UMUTeam at MeOffendEs 2021: Ensemble Learning for Offensive Language Identification using Linguistic Features, Fine-grained Negation, and Transformers

José Antonio García-Díaz[1][0000−0002−3651−2660],
Salud María Jiménez-Zafra[2][0000−0003−3274−8825], and
Rafael Valencia-García[1][0000−0003−2457−1791]

[1] Facultad de Informática, Universidad de Murcia,
Campus de Espinardo, 30100, Spain
{joseantonio.garcia8,valencia}@um.es
[2] Computer Science Department, SINAI, CEATIC,
Universidad de Jaén, 23071, Spain
sjzafra@ujaen.es

Abstract. This paper presents the participation of the UMUTeam in the MeOffendEs shared task at IberLEF 2021. This task involves the identification and categorisation of offensiveness in Spanish comments from different social networks (YouTube, Instagram and Twitter), and Mexican Spanish tweets. Specifically, four subtasks were proposed: the first one on multi-class classification of offensiveness types, the second one also concerning multi-class classification but with contextual information, the third one on a binary classification of texts as offensives or non-offensives, and the last one also regarding a binary classification but with metadata. Subtasks 1 and 2 focus on generic Spanish, and subtasks 3 and 4 on Mexican Spanish. We have participated in the four subtasks with the aim of promoting the automatic identification of offensiveness in Spanish variants. Our proposal for solving these subtasks is based on the combination of linguistic features (including fine-grained negation features) and embeddings using transformers and ensemble learning. We ranked in second place in subtask 1 with a micro-averaged F1-score of 87.8289%, first in subtask 2 with a micro-averaged F1-score of 87.8289%, fifth in subtask 3 with a macro-averaged F1-score of 67.0588%, and first in subtask 4 with a macro-averaged F1-score of 66.9449%.

Keywords: Offensiveness · Feature Engineering · Negation processing · Transformers · Ensemble learning · Natural Language Processing

IberLEF 2021, September 2021, Málaga, Spain.

1 Introduction

This work describes the participation of the UMUTeam in the shared task MeOffendEs 2021 [22], organised in the IberLEF 2021 workshop [21] and aimed at the identification of offensive language and its categories. A text is considered offensive if it contains hurtful, derogatory or obscene comments made by one person to another person [25]. The use of offensive language in social media has increased in recent years. Some users make use of the freedom of expression offered by these media to communicate in an offensive way. This poses a major problem for society as offensive comments can cause significant harm to the people they are directed at, such as depression or suicide.

The aim of MeOffendEs 2021 is to promote the development of tools to detect and recognise offensive language and its categories in Spanish and Mexican Spanish. Specifically, they are proposed four subtasks:

- *Subtask 1: Non-contextual multiclass classification for generic Spanish.* Identify the type of offensiveness used in each of the given comment: non-offensive (NO), non-offensive but with inadequate language (NOM), offensive and target is a person (OFP) or, offensive and target is a group of people or collective (OFG).
- *Subtask 2: Contextual multiclass classification for generic Spanish.* Use metadata as an additional source of information to identify the type of offensiveness: NO, NOM, OFP or OFG.
- *Subtask 3: Non-contextual binary classification for Mexican Spanish.* Classify each of the given tweet as offensive or non-offensive.
- *Subtask 4: Contextual binary classification for Mexican Spanish.* Use metadata as an additional source of information to classify each tweet as offensive or non-offensive.

The remainder of this manuscript is organised as follow. First, in Section 2, a short overview on workshops regarding offensiveness and Spanish corpora is presented. Next, in Section 3, we give some insights regarding the datasets that were made available to the participants. Following, in Section 4, the methodology of our proposal is described. In Section 5, we show the results achieved by our team and compare them with those obtained by the rest of participants. Finally, the conclusions and future research directions are shown in Section 6.

