Procesamiento del Lenguaje Natural, Revista nº 67, septiembre de 2021, pp. 183-194 recibido 05-07-2021 revisado 12-07-2021 aceptado 16-07-2021

Overview of MeOffendEs at IberLEF 2021: Offensive Language Detection in Spanish Variants

Resumen de la tarea MeOffendEs en IberLEF 2021: Detección de lenguaje ofensivo en las variantes del español

Flor Miriam Plaza-del-Arco[1], **Marco Casavantes**[2], **Hugo Jair Escalante**[2], **M. Teresa Martín-Valdivia**[1], **Arturo Montejo-Ráez**[1], **Manuel Montes-y-Gómez**[2], **Horacio Jarquín-Vásquez**[2], **Luis Villaseñor-Pineda**[2,3]
[1]Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
[2]Laboratorio de Tecnologías del Lenguaje (INAOE), Mexico
[3]Centre de Recherche GRAMMATICA (EA 4521), Université d'Artois, France
{fmplaza, maite, amontejo}@ujaen.es
{hugojair, mmontesg, villasen}@inaoep.mx

Abstract: This paper is an overview of MeOffendES 2021, organized at IberLEF 2021 and co-located with the 37th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2021). The main purpose of MeOffendEs is to promote research on the detection of offensive language in Spanish variants. The shared task involve four subtasks, the first two correspond to the identification of offensive language categories in generic Spanish texts from different social media platforms, while subtasks 3 and 4 are related to the identification of offensive language targeting the Mexican variant of Spanish. Two annotated datasets on offensive language have been released to the Natural Language Processing community. MeOffendes attracted a large number of participants: a total of 69 signed up to participate in the task, 12 submitted official runs on the test data, and 10 submitted system description papers. Corpora and results are available at the shared task website at `https://competitions.codalab.org/competitions/28679`.
Keywords: MeOffendEs, detección de lenguaje ofensivo, procesamiento del lenguaje natural, clasificación de textos.

Resumen: Este artículo presenta la tarea MeOffendES 2021, organizada en iberLEF 2021 junto a la 37ª Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2021). El objetivo principal de MeOffendEs es promover la detección del lenguaje ofensivo en las variantes del español. La tarea compartida implica cuatro subtareas, las dos primeras corresponden a la identificación de categorías de lenguaje ofensivo en textos genéricos en español extraídos de diferentes redes sociales, mientras que las subtareas 3 y 4 están relacionadas con la identificación de lenguaje ofensivo dirigido a la variante mexicana del español. Para la competencia se han puesto a disposición de la comunidad del Procesamiento del Lenguaje Natural dos conjuntos de datos anotados con lenguaje ofensivo. MeOffendes ha atravído a un gran número de participantes: un total de 69 se inscribieron para participar en la tarea, 12 presentaron resultados oficiales sobre los datos de evaluación y 10 presentaron artículos describiendo su sistema. Los conjuntos de datos y los resultados oficiales están disponibles en el sitio web de la tarea compartida: `https://competitions.codalab.org/competitions/28679`.
Palabras clave: MeOffendEs, offensive language detection, natural language processing, text classification.

1 Introduction

Offensive language detection is part of the broader domain of text classification, and closely related to the plethora of subjective language classification tasks (Wiebe et al., 2004), including sentiment analysis (Medhat,

ISSN 1135-5948. DOI 10.26342/2021-67-16

Hassan, and Korashy, 2014), emotion detection (Canales and Martínez-Barco, 2014) and hate speech detection (MacAvaney et al., 2019). Offensive language is more individual-oriented, as offenses are intended to upset or to embarrass someone by means of insults or impolite expressions. The growing participation of people in social media has raised the problem of a frequent use of these communication channels as an uncontrollable means to publish rude messages against others. It is not difficult to realize the large number of works regarding this topic and focusing on different languages or targeted communities.

This interest, which extends through more than a decade so far, has motivated several evaluation campaigns, being the more recent the MEX-A3T task at IberLEF 2020 (Aragón et al., 2020b), Offenseval at SemEval 2020 (Zampieri et al., 2020), or OSACT4 shared task at AOSCT 2020 (Mubarak et al., 2020).

Regarding the approaches applied to tackle the challenge of detecting offensive texts, they share many common methods and algorithms usually explored in text classification tasks, from early lexical based approaches (Razavi et al., 2010) to current deep learning based ones (Plaza-del Arco et al., 2021).

With the aim of promoting research in the detection of offensive language for Spanish and its Mexican variant, we introduced the MeOffendEs task at IberLEF 2021 (Montes et al., 2021) with four subtasks. The first two subtasks involve a novel dataset for offensive language research in general Spanish (OffendEs). The dataset contains users comments in response to posts from well-known influencers in different social media platforms (Twitter, YouTube and Instagram). Comments are annotated on four different classes that involve non-offensive and offensive categories. Participants had to develop solutions for identifying those categories from comments. Additional information is provided in the dataset, including the influencer ID and the social media source of the post. Systems have to face with different challenges in these subtasks: *(i)* different language registers from three different social media platforms *(ii)* multi-class classification on non-offensive and offensive categories, and *(iii)* multi-output prediction based on annotators agreement. A total of five teams submitted their prediction systems for subtask 1, and two teams for subtask 2. Finally, we have received four description papers from the participants in these subtasks. Overall, the systems used by the participants explore state-of-the-art classification models including traditional neural networks and Transformer architectures. The results obtained by the participants motivates to further research on the offensive language detection in different social media platforms and on the identification of different offensive language categories.

In addition to the previous subtasks, MeOffendEs involves two subtasks on offensive language detection targeting the Mexican variant of Spanish (subtasks 3 and 4). These are a continuation of previous efforts in trying to detect aggressiveness and offensiveness in tweets in the context of the IberLEF (Aragón et al., 2020a; Aragón et al., 2019) and IberEval forums (Álvarez-Carmona et al., 2018) for the same variant of Spanish. As in previous editions, participants had to develop solutions to recognize offensiveness in tweets. This time, however, we released additional information accompanying tweets, with the hope that such information could be beneficial for improving the recognition performance. Also, these subtasks consider a new dataset build upon those used in past editions. The main difference being the annotation process: for the IberLEF2021 campaign tweets were carefully analyzed and labeled by a committee of annotators (see Section 3.2). Our goal was to provide a curated dataset that could result in more reliable conclusions.

As in previous editions, the aim of subtasks 3 and 4 were to motivate research on the analysis of offensive language in Mexican Spanish. A language whose characteristics and variations make it unique in its kind and different to other languages, making it necessary to have tailored techniques for this language. Likewise, cultural aspects like the use of language with sexual connotation for non offensive communications make it particularly challenging in terms of disambiguation. This phenomenon is not present in the languages considered in other evaluation forums and therefore it has not been studied elsewhere.

Subtasks 3 and 4 attracted a considerable number of participants, most of them implementing solutions based on cutting edge language modeling tools available. In general terms, performance of solutions was lower than that achieved in previous years (see,

e.g., (Aragón et al., 2020a)), this could be due to the more careful annotation process. For subtask 3, which does not consider any additional metadata, most participants outperformed the baseline, whereas for subtask 4, no team outperformed[1] it. Contrary to what we were expecting, subtask 3 received more submissions even when additional data was provided for subtask 4. Overall, results are encouraging and motivate further research. As with the other subtasks, we will keep the leaderboard of the competition open, so that users can keep making submissions despite the competition is over.

The remainder of this paper is organized as follows. Section 2 introduces the four subtasks that are part of MeOffendEs. Then, Section 3 describes the corpora considered for the different subtasks. Next, the solutions proposed to approach the posed problems and their results are reviewed in Section 4. Finally, conclusions are presented in Section 5.

2 Task description

This section describes in detail the four subtasks that are part of the MeOffendEs competition at IberLEF 2021.

The whole shared task was implemented in the CodaLab platform[2] and every subtask comprised two phases: (1) a development phase in which participants had access to labeled trial and training data, and where they could make submissions to test the platform; and (2) a final phase in which unlabeled test data was released and participants uploaded the predictions of their systems to the platform. The evaluation and raking for the official results used the performance score measured in the test phase. The different subtasks on different corpora (see Section 3), and their evaluation measures, are described below:

- **Subtask 1: Non-contextual multiclass classification for generic Spanish**. Participants had to classify comments into the four different categories associated with the OffendEs corpus (see subsection 3.1). No information about the comment (source or influencer ID) is provided. Participants can optionally submit confidence values to predictions (as a probability for each class, so they all sum 1.0) for the four considered categories, in order to evaluate the agreement of predictions with confidence of ten human annotators. For evaluation we considered the micro-averaged precision, recall and f_1 measures. In cases where participants submit confidence values (between 0 and 1) to their outputs, Mean Squared Error (MSE) is applied (with error value equal to one for wrongly predicted classes).
- **Subtask 2: Contextual multiclass classification for generic Spanish**. Same problem as subtask 1, but metadata (information about targeted users and the related social media) is provided to participants. Participants had access to information associated with posts: social media source, influencer genre and influencer name. The same evaluation measures as subtask 1 were taking into account for this subtask.
- **Subtask 3: Non-contextual binary classification for Mexican Spanish.** Participants must classify tweets as offensive or non-offensive in the OffendMEX corpus, this is, a binary text classification problem. For evaluation we considered the precision, recall and f_1 measure with respect to the offensive class.
- **Subtask 4: Contextual binary classification for Mexican Spanish.** Same problem as subtask 3, but metadata about each tweet was provided to participants. For this subtask, participants had access to information associated with the tweets and their authors like: date, retweet count, followers count, etc. The aim of including this information was to determine to what extent contextual information of tweets and users is useful for improving the detection performance. The same evaluation measures as subtask 3 were used for this one.

3 Datasets and baselines

3.1 OffendES

For subtasks 1 and 2 we have released a novel dataset for offensive language research

[1] Please note that baselines were very competitive as described in Section 3.2.1

[2] `https://competitions.codalab.org/competitions/28679`

in general Spanish (OffendEs). Focusing on young influencers from the well-known social platforms of Twitter, Instagram, and YouTube, we have collected a corpus composed of 47,128 Spanish comments manually labeled on offensive pre-defined categories. A subset of the corpus is labeled with three annotators while another subset is labeled with ten annotators. The latter attaches a degree of confidence to each label computed as the ratio of annotators that agreed on the majority label over the total number of annotators, so both multiclass classification and multioutput regression studies can be carried out. For the competition, we have selected 30,416 posts from the total. The posts are labeled with the following categories:

- Offensive, target is a person (OFP). Offensive text targeting a specific individual.

- Offensive, target is a group of people or collective (OFG). Offensive text targeting a group of people belonging to the same ethnic group, gender or sexual orientation, political ideology, religious belief or other common characteristic.

- Offensive, target is different from a person or a group (OFO). Offensive text where the target does not belong to any of the previous categories, e.g., an organization, an event, a place, an issue.

- Non-offensive, but with expletive language (NOE). A text that contains rude words, blasphemes or swearwords but without the aim of offending, and usually with a positive connotation.

- Non-offensive (NO). Text that is neither offensive nor contains expletive language.

We consider a post as offensive when language is used to commit an explicit or implicitly directed offense that may include insults, threats, profanity or swearing.

Additional to the text of the comment, two features were also provided as "contextual" information: the name of the social platform where that comment was posted to, and the gender addressee of the comment, i.e. the targeted user.

Finally, different sets have been released for the competition. During the pre-evaluation phase training and development (Dev) sets were provided to the participants and for the evaluation phase the test set was release, Table 1 shows the number and percentage of posts corresponding to each set by the above categories.

Label	Training	Development	Test
NO	13,212	64	9651
NOE	1,235	22	2340
OFP	2,051	10	1404
OFG	212	4	211
Total	16,710	100	13,606

Tabla 1: Distribution of the OffendES categories by subset (Training, Dev, Test) in MeOffendES subtasks 1 and 2.

3.1.1 Baseline

To evaluate the non-contextual multiclass classification task on the OffendEs dataset, we implemented a straightforward baseline system based on a bag-of-words of unigrams, bigrams and trigrams and an linear SVM classifier. For the multiouput regression task we use a multiouput regressor along with the Epsilon-Support Vector Regression. No preprocessing has been applied to the text, nor has a hyperparameter search been performed. We refer to these baselines as *baseline-svm*.

3.2 OffendMEX

For subtasks 3 and 4 we have released a novel dataset in Mexican Spanish collected from Twitter and manually labeled for offensiveness. The resource is formed by tweets labeled with binary and multiclass categories with the following types of offensiveness according to a recent categorization (Díaz-Torres et al., 2020): offensive, aggressive and vulgar (but not offensive). Although the inherent problem is a multi label classification one (e.g., a tweet can be offensive but not vulgar), we are approaching the underlying binary-classification task of interest: distinguishing offensive from non-offensive tweets. Nevertheless, we released all of the labels for the training data as additional information that participants can exploit when developing their solutions. Such information was not provided in the test set partition.

Additionally, for subtask 4, we distributed metadata information associated with tweets in the corpus, these include: date, retweet count, favourite count, reply status, quote status; and metadata derived from users, including: verification status, followers count, listed count, favourites count, status count,

date the account was created, among a few others associated to the user profile and image. Detailed information on the considered metadata can be found in the corresponding API documentation (Twitter, 2021).

Partition	Tweets	Off.	No Off.
Trial	76	41	35
Training	5,060	1,381	3,679
Test	2,183	600	1,583
Total	7,319	2,022	5,287

Tabla 2: Number of tweets in the OffendMex corpus.

Table 2 summarizes the OffendMex data set used for subtasks 3 and 4. The Trial partition was rather small, as the idea was to use such partition for testing the submission system. Training and test partitions are larger and present an approximate class imbalance ratio of 2,6 favoring the non-offensive class.

3.2.1 Baselines

In order to approach the Contextual and Non-contextual binary classification for Mexican Spanish, we implemented two popular approaches that have shown to be hard to beat in both subtasks: *i*) a Bidirectional Gated Recurrent Unit (Bi-GRU) neural network for the Non-Contextual binary classification, and *ii*) a XGBoost + BETO ensemble for the Contextual binary classification.

