Spanish FatPhoCorpus 2023: Combating Fatphobia in Social Media in Spanish using Transformers

José Antonio García-Díaz[1], Ronghao Pan[1], Salud María Jiménez-Zafra[2] and Rafael Valencia-García[1]

[1]Departamento de Informática y Sistemas, Facultad de Informática, Universidad de Murcia
[2]Computer Science Department, SINAI, CEATIC, Universidad de Jaén

Abstract
Social media can aggravate body dissatisfaction and weight discrimination. Fat-shaming content is widespread on social media and targets people of different weights. Fatphobia and its harmful effects affect not only the general public but also children, causing lasting psychological damage. In this work, we make two contributions to the detection of fatphobia in social networks. On the one hand, we compile the Spanish FatPhoCorpus 2023, a multiclass dataset that allows the identification of hate speech, offensive content, and hopeful messages related to body image. On the other hand, we evaluate this novel dataset with a baseline of several pre-trained models based on Transformers. The best result is obtained with multilingual TwHIN, which achieves a macro average F1-score of 59.158%.

Keywords
Fatphobia detection, Hate speech detection, Hope speech detection, Natural Language Processing

1. Introduction

Body image dissatisfaction and fatphobia are two concerns that are increasingly prevalent in today's generation. Body image is defined as a set of self-evaluations regarding one's physical appearance [1] and fatphobia, in particular, is a general reactionary attitude involving a form of dislike, fear, and intolerance toward fatness [2].

Social media exposure has led to higher levels of body image concerns. Indeed, previous research has shown evidence of this. [3] attribute the prevalence of body dissatisfaction to the rise of "bedroom culture" and social media. Millennials are increasingly invested in the virtual world, creating an idealized online person [4].

Weight discrimination is one of the most common forms of discrimination. It is reported in percentages comparable to those of racism [5]. Social media are often used to spread fat-shaming content that condemns a person for their body shape and size. Celebrities are often targets of weight bias. For example, megastar Rihanna was recently been the victim of fat-shaming attacks [6]. In addition, among children, weight bullying is one of the most common types of bullying [7]. Often, this type of bullying is a stressor that children will face for years, from infancy to adulthood.

It has been shown that fat-shaming often leads to decreased self-esteem and can cause psychological damage [6], so this type of content should be pursued. This work aims to combat the body image hate and rejection in social networks by detecting it in texts using Natural Language Processing (NLP). For this purpose, we present the Spanish FatPhoCorpus 2023, a corpus related to fatphobia, in order to promote the development of systems to detect this phenomenon and to fight against the hatred and rejection that people suffer. It is a multi-class corpus in which 4 classes have been considered: hate, offensive, hope, and none. It will help to determine whether a text contains hate speech, is offensive or hopeful in relation to body image, or whether it does not reflect any of the above.

The rest of the paper is organized as follows. First, in Section 2 presents an overview of the state of the art on the three main topics of the corpus, offensiveness, hate and hope. Next, Section 3 describes the methodology used to combat the phenomenon of fatphobia using NLP. It consists of the compilation and annotation of the Spanish FatPhoCorpus 2023, and the proposal of an automatic detection system based on transformers. Then, the results obtained in the experiment are presented and discussed in the Section 4. Finally, we end with the conclusions in Section 5.

2. State of the art

This section describes the state of the art concerning fatphobia (see Section 2.1) and the different labels considered in our study, namely offensiveness (see Section

SEPLN-2024: 40th Conference of the Spanish Society for Natural Language Processing. Valladolid, Spain. 24-27 September 2024.
joseantonio.garcia8@um.es (J. A. García-Díaz); ronghao.pan@um.es (R. Pan); sjzafra@ujaen.es (S. M. Jiménez-Zafra); valencia@um.es (R. Valencia-García)
0000-0002-3651-2660 (J. A. García-Díaz); 0009-0008-7317-7145 (R. Pan); 0000-0003-3274-8825 (S. M. Jiménez-Zafra); 0000-0003-2457-1791 (R. Valencia-García)

2.2), hate speech (see Section 2.3), and hope speech (see Section 2.4).

2.1. Fatphobia

As far as we know, there is not much work on fatphobia from an NLP perspective. One of the papers identified was conducted by [8], in which they study the effect of shutting down two Reddit channels that incite hate: (i) `r/fatpeoplehate`, a subreddit dedicated to posting pictures of overweight people to ridicule them and (ii) `r/CoonTown`, a racist subreddit with violent hate speech against African Americans. To do this, they compiled a dataset of all posting activity on Reddit in 2015 and used the texts from the two aforementioned subreddits to build a lexicon of hate words, which they used to examine the user-level effects of the ban.

Other existing work deals with fatphobia as a category of hate speech or prejudice. [9] provided HateBR, a corpus of Brazilian Portuguese Instagram comments for the detection of offensive language and hate speech. The corpus consists of 7000 annotated documents. Later, the authors extended the work by generating a specialized lexicon manually extracted by a linguist from the HateBR corpus [10]. This Brazilian Portuguese lexicon was annotated with contextual information, and translated and adapted by native speakers into Turkish, German, French, English and Spanish. This context-aware lexicon is called "MOL-Multilingual Offensive Lexicon". In addition, the authors conducted experiments using the generated corpus and lexicon for the identification of hate speech in Brazilian Portuguese. [11] organized the shared task HUHU in the IberLEF 2023 workshop [12], on the Detection of Humor Spreading Prejudice in Twitter. They proposed a framework to study how humor is used to discriminate against minorities and analyze its interaction with the level of prejudice expressed against specific groups. For this, it is provided a corpus of prejudiced tweets in Spanish annotated with the presence of humor, their degree of prejudice and the targeted groups, being overweight people (fatphobia) one of them.

It is worth noting that previous work has treated fatphobia as a category within hate speech. Our work differs from previous work in that we treat fatphobia as a major theme to identify hateful, offensive or hopeful messages. In addition, we focus on the Spanish language.

2.2. Offensiveness

Offensiveness is the fact of being rude in a way that makes someone to feel upset or angry because it shows a lack of respect. A text is considered offensive if it contains any form of unacceptable language, that is, whether it contains insults, threats, or bad language [13]. Offensive language is usually divided into three classes: (i) "profanity", i.e. the of use of swear words (e.g. fuck), but without the intention to insult someone; (ii) "insult", when there is a clear intention to offend someone with disrespect and contempt; and (iii) "abuse", when a person is insulted by using a special type of degradation that is considered representative of a group by negatively attributing to him/her a quality related to a universal, pervasive or immutable characteristic [14].

Offensive language detection has become one of the most popular research areas in the field of NLP, as it is increasingly observed in social media. It is often formulated as a binary (offensive/non-offensive) or multi-categorization task, considering the types of offensiveness and the targets of the offensive speech [15].

Several shared tasks have been organized to promote the detection of this type of discourse, such as GermEval 2018 [16] and GermEval 2019 [17] on German tweets, OffensEval 2019 [14] on English texts, OffensEval 2020 [18] on texts written in Arabic, Danish, English, Greek and Turkish, and MeOffendES 2021 [15] on texts written in Spanish variants.

2.3. Hate speech

Hate speech is the language that targets a person or group with the intention of causing harm or social disruption [19]. This targeting is usually done on the basis of some characteristic such as race, gender, sexual orientation, nationality or religion [20]. It is a class of offensive language, specifically "abuse", and is sometimes referred to as abusive language [21].

The negative impact of the spread of hate content has led to an increasing number of researchers focusing on this issue. Most studies focus on the detection of hate content in general [13], the identification of racism [22], the detection of misogyny [23], and the identification of xenophobia [24]. In fact, there are many shared tasks on the identification of hate speech, such as HASOC 2019 [25] and HASOC 2020 [26], and shared tasks on specific topics such as sexism, toxicity or racism. For example, the AMI shared task on the automatic identification of misogyny at IberEval 2018 [27] and Evalita 2018 [28], the HatEval shared task on the detection of hate speech against immigrants and women [29], and some shared task organized in the framework of the IberLEF workshop [30, 31, 12] on (i) the identification of sexism in social networks, as in EXIST 2021 [32] and 2022 [33], (ii) the detection of toxicity in DETOXIS 2021 [34]), (iii) the detection and classification of racial stereotypes in DETESTS 2022 [35], and (iv) the identification of hate speech towards the LGBTQ+ population in HOMO-MEX 2023 [36].

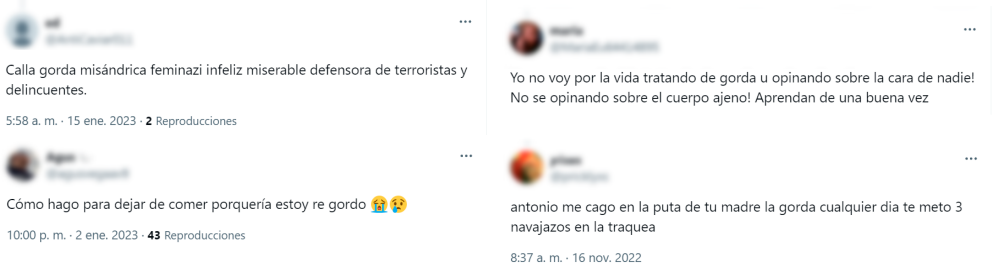

Figure 1: Four examples of the dataset for tweets labeled Hate (top left), Hope (top right), None (bottom left), and Offensive (bottom right).

2.4. Hope speech

Hope speech is the type of speech that is able to relax a hostile environment [37] and that helps, gives suggestions and inspires people for good when they are in times of illness, stress, loneliness or depression [38].

In the current digital age, social media provides a place for people to freely express their views and opinions. They have become critical for people from minority groups seeking help and support online [39, 12]. People have turned to these media to satisfy their informational, emotional, and social needs as they seek to connect with others, experience a sense of social inclusion, and cultivate a sense of belonging through active participation in online communities. The presence of these factors has a profound impact on physical and psychological well-being as well as mental health [40, 41]. This has led to recent research exploring positive content, such as messages of hope, and the promotion of constructive activities in pursuit of equality, diversity, and inclusion.

There has been a remarkable increase in attention to this topic since 2021 and several workshops have been organized to address the challenge of detecting hope speech in English, Tamil, Malayalam, Bulgarian, Hindi and Spanish, such as LT-EDI-EACL2021 [42], LT-EDI-ACL2022 [43], LT-EDI-RANLP2023 [44] and the shared task HOPE in IberLEF 2023 [12].

3. Methodology

In the following section, the compilation and annotation process of the Spanish FatPhoCorpus 2023 is described in Section 3.1 and the pipeline for the evaluation of the dataset is presented in Section 3.2.

3.1. Spanish FatPhoCorpus 2023 compilation and annotation

The UMUCorpusClassifier tool [45] was used to compile the dataset. We started by querying for specific keywords on X (formerly Twitter). The keywords are gordofobia (fatphobia), sobrepeso (overweight), gordo/a (fat), obeso (obese), obesidad (obesity), foca (seal), glotón (glutton), glotonería (gluttony), "talla grande" (plus size), anorexia, flaco (skinny), "desorden alimenticio" ('eating disorder), delgado (thin), delgadez (thinnes), and calorías (calories). The dataset is compiled from all countries where Spanish is spoken. This can be challenging due to differences in cultural factors between Spanish-speaking countries, resulting in different interpretations of adjectives such as "gordo" (fat) or "flaco" (thin). In the first iteration, 39,787 tweets were collected from June 2020 to July 2023, focusing on the specified keywords.

The annotation phase was performed by the research team and an external collaborator. We considered the following labels: (i) *Hate*, if the text contains hate speech towards people because of their body, (ii) *Hope*, if the text tends to help, relax, and inspire people for their body weight perception, (iii) *Offensive*, if the text contains insults and profanity but no fatphobia, and (iv) *None*, if none of the rest.

Each tweet was annotated by four members, being two men and two women, members of our research group. The annotators do not have specific knowledge of the area, but rather linguistics and computer science. The inter-annotator agreement based on Krippendorff's alpha is 0.712. The final label is assigned based on the mode of the annotations. Ties were resolved in internal meetings.

The analysis of the dataset revealed that some of the offensive tweets contain homophobia, racism or transphobia in addition to fatphobia. Furthermore, some tweets contain irony and figurative language, while others are self-deprecating. Fourth, we observed a lot of internalized fatphobia, as several users wrote that they are very afraid of being fat because they associate being fat with being ugly. However, the main finding is that several tweets used "thin" or "fat" in a familiar way to refer to partners and pets in an affective way and these tweets were labeled as None.

To select the number of tweets that are not related

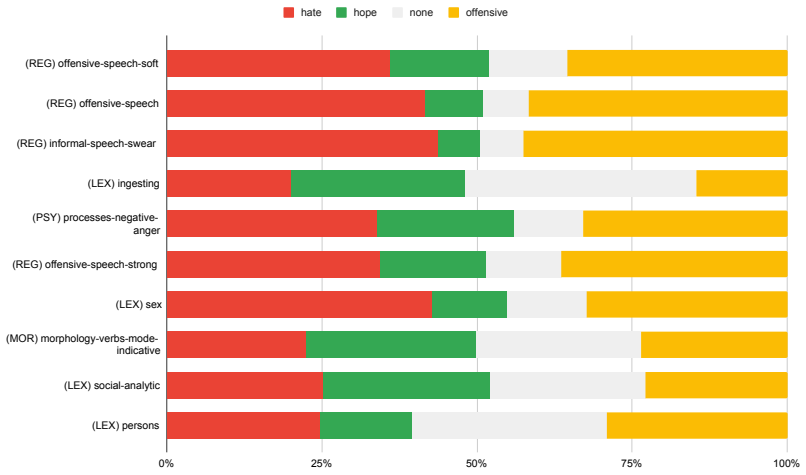

Figure 2: Top-10 Information Gain of the dataset per label and linguistic feature. The values are normalized so that each row represents a value of 100%.

to offensive, hate, or hope, we use agglomerative clustering with the tweets labeled None. The idea is to get a large but representative number of tweets of different types. To do this, we extract their contextual sentence embeddings, obtained by hackathon-pln-es/paraphrase-spanish-distilroberta, and create five clusters. We then randomly selected one tweet for each cluster until the cluster with the fewest instances was empty. This process reduced the number of tweets marked as None to 4,570.

In the final version of the dataset, a total of 6,145 tweets were collected, annotated, and verified. The remaining tweets were removed because they were too short or it was not clear how to annotate them. These tweets are divided into training, validation, and test sets using a 70-15-15 ratio, with stratification to maintain the balance of labels between the splits. Table 1 shows the statistics for each label and split. Note that hope language is the label with the lowest representation, while the remaining labels have comparable proportions.

Table 1
Dataset distribution for label and split.

Label	Train	Val	Test	Total
Hate	613	131	132	876
Hope	58	13	13	84
None	3,221	672	677	4,570
Offensive	430	92	93	615
Total	4,322	908	915	6,145

Figure 1 shows four examples from the dataset, one for each label. The top left example is for a text labeled as *hate-speech*, which contains fatphobia but also misogyny. The top right example promotes hope. It is a statement from a person who refuses to give its opinion about other people's bodies. The bottom left example is labeled as *None*, and it is about a person who regrets eating junk food and finds it hard to stop. The bottom right example is labeled as *Offensive*. It is about a person who curses and insults another man and threatens him with physical violence.

To analyze the corpus, we used the UMUTextStats tool [46] to obtain linguistic features to calculate the information gain for each label. As Figure 2 shows, soft offensive language and swear words, psychological processes related to anger, and lexis related to sex show a correlation with hate speech, and to a lesser extent, with offensive language. The only exception is the food-related terminology, which is less common in documents labeled as *Hate* but is more common in hope speech and fatphobia unrelated tweets. In addition, we observe that analytical thinking is used almost equally in all labels.

Finally, the dataset is now available to the scientific community[1]. However, according to the X's guidelines[2], we will only publish the tweet IDs to protect the users' right to be forgotten.

[1] https://pln.inf.um.es/corpora/hate-speech/hate-fatphobia-2024.zip

[2] https://developer.twitter.com/en/developer-terms/policy

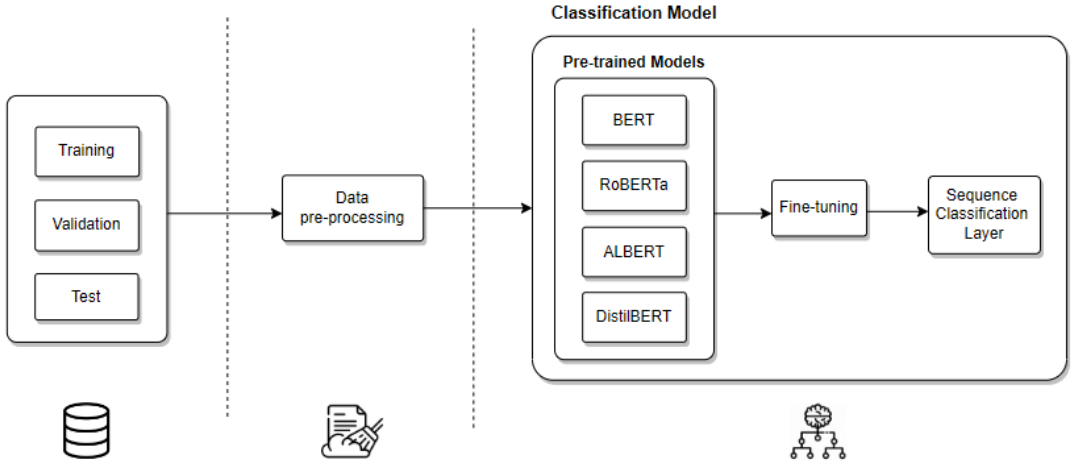

Figure 3: System architecture for the evaluation of the Spanish FatPhoCorpus 2023.

3.2. System architecture

To evaluate the dataset we rely on different pre-trained models for this classification task. For this purpose, the system shown in Figure 3 is developed. In short, the pipeline can be described as follows. First, the data pre-processing stage is performed on the dataset to produce a proper and clean format, since the quality of the input data directly affects the performance and generalizability of the model. In this case, we perform a data cleaning stage where acronyms are expanded, elongations, and digits are removed along with hyperlinks, hashtags, quotation marks, and other punctuation symbols. Finally, we evaluated different pre-trained models based on Transformers and with different architectures, such as BERT, RoBERTa, ALBERT and DistilBERT to classify different categories using the fine-tuning approach. In this case, since all the evaluated models are of the encoder type, we have added a sequence classification layer to perform the fine-tuning process.

The fine-tuning process is performed using hyperparameter optimization stage to obtain the learning rate (uniformly sampled between 1e-5 and 5e-5), the number of epochs (between 1 and 5), the batch size ([8, 16]), the warm-up steps ([0, 250, 500, 1000]), and the weight decay (uniformly sampled between 0.0 and 0.3). A total of 10 runs are evaluated for each LLM. Table 2 shows the best hyperparameters for each LLM. The lower number of training epochs is achieved by BETO and TwHIN. Both without no warm-up steps. Other models, including lightweight models such as ALBETO and DistilBETO, require a larger number of epochs (5 and 4 respectively). Another finding is that the number of warm-up steps is usually small, with no steps for ALBETO, BERTIN, DistilBETO and TwHIN.

Table 2
Hyperparameter tuning of the seven evaluated LLMs. The parameters evaluated are the learning rate (lr), number of training epochs (e), the batch size (bs), the warm-up steps (ws), and the weight decay (wd).

LLM	lr	e	bs	ws	wd
ALBETO	2.2e-05	5	8	0	0.21
BERTIN	1.1e-05	5	16	0	0.11
BETO	4e-05	3	8	500	0.23
DistilBETO	3e-05	4	16	0	0.26
MarIA	4.4e-05	4	8	250	0.18
TwHIN	2.6e-05	3	8	0	0.17
RoBERTuito	2.1e-05	5	16	500	0.16

4. Results and discussion

In the experiments, a simple process is performed by fine-tuning several pre-trained models based on Transformers. The models evaluated are of the encoder type, i.e. means that they use only the encoder component of a Transformers model. These models are often characterized by having “bidirectional” attention, which is why they are also called “auto-encoding” models. The pre-training of these models typically involves corrupting a given sentence in some way, for example, by masking random words in it (Masked Language Modeling), and then instructing the model to find or reconstruct the orig-

inal sentences. Therefore, depending on the language of the pre-training corpus, the models can be monolingual or multilingual.

The monolingual models evaluated are: (i) BETO [47], a Spanish BERT model trained on the Spanish Unannotated Corpora; (ii) MarIA [48], which is a pre-trained model based on RoBERTa and trained exclusively on Spanish texts collected from web crawling of the National Library of Spain; (iii) BERTIN [49], another model based on RoBERTa and trained on the Spanish part of the mC4 dataset; (iv) ALBETO [50], a pre-trained model based on ALBERT (a lightweight version of BERT), pretrained only on Spanish documents; (v) DistilBETO [50], is a model trained using distillation techniques to transfer the weights of BETO to a new model with fewer layers and reduced complexity; (vi) RoBERTuito [51], a pre-trained model based on RoBERTa for the analysis of social media text in Spanish, and trained according to RoBERTa guidelines on 500 million tweets. In terms of multilingual models, TWHIN [52] was evaluated, which is a multilingual model trained on X covering over 100 languages.

In this case, since these are encoder-type models, we have added a sequence classification layer to perform fine tuning for a classification task. This layer consists of dense structure with as many neurons as there are output classes to classify. To evaluate the classifiers, we use the macro F1-score as the main benchmark metric, which weighs the precision and recall of each class and combines the results without considering the class imbalance. In this way, we select the best model that performs equally well for each label. Another reference metric is the weighted F1-score (W-F1), which is an evaluation measure used in classification problems, especially when dealing with unbalanced datasets, where some classes may have many more examples than others. Unlike the macro F1-score, this metric takes into account the class imbalance by assigning different weights to each class based on their frequency in the dataset.

Table 3 shows the results of all the pre-trained models. In general, all the evaluated models performed equally well in terms of macro and weighted F1-score. The results using the weighted precision, recall and F1-score are higher due to the class imbalance and the reliability of all models with the None label. Among the evaluated models, the multilingual TwHIN is the best performing model, with a weighted F1-score of 83.234% and a macro F1-score of 59.158%. The monolingual models, BETO and MarIA, based on the BERT and RoBERTa architectures respectively, also performed similarly, with a macro F1 score of 0.603% improvement of MarIA over BETO. Furthermore, it is also worth noting that ALBETO (the lightweight version of BETO) outperformed distilled BETO (macro F1-score of 52.159% vs 48.071). In this sense, ALBETO gave similar results to BETO, with a significantly reduced training time.

Next, we report the classification report of the best model, TwHIN. The Table 4 shows the precision, recall, and F1-score with the test set. It can be seen that TwHIN performs quite well in predicting hate speech with an F1-score of 64.945%, but the model is less reliable in predicting hope and offensive content from a fatphobic perspective. Furthermore, all labels behave similarly in terms of precision and recall.

We used TwHIN to perform the error analysis because it is the fine-tuned model that produced the highest macro F1-score over the test split. Figure 4 shows the confusion matrix for this model, which allows us to identify cases where the model makes incorrect predictions. We can see that the model misclassified 27 instances of hate speech as offensive speech. Analyzing these cases, it was found that when the level of hate is very high, our model tends to identify it as offensive. It was also observed that the analyzed model has difficulties to handle hope speech as it confused 5 instances with hate and 4 with none. The case of offensive speech is also confused with hate (29 cases) and none (21 cases).

It is worth noting that in an earlier version of the annotation processes, we identify a large number of posts coming from different Spanish-speaking countries, including Spain, Argentina, Mexico, and so on. In this sense, there are adjectives with ambiguous meanings, such as “flaco”, which is an informal way for Argentines to refer to a man or a woman in an affectionate way, even if they are not thin. Other examples are the word “gorda” in text 3 and the word “flaco”. In the first version of the dataset, these tweets were classified as *Hope* (leading to false positives), when in fact, they should have been classified as *None*. In addition to the cultural background of different Spanish-speaking countries, we identified another challenge related to the fact that people use informal language on social networks, in addition to not using punctuation or accents correctly. This is particularly serious in the case of Spanish, as Spanish does not change the order of words when moving from declarative to interrogative sentences. Often, this meaning can be taken out of context with considerable human effort, but it makes it very difficult for NLP tools to understand the language.

Table 5 shows some examples of misclassifications made by TwHIN that summarize the challenge of identifying hate, hope, and aggressive fatphobic comments. The first example contains a message suggesting that normalizing the message of being okay with one’s body is counterproductive. This example is difficult to identify because it uses polite language. The second example is considered offensive by the model because it contains many derogatory terms, but it also contains terms related to fatphobia as well as homophobia. The third example, annotated as hope speech, is incorrectly identified by

Table 3
Benchmarking of the various pre-trained models. The reported metrics for each model include weighted precision (W-P), weighted recall (W-R), weighted F1-score (W-F1), and macro F1-score (M-F1).

Model	W-P	W-R	W-F1	M-F1
ALBETO	80.526	80.219	80.266	52.159
BERTIN	81.943	81.093	81.087	52.188
BETO	80.815	80.219	80.433	52.169
DistilBETO	79.229	79.891	79.550	48.071
MarIA	82.637	78.907	80.141	52.772
RoBERTuito	82.495	82.951	82.699	57.291
TwHIN	**83.454**	**83.060**	**83.234**	**59.158**

Table 4
Classification report of the precision (P), recall (R) and F1-score (F1) of the uncased version of TwHIN.

	P	R	F1
Hate	63.309	66.667	64.945
Hope	40.000	30.769	34.783
None	93.853	92.467	93.155
Offensive	42.424	45.161	43.750
Macro avg	59.897	58.766	59.158
Weighted avg	83.454	83.060	83.234

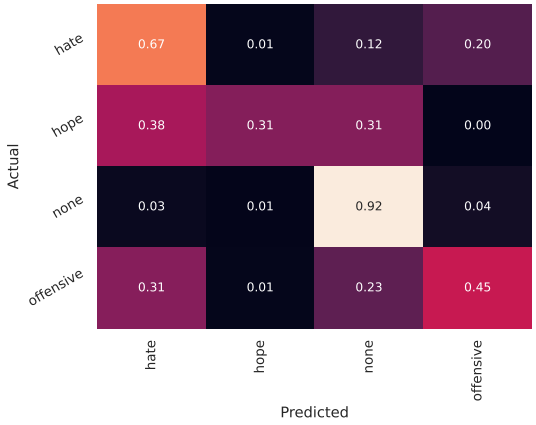

Figure 4: Confusion matrix of the TwHIN model

TwHIN as hate speech. This example contains difficult words such as because it contains a slang from Chile and Peru (*weon*) that refers to stupid and lazy people and because it contains negation cues. The fourth example was identified as None. This is correct because it uses the word *flaco* as colloquial way. However, it uses the adjective *Mongolian*, which is used as an insult in Spain. However, this word contains a typo so it is possible that TwHIN does it not consider an offensive word. The last example is hate speech, which was misclassified classified as hope speech. The author is complaining about the so-called crystal generation regarding issues of racism, fatness and homophobia. In this case, the problem is that the author it does only states a premise but the conclusion is not clear.

Finally, to get some insight into the errors made by TwHIN, we get the subset of the test field that was misclassified and obtain the linguistic features (see Figure 5). We observed that the errors contain morphological features, such as interjections, the use of verbs in the imperative and in the infinitive; besides, we also observed features related to negativity and that the number of words and syllables is also relevant.

5. Conclusions

In this paper we have described the compilation and annotation process of the Spanish FatPho Corpus 2023, a dataset for the detection of fatphobic comments in social networks in Spanish. This work has considered four types of comments, according to whether they contain hate speech, hope speech, offensiveness, or none of the above. The resulting dataset has been evaluated with different pre-trained models based on Transformers, obtaining the best result TwHIN, a multilingual general purpose model trained with tweets, with a macro F1 score of 59.158%.

As future work, we will expand the dataset. One limitation we found during the evaluation was the difficulty in properly identifying hope and offensive fatphobic tweets. This is particularly relevant in the case of hope speech, where the dataset contains only 84 texts. Furthermore, we found that background and cultural differences are very important for detecting fatphobia in social networks. Therefore, we propose the annotation of the Spanish-speaking country of the authors and evaluate the cultural differences when it comes to finding similarities and differences between different Spanish-speaking countries. In this sense, we will extend the annotation of the dataset

Table 5
Error Analysis with some examples of misclassifications performed by TwHIN. We include the text and a literal translation of the text to English using the Google Translation Service. We include the labels Hate (Ha), None (No), Offensive (Of) and (Ho) for Truth (T) and Prediction (P).

#	**Text**	**T**	**P**
1	No es por ser gordofobico nada pero amigo en usa tratan de normalizar el "se tu mismo" oh ese estilo de cosas?. Recomiendo ver el video de tri-line hacerca de ese tema pero en mi sincera opinion,no esta mal ser gordo pero es mejor bajar de peso. (*It is not because I am fatphobic at all, but my friend in the USA they try to normalize "be yourself" oh that style of things? I recommend watching the Tri-Line video on this topic, but in my honest opinion, it is not bad to be fat, but it is better to lose weight.*)	Ha	No
2	Nunca nos olvidaremos a la derecha corrupta llevo a la ruina al país en cabeza del gordo marica q el uribismo puso ahí unas ratas (*We will never forget that the corrupt right led the country to ruin at the head of the fat faggot that Uribismo put some rats there*)	Ha	Of
3	Esta vez no estoy de acuerdo con su comentario, y debo decir que es bastante weon el comentario, la obesidad se da no solo por comer mucho o una mala alimentacion, deberia informarse antes de decir esa estupidez, y se gano un seguidor menos. (*This time I do not agree with his comment, and I must say that the comment is quite bad, obesity is not only caused by eating a lot or bad diet, you should inform yourself before saying that stupid thing, and you gained one less follower.!*)	Ho	Ha
4	Pero flaco ustedes solamente pagan cuota de socio, con eso estás adentro, nosotros abono para NUESTRA cancha, y si no tenés abono y sos socio, tenés q pagar la entrada para ir al kempes, mogolico (*But "flaco", you only pay the membership fee, with that you are in, we have a subscription for OUR field, and if you don't have a subscription and you are a member, you have to pay the entrance fee to go to the Kempes, you idiot.*)	Of	No
5	La generación de cristal es bastante rara. Se molestan cuando uno le dice negro, flaco o gordo a alguien, o cuando se les dice que solo existe hombre o mujer en términos biológicos... (*The crystal generation is quite rare. They get upset if you call someone black, thin, or fat, or if you tell them that there is only one biological male or female...*)	Ha	Ho

by including volunteers from different Spanish-speaking countries. Finally, regarding the evaluation of the models, we will evaluate the feature integration to combine the strengths of different pre-trained models based on Transformers.

Acknowledgments

This work has been partially supported by projects LaTe4PoliticES (PID2022-138099OB-I00) funded by MICIU/AEI/10.13039/501100011033 and the European Regional Development Fund (ERDF)-a way of making Europe, LT-SWM (TED2021-131167B-I00) funded by MICIU/AEI/10.13039/ 501100011033 and by the European Union NextGenerationEU/PRTR, CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), and SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government, FedDAP (PID2020-116118GA-I00) and Trust-ReDaS (PID2020-119478GB-I00) supported by MICINN/AEI/10.13039/501100011033, and "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia. The research work conducted by Salud María Jiménez-Zafra has been supported by Action 7 from Universidad de Jaén under the Operational Plan for Research Support 2023-2024. Mr. Ronghao Pan is supported by the Programa Investigo grant, funded by the Region of Murcia, the Spanish Ministry of Labour and Social Economy and the European Union - NextGenerationEU under the "Plan de Recuperación, Transformación y Resiliencia (PRTR)".

References

[1] T. F. Cash, Body image: Past, present, and future, 2004.

[2] B. B. E. Robinson, L. C. Bacon, J. O'reilly, Fat phobia: Measuring, understanding, and changing anti-fat attitudes, International Journal of Eating Disorders 14 (1993) 467–480.

[3] C. N. Wagner, E. Aguirre Alfaro, E. M. Bryant, The relationship between instagram selfies and body image in young adult women, First Monday 21 (2016).

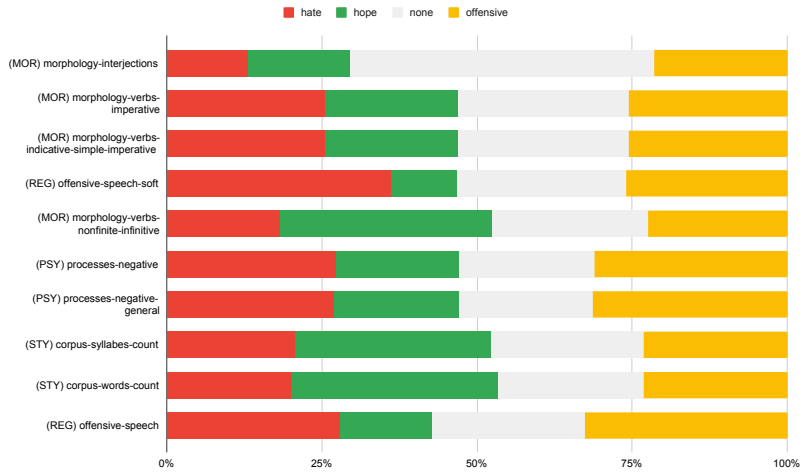

Figure 5: Top-10 Information Gain of the wrong classifications performed of TwHIN per label and linguistic feature. The values are normalized so that each row represents a value of 100%.

[4] A. L. Gonzales, J. T. Hancock, Mirror, mirror on my facebook wall: Effects of exposure to facebook on self-esteem, Cyberpsychology, behavior, and social networking 14 (2011) 79–83.

[5] L. C. Stoll, Fat is a social justice issue, too, Humanity & Society 43 (2019) 421–441. URL: https://doi.org/10.1177/0160597619832051.

[6] J. Aziz, Social media and body issues in young adults: an empirical study on the influence of Instagram use on body image and fatphobia in catalan university students, Master's thesis, Universitat Pompeu Fabra (UPF), 2017.

[7] R. Puhl, Y. Suh, Health consequences of weight stigma: implications for obesity prevention and treatment, Current obesity reports 4 (2015) 182–190.

[8] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, E. Gilbert, You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech, Proceedings of the ACM on human-computer interaction 1 (2017) 1–22.

[9] F. Vargas, I. Carvalho, F. Rodrigues de Góes, T. Pardo, F. Benevenuto, HateBR: A large expert annotated corpus of Brazilian Instagram comments for offensive language and hate speech detection, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7174–7183. URL: https://aclanthology.org/2022.lrec-1.777.

[10] F. Vargas, I. Carvalho, T. A. S. Pardo, F. Benevenuto, Contextual-aware and expert data resources for brazilian portuguese hate speech detection, Research Square (2023). URL: https://doi.org/10.21203/rs.3.rs-2050376/v2.

[11] R. L. Tamayo, B. Chulvi, P. Rosso, Everybody hurts, sometimes overview of hurtful humour at iberlef 2023: Detection of humour spreading prejudice in twitter, Procesamiento del Lenguaje Natural 71 (2023) 383–395.

[12] S. M. Jiménez-Zafra, M. Á. Garcia-Cumbreras, D. García-Baena, J. A. Garcia-Díaz, B. R. Chakravarthi, R. Valencia-García, L. A. Ureña-López, Overview of hope at iberlef 2023: Multilingual hope speech detection, Procesamiento del Lenguaje Natural 71 (2023) 371–381.

[13] F. M. Plaza-del Arco, M. D. Molina-González, L. A. Urena-López, M. T. Martín-Valdivia, Comparing pre-trained language models for spanish hate speech detection, Expert Systems with Applications 166 (2021) 114120.

[14] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, 2019, pp. 75–86.

[15] F. M. Plaza-del Arco, M. Casavantes, H. J. Escalante, M. T. Martín-Valdivia, A. Montejo-Ráez, M. Montes, H. Jarquín-Vásquez, L. Villaseñor-Pineda, et al., Overview of meoffendes at iberlef 2021: Offensive language detection in spanish variants, Proce-

samiento del Lenguaje Natural 67 (2021) 183–194.
[16] M. Wiegand, M. Siegel, J. Ruppenhofer, Overview of the germeval 2018 shared task on the identification of offensive language, in: 14th Conference on Natural Language Processing KONVENS 2018, 2018, pp. 1–10.
[17] J. M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, Overview of germeval task 2, 2019 shared task on the identification of offensive language, in: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), German Society for Computational Linguistics & Language Technology und ..., 2019, pp. 352–363.
[18] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), in: Proceedings of the Fourteenth Workshop on Semantic Evaluation, 2020, pp. 1425–1447.
[19] J. A. García-Díaz, S. M. Jiménez-Zafra, M. A. García-Cumbreras, R. Valencia-García, Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers, Complex & Intelligent Systems 9 (2023) 2893–2914.
[20] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the fifth international workshop on natural language processing for social media, 2017, pp. 1–10.
[21] H. Sohn, H. Lee, Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations, in: 2019 International Conference on Data Mining Workshops (ICDMW), IEEE, 2019, pp. 551–559.
[22] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 1668–1678.
[23] J. A. García-Díaz, M. Cánovas-García, R. C. Palacios, R. Valencia-García, Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings, Future Gener. Comput. Syst. 114 (2021) 506–518. URL: https://doi.org/10.1016/j.future.2020.08.032. doi:10.1016/j.future.2020.08.032.
[24] F.-M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña López, M. T. Martín-Valdivia, Detecting misogyny and xenophobia in spanish tweets using language technologies, ACM Trans. Internet Technol. 20 (2020). URL: https://doi.org/10.1145/3369869. doi:10.1145/3369869.
[25] T. Mandl, S. Modha, P. Majumder, D. Patel, M. Dave, C. Mandlia, A. Patel, Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages, in: Proceedings of the 11th forum for information retrieval evaluation, 2019, pp. 14–17.
[26] T. Mandl, S. Modha, A. Kumar M, B. R. Chakravarthi, Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german, in: Forum for Information Retrieval Evaluation, 2020, pp. 29–32.
[27] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., IberEval SEPLN 2150 (2018) 214–228.
[28] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, Overview of the evalita 2018 hate speech detection task, in: EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, volume 2263, CEUR, 2018, pp. 1–9.
[29] V. Basile, C. Bosco, E. Fersini, N. Debora, V. Patti, F. M. R. Pardo, P. Rosso, M. Sanguinetti, et al., Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter, in: 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 54–63.
[30] J. Gonzalo, M. Montes-y Gómez, P. Rosso, Iberlef 2021 overview: Natural language processing for iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop, 2021, pp. 1–15.
[31] J. Gonzalo, M. Montes-y Gómez, F. Rangel, Overview of iberlef 2022: Natural language processing challenges for spanish and other iberian languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022), CEUR, 2022, pp. 1–12.
[32] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, T. Donoso, Overview of exist 2021: sexism identification in social networks, Procesamiento del Lenguaje Natural 67 (2021) 195–207.
[33] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, A. Mendieta-Aragón, G. Marco-Remón, M. Makeienko, M. Plaza, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2022: sexism identification in social networks, Procesamiento del Lenguaje Natural 69 (2022) 229–240.
[34] M. Taulé, A. Ariza, M. Nofre, E. Amigó, P. Rosso, Overview of detoxis at iberlef 2021: detection of toxicity in comments in spanish, Procesamiento del lenguaje natural 67 (2021) 209–221.
[35] A. Ariza-Casabona, W. S. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, P. Rosso, Overview of detests at iberlef 2022: Detection

and classification of racial stereotypes in spanish, Procesamiento del lenguaje natural 69 (2022) 217–228.

[36] G. Bel-Enguix, H. Gómez-Adorno, G. Sierra, J. Vásquez, S. T. Andersen, S. Ojeda-Trueba, Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish speaking lgbtq+ population, Procesamiento del lenguaje natural 71 (2023) 361–370.

[37] S. Palakodety, A. R. KhudaBukhsh, J. G. Carbonell, Hope speech detection: A computational analysis of the voice of peace, arXiv preprint arXiv:1909.12940 (2019).

[38] B. R. Chakravarthi, HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.

[39] D. García-Baena, M. Á. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, Hope speech detection in spanish: The lgbt case, Language Resources and Evaluation (2023) 1–28.

[40] D. N. Milne, G. Pink, B. Hachey, R. A. Calvo, Clpsych 2016 shared task: Triaging content in online peer-support forums, in: Proceedings of the third workshop on computational linguistics and clinical psychology, 2016, pp. 118–127.

[41] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. P. McCrae, M. Á. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, et al., Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the second workshop on language technology for equality, diversity and inclusion, 2022, pp. 378–388.

[42] B. R. Chakravarthi, V. Muralidaran, Findings of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Kyiv, 2021, pp. 61–72. URL: https://aclanthology.org/2021.ltedi-1.8.

[43] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. McCrae, M. Á. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, D. García-Baena, J. García-Díaz, Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 378–388. URL: https://aclanthology.org/2022.ltedi-1.58. doi:10.18653/v1/2022.ltedi-1.58.

[44] P. Kumaresan, B. R. Chakravarthi, S. Cn, M. Á. García, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, M. Hardalov, I. Koychev, P. Nakov, D. García-Baena, K. Kumar Ponnusamy, B. Preston, Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, 2023, pp. 47–53. doi:10.26615/978-954-452-084-7_007.

[45] J. A. García-Díaz, Á. Almela, G. Alcaraz-Mármol, R. Valencia-García, Umucorpusclassifier: Compilation and evaluation of linguistic corpus for natural language processing tasks, Procesamiento del Lenguaje Natural 65 (2020) 139–142.

[46] J. A. García-Díaz, P. J. Vivancos-Vicente, A. Almela, R. Valencia-García, Umutextstats: A linguistic feature extraction tool for spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 6035–6044.

[47] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, J. Pérez, Spanish pre-trained bert model and evaluation data, in: PML4DC at ICLR 2020, 2020.

[48] A. G. F. no, J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, M. Villegas, Maria: Spanish language models, Procesamiento del Lenguaje Natural 68 (2022). URL: https://upcommons.upc.edu/handle/2117/367156#.YyMTB4X9A-0.mendeley. doi:10.26342/2022-68-3.

[49] J. D. la Rosa, E. G. Ponferrada, M. Romero, P. Villegas, P. G. de Prado Salas, M. Grandury, Bertin: Efficient pre-training of a spanish language model using perplexity sampling, Procesamiento del Lenguaje Natural 68 (2022) 13–23. URL: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403.

[50] J. Cañete, S. Donoso, F. Bravo-Marquez, A. Carvallo, V. Araujo, Albeto and distilbeto: Lightweight spanish language models, in: Proceedings of the 13th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022.

[51] J. M. Pérez, D. A. Furman, L. Alonso Alemany, F. M. Luque, RoBERTuito: a pre-trained language model for social media text in Spanish, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7235–7243. URL: https://aclanthology.org/2022.lrec-1.785.

[52] X. Zhang, Y. Malkov, O. Florez, S. Park,

B. McWilliams, J. Han, A. El-Kishky, Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter, arXiv preprint arXiv:2209.07562 (2022).