2 Background information

Offensive language detection is a task of recent interest due to the proliferation of this type of language in social media. One of the strategies used to stop messages with offensive content is to report these messages, but doing this manually is not feasible due to the large amount of information that is published daily on the Web. Therefore, research efforts are being invested to automate this process. Studies on offensive language have focused on hate speech [18, 5], cyberbulling [6, 2] and aggression [16]. In fact, we can find a large set of shared tasks about this topic, such as the 2018 and 2019 editions of the GermEval Shared Task on the Identification of Offensive Language [25, 24], the 2019 and 2020 editions of the OffensEval shared task on Identifying and Categorising Offensive Language in Social Media [26, 27], the AMI shared task on Automatic Misogyny Identification at IberEval 2018 [8] and Evalita 2018 [7], the 2019 and 2020 editions of the HASOC track on Hate Speech and Offensive Content Identification [20, 19], the HatEval shared task on the Detection of Hate Speech against Immigrants and Women [3], the MEX-A3T track at IberLEF 2019 on Authorship and Aggressiveness Analysis [1], and the 2018 and 2020 editions of the TRAC shared task on Trolling, Aggression and Cyberbullying [15, 17].

Of the tasks previously mentioned only AMI and HatEval provided datasets on generic Spanish and MEX-A3T on Spanish Mexican, which are the languages under studied in the MeOffendEs task. On the one hand, the two AMI datasets consist of documents written in English, Spanish and Italian and are annotated according to three levels: misogyny (misogyny or not misogyny), misogynistic category (discredit, derailing, dominance, sexual harassment and threats of violence, and stereotype and objectification) and target (individuals or groups). In the AMI shared task of IberEval and Evalita, two tasks were proposed: a binary classification on misogyny identification and a categorisation of misogynistic behaviours and targets. On the other hand, the HatEval dataset is composed of tweets written in Spanish and English related to hate-speech towards women and immigrants. Similar to AMI, two tasks were proposed in HatEval: a binary hate speech detection against immigrants and women, and an aggressive behaviour and target classification in which the participants were encouraged to discern between aggressive or not aggressive messages, to later identify if the victim of the harassment is a person or a collective. Finally, MEX-A3T proposed a binary aggressiveness detection track focused on identifying aggressive tweets written in Mexican Spanish. The MEX-A3T dataset was compiled from Mexico City and contains documents with offensive, vulgar, and aggressive language.

3 Datasets

Subtask 1 and subtask 2 are multi-classification, in which documents were tagged as non-offensive (NO), non-offensive but with inadequate language (NOM), and offensive, discerning whether the target is a person (OFP) or a group (OFG). According to the description provided by the organisers, the dataset was compiled from multiple social networks, including YouTube, Instagram and Twitter. The dataset was provided to the participants in two sets, depending on whether the documents were labelled by three or ten annotators. It is worth mentioning that the organisers of the task divided the corpus into three splits, namely, train, development, and test. However, the official development set consisted only in 100 examples and, therefore, as the labels were not balanced, some classes get under represented. Therefore, we decided to merge train and dev, and generate two custom splits into a partition of 80-20. The original splits can be downloaded at https://github.com/pendrag/MeOffendEs. The dataset distribution for subtasks 1 and 2 are depicted on Table 1, in which we can observe that the label OFG is the one with fewer instances.

Table 1. Datasets distribution for subtasks 1 and 2

label	total	train	validation	test
NO	13276	10621	2655	-
OFP	2073	1658	415	-
NOM	1245	996	249	-
OFG	216	173	43	-
Total	16810	13448	3362	13606

In case of generic Spanish, in subtask 2, some contextual information were provided along with the dataset, regarding the author of the document. This metadata includes information about the social media in which the comment was posted, the name of the channel or the main user involved (also known as the influencer), and its gender.

For subtasks 3 and 4, concerning Spanish Mexican, the dataset consisted into documents labelled as *offensive* or *non-offensive*. According to the organisers, this dataset was compiled from Twitter and labelled at first place as: *offensive*, *aggressive*, and *vulgar but non offensive*, but finally merged as binary class. The distribