribution of tweets remained as shown in Table 4.2.

From the columns denoted with emojis 😆 and 😐, representing humorous and non-humorous instances respectively, it can be noticed an important imbalance. On the other hand, it is important to note that a single tweet might 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.

Source	😆	😐	G_1	G_2	G_3	G_4
Crawled	607	2323	1652	791	753	169
HAHA	518	1	328	66	89	100
Total	1125	2324	1980	857	842	269

Table 4.2: Final dataset statistics.

Conducted 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 4.1, which focuses specifically on pairwise relations.

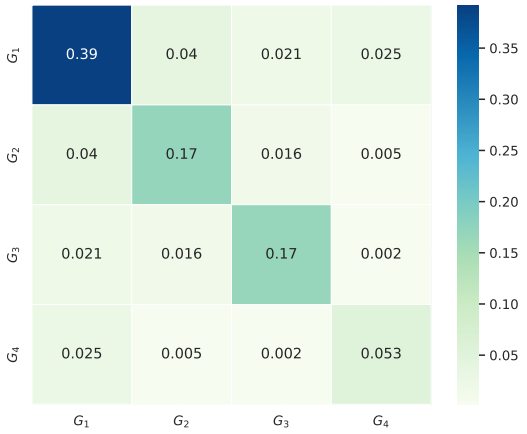

Figure 4.1: Pairwise co-occurrence of minorities being the target of prejudice.

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

Regarding the proportion of humorous and non-humorous messages targeting each minority, it can be noticed a consistent pattern of unbalance 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, as illustrated in Figure 4.2.

Finally, it was investigated how levels of prejudice, as measured by Equation 4.2, are distributed in both positive and negative cases of humor. The distribution of these prejudice levels, as depicted in Figure 4.3, reveals that there is a shift towards more hurtful messages among tweets that convey jokes. This observation offers empirical evidence that humor 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 humor in reinforcing and perpetuating prejudice (Miller et al., 2019).

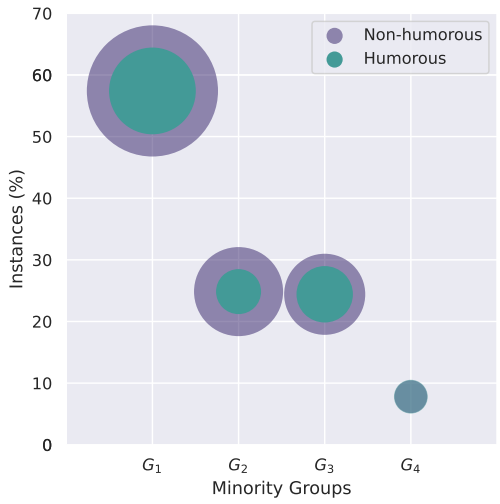

Figure 4.2: Humor vs Non-humor for each minority group.

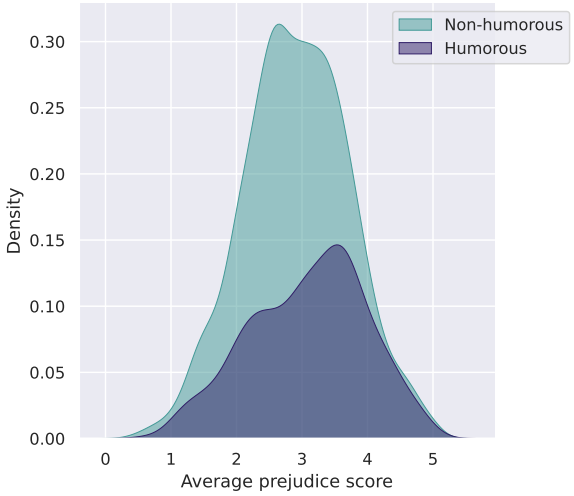

Figure 4.3: Degree of prejudice in humorous/non-humorous texts.

4.2 HUHU: Detection of Humor Spreading Prejudice in Twitter

HUHU was held as part of the Iberian Languages Evaluation Forum (IberLEF) in its 2023 edition. IberLEF[2] is a shared evaluation campaign for NLP systems in Spanish and other Iberian languages. In an annual cycle that starts in December (with the call for task proposals) and ends in September (with an IberLEF meeting collocated with SEPLN[3]), several challenges are run with large international participation from research groups in academia and industry. In HUHU 46 teams out of the 77 registered, made at least one submission,

[2] https://sites.google.com/view/iberlef-2023/home
[3] http://www.sepln.org/

including participation from Spain, Mexico, Portugal, China, and India.

Three distinct subtasks were explored to assess the HUrtful HUmor (HUHU) observed in the dataset and the dimensions of prejudice. In this Section, each of them is described along with the corresponding metrics employed to assess the performance of the proposed systems.

4.2.1 HUrtful HUmor Detection

The first subtask consisted in determining whether a prejudicial tweet is intended to cause humor. Participants had to distinguish between tweets that use humor to express prejudice and tweets that express prejudice without humor. Systems were evaluated and ranked employing the F1-score over the positive class defined as:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

4.2.2 Prejudice Target Detection

Subtask 2a, considering the minority groups analyzed, had the aim of identifying 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 4.1.2.

$$\bar{F_1} = \frac{1}{4} \sum_k F_1^{(k)}$$

4.2.3 Degree of Prejudice Prediction

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

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

where n is the size of the dataset, $\hat{y}_i$ is the value prejudice estimated for an instance i, and y_i its corresponding ground truth value.

4.3 Experimental Framework and Baseline Models

Three baseline models to establish a comparative framework were examined. These models encompassed different approaches, one utilizing a classic machine learning approach and the other leveraging state-of-the-art transformer architectures.

The first baseline model utilizes 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, it was incorporated a fine-tuning version of the BLOOM model (Scao et al., 2022), a multilingual model, on its bloom-1b1 variant ($\sim$ one billion of parameters).

During the fine-tuning process of the transformer-based models, it was employed the RMSprop algorithm (Hinton et al., 2012) for parameter optimization following the strategy proposed in the Universal Language Model FineTuning (ULMFiT) (Howard and Ruder, 2018b). That is, for each layer of the PLM a different learning rate was set up, increasing it using a multiplier while the neural network gets deeper. This multiplier changes by a 0.1 factor from a layer L_i to another L_{i+1}. This dynamic learning rate was used to keep most information from the pre-training at shallow layers and biasing the deeper ones to learn about the specific tasks. In previous works related to hate speech, humor recognition, and other ways of figurative language (Labadie et al., 2021a; Palomino and Ochoa-Luna, 2020; Labadie et al., 2021b), this strategy has yielded a performance improvement. In the case of the BETO model a batch size for fine-tuning of 32 examples was employed and 16 for BLOOM.

In addition to the aforementioned models, a more naive approach for the classification task was explored. 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.

4.3.1 Dataset Partitioning

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

When constructing the training and test datasets, we ensured they closely reflected the distributions observed in the overall corpus. This was 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 4.3. The training set was composed just of instances containing prejudice towards one or two minorities simultaneously at most, given the reduced number of tweets with three or more targets.

Source	😆	😐	G_1	G_2	G_3	G_4
Train	869	1802	1292	607	664	214
Test	256	522	688	250	178	55

Table 4.3: Distribution into training and test set.

It was also ensured that the distributions depicted in Figure 4.3 were preserved in training-test partitioning, including the skewness towards more hurtful content for jokes. The final dataset (Labadie et al., 2023) is available online as part of the Zenodoo community. [4]

4.4 HUHU Analysis

The study conducted as part of this research starts with a comparative review on the models employed by participants in the competition and how their approaches differ. Most of

[4] https://zenodo.org/record/7967255

the presented systems involved preprocessing steps and employed traditional ML and DL models, specifically transformer-based architectures (Table A.4, Table A.5 and Table A.6 in Appendix A.4 show the top-ranked systems along with the results of the proposed baselines for each explored subtask[5]). In this section, observations related to these aspects are exposed.

4.4.1 Preprocessing

As part of their preprocessing strategy, several teams employed various techniques such as converting all tweets to lowercase, lemmatizing 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 process, discarding any remaining unmasked instances. Moreover, some 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 tokenize the tweets directly (Kaoshik and Kather, 2023; Peng and Lin, 2023; Inácio and Oliveira, 2023).

4.4.2 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 et al., 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 of 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) utilized 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 et al. (2023) adopted a similar strategy but using Decision Trees Regressors and SVMs instead. Finally, Inácio and Oliveira (2023) and Sacristán et al. (2023) employed contextual embeddings coming from Large Language ModelLLMs (LLMLLMs) to feed SVMs in subtask 1.

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

On the other hand, some systems solely relied on pre-trained and fine-tuned LLMs based on transformer architectures, like the best-performing system in the regression subtask proposed by the team M&C, which consisted in a simple fine-tuning of RoBERTa model. 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 et al., 2021), RoBERTuito Cased, BERT Cased 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.

4.5 Discussion

As stated in the beginning of this Chapter, a near-balanced number of submissions was observed using both DL and traditional ML architectures, leading to a pretty fair analysis. Figure 4.4 shows the distribution of F1-scores for both subtask 1 (left) and subtask 2a (right) on DL and ML approaches.

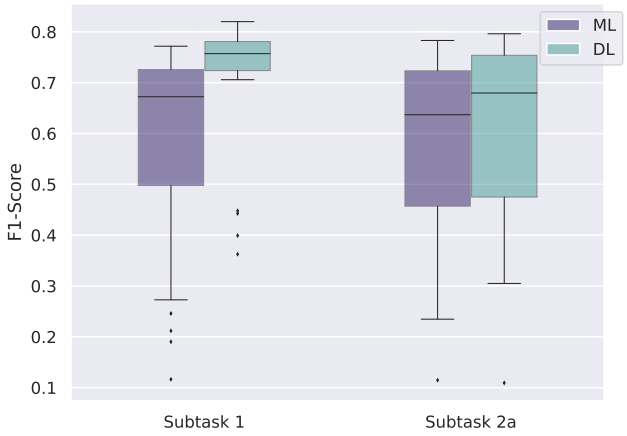

Figure 4.4: ML and DL systems performance for subtask 1 and subtask 2a.