For the Bi-GRU neural network baseline, all text was pre-processed by removing special characters and stopwords (with the exception of personal pronouns); in order to enrich the vocabulary, all hashtags were segmented by words (*e.g.* #EsDeLesbianas - es de lesbianas), and all emojis were converted into words (*e.g.* ☺ - 'cara sonriente'). As input features pre-trained Spanish FastText (Grave et al., 2018) embeddings were used, and a fully-connected softmax layer handle the class probabilities. Alternatively, for the XGBoost + BETO ensemble baseline the data pre-processing steps consisted of converting the text to lowercase and stripping it of emojis. This ensemble involves two stages. In the first stage the messages were classified considering only their textual content using BETO (Cañete et al., 2020). Subsequently, in the second stage BETO predictions were concatenated to the three most discriminative metadata features (*Tweet favorite count*, *User listed count* and *Default profile*) to form new vectors, handled at the end by a XGBoost classifier (Chen and Guestrin, 2016). We refer to these two baselines as *baseline-dl* below.

In addition to the previous baselines we evaluated the performance of a rather straightforward baseline based on a bag-of-unigrams-bigrams-trigrams and an linear SVM classifier, where a similar preprocessing as above was applied. The goal of this baseline was to evaluate the added value metadata when using a linear model, and to assess the margin of benefits of approaches over a direct baseline method. We will refer to these baselines as *baseline-bow*.

4 Participant approaches and results

4.1 Subtask 1

This subtask, as introduced previously, proposes a pure multi-class text classification problem or a multi-output one. Here, a brief description of participants' systems is provided.

NLP-CIC team (Aroyehun and Gelbukh, 2021) used the multilingual model XLM-RoBERTa pre-trained on Twitter texts and Sentiment Analysis data. They show that Sentiment Analysis and the social domain adaption is beneficial for the problem of offensive language detection. The system ranked the first position in the competition.

UMUTeam (Garcá-Díaz, Jiménez-Zafra, and Valencia-García, 2021) explored a wide range of features and how to combine them in a final multi-layer perceptron (MLP) with several tentative configurations. The features considered were lexical features, negation features, word and sentence embeddings from different embedding algorithms (fastText , word2vec, gloVe and a Spanish version of BERT). Word embeddings were evaluated isolated from the rest of features using convolutional networks and two well-known recurrent architectures like LSTM and Bi-GRU, although MLP was the one showing the best behavior. In general, these features were further pre-processed, with MinMax scaler for linguistic ones and Robust scaler for negation features. All these features as filtered using mutual information. Also, several approaches to combine the total number of features generated were evaluated, including

majority voting, weighted voting and logistic regression. Different kinds of shape and different number of layers, number of neurons, dropout probabilities, batch sizes and activation functions defined a varied number of experiments in order to identify the best configuration for system hyperparameters. From official results it can be drawn that a combination of BERT-based encodings (pre-trained and fine-tuned), with sentences embeddings and lexical and negation features became the best solution. When linguistic features were removed, the system obtained the second position in the competition.

The GDUFS_DM team applied sequence classification system fine-tuned on a pre-trained BERT model and composing the final encodings for the text from a max pooling of the sentence encondings from all layers and token encondings from last layer. Two additional techniques were integrated in the final system: pseudo-labeling and focal loss. The former technique consists of a two-stage training, were test labels are predicted and re-entered into the learning process to produce a larger training set. Focal loss was used as a way to correct class imbalance. The system ranked in the third position in the competition.

Marta Navarrón and Isabel Segura (García and Bedmar, 2021) explored different deep learning models including Long-Short Term Memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT). The best results was archived by the BERT model. The system ranked in the fourth position in the competition.

4.2 Subtask 2

UMUTeam was the only team in submitting results to this subtask. They applied the same system to add to the set of features applied one-hot encodings of contextual columns (gender and media). Robust scaler was also applied to these two features, as done with negation ones. Compared to what was obtained in subtask 1, the integration of contextual information contributed to a small, but consistent improvement in final scores.

4.3 Analysis of subtasks 1 and 2

Table 3 and Table 4 provides a summary of the official results for subtasks 1 and 2 in terms of micro-average and macro-average of Precision, Recall and F_1 scores, respectively. Regarding the multiclass classification setting, it can be notice that all the teams outperformed our *baseline-svm* which shows the success of the neural network models employed by the participants compared to classical machine learning algorithms. However, for the multioutput regression setting, two of the four teams outperformed the SVM regressor baseline, which shows the success of the classical learning algorithm in this setting. For the non-contextual multiclass classification, it can be seen that the scores of the first three teams are very close. This closeness in performance could be due to the fact that most of these top ranked participants relied on similar pretrained models in their solutions (Spanish BERT model, except for NLP-CIC, who fine-tuned a multilingual RoBERTa model). But greater differences can be observed when looking at the MSE error. The lower MSE value is, the closer is the system to the behaviour of human annotators. In that case, XML-RoBERTa almost reduces to a half the error of the second system in the ranking. Finally, both F_1 scores and MSE errors are consistent in terms of ranking order.

For the second subtask, only one team evaluate their system. We can observe that the contextual information did not improve performance, in terms of F_1 score, to that obtained by their system in subtask1. But regarding MSE, including those additional features (social media platform and gender of the targeted user) do led to a system closer to human annotator behaviour.

Subtask 1: Non-contextual classification				
Team	**P**	**R**	F_1	**MSE**
NLP-CIC	0.8815	0.8815	0.8815	0.0231
UMUTeam	0.8782	0.8782	0.8782	0.0411
GDUFS_DM	0.8732	0.8732	0.8732	0.0672
Marta_Isabel	0.8416	0.8417	0.8416	0.0697
baseline-svm	0.8285	0.8285	0.8285	0.0615
Subtask 2: Contextual classification				
Team	**P**	**R**	F_1	**MSE**
UMUTeam	0.8782	0.8782	0.8782	0.0409

Tabla 3: Subtasks 1 and 2 official ranking. Results are in terms of Micro-precision, Micro-Recall and Micro-F_1 scores.

Subtask 1: Non-contextual classification				
Team	**P**	**R**	F_1	**MSE**
NLP-CIC	0.7679	0.7093	0.7324	0.0231
UMUTeam	0.7861	0.6919	0.7301	0.0411
GDUFS_DM	0.7565	0.7002	0.7239	0.0672
Marta_Isabel	0.5781	0.5451	0.5595	0.0697
baseline-svm	0.6278	0.4831	0.5236	0.0615
Subtask 2: Contextual classification				
Team	**P**	**R**	F_1	**MSE**
UMUTeam	0.7879	0.6921	0.7308	0.0409

Tabla 4: Subtasks 1 and 2 official ranking. Results are in terms of Macro-precision, Macro-Recall and Macro-F_1 scores.

4.4 Subtasks 3 and 4: Offensive language identification in Mexican Spanish

We now analyze the results obtained by participants of subtasks 3 and 4. For the former, a total of 10 teams submitted runs for the final phase and were considered for the official leaderboard. In addition, two other teams submitted runs but these did not qualify for the official ranking. For subtask 4 we received the submissions from three different teams. This was somewhat disappointing as we were expecting participants to exploit the metadata provided with the dataset.

Subtask 3: Non-contextual classification			
Team	P	R	F_1
CIMAT-MTY-GTO	0.7600	0.6533	0.7026
NLP-CIC	0.7550	0.6407	0.6932
DCCD-INFOTEC	0.6733	0.6966	0.6847
CIMAT-GTO	0.6633	0.6958	0.6792
UMUTeam	0.6650	0.6763	0.6706
Timen	0.6000	0.6081	0.6040
CIC-IPN	0.5350	0.6874	0.6017
Baseline-bow	0.6040	0.5517	0.5767
Baseline-dl	0.7192	0.4100	0.5222
xjywing	0.8883	0.3417	0.4935
aomar	0.8750	0.3239	0.4728
CEN-Amrita	0.9183	0.3143	0.4683
Subtask 4: Contextual classification			
Team	P	R	F_1
Baseline-dl	0.6629	0.6983	0.6801
UMUTeam	0.6683	0.6705	0.6694
CIC-IPN	0.5383	0.6843	0.6026
Baseline-bow	0.6062	0.5517	0.5777
Timen	0.4233	0.4456	0.4341

Tabla 5: Final results for the Contextual and Non-contextual binary classification for Mexican Spanish

Table 5 provide a summary of the official results for subtasks 3 and 4. For the former, i.e., non-contextual binary classification, it can be seen that there were only 3 teams that did not beat the baselines associated with the task. The best performance was obtained by the CIMAT-MTY-GTO team with a relative improvement over the *bow* and *dl* baselines of 21 % and 34 %, respectively. Followed closely by the next 4 teams in the ranking. This closeness in performance could be due to the fact that most of these top ranked participants relied on similar pretrained models in their solutions, see Table 6.

Interestingly, *baseline-bow* outperformed the one based on the Bi-GRU. This could be due to the fact that the latter model was trained only on the available data, which may be of limited size given the complexity of GRU models. Other participants outperformed the baselines because they used external resources and pretrained models (see Table 6).

Regarding task 4, from Table 5 it can be seen that no team outperformed *baseline-dl*. This is due in part to the competitiveness of the baseline model, but also, to the fact that participants did not do any special processing for the provided metadata (see below). Still, two out of the three teams were close to the baseline. On the other hand, only one team did not outperformed the *baseline-bow*. The improvement of *baseline-dl* over *baseline-bow* could be due to the fact that the former used a representation based on a transformer, opposed to the Bi-GRU baseline considered for subtask 3.

Finally, when comparing the results of *baseline-bow* in both tasks, it is observed that the added value of metadata in subtask 4 only yield a negligible improvement. This confirms that the sole inclusion the features is not enough to improve performance.

4.4.1 Systems descriptions

Table 6 summarizes the contributions from teams that participated in subtasks 3 and 4, it shows the adopted models and highlights any novel aspect of the different approaches. In the following, we outline the main findings from the methodologies proposed to approach subtasks 3 and 4.

- **Transformer-based solutions were common.** Most teams relied on pretrained transformers for Spanish in the modeling process, we assume this was with the purpose of alleviating the small sample problem. This is in line with trends in

general NLP, and in general it was very helpful: most of top ranked participants used transformers. Despite these results, we think that more specialized mechanisms could help to boost performance when using transformers.

- **Advanced linguistic features were not considered in most approaches.** Only a couple of teams (UMUTeam and CIC-IPN teams) proposed solutions that included linguistic analyses, and another team relied on a genetic programming formulation (DCCD-INFOTEC). It is interesting that their performance was competitive, even when no transformer was used. It may be expected that an adequate combination of linguistic and data-driven features could result in better performance.
- **External data.** External data for further fine tuning transformer models was adopted by top ranked participants. Suggesting this is a promising way for further improving the performance of transformers.
- **No special treatment for processing metadata for subtask 4.** It was somewhat disappointing that participants of subtask 4, did not took full advantage of metadata. These features were concatenated to the other input spaces and feed to classification models. We are still confident that recognition performance can be improved when these features are used effectively.

4.4.2 Analysis

In order to further analyze the participants' results, we performed an analysis on the complimentariness and diversity of the predictions from the different teams. For this analysis we used the last run from every team in the different subtasks. We measured the diversity of predictions by using the Coincident Failure Diversity (CFD) measure (Tang, Suganthan, and Yao, 2006). This measure expresses with a number in $[0, 1]$ the extend to which the errors made by different classification system overlap (the higher the value the less correlated the errors are). On the other hand, to measure the complimentariness of predictions, we calculated the Maximum Possible Accuracy (MPA), this is the accuracy that can be obtained if we consider a tweet correctly classified when at least one of the considered systems classified it correctly.

Table 7 shows the CFD and MPA values obtained for the considered runs from subtasks 3 and 4, for completion we also report the best accuracy (BA) obtained by any of the considered systems. From these results, it is possible to observe that the MPA in both subtasks is considerably greater than the BA in both subtasks, suggesting that the participants systems and approaches are somewhat complementary to each other: performance could improve by ≈16 % and ≈11 % if the predictions from the available systems were combined optimally for subtasks 3 and 4, respectively. The values of the CFD measure show that there is a high diversity among errors in the predictions of both subtasks. Most notably for subtask 3. Results are consistent with the number of participating teams in both subtasks.

As a result of applying the MPA metric, it was possible to identify those common errors across all systems[3]. In fact, there are only 34 tweets that no system could classify correctly. Nine of them are offensive tweets that were classified as non-offensive. Below we present some of these tweets (a rough translation to English is provided), where we can identify offenses with no vulgar or profane words (*e.g.* "gata"), the use of out of the training vocabulary words (*e.g.* "whore"), the masking of potentially offensive words (*e.g.* "pen...."), as well as some ironic comments.

- Nada peor que whore y gorda #sorrynotsorry. (ENG: *Nothing worse than whore and fat #sorrynotsorry*)
- Yo no te quise decir gata, pero bueno. Eres una gata (ENG: *I did not mean to say gata. But well. You are a gata.*)
- Soy yo o @USUARIO está bien pen.... hace dramas se pone loca y pierde jajaja @ExatlonMx. (ENG: *It is me or @USER is pen... she gets crazy and loses hahah @ExatlonMx.*)
- Básicamente, el feminismo se trata de feas peleando por los derechos de las guapas. (ENG: *Basically, feminism is about the ugly girls fighting for the rights of the pretty ones.*)

A couple of mistakenly classified non-offensive tweets are the following:

- @USUARIO Vas en micro, camina, se suben unos HDP y les quitan todo a todos En un taxi, te pueden secuestrar... En el metro hay n carteristas. (ENG: *@USER you are going in bus, it moves, some HPD get in, they steal everyone. In a can you can be kipnaped.... in the subay there are n pinpockets.*)

[3] *NOTE: In this section we include examples of language that may be offensive to some readers, these do not represent the perspectives of the authors.*

Systems considered in the official ranking.			
Team	Novel elements	Transformer / Model	Reference
CIMAT-MTY-GTO	External data was from hate-speecha detection and sentiment analysis was used to augment the training set.	Ensemble of BERT models for Spanish (BETO)	(Gómez-Espinosa, Muñiz-Sanchez, and López-Monroy, 2021)
NLP-CIC	The model was trained with both tweets and sentiment analysis data in Spanish.	XLM-RoBERTa	(Aroyehun and Gelbukh, 2021)
DCCD-INFOTEC	A combination of different models trained for humor, aggressiveness and misogyny detection, in addition to models trained on the provided training set and a reverse version of it.	EvoMSA (genetic programming based model)	(Calderón, Tellez, and Graff, 2021)
CIMAT-GTO	The models were trained taking advantage of the auxiliary sentence for the transformer. Two methods for obtaining auxuliary sentences were proposed.	Ensemble of BERT models for Spanish (BETO)	(Sánchez-Vega and López-Monroy, 2021)
UMUTeam*	A variety of linguistic features, including negation were considered and combined with learned representations.	Ensembles of models based on linguistic and learned features (embeedings).	(Garcá-Díaz, Jiménez-Zafra, and Valencia-García, 2021)
CIC-IPN*	A diversity of configurations were tested, a model pretrained on tweets and sentiment analysis data obtained the best performance.	XLM-RoBERTa	(Huerta-Velasco and Calvo, 2021)
CEN-Amrita	Better results were obtained withe the Bi-LSTM model	Bidirectional LSTM and BERT (bert-base-multilingual-cased)	(Sreelakshmi, Premjith, and Soman, 2021)
Additional models			
QuSwe1d0n	The representation obtained from the transformer was feed to a CNN based model.	XLM and CNN	(Qu, Que, and Shuangjun, 2021)
YNU_qyc	The output of the transformer was feed to an LSTM model, a K-folding ensemble scheme was adopted.	XLM-RoBERTa and LSTM	(Qu, Yang, and Wang, 2021)

Tabla 6: Summary of system descriptions that participated in subtasks 3 and 4. * Indicates this team participated in both tasks. We separate systems that qualified for the official results and additional systems.

Subtask	BA	MPA	CFD	NoT
3	0.8277	0.9844	0.6073	10
4	0.8185	0.9271	0.2685	3

Tabla 7: Comparison of MPA and CFD results between the Contextual and Non-contextual binary classification. The Best Accuracy (BA) obtained by the participating teams in each subtask, was used to compare the complementarity obtained with the MPA metric; NoT stands for Number of Teams.

- Sus pinches relaciones empiezan con un Invita a tus amigas las más putas y piden fidelidad, malditos ilusos. (ENG: *Your damn relationships start with an invite to your friends, the most promiscuous and you ask fidelity, fkng dreamer.*)

This analysis suggests the most difficult instances are those that require further linguistic analysis. This evidences the inherent challenges of this variation of language and the detection of offensiveness in text. Motivating further research on this subject.

Another important aspect to mention is that the corpus used this year took a subset of last year's data and was relabeled with the guide proposed by (Díaz-Torres et al., 2020). For this task, a group of labelers of different ages was selected (3 adults, 6 youth) and balancing the number of males and females (4 females, 5 males). This new relabeling was the main decrease in the reported results (baseline decreased by 0.19 points compared

to last year). Diversity in the group of labelers (generational change as well as gender) increased the variations present in both the training set and the test set. The creation of robust systems for this task must consider these scenarios both during the training phase and to provide a confidence rating of the prediction made by the automatic method.

5 Conclusion

The MeOffendEs shared task at IberLEF attempts to continue to the research in offensive language detection in Spanish. A new dataset on generic Spanish has been prepared for this edition, as a companion collection to the existing one on Mexican Spanish, enabling intensive experimentation over a large number of messages from different social media platforms. This evaluation campaign allowed participants to test their systems on this classification task. Different algorithms, features, techniques and configurations were explored, reporting the effectiveness of the approaches and contributing to the advance of offensive language detection systems.

A total of 69 participants registered to the MeOffendEs shared task. However, only 12 teams participated in the final phase of the challenge. Interesting findings and conclusions have been drawn and very competitive approaches are now available to approach the 4 proposed subtasks. Given the interest from the community we are keeping the challenge website open so that anyone interested in trying their own methods can do it at any time.

Acknowledgements

We would like to thank CONACyT for partially supporting this work under grants CB-2015-01-257383 and the Thematic Networks program (Language Technologies Thematic Network). Hugo Jair Escalante is supported by CONACyT under project grant CONACYT CB-S-26314.

This work is also partially supported by the grant P20_00956 (PAIDI 2020) from Andalusian Regional Government, a grant from European Regional Development Fund (FEDER), the LIVING-LANG project [RTI2018-094653-B-C21], and the Ministry of Science, Innovation and Universities (scholarship [FPI-PRE2019-089310]) from the Spanish Government.

References

Álvarez-Carmona, M. Á., E. Guzmán-Falcón, M. Montes-y-Gómez, H. J. Escalante, L. Villaseñor-Pineda, V. Reyes-Meza, and A. Rico-Sulayes. 2018. Overview of MEX-A3T at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In P. Rosso, J. Gonzalo, R. Martínez, S. Montalvo, and J. C. de Albornoz, editors, *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018) co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018), Sevilla, Spain, September 18th, 2018*, volume 2150 of *CEUR Workshop Proceedings*, pages 74–96. CEUR-WS.org.

Aragón, M. E., M. Á. Álvarez-Carmona, M. Montes-y-Gómez, H. J. Escalante, L. Villaseñor-Pineda, and D. Moctezuma. 2019. Overview of MEX-A3T at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, J. Carrillo-de-Albornoz, S. Montalvo, L. Chiruzzo, S. Collovini, Y. Gutiérrez, S. M. J. Zafra, M. Krallinger, M. Montes-y-Gómez, R. Ortega-Bueno, and A. Rosá, editors, *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF-SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421 of *CEUR Workshop Proceedings*, pages 478–494. CEUR-WS.org.

Aragón, M. E., H. J. Jarquín-Vásquez, M. Montes-y-Gómez, H. J. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, J. P. Posadas-Durán, and G. Bel-Enguix. 2020a. Overview of MEX-A3T at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In M. Á. G. Cumbreras, J. Gonzalo, E. M. Cámara, R. Martínez-Unanue, P. Rosso, S. M. J. Zafra, J. A. O. Zambrano, A. Miranda, J. P. Zamorano, Y. Gutiérrez, A. Rosá, M. Montes-y-Gómez, and M. G. Vega, editors, *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Pro-*

cessing (SEPLN 2020), Málaga, Spain, September 23th, 2020, volume 2664 of *CEUR Workshop Proceedings*, pages 222–235. CEUR-WS.org.

Aragón, M., H. Jarquín, M. M.-y. Gómez, H. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, G. Bel-Enguix, and J. Posadas-Durán. 2020b. Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In *Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain*.

Aroyehun, S. T. and A. Gelbukh. 2021. Evaluation of intermediate pre-training for the detection of offensive language. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

Calderón, J. J., E. S. Tellez, and M. Graff. 2021. Dccd-infotec at meoffendes@iberlef21 subtask 3: A transfer learning approach based on evomsa's stacked generalization. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

Canales, L. and P. Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43.

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

Chen, T. and C. 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, page 785–794, New York, NY, USA. Association for Computing Machinery.

Díaz-Torres, M. J., P. A. Morán-Méndez, L. Villasenor-Pineda, M. Montes-y Gómez, J. Aguilera, and L. Meneses-Lerín. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a Mexican Spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 132–136, Marseille, France, May. European Language Resources Association (ELRA).

Garcá-Díaz, J. A., S. M. Jiménez-Zafra, and R. Valencia-García. 2021. Umuteam at meoffendes 2021: Ensemble learning for offensive language identification using linguistic features, fine-grained negation and transformers. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

García, M. N. and I. S. Bedmar. 2021. Detecting offensiveness in social network comments. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

Gómez-Espinosa, V., V. Muñiz-Sanchez, and A. P. López-Monroy. 2021. Transformers pipeline for offensiveness detection in mexican spanish social media. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, page .

Huerta-Velasco, D. A. and H. Calvo. 2021. Using lexical resources for detecting offensiveness in mexican spanish tweets. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

MacAvaney, S., H.-R. Yao, E. Yang, K. Russell, N. Goharian, and O. Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Medhat, W., A. Hassan, and H. Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.

Montes, M., P. Rosso, J. Gonzalo, E. Aragón, R. Agerri, M. Álvarez Carmona, E. Álvarez Mellado, J. Carrillo-de Albornoz, L. Chiruzzo, L. Freitas, H. Gómez Adorno, Y. Gutiérrez, S. Lima,

S. M. Jiménez-Zafra, F. M. Plaza-del-Arco, and M. Taulé, editors. 2021. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*.

Mubarak, H., K. Darwish, W. Magdy, T. Elsayed, and H. Al-Khalifa. 2020. Overview of osact4 arabic offensive language detection shared task. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 48–52.

Plaza-del Arco, F. M., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2021. Comparing pretrained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Qu, S., Q. Que, and Shuangjun. 2021. Noncontextual binary classification for mexican spanish with xlm and cnn. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

Qu, Y., Y. Yang, and G. Wang. 2021. Ynu_qyc at meoffendes@iberlef 2021:the xlm-roberra and lstm for identifying offensive tweets. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

Razavi, A. H., D. Inkpen, S. Uritsky, and S. Matwin. 2010. Offensive language detection using multi-level classification. In A. Farzindar and V. Kešelj, editors, *Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sánchez-Vega, F. and A. P. López-Monroy. 2021. Cimat-gto at meoffendes 2021: Bert's auxiliary sentence focused on word's information for offensiveness detection. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

Sreelakshmi, K., B. Premjith, and K. P. Soman. 2021. Transformer based offensive language identification in spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021)*, CEUR Workshop Proceedings. CEUR-WS.org.

Tang, E. K., P. N. Suganthan, and X. Yao. 2006. An analysis of diversity measures. *Mach. Learn.*, 65(1):247–271.

Twitter. 2021. Tweets – twitter developers. `https://developer.twitter.com/`. Accessed: 2021-06-30.

Wiebe, J., T. Wilson, R. Bruce, M. Bell, and M. Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Zampieri, M., P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, and Ç. Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.