From here, it can be seen that most DL-based systems, as well as the average within this category, exhibited superior performance for humor recognition task. However, the situation differs when considering subtask 2a, which involves identifying the targeted group in tweets. In this case, there is considerable reliance on specific terms related to the ground truth minority. Consequently, even straightforward techniques for text representation like Bag of Words (BoW) can yield nearly as precise predictions as the more intricate modeling approaches employed by transformer-based models.

Regarding subtask 2b in Figure 4.5, where the Kernel Density Estimation of the systems' performance is depicted, it can be observed a greater representation of ML systems, specifically

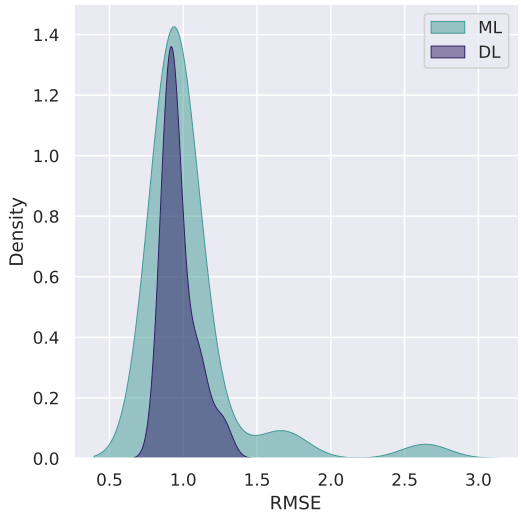

Figure 4.5: ML and DL systems performance for subtask 2b.

just 17 of the 55 documented submissions were based on DL. Nevertheless, systems under this paradigm (DL) again presented a more regular and shifted distribution towards lower RMSE values, i.e., a better performance. In fact, it must be pointed out that the top three ranked systems were based on transformer-based architectures.

To explore the errors made by the systems in subtask 1, test instances were categorized 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)[6]. From this schema, a significant relationship (Pearson-$\chi^2 = 238.545$, df $= 3$, $p < 0.001$) was noticed between the difficulty level and whether the text was a joke. Among the non-humorous texts, approximately 65% of instances were classified as *easy* or *very easy*. However, this percentage dropped significantly for the humorous class to only 7.8% of instances. These findings are summarized in Table 4.4.

Difficulty	Non humor	Humor	Humor G_1	Humor G_2	Humor G_3	Humor G_4
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 4.4: Percentage of correctly classified instances according to its difficulty category.

In addition, it was analyzed whether a particular target group introduced any difficulty in the systems' ability to discriminate between humorous and non-humorous instances. As a result, no significant relation was found between the difficulty level and the target group mentioned in instances when not conveying funny messages. However, when analyzing jokes specifically, the results of the Pearson-χ^2 test showed a significant relationship between the

[6]see some examples for each category in Appendix A.5

degree of difficulty in recognizing humor and the groups mentioned in the text (Pearson-χ^2 = 34.071, df = 16, $p < 0.005$).

As shown in Table 4.4, the most challenging aspect for humor recognition is related to instances mentioning the LGBTI+ group (Humor G_2). This finding indicates that jokes referring to this particular group tend to be more difficult for the systems to be recognized as humorous compared to other instances in this particular dataset. These findings shed light on the complexities involved in humor recognition, especially when it comes to sensitive topics or target groups. This highlights the importance of considering the impact of context and target groups when developing systems for humor analysis and understanding, especially for models whose performance is closely related to the lexicon employed in textual messages.

For subtask 2a, it was examined the overall capacity of systems to identify the mention of each minority group in the instances. As shown in Table 4.5, systems had more difficulty recognizing every targeted group when more than one was mentioned in the same tweet. The Mann-Whitney Test was applied, and all scores differed significantly ($p < 0.001$).

It was also explored if the presence of humor introduced any significant impact on the subtask 2a, that is to say, the recognition of the target group. The Mann-Whitney Test indicates that the LGBTI+ group (G_2) is better recognized in humorous texts than in non-humorous, and for the over-weighted people group (G_4) is observed the converse situation.

Moreover, at the level of each instance, a significant Spearman correlation was spotted between the percentage of systems correctly recognizing the target group and the level of prejudice (mean) of this instance. This correlation indicates that systems had an increased ability to recognize all target groups in tweets being judged more prejudicial by annotators (see Table 4.5).

Target	Non-humor	Humor	Single group	Multiple groups	Correlation with prejudice degree
G_1	38.61	38.09	45.01	30.08	0.310**
G_2	41.39	46.44	52.06	31.56	0.499**
G_3	45.88	46.67	52.21	38.40	0.561**
G_4	53.84	51.72	55.00	50.77	0.354**

**Correlation is significant at the 0.01 level.

Table 4.5: Percentage of teams correctly identifying the target group.

Regarding subtask 2b, the best-ranked submission predictions were analyzed and computed the differences between the prejudice scores given by the annotators and their predicted scores. In Figure 4.6, positive values represented cases of overestimating the degree of prejudice, and negative values the opposite.

The potential effects of the presence of humor and the number of targeted minority groups on the performance of the best system were explored in terms of the aforementioned differences. 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), parametric analysis was conducted to explore the interaction effect of these two variables.

By performing the ANOVA, a significant interaction between the two independent vari-

ables (number of groups mentioned x humor vs non humor) was observed ($F(1) = 15.008$, $p < 0.001$) as well as 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 humor and non-humorous texts in terms of this underestimation (see Figure 4.6).

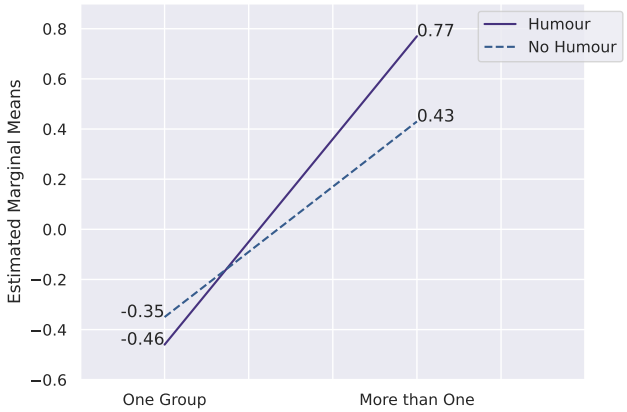

Figure 4.6: Prejudice degree according to humor presence and number of targeted groups on instances.

In addition, the study also explored the estimation of the degree of prejudice in tweets that mentioned specific target groups. The objective was to determine if specific and isolated[7] groups introduced significant differences in the level of prejudice predicted by the systems.

The results indicated that while the interaction between humor and prejudice was not found to be significant, there was a significant effect observed by the group mentioned in the tweet with respect to the degree of prejudice ($F(2) = 24.446$, $p < 0.001$). This means that the target groups mentioned in the tweets had a noteworthy impact on the perceived level of prejudice present in the messages. The results showed that the model tended to overestimate the degree of prejudice against the LGBTI+ group (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 overweight people always appeared alongside another target group; thus, it was not included in this particular analysis.

4.6 Conclusions

This chapter focused on exploring different aspects of hurtfulness in humor and how Artificial Intelligence systems, particularly ML-based and DL-based models, deal with this phenomenon in three related tasks. Specifically, in the first analyzed issue - *HUrtful HUmor Detection* -, it was observed that transformer-based models obtain a higher performance with respect to traditional machine learning algorithms in identifying instances of hurtful humor. Concerning the *Degree of Prejudice Prediction*, a similar phenomenon across the different analyzed systems

[7]It must be noticed that instances targeting only one group were considered

was identified, resulting in the best estimators RoBERTa-based models. In contrast, for *Prejudice Target Detection*, despite the best-performed system being based on ensembling multiple transformers models, traditional machine learning techniques were not far behind in terms of performance. In this case, the performance of both paradigms was quite balanced due to the highly lexicon-dependent characteristic of this task.

From the error analysis, we can infer that it is evident that humor recognition is still a challenge for most systems, especially when humor is used to convey prejudice. Notably, a skewed distribution in the difficulty categories is observed in relation to the systems' predictions, with a prevalence towards more complex groups within humorous instances. This observation indicates difficulties in identifying instances of humor. Interestingly, it also was found that the degree of prejudice played a crucial role in aiding systems to recognize the target group: when humans label a tweet as more prejudicial, the systems recognize better the victim of this prejudice. Furthermore, during the corpus construction, a tendency was observed where the perception of prejudice increased when the targeted group was also an object of mockery. This phenomenon was reflected in the skewness observed in the prejudice degree distribution for the humor class concerning the non-humor class. The latter gives some pieces of evidence of the perspectivism and subjectivity with which the annotation process of this kind of data is approached.

Chapter 5

Perspectivism in the Annotation of Sexist Jokes

As stated at the beginning of this thesis, NLP tasks whose instances labels depend on social and background-knowledge, rarely exceed moderate levels of inter-rater agreement in corpora (Landis and Koch, 1977) without restrictive specifications or training that limit the annotators' freedom of expression (Kiela et al., 2020; Kocoń et al., 2021). In the same way, some works have shown that ML models' quality is directly related to the inter-rater agreement in the dataset in which these models are trained on (Richie et al., 2022; Waseem, 2016).

This Chapter[1] delves into the analysis of disagreement arising during datasets construction, a crucial attribute of highly subjective tasks like the one explored in Chapter 4, where it was observed a tendency to assume jokes mentioning certain groups as more prejudicial by observers in the annotation process of the HUHU dataset. Moreover, in that study, when annotators were tasked to identify the presence of prejudice related to a given target group, a notable variability emerged from one group to another in terms of the Inter-Rater Agreement (IRA), as depicted in Table 4.1. This realization prompted the proposition to further analyze the existence of divergent perspectives when annotating sexist jokes (the one supposed to be more controversial according to Table 4.1).

This Chapter examines the third research question of this thesis: *How could NLP deal with the different perspectives in a given society on phenomena such as humor and prejudice at the annotation level?*. To this end, the nature of these perspectives inherent in such annotation tasks is studied aiming to determine whether these viewpoints align with social stances or stem from individual positions, considering that a suitable size for the annotator group could be delimited for the first scenario. The latter comes under the hypothesis that from a certain point, adding more annotators does not increase the disagreement in an annotation task, which is tested in this Chapter.

To facilitate this investigation, a subset of 210 sexist jokes was selected from the HUHU dataset (see on Section 4.1). These instances were subjected to re-annotation by a panel of 76 annotators, whose sexist attitudes and ideology were previously measured via an anonymous questionnaire. Several statistical analyses were made to test the influence of attitudes in the annotation and identify the different perspectives among the annotators groups.

[1] A more elaborated version of this research has been presented to the 2nd Workshop on Perspectivist Approaches to NLP and is currently in press (Chulvi et al., 2023)

5.1 The Problem of Different Perspectives

In modern computational linguistics, the standardized annotation process of a corpus includes different techniques to classify a single piece of language in a given taxonomy. It implies training annotators (Gomez et al., 2019; Suryawanshi et al., 2020), multiple classification subjects, measures of inter-rater agreement, harmonization (Kiela et al., 2020), aggregation by the majority, and construction of a "gold standard" corpus representing potential probability distributions which are used to interpolate real-world scenarios by machine learning.

The first evidence of the relevance of different perspectives in the annotation task is the recent research of Sap et al. (2022) showing strong associations between annotator identity and beliefs and their toxicity ratings. Specifically, their results show that more conservative annotators and those who scored highly on a racist beliefs scale were less likely to rate anti-black language. Also, considering how annotation is related to the social position the annotator holds, Akhtar et al. (2019, 2021) leverage different opinions emerging from groups of annotators to study how polarized instances affect the performance of the classifiers. Considering binary classification tasks, they introduce a novel measure of the polarization of opinions able to identify which instances in a dataset are more controversial. In a pilot study about xenophobia arguments in the context of Brexit, the annotation process was organized to contrast the annotation done by three people with an immigrant background (target group) vs the one done by three people with a mainstream background as a control group. Using their polarization index[2], the authors show how in several tweets, all the members of the target group (immigrants) marked the message as racist and hateful, while the members of the control group marked it as conveying no hate or racism. Interestingly, they only found a few tweets (1.13%) on which all the annotators agreed that they contained hateful messages.

Implicitly, in this work, the authors assume a similar perspective to the one defended in this chapter, i.e., the nature of the disagreement is social and sustained by a social conflict, but they do not provide any empirical measure of annotators' attitudes. Their results suggest that consensus-based methods to create gold standard data (as it was done for HUHU) are not necessarily the best choice when dealing with what they call highly subjective phenomena.

The results of these researches are in line with the sift paradigm introduced in the *Perspective Data Manifesto*[3]. In this theoretical framework, this new paradigm presents a more rigorous approach to handling subjective tasks within the realm of NLP. It advocates for releasing datasets in pre-aggregated formats and emphasizes the importance of constructing new evaluation metrics encompassing diverse perspectives from different backgrounds. The incorporation of perspectivism in NLP research continues to expand, as highlighted in a recent review (Uma et al., 2021) presenting as a significant concern: the potential labeling bias introduced due to the cultural backgrounds of annotators (Sap et al., 2019; Waseem, 2016).

5.2 Attitudes in the Annotation Task

In binary classification tasks, annotators rely on their personal attitudes and beliefs to make decisions, as shown in the result of Sap et al. (2022). In the context of the current annotation

[2]To measure polarization for a message with annotations from two annotator groups, the authors use the normalized χ^2 statistics in their approach. This statistic tests the independence of annotation distributions against a uniform distribution, with a uniform distribution indicating total disagreement. Normalizing χ^2 yields values between 0 (total disagreement) and 1 (perfect agreement).

[3]https://pdai.info/

task, the relevant attitude is the level of sexism of annotators.

Traditionally, sexism has been understood as encompassing discriminatory attitudes towards women, including both overt and subtle forms (Swim and Hyers, 2009). The ambivalent sexism theory, proposed by Glick and Fiske (1996, 2011), introduced the concept of two interrelated components: hostile sexism (overtly negative attitudes towards women) and benevolent sexism (seemingly positive attitudes that are actually discriminatory). Despite their contrasting tones, these components are positively correlated and work together to sustain gender inequalities (Barreto and Doyle, 2023). Another related concept is neosexism or modern sexism, similar to modern racism, which involves denying ongoing discrimination, opposing women's demands, and rejecting policies for enhancing women's societal standing (Tougas et al., 1995; Swim et al., 1995).

In a recent review on ambivalent sexism, Barreto and Doyle (2023) point out future directions in the study of sexism due to the rapid developments in societal norms and attitudes towards sex, gender, and sexuality across many countries. Notably, despite the growing research highlighting a rise in men with self-proclaimed anti-feminist agendas (Blais and Dupuis-Déri, 2012; Blais, 2021; O'Donnell, 2021), the authors don't explicitly delve into investigating the relationship between hostile sexism and anti-feminist attitudes. Exploring this interaction becomes pertinent because a distinct form of intense hostility towards women is using anti-feminist frameworks while endorsing certain feminist policies such as equality (Off, 2023).

Chulvi et al. (2023) propose to denominate this new latent attitude, *Hostile neosexism*. In contrast to existing scales like modern sexism or neosexism, *Hostile neosexism* exhibits a higher degree of hostility towards women. At its core, *Hostile neosexism attitude* claims that societal changes driven by feminism inherently disadvantage men as a group. Despite the hostile sexism subscale (Glick and Fiske, 1996) was primarily driven by the idea that men's dominance over women is both appropriate and desirable, some items of this subscale connect well with the idea that nowadays there is no reason for feminist demand and that the feminist movement overreacts (see items 3, 4 and 5 in Appendix B.1). In the frame of this research, the attitudes of annotators in a preliminary version of the *Hostile neosexim scale* have been measured via an anonymous questionnaire.

5.3 Study Design and Data Annotation

This study relied on a manually selected set of 210 jokes conveying prejudice against women from the HUHU corpus (Section 4.1). During the annotation process of the HUHU dataset, as previously described, three annotators assessed each instance for the presence of humor and prejudice, following a criterion for annotation based on the relative majority agreement of the annotators, with a threshold of 2 out of 3. For the present study, jokes that convey different kinds of prejudice against women were selected and classified into five categories to describe the dataset's content. The five categories are listed below, along with their corresponding representation percentage concerning the total 210 instances, accompanied by an illustrative example of a joke for each category.

1. **Present women as dummies, only concerned about their bodies or about money** (40% of the dataset), e.g.: "If Socrates had been a woman, he would have said: "I just know that I don't know what to wear".

2. **Feature women as possessive, complicated and dominant** (22.5%), e.g.: "Women

get angry for 5 reasons: 1) For everything 2) For nothing 3) Because they do 4) Because they don't 5) Just in case".

3. **Say that they are gossips and enemies among themselves** (2.5%), e.g.: "If women governed, there would be no World War III, only little groups of countries badmouthing and smiling at each other".

4. **Introduce them as malicious, sluts or justifies violence** (12%), e.g.: "Women are like bags of ice, with a few punches they loosen up".

5. **Anti-feminist jokes** (23%), e.g.: "I have just been informed that Spanish troops on the war front are being brutally offended by macho and patriarchal attitudes on the part of the Russian army. It is a disgrace that this is still happening in the 21st century".

A total of 76 psychology students, comprising 76.3% women and 23.7% men, participated in the experiments as a practical exercise component during their first year of the bachelor's degree. This activity spanned two hours and was conducted in a distraction-free environment. To ensure anonymity, each student was assigned a confidential number. They were granted access to an Excel document for labeling the jokes. The annotation process entailed three tasks. In task 1, participants read the 210 jokes and categorized them as either containing sexism (prejudice against women) or not. Additionally, they were required to indicate if the text is intended to be humorous (task 2). For task 3, annotators were asked to rate the offensiveness of the prejudice present in the text on an ordinal scale: 0 (not at all), 1 (slightly), 2 (somewhat), and 3 (very much).

After completing the annotation tasks, students utilized their confidential numbers to respond to a questionnaire. This questionnaire encompassed the *Hostile neosexism scale* and a question regarding their ideology. This question about ideology was included because political conservatism has been found to explain more variance in *ambivalent sexism* than gender (de Geus et al., 2022; Hellmer et al., 2018).

To measure annotators' attitudes in *Hostile Neosexism*, a short scale denominated *Brief Hostile Neosexism Scale* was used. It was composed of six items: three of them (4 to 6) are part of the *Hostile Sexism subscale* of the *Ambivalent Sexism Scale* from Glick and Fiske (1996) and the other three (1 to 3) are new items created *ah-hoc* to measure anti-feminist attitude in the framework of this research (See questionnaire in Appendix B.1). To measure ideology, annotators were asked to indicate their position by answering the following question: If you had to define your political orientation, where would you place yourself on this scale? The answer must be expressed on a 7-point Likert-type scale where 1 was "left" and 7 was "right". The voluntary participation and data anonymity were guaranteed following the European Code of Conduct for Research Integrity[4].

5.4 Experimental Framework

As stated before, this work evaluates the influence of attitudes on the annotation process. To derive annotators' latent attitudes, an Item Factor Analytic approach is employed, which is an extension of classical linear factor analysis and particularly suitable for addressing categorical

[4]https://allea.org/code-of-conduct/

variables. Specifically, within the framework of Item Response Theory (IRT) (de Ayala, 2009), it is adopted the two-parameter normal ogive (2PNO) formulation (Samejima, 1969):

$$Pr(X_{ik} = c|\theta_i, \boldsymbol{\gamma}_k, \lambda_k) = \Phi\left(\lambda_k\theta_i - \gamma_{k,c}\right) - \Phi\left(\lambda_k\theta_i - \gamma_{k,c+1}\right) \tag{5.1}$$

where $\Phi(\cdot)$ is the normal cumulative function. Here the probability of observing a given category $c = 1, \ldots, C$, for unit $i = 1, \ldots, N$ and item $k = 1, \ldots, K$, is modeled in terms of the latent trait θ_i, the factor loading λ_k and a vector of ordered threshold $\boldsymbol{\gamma}_k$. To estimate the model parameters, it was embraced a fully Bayesian approach that incorporated the handling of missing values (Fontanella et al., 2016).

Regarding the study of the inter-rater agreement in the task of annotating sexism, because the sampled data comes from the HUHU dataset, it was observed that in the binary annotation scheme, most of the texts are categorized as jokes conveying prejudice against women by annotators, with 81% of the annotations falling into this category. This skewed data distribution leads to a low level of agreement among different raters when using traditional inter-rater agreement measures such as Fleiss' κ or Kripendorf's α, the same as presented in the analysis in the full dataset construction (see Section 4.1.1). This discrepancy arises from the paradoxical situation where the observed agreement appears to be very high, while the chance-corrected agreement is actually low (Di Eugenio and Glass, 2004). To address this issue, the Gwet's AC_1 measure of inter-rater agreement is employed again.

5.4.1 Social or Individual Disagreement in the Perspectivism Paradigm?

As pointed out in the beginning of this Chapter, when dealing with the perspectivism paradigm, i.e., and considering the existence of different perspectives, the first question to solve is whether they are individual or social positions. In the second case, it is possible to delimitate a suitable size of the annotator's group representing the different perspectives in a task. Nevertheless, if the disagreement is an individual phenomenon, this would not be possible. This research tries to test the hypothesis that the disagreement has a social nature, and from a certain point adding more annotators does not increase the disagreement between annotators.

In order to explore the impact of the number of annotators on the inter-rater agreement, samples were generated without replacement from the pool of 76 annotators for each instance within the dataset. The size of these samples initially spans from 3 to 45 annotators. To ensure statistical reliability, a total of 10,000 iterations were conducted for each sample size.

The results of this analysis are presented in Figure 5.1. In particular, Figure 5.1 (a) depicts the mean and 95% confidence interval for each sample size. To determine the optimal annotator sample size that leads to stabilization in the variability of the Gwet's AC_1 coefficient, the knee-point method was employed (Kaplan, 2023). This method is commonly used to identify the point at which a graph exhibits a significant change in slope. In this study, the knee-point method was applied to the amplitude of the confidence intervals (see Figure 5.1(b)). Through the application of the knee point method, an annotator sample size of $n = 12$ was determined to be the point of stabilization for AC_1 variability, indicating that further increases in the number of annotators do not yield significant modification in agreement.

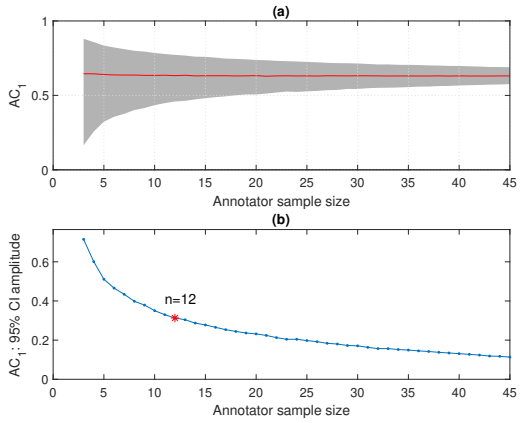

Figure 5.1: Simulation results: (a) Mean and 95% confidence interval of Gwet's AC_1 coefficient; (b) Amplitude of the 95% confidence interval of Gwet's AC_1 coefficient and knee-point.

5.4.2 Attitudes Influence on Inter-Rater Agreement

A Bayesian exploratory IRT analysis was employed, following the approach described in (Fontanella et al., 2019), to evaluate the construct validity of the scale outlined in Section 5.3. The results of the analysis indicated that the scale exhibits unidimensionality, supporting its validity as a measurement tool for the intended construct. Therefore, a unidimensional 2PNO model (Equation 5.1) was exploited to estimate the *Hostile neosexism* attitude of the annotators, taking into account the influence of their gender and ideology as relevant features. The estimated values for the model parameters can be found in Table 5.1. The factor loadings indicate the weight of the corresponding items in the derivation of the latent trait scores, while the location values give insights on the level of consolidation of the corresponding *Hostile neosexism* attitude: lower values correspond to a belief that gains more support in the sample (Villano et al., 2017). It must be noticed that for the regression parameter estimates, the only covariate that seems to impact the *Hostile neosexism* attitude significantly is endorsing the right ideology.

To assess the influence of the *Hostile neosexism* attitude on the level of agreement, it was contrasted the inter-rater agreement among the $n = 12$ annotators in three subgroups: a homogeneous group with the lowest scores on the *Hostile neosexism* attitude, a homogeneous group with the highest scores, and a mixed group with six annotators positioned at the lower end of the *Hostile neosexism* and six annotators positioned at the higher end. The observed and expected agreements and the Gwet's AG_1 coefficients for all the 76 annotators and for the three subgroups are shown in Table 5.2. The results demonstrate a clear distinction in the level of agreement among the annotators with lower *Hostile neosexism* attitude compared to the other groups. On the other hand, the agreement within the mixed group is similar to that observed in the overall population of annotators, indicating a comparable level of consensus among individuals with varying levels of *Hostile neosexism* attitude.

A second sub-sampling strategy was developed to test the influence of attitudes on the level of agreement. A simulation was conducted with a sample size of $n = 12$, and the sample units were randomly selected from sub-populations characterized by scores on the latent trait below the first quartile (*Low Hostile Neosexism*), above the third quartile (*High Hostile Neosexism*),

	posterior mean	95% credible interval
	Factor loadings	
Item6	1.674	(1.054, 2.671)
Item1	1.137	(0.745, 1.572)
Item3	1.059	(0.706, 1.433)
Item4	1.408	(0.950, 1.996)
Item5	1.271	(0.826, 1.821)
Item2	0.717	(0.404, 1.034)
	Location value$^{(a)}$	
Item1	-0.971	(-1.320, -0.619)
Item3	-0.640	(-0.948, -0.333)
Item4	-0.536	(-0.886, -0.207)
Item2	0.444	(0.157, 0.733)
Item5	0.682	(0.390, 0.989)
Item6	1.131	(0.761, 1.541)
	Regression coefficients	
intercept	-0.628	(-1.256, 0.000)
male	0.351	(-0.244, 0.946)
left$^{(b)}$	-0.500	(-1.217, 0.215)
moderate left$^{(b)}$	0.000	(-0.753 , 0.733)
right$^{(b)}$	0.744	(0.004, 1.513)

(a) average of the threshold values for a given item

(b) baseline: centre

Table 5.1: Hostile neosexist scale: parameter estimates.

	n	observed agreement	expected agreement	Gwet's $AC1_1$
All annotators	*76*	*0.741*	*0.298*	*0.631*
Lowest Hostile Neosexism	12	0.828	0.209	0.782
Highest Hostile Neosexism	12	0.722	0.306	0.599
Mixed Hostile Neosexism	12	0.751	0.273	0.658

Table 5.2: Inter-rater agreement comparison for samples with different Hostile Neosexism attitudes.

and evenly distributed between the two sub-populations (*Mixed Hostile Neosexism*). From each group, 10,000 samples were selected without replacement. The findings (see Figure 5.2) provide further evidence of the influence of attitude on the level of agreement in the annotation process.

Following the two strategies, it was found that again the level of agreement decreases among the *Mixed Hostile Neosexism* group but also among *High Hostile Neosexism*. The decline in agreement among mixed groups is understandable but would not be expected among homogeneous groups high in *Hostile Neosexism*.

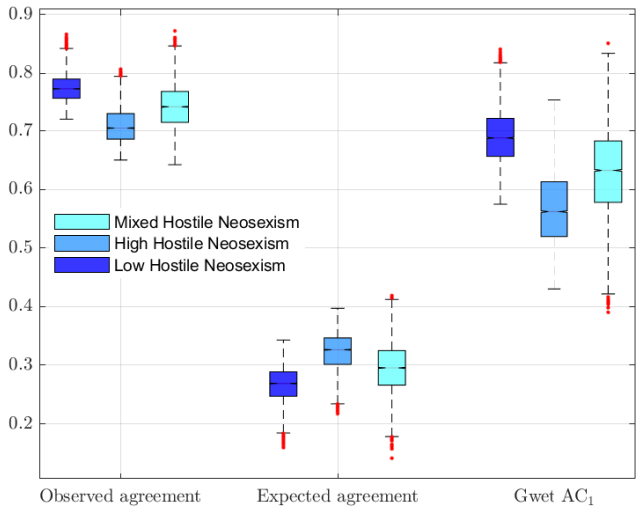

Figure 5.2: Observed and Expected inter-rater agreements and Gwet's AC_1 for samples of $n = 12$ annotators with low, high, and mixed scores on the Hostile Neosexism attitude.

This unexpected result provides insights for further research where exploring the inconsistency between attitude and behavior would be appropriate. Annotating instances from a corpus is a behavior in the sense that the subject makes a decision and acts according to this decision. Then the different results obtained in the group with High Hostile Neosexism and in the group with Low Hostile Neosexism could be investigated in the framework of the relation between attitudes and behaviors. The relation between these two concepts is a complex matter that doesn't have a clear-cut answer in the realm of social psychology literature (for a comprehensive understanding see the review by Eagly and Chaiken (1993)).

For instance, in early steps of the discipline Campbell (1963) argued that individuals with negative attitudes towards minority groups might be reluctant to express those attitudes publicly due to societal norms of tolerance and politeness. The inconsistencies between attitudes and behaviors can be attributed to false conformity with social norms or to a lack of clear social norms for the subject. Related to our data, it is clear that the social norm in the context of this annotator, students of Psychology in an Western country such as Spain, is more to be low in *Hostile neosexism*. Then, it will not be surprising that in the group of annotators that manifests higher scores in *Hostile neosexism*, the inconsistencies between attitude and behavior were greater, impacting the inter-rater agreement.

5.5 Conclusions

This chapter delved into a deep examination of the data annotation issue, particularly in the context of datasets like the one introduced in Chapter 4, where instances are assessed along controversial dimensions, such as prejudicial humor. Within this study's framework, a methodology was developed to tackle several common challenges that arise when attempting to translate the concept of perspectivism into a coherent annotation strategy. Firstly, with regards to the inclusion of a larger number of observers in the annotation process and its

impact on inter-rater agreement (or disagreement), it was evident that the nature of such disagreement in the annotation is more a social phenomenon than an individual one. This is why, beyond a certain point, adding more individuals does not significantly increase the level of disagreement. Furthermore, it was ascertained that diverse perspectives stem from individual attitudes, but there is also a need for more research that explores the not-always-consistent behaviors exhibited by annotators in alignment with these attitudes.

Chapter 6

Conclusions and Future Work

This master's thesis has tackled significant challenges in humor recognition while also exploring implications and methodologies stemming from including elements like sexist humor or other subjective communicative devices within annotation procedures. This consideration is especially vital when approaching the linguistic application of humor recognition from a computational perspective and employing Machine Learning approaches.

The research work carried out in the framework of this thesis was comprised of three parts. The first one aimed to answer **RQ 1.** *How robust are transformers models when dealing with translated humorous messages?* where the impact of translations on the semantics of humorous messages was assessed to determine whether a prior vanishing of funniness was a consequence of the translation. Subsequently, some strategies were investigated to mitigate challenges posed by cross-language scenarios, leading to an empirical analysis of the models' robustness for humor recognition when facing these translations compared to the original messages.

It was found that the translation of humorous texts comes with different linguistic implications that make pre-trained transformer-based models less adept at recognizing humor in cross-language scenarios. The main concerns were related to contextual information, background knowledge dependency, and lexical attributes of the language. This vanishing becomes more pronounced when dealing with creative forms of humor, such as wordplays involving phonetics, word polysemy, and phrasal ambiguity. Nonetheless, neural machine translation demonstrated the ability to preserve humorous semantics individually. Also, despite the referred humor recognition vanishing, when the samples are translated and assessed directly in the language of the models' fine-tuning process, humor is better recognized in a cross-language scenario.

In this line, future works may extend the conducted analysis towards a broader spectrum of languages and translations provided by readily available machine translation systems to ensure reproducibility. Additionally, since almost every top-ranked system proposed in the shared tasks related to the explored datasets employed transformer-based architectures (see Section 2.1), it would be reasonable to evaluate them on the experiments presented in this study to obtain more empirical evidence. Furthermore, two potential strategies deserve to be investigated further, considering the pivotal role of cultural and contextual knowledge in influencing performance. Firstly, examining how mitigating topic bias in datasets could aid models in addressing the cross-domain phenomenon (a challenge arising when extending knowledge from one dataset to another). Secondly, a strategy involving partially updating

the model knowledge could be pursued. This would entail identifying key examples as domain concepts from new datasets and incorporating them during the model's fine-tuning process.

The second main investigation was centered around addressing **RQ 2.** *Does hurtfulness pose challenges in humor recognition when jokes are nuanced with prejudicial messages?*. This investigation was conducted in the context of the proposed shared task at IberLEf 2023 on *HUrtful HUmor (HUHU): Detection of Humour Spreading Prejudice in Twitter*. This was the first shared task in the Spanish language focusing on studying humor in prejudicial messages against: (i) *women and feminists*, (ii) *the LGBTI+ community*, (iii) *immigrants and racially discriminated people*, and (iv) *over-weighted people*.

In the framework of this shared task, different aspects of hurtfulness in humor were explored, and how Artificial Intelligence systems, particularly ML-based and DL-based models, address this phenomenon in three related subtasks. Specifically, with respect to the first issue on *HUrtful HUmor Detection*, it was observed that transformer-based models exhibited higher performance than traditional machine learning algorithms in identifying instances of hurtful humor. Likewise, a similar phenomenon across the different analyzed systems was identified for the third aspect on *Degree of Prejudice Prediction*. In contrast, for *Prejudice Target Detection*, despite the most effective system being based on ensembling multiple transformers models, traditional machine learning techniques were not far behind in terms of performance. This equilibrium in performance between the two paradigms can be attributed to the task's heavy reliance on lexicon-dependent characteristics.

Derived from the error analysis, it became evident that humor recognition is still a challenge for most systems. Upon categorizing each instance based on the difficulties encountered by the proposed systems in classifying them as humorous or not, a skewed distribution on the categories of difficulty was observed, with a prevalence towards more complex groups within humorous instances. This observation indicates difficulties in identifying instances of humor.

Another important observation resulting from the analysis was a tendency during the corpus construction where the perception of prejudice increased when the targeted group was also an object of mockery. This phenomenon was reflected in a skewness observed in the prejudice degree distribution for the humor class w.r.t. the non-humor class. Moreover, when annotators were asked to identify the presence of prejudice to a given target group, a notable variability emerged from one group to another in terms of IRA. The latter gave some pieces of evidence of the perspectivism and subjectivity with which the annotation process of this kind of data is approached.

As a further step in investigating these phenomena, it would be interesting to incorporate a broader set of training examples into the dataset. These additional examples could be sourced from a more diverse set of accounts while adhering to the same data mining framework to avoid biases during the training of ML systems. Additionally, attention should be directed towards alleviating the imbalance between the positive and negative humor classes. Furthermore, an extension of the analysis could encompass the evaluation of the performance of large language models, such as LLama2 (Touvron et al., 2023), ChatGPT and GPT-4 (Liu et al., 2023), etc. Including these models, alongside exploring alternative techniques like in-context learning and zero-shot learning, could greatly enrich the results. This extension would also facilitate the assimilation of insights from new resources that are now available, including the ones mentioned above.

Finally, to address **RQ 3.** *How could NLP take into account the different perspectives in a given society on phenomena such as humor and prejudice at the annotation level?*, and taking

some insights from constructing the HUHU dataset, the third study assessed issues related to annotating data along highly subjective dimensions.

This exploration led to the formulation of a methodology designed to tackle prevalent challenges that emerge when attempting to translate the concept of perspectivism into a coherent annotation strategy. This strategy duly considers diverse viewpoints from individuals with varying attitudes, encompassing their contextual background and knowledge.

When we analyzed how involving a larger number of observers in the annotation process affected inter-rater agreement or disagreement, it became clear that this disagreement primarily stemmed from a social phenomenon rather than an individual one. The latter is aligned with the idea that a person can represent a way of thinking (perspective). This is why, beyond a certain point, adding more individuals did not significantly increase the level of disagreement.

Also, it was ascertained that diverse perspectives stem from individual attitudes, but there is also a need for more research that explores the inconsistent or consistent behaviors exhibited by annotators aligned with these attitudes. The latter, according to the low levels of agreement observed in the experiment within homogeneous groups of individuals (*High Hostile Neosexism*). Hence as a future line of research, this work should include an analysis of the possible relation between attitudes and behaviors and how these two elements could produce the perspective that a subject represents. On the other hand, it is important to acknowledge that the extended annotator team was composed solely of psychology students; this is why even when this group exhibited multiple perspectives according to observers' attitudes, a wider and heterogeneous group of observers should be included in further studies. Finally, to avoid the constraint of including only sexist jokes in the study, we will consider a more complex scenario in a further analysis.

References

María Carmen Aguirre and Angel Cadena. 2023. Using Vector Embeddings and Feature Vectors to Humor Identification. In *IberLEF@SEPLN*, CEUR Workshop Proceedings. CEUR-WS.org.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A New Measure of Polarization in the Annotation of Hate Speech. In *AI2019 – Advances in Artificial Intelligence*. Springer International Publishing.

Sohail Akhtar, Valerio Basile, and Vivivana Patti. 2021. Whose Opinions Matter? Perspective-aware Models to Identify Opinions of Hate Speech Victims in Abusive Language Detection. *CoRR*.

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of Hate Speech in Social Networks: A Survey on Multilingual Corpus. In *6th International Conference on Computer Science and Information Technology*, volume 10.

Issa Annamoradnejad. 2021. ColBERT: Using BERT Sentence Embedding for Humor Detection. In *IberLEF@SEPLN*, volume abs/2004.12765.

Bosser Anne-Gwenn, Ermakova Liana, Dupin de Saint-Cyr Florence, De Loor Pierre, Charpennay Victor, Pépin-Hermann Nicolas, Alcaraz Benoit, Autran Jean-Victor, Devillers Alexandre, Grosset Juliette, Hénard Aymeric, and Marchal-Bornert Florian. 2022. Poetic or Humorous Text Generation: Jam event at PFIA2022. In *Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings.

Iván Árcos and Jaime Pérez. 2023. Detecting HUrtful HUmour on Twitter using Fine-Tuned Transformers and 1D Convolutional Neural Networks. In *IberLEF@SEPLN*. CEUR-WS.org.

Havid Ardi, Muhd Al Hafizh, Iftahur Rezqy, and Raihana Tuzzikriah. 2022. Can Machine Translations Translate Humorous Texts? *Humanus: Jurnal Ilmiah Ilmu-ilmu Humaniora*, 21.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel 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.

Hakima Arroubat. 2022. Wordplay Location and Interpretation with Deep Learning Methods. In *Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings.

Salvatore Attardo. 2002. Translation and Humour. *The Translator*, 8.

Salvatore Attardo. 2017. *The General Theory of Verbal Humor*, chapter 10. Routledge.

Manuela Barreto and David Matthew Doyle. 2023. Benevolent and Hostile Sexism in a Shifting Global Context. *Nat Rev Psychol*, 2.

V. Basile, T. Caselli, A. Balahur, and L. Ku. 2022. Editorial: Bias, Subjectivity and Perspectives in Natural Language Processing. *Frontiers in Artificial Intelligence*, 5.

Valerio Basile. 2020. It's the End of the Gold Standard as we Know it. On the Impact of Pre-aggregation on the Evaluation of Highly Subjective Tasks. In *DPAI*.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In *Italian Conference on Computational Linguistics*.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017a. 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)*. Association for Computational Linguistics.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017b. DataStories at SemEval-2017 Task 6: Siamese LSTM with Attention for Humorous Text Comparison. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics.

Michael Billig. 2005. *Laughter and Ridicule: Toward a Social Critique of Humour*. Sage.

Melissa Blais. 2021. The Impact of Masculinist Counter-framing on the Work of Meaning-making of Violence Against Women. *Interface*, 13.

Melissa Blais and Francis Dupuis-Déri. 2012. Masculinism and the Antifeminist Countermovement. *Social Movement Studies*, 11.

María J. Blanca, Rafael Alarcón, Jaume Arnau, Roser Bono, and Rebecca Bendayan. 2017. Non-normal Data: Is ANOVA Still a Valid Option? *Psicothema*, 29(4):552–557.

Hugo Albert Bonet, Aina Magraner Rincón, and Alba Martínez 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.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Donald T. Campbell. 1963. Social Attitudes and Other Acquired Behavioral Dispositions. In *Psychology: A study of a science. Study II. Empirical substructure and relations with other sciences. Investigations of man as socius: Their place in psychology and the social sciences*, volume 6. McGraw-Hill.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! Keyword Extraction from Single Documents Using Multiple Local Features. *Information Sciences*, 509.

Santiago Castro, Luis Chiruzzo, and Aiala Rosá. 2018. Overview of the HAHA Task: Humor Analysis Based on Human Annotation at IberEval 2018. In *IberEval@ SEPLN*.

Andrew Cattle and Xiaojuan Ma. 2017. SRHR at SemEval-2017 Task 6: Word Associations for Humour Recognition. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics.

Dushyant Singh Chauhan, Gopendra Vikram Singh, Aseem Arora, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A Sentiment and Emotion Aware Multimodal Multiparty Humor Recognition in Multilingual Conversational Setting. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. Association for Computing Machinery.

Luis Chiruzzo, Santiago Castro, Mathias Etcheverry, Diego Garat, Juan José Prada, and Aiala Rosá. 2019. Overview of HAHA at Iberlef 2019: Humor Analysis Based on Human Annotation. In *IberLEF@SEPLN*.

Luis Chiruzzo, Santiago Castro, Santiago Góngora, Aiala Rosa, J. A. Meaney, and Rada Mihalcea. 2021. Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. *Procesamiento del Lenguaje Natural*, 67.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Association for Computational Linguistics.

Berta Chulvi, Lara Fontanella, Roberto Labadie, and Paolo Rosso. 2023. Social or individual disagreement? Perspectivism in the annotation of sexist jokes. In *PERSPECTIVES 2023. Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop 2023*. CEUR-WS.org.

André Clemêncio, Ana Alves, and Hugo Gonçalo Oliveira. 2019. Recognizing Humor in Portuguese: First Steps. In *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, Proceedings, Part II*. Springer-Verlag.

Javier Cruz, Lucas Elvira, Miguel Tabernero, and Isabel Segura-Bedmar. 2023. In Unity, There Is Strength: On Weighted Voting Ensembles for Hurtful Humour Detection. In *IberLEF@SEPLN*. CEUR-WS.org.

R. J. de Ayala. 2009. *The Theory and Practice of Item Response Theory*. The Guilford Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, volume 1. Association for Computational Linguistics.

Barbara Di Eugenio and Michael Glass. 2004. The Kappa Statistic: A Second Look. *Computational Linguistics*, 30.

Thomas J DiCiccio and Bradley Efron. 1996. Bootstrap confidence intervals. *Statistical science*, 11(3):189–228.

David Donahue, Alexey Romanov, and Anna Rumshisky. 2017. HumorHawk at SemEval-2017 Task 6: Mixing Meaning and Sound for Humor Recognition. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics.

Alice. H Eagly and Shelly Chaiken. 1993. *The Psychology of Attitudes*, chapter The Impact of Attitudes on Behaviors. Harcourt brace Jovanovich college publishers.

Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023. Overview of JOKER - CLEF-2023 track on Automatic Wordplay Analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association*.

Liana Ermakova, Tristan Miller, Fabio Regattin, Anne-Gwenn Bosser, Claudine Borg, Élise Mathurin, Gaëlle Le Corre, Sílvia Araújo, Radia Hannachi, Julien Boccou, Albin Digue, Aurianne Damoy, and Benoît Jeanjean. 2022a. Overview of JOKERCLEF 2022: Automatic Wordplay and Humour Translation Workshop. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Springer International Publishing.

Liana Ermakova, Tristan Miller, Fabio Regattin, Anne-Gwenn Bosser, Claudine Borg, Élise Mathurin, Gaëlle Le Corre, Sílvia Araújo, Radia Hannachi, Julien Boccou, et al. 2022b. Overview of JOKER CLEF 2022: Automatic Wordplay and Humour Translation workshop. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 447–469. Springer.

Bobak Farzin, Piotr Czapla, and Jeremy Howard. 2019. Applying a Pre-trained Language Model to Spanish Twitter Humor Prediction. In *IberLEF@SEPLN*.

Lara Fontanella, Sara Fontanella, Pasquale Valentini, and Nickolay Trendafilov. 2019. Simple Structure Detection Through Bayesian Exploratory Multidimensional IRT Models. *Multivariate Behavioral Research*, 54.

Lara Fontanella, Pasquale Villano, and Marika Di Donato. 2016. Attitudes Towards Roma People and Migrants: A Comparison Through a Bayesian Multidimensional IRT Model. *Quality and Quantity*, 50.

Thomas E. Ford, Christie F. Boxer, Jacob Armstrong, and Jessica R. Edel. 2008. More than "just a joke": The Prejudice-releasing Function of Sexist Humor. *Personality and Social Psychology Bulletin*, 34(2):159–170.

Thomas E. Ford and Mark A. Ferguson. 2004. Social Consequences of Disparagement Humor: A Prejudiced Norm Theory. *Personality and Social Psychology Review*, 8.

Simona Frenda, Alessandra Cignarella., Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The Unbearable Hurtfulness of Sarcasm. *Expert Systems with Applications (ESWA)*, 193.

Sigmund Freud. 1960. *Jokes and their Relation to the Unconscious*. Norton.

Pablo Sánchez García and Constantino Martínez de la Rosa. 2023. Dimensionality Reduction Techniques to Detect Hurtful Humour. In *IberLEF@SEPLN*. CEUR-WS.org.

José Antonio García-Díaz and Rafael Valencia-García. 2021. UMUTeam at HAHA 2021: Linguistic Features and Transformers for Analysing Spanish Humor. The What, the How, and to Whom. In *IberLEF@SEPLN*, volume 2943 of *CEUR Workshop Proceedings*. CEUR-WS.org.

José Antonio García-Díaz and Rafael Valencia-García. 2023. UMUTeam at HUHU 2023: Detecting Prejudices in Humour Using Ensemble Learning and Knowledge Integration. In *IberLEF@SEPLN*. CEUR-WS.org.

Roosmarijn de Geus, Elizabeth Ralph-Morrow, and Rosalind Shorrocks. 2022. Understanding Ambivalent Sexism and its Relationship with Electoral Choice in Britain. *British Journal of Political Science*, 52.

Bilal Ghanem, Jihen Karoui, Farah Benamara, Paolo Rosso, and Véronique Moriceau. 2020. Irony Detection in a Multilingual Context. In *European Conference on Information Retrieval*. Springer.

Michelle Girvan and Mark EJ Newman. 2002. Community Structure in Social and Biological Networks. *Proceedings of the national academy of sciences*, 99.

Peter Glick and Susann. T. Fiske. 1996. Ambivalent Sexism Inventory: Differentiating Hostile and Benevolent Sexism. *Journal of Personality and Social Psychology*, 70.

Peter Glick and Susann. T. Fiske. 2011. Ambivalent Sexism Revisited. *Psychology of women quarterly*, 35.

Raul Gomez, Jaume Gibert, Lluís Gómez, and Dimosthenis Karatzas. 2019. Exploring Hate Speech Detection in Multimodal Publications. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467.

Maarten Grootendorst. 2020. KeyBERT: Minimal Keyword Extraction with BERT.

Karish Grover and Tanishq Goel. 2021. HAHAIberLEF2021: Humor Analysis using Ensembles of Simple Transformers. In *IberLEF@SEPLN*.

Aishwarya Gupta, Avik Pal, Bholeshwar Khurana, Lakshay Tyagi, and Ashutosh Modi. 2021. HumorIITK at SemEval-2021 Task 7: Large Language Models for Quantifying Humor and Offensiveness. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics.

Kilem Li Gwet. 2008. Computing inter-rater Reliability and its Variance in the Presence of High Agreement. *British Journal of Mathematical and Statistical Psychology*, 61.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *CoRR*, abs/2006.03654.

Khal Hellmer, Johanna T. Stenson, and Kristi M. Jylhä. 2018. What's (not) Underpinning Ambivalent Sexism?: Revisiting the Roles of Ideology, Religiosity, Personality, Demographics, and Men's Facial Hair in Explaining Hostile and Benevolent Sexism. *Personality and Individual Differences*, 122.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.*, 9.

Nabil Hossain, John Krumm, Michael Gamon, and Henry 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.

Jeremy Howard and Sebastian Ruder. 2018a. 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)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018b. 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.

Marcio Lima Inácio and Hugo Gonçalo Oliveira. 2023. Attempting to Recognize Humor via One-Class Classification. In *IberLEF@SEPLN*. CEUR-WS.org.

Adilzhan Ismailov. 2019. Humor Analysis Based on Human Annotation Challenge at IberLEF 2019: First-place Solution. In *IberLEF@SEPLN*.

James Jones. 1972. *Prejudice and Racism*. Addinson-Wesley.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics.

Jay Kaoshik and Sharmila Banu 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.

D. Kaplan. 2023. Knee Point. `https://www.mathworks.com/matlabcentral/fileexchange/35094-knee-points`. [Online; accessed June 7, 2023].

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20. Curran Associates Inc.

Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J. Inman. 2021. 1D Convolutional Neural Networks and Applications: A survey. *Mechanical Systems and Signal Processing*, 151.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese Neural Networks for One-shot Image Recognition. In *ICML deep learning workshop*, volume 2. Lille.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław Kazienko. 2021. Offensive, Aggressive, and Hate Speech Analysis: From Data-centric to Human-centered Approach. *Information Processing & Management*, 58.

Roberto Labadie, Daniel Castro, and Reynier Ortega. 2021a. Deep Modeling of Latent Representations for Twitter Profiles on Hate Speech Spreaders Identification Task—Notebook for PAN at CLEF 2021. In *CLEF 2021 Labs and Workshops, Notebook Papers*, volume 2936. CEUR-WS.org.

Roberto Labadie, Berta Chulvi, and Paolo Rosso. 2023. HUrtful HUmour (HUHU): Detection of Humour Spreading Prejudice in Twitter. *Procesamiento del Lenguaje Natural (SEPLN)*, 71.

Roberto Labadie, Mariano Jason Rodriguez, Reynier Ortega, and Paolo Rosso. 2021b. RoMa at SemEval-2021 Task 7: A Transformer-based Approach for Detecting and Rating Humor and Offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33.

Lukas Lange, Heike Adel, and Jannik Strötgen. 2021. Boosting Transformers for Job Expression Extraction and Classification in a Low-Resource Setting. In *IberLEF@SEPLN*, CEUR-WS.org.

Walter Lipmann. 1922. *Public Opinion*. New York:Harcourt Brace.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, et al. 2023. Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Peter Alan Low. 2011. Translating Jokes and Puns. *Perspectives*, 19.

Jihang Mao and Wanli Liu. 2019. A BERT-based Approach for Automatic Humor Detection and Scoring. In *IberLEF@SEPLN*.

J. A. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid 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.

Lucía Inés Merlo. 2022. *When Humour Hurts: A Computational Linguistic Approach*. Bachelor's thesis, Universitat Politècnica de València.

Lucía Inés Merlo, Berta Chulvi, Reynier Ortega-Bueno, and Paolo Rosso. 2023. When Humour Hurts: Linguistic Features to Foster Explainability. *Procesamiento del Lenguaje Natural*, 70.

Rada Mihalcea and Carlo Strapparava. 2005. Making Computers Laugh: Investigations in Automatic Humor Recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Stuart S. Miller, Conor J. O'Dea, Tiffany J. Lawless, and Donald A. Saucier. 2019. Savage or satire: Individual differences in perceptions of disparaging and subversive racial humor. *Personality and Individual Differences*, 142.

Tristan Miller. 2019. The Punster's Amanuensis: The Proper Place of Humans and Machines in the Translation of Wordplay. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*. Incoma Ltd., Shoumen, Bulgaria.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. *Proceedings of the 28th ACM Conference on Hypertext and Social Media*.

Terufumi Morishita, Gaku Morio, Hiroaki Ozaki, and Toshinori Miyoshi. 2020. 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*. International Committee for Computational Linguistics.

Kelly M. O'Donnell. 2021. Incel Mass Murderers: Masculinity, Narrative, and Identity. *Ohio Communication Journal*, 59.

Gefjon Off. 2023. Complexities and Nuances in Radical Right Voters' (Anti)Feminism. *Social Politics: International Studies in Gender, State & Society*.

Reynier Ortega-Bueno, Berta Chulvi, Francisco Rangel, Paolo Rosso, and Elisabetta Fersini. 2022. Profiling Irony and Stereotype Spreaders on Twitter (IROSTEREO). Overview for PAN at CLEF 2022. volume 3180. CEUR-WS.org.

Reynier Ortega-Bueno, Carlos Enrique Muñiz-Cuza, José Eladio Medina-Pagola, and Paolo Rosso. 2018. UO_UPV: Deep Linguistic Humor Detection in Spanish Social Media. In *IberEval@SEPLN*.

Reynier Ortega-Bueno, Paolo Rosso, and José Eladio Medina-Pagola. 2019. UO_UPV2 at HAHA 2019: BiGRU Neural Network Informed with Linguistic Features for Humor Recognition. In *IberLEF@SEPLN*.

José Ortiz-Bejar, Vladimir Salgado, Mario Graff, Daniela Moctezuma, Sabino Miranda-Jiménez, and Eric Sadit Tellez. 2018. INGEOTEC at IberEval 2018 Task HaHa: μTC and EvoMSA to Detect and Score Humor in Texts. In *IberEval@SEPLN*.

Daniel Palomino and José Ochoa-Luna. 2020. Palomino-Ochoa at SemEval-2020 Task 9: Robust System Based on Transformer for Code-Mixed Sentiment Classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics.

Chao Pang, Xiaoran Fan, Weiyue Su, Xuyi Chen, Shuohuan Wang, Jiaxiang Liu, Xuan Ouyang, Shikun Feng, and Yu Sun. 2021. abcbpc at SemEval-2021 Task 7: ERNIE-based Multi-task Model for Detecting and Rating Humor and Offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics.

John C. Paolillo. 2007. How Much Multilingualism?: Language Diversity on the Internet. In *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press.

Silvia Paun, Ron Artstein, and Massimo Poesio. 2022. *Statistical Methods for Annotation Analysis*. Springer Cham, Switzerland.

Minna Peng and Nankai Lin. 2023. Cross-task Interaction Mechanism for Humour Prejudice Detection. In *IberLEF@SEPLN*. CEUR-WS.org.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco 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. European Language Resources Association.

Diana-Elena Popa. 2005. Jokes and Translation. *Perspectives: Studies in Translatology*, 13.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2017. SemEval-2017 Task 6: Hashtag-Wars: Learning a Sense of Humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From Hmor Recognition to Irony Detection: The Figurative Language of Social Media. *Data & Knowledge Engineering*, 74. Applications of Natural Language to Information Systems.

Russell Richie, Sachin Grover, and Fuchiang (Rich) Tsui. 2022. Inter-annotator Agreement is not the Ceiling of Machine Learning Performance: Evidence from a Comprehensive Set of Simulations. In *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2015. Reasoning about Entailment with Neural Attention. *CoRR*, abs/1509.06664.

Mariano Rodriguez, Reynier Ortega-Bueno, and Paolo Rosso. 2021. RoMa at HAHA-2021: Deep Reinforcement Learning to Improve a Transformed-based Model for Humor Detection. In *IberLEF@SEPLN*, volume 2943 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mónica Romero-Sánchez, Hugo Carreterio-Dios, Jesus L. Megías, Miguel Moya, and Thomas E. Ford. 2017. Sexist Humor and Rape Proclivity: The Moderating Role of Joke Teller Gender and Severity of Sexual Assault. *Violence against women*, 23.

Alon Rozental, Dadi Biton, and Ido Blank. 2020. Amobee at SemEval-2020 Task 7: Regularization of Language Model Based Classifiers. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics.

David Borregón Sacristán, Antonio Pérez Mu noz, and Lucas Sebastián Peris. 2023. Building Robust Models for Detecting Offensive Content and Quantifying Prejudice in Online Platforms. In *IberLEF@SEPLN*. CEUR-WS.org.

Fumi Samejima. 1969. *Estimation of a Latent Ability Using a Response Pattern of Graded Scores*. Psychometric Society.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *CoRR*, abs/1910.01108.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Ignacio Sastre, Alexis Baladón, Mauricio Berois, Fernanda Cánepa, Agustín Lucasa, Santiago Castro, Santiago Góngora, and Luis Chiruzzo. 2023. RETUYT-InCo Submission at HUHU 2023: Detecting Humor and Prejudice through Supervised Methods. In *IberLEF@SEPLN*. CEUR-WS.org.

Brandon M Savage, Heidi L Lujan, Raghavendar R Thipparthi, and Stephen E DiCarlo. 2017. Humor, Laughter, Learning, and Health! A Brief Review. *Advances in physiology education*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR*, abs/2211.05100.

Emanuel Schmider, Matthias Ziegler, Erik Danay, Luzi Beyer, and Markus Buehner. 2010. Is It Really Robust? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6:147–151.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing*, 45.

Bingyan Song, Chunguang Pan, Shengguang Wang, and Zhipeng Luo. 2021a. DeepBlueAI at SemEval-2021 Task 7: Detecting and Rating Humor and Offense with Stacking Diverse Language Model-Based Methods. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*. Association for Computational Linguistics.

Kwangok Song, Kyle M. Williams, Diane L. Schallert, and Alina Adonyi Pruitt. 2021b. Humor in Multimodal Language Use: Students' Response to a Dialogic, Social-networking Online Assignment. *Linguistics and Education*, 63.

Guillem García Subies, David Betancur Sánchez, and Alejandro Vaca. 2021. BERT and SHAP for Humor Analysis based on Human Annotation. In *IberLEF@SEPLN*, volume 2943 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal Meme Dataset (MultiOFF) for Identifying Offensive Content in Image and Text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*. European Language Resources Association (ELRA).

Janet K. Swim, Kathryn J. Aikin, Wayne S. Hall, and Barbara A. Hunter. 1995. Sexism and Racism: Old-fashioned and Modern Prejudices. *Journal of Personality and Social Psychology*, 68.

Janet K. Swim and Lauri L. Hyers. 2009. Sexism. In *Handbook of Prejudice, Stereotyping, and Discrimination*. Psychology Press.

David Tomás, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. 2022. Transformer-based Models for Multimodal Irony Detection. *Journal of Ambient Intelligence and Humanized Computing*.

Joseph Tomasulo, Jin Wang, and Xuejie Zhang. 2020. YNU-HPCC at SemEval-2020 Task 7: Using an Ensemble BiGRU Model to Evaluate the Humor of Edited News Titles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics.

Francine Tougas, Rupert Brown, Ann M. Beaton, and Stéphane Joly. 1995. Neosexism Scale. *Personality and Social Psychology Bulletin*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Villy Tsakona. 2017. Genres of Humor. In *The Routledge handbook of language and humor*. Routledge.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72.

Camilla Vásquez and Erhan Aslan. 2021. "Cats Be Outside, How About Meow": Multimodal Humor and Creativity in an Internet Meme. *Journal of Pragmatics*, 171.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17. Curran Associates Inc.

Paola Villano, Lara Fontanella, and Marika Di Donato. 2017. Stereotyping Roma People in Italy: IRT Models for Ambivalent Prejudice Measurement. *International Journal of Intercultural Relations*, 57.

Lianxi Wang, Xiaotian Lin, Nankai Lin, Yingwen Fu, Kaiying Wu, and Jiajun Wu. 2021. Humor Analysis in Spanish Tweets with Multiple Strategies. In *IberLEF@SEPLN*, volume 2943 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Minghan Wang, Hao Yang, Ying Qin, Shiliang Sun, and Yao Deng. 2020. Unified Humor Detection Based on Sentence-pair Augmentation and Transfer Learning. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation.

Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. 2019. Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. In *The World Wide Web Conference*.

Zeerak Waseem. 2016. Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*. Association for Computational Linguistics.

Ka Wong, Praveen Paritosh, and Lora 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.

Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu. 2021. MUMOR: A Multimodal Dataset for Humor Detection in Conversations. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I*. Springer-Verlag.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019a. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zixiaofan Yang, Lin Ai, and Julia Hirschberg. 2019b. Multimodal Indicators of Humor in Videos. In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*.

Patrick Zabalbeascoa. 2005. Humor and Translation - an Interdiscipline. *Humor-International Journal of Humor Research*, 18.

Appendix A

HUHU Shared Task

A.1 HUHU Dataset Construction

The following Table shows the keywords determined from the initial macro-class to filter instances obtained from Twitter.

rollo homosexual	moro	inmoral	feminista
homosexuales abren	negro hijo puta	volver homosexuales	malditos
putos negros mierda	los lgbtqwerty	feministaradical	entre gay
homosexuales desviados	homofobo	violadas	homosexuales culos
putos maricones	extremistas	xenofobia	feminicidio
maricon mierda	negros mierda est	gay tratara	gays
ser homfobo	homosexuales cada	trans	trolos homosexuales
feas	ilegales	negros hijos puta	violaciones
ser homosexual	asquerosos gays	machistas	odio negros
negros	patriarcal	odio	homfobo
misoginas	feministas solo tener	bisexuales	maricones
homosexual pesar	ser gay ser	saludo homofobo	feminismo
tetas	hijo puta	heterosexualeslivesmatter	inmigrantes
negros hijos	puto maricon mierda	mujeres solo	homosexuales dan
homosexual ser	mujeres	inmigrantes mierda	lgbti
queers	lesbico	maricon	feminazi
ser gay	moro mierda	homosexuales	lgbtqwerty
gord	gay reprimido	puta	putos negros odio
gorda	sea lesbina	aborto	moros
putas	gordo	negros negros putos	homofobo para
negros mierda	homosexual	putos homosexuales	homosexuales en
gorda	homosexual que	homosexual haciendo	humornegro
putos	homosexuales los	menas	feminazis
homosexual sera	feministas solo	gays asquerosos	mujeres solo sirven
feministas	foca	chsitesracistas	homosexual para

el homosexual	mujer hace	homosexualidad gente	chistes sexistas
ser puto	acerquen gays	negro	lesbianismo
mierda hijos puta	malparidos	ser mujer	putos maricones mierda
mierda homosexual	maricones mierda	putos negros	negros negros
travesti	negros odio negros	homosexualidad	para homosexuales
mierda putos	mujer cocina	homoalergica	puto
putos moros	chiste	lgbt	homosexuales brutal
porque homosexual	putos inmigrantes	maltratadora	gay ser
machista	humor	culo	gay
lesbianas	homosexual todos	homofbica	mierda puto mierdas
homosexuslismo	ser feminazi	feminaci	humorracista
odia	inmigrantes ilegales	desviados queers	gordas
todas feminazis	putos chinos	los homosexuales	estos gays
travestismo	dos gays	lesbiana	lgbt antilgbt
gitanos	chinos	queer	asco

Table A.1: Keywords employed for tweets filtering according to macro-classes.

A.2 Guidelines for Second Annotation Step

You will be annotating both humorous and non-humorous tweets that contain expressions of prejudice directed towards specific groups. These groups are: 1) women or feminists, 2) individuals with a non-heteronormative sexual orientation (LGTBi+), 3) immigrants or individuals who are racially categorized based on their ethnicity or religion, and 4) people who are overweight.

Prejudice is a universal phenomenon that assumes different labels based on the group it targets. The various terminologies used to describe prejudice when applied to these groups are presented in Table A.2.

Target Group	Type of Prejudice
Women and Feminists	Sexism or Misogyny
LGBTI+ Community	Homophobia, Transphobia or LGTBphobia
Immigrants and Racially Discriminated People	Xenophobia, Racism, or Religious Intolerance
Over-Weighted People	Gordofobia or Obesophobia

Table A.2: Target group and type of prejudice for annotation instructions.

All of these forms of prejudice involve two main elements: firstly, the generalization of an entire group, and secondly, the conveyance of negative content. This negative content is typically communicated in two ways: contempt (underestimating or mocking the group) or hostility (attacking the group).

Both contempt and attack can manifest in varying degrees. Your task involves evaluating (based on your own perspective) the extent of prejudice present in each message. It's important to note that this task is subjective, as everyone holds their own opinions. Thus, we request you to indicate the degree to which each message portrays an image that belittles

or targets the specified group on a scale of 1 to 5. In this scale, 1 signifies "a little," and 5 signifies "a lot."

A.3 HAHA 2021 Dataset

Prejudiced Target	HAHA category
women and feminists	body shaming-women, women, professions-women, age-women, men-women, technology-women, religion-women, family/relationships-women, ethnicity/origin-women
LGBTI+ community	family/relationships, lgbt, lgbt-professions, body shaming-lgbt, ethnicity/origin-lgbt, lgbt-religion
immigrants and racially discriminated people	ethnicity/origin, ethnicity/origin-professions, ethnicity/origin-women, ethnicity/origin-substance use, ethnicity/origin-body-shaming, age-ethnicity/origin, ethnicity/origin-health
over-weighted people	body shaming, body shaming-self-deprecating, age-body shaming, body shaming-men, body shaming-professions, body shaming-lgbt

Table A.3: Proposed matching between the categories from HAHA 2021 dataset and the prejudiced minorities studied in Chapter 4

A.4 HUHU Results

Table A.4, Table A.5 and Table A.6 show the top-ranked systems along with the results of the proposed baselines for each subtask of the HUHU shared task. The full Rankings can be found on the shared task web page at: https://sites.google.com/view/huhuatiberlef23/results.

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
BETO		*0.759*
SVM-3gram-char		*0.679*
allTrue		*0.492*

Table A.4: Top-ranked systems for subtask 1: HUrtful HUmour Detection.

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 A.5: Top-ranked systems for subtask 2a: Prejudice Target Detection.

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
FENRIRFENIX	1	0.895
LaVellaPremium	2	0.898
Climent	1	0.899
SVM-3gram-char		*0.907*
BLOOM-1b1		*0.915*

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

A.5 Results Analysis

In Table A.7 are shown examples of each category of difficulty according to the submitted systems to the HUHU competition.

Difficulty Category	Humor	Non Humor
Very easy	No existe	No voy a volver a usar el término "feminazi". Creo que el término "feminista" ya es bastante insultante.
Easy	¿En qué se parecen las mujeres a las nubes?. En que cuando se van se queda un día estupendo. #Chistes #Humor #Ane...	La minoría feminazi quieren que todas las mujeres sean zombies no pensantes de ultra izquierda por eso es que odian e insultan a las mujeres de verdad
Difficult	¿Saben por qué todos los negros que vemos son rápidos? Porque los lentos están en la cárcel.	todas putas menos mi madre y mi abuela
Very difficult	– Papá. – Dime, hijo. – Hija. Soy de género fluido, hoy me siento mujer. – Dime, hija. – Ah no, me siento hombre, pero negro y bisexual. – Dime, hijo. – Espera, no, me siento una mujer trans vietnamita. – Hija, ¿vas a dejar de hacer la? – ¡Tránsfobo de, te odio!	La homofobia ya no es un problema. Ahora los homosexuales dirigen TV, hay grandes corporaciones dándoles alabanza, gozan de privilegios por ser "oprimidos". Salir del clóset no es un acto heroico sino de comodidad. Quien lo hace disfruta del gay privilege from day one.

Table A.7: Top-ranked systems for subtask 2b.

Appendix B

Annotation Perspectivism

B.1 Annotators Attitudes and Ideology

The items from the questionnaire utilized to evaluate the attitudes and ideology of each annotator engaged in the study described in Chapter 5 are listed below:

1. Some of the demands of the feminist movement seem to me to be a bit exaggerated.
2. I sometimes feel that our society pays too much attention to the rights of certain minorities.
3. In the name of equality, many women try to gain certain privileges.
4. Many women interpret innocent comments and actions as sexist.
5. Women are easily offended.
6. Women exaggerate the problems they suffer because they are women.

Appendix C

Scientific Contributions

The following contributions have been made as partial or complete results of the conducted research within the framework of this thesis:

- Labadie-Tamayo, R., Ortega-Bueno, R., Rosso, P. , Rodriguez-Cisnero, M., On the Poor Robustness of Transformer Models in Cross-Language Humor Recognition. Procesamiento del Lenguaje Natural (SEPLN), num. 70
- Main organizer of the *"HUrtful HUmour (HUHU): Detection of humor spreading prejudice in Twitter"* shared task held at IberLEF 2023.
- The dataset constructed for HUHU has been made publicly available to the NLP research community at https://zenodo.org/record/7967255.
- Labadie-Tamayo, R., Chulvi, B., Rosso, P. , Everybody Hurts, Sometimes. Overview of HUrtful HUmour at IberLEF 2023: Detection of Humour Spreading Prejudice in Twitter. Procesamiento del Lenguaje Natural (SEPLN), num. 71
- Chulvi, B., Fontanella, L., Labadie-Tamayo, R. Rosso, P., Social or Individual Disagreement? Perspectivism in the Annotation of Sexist Jokes. In PERSPECTIVES 2023. Proceedings of the Perspectives on the Evaluation of Recommender Systems Workshop 2032. CEUR-WS.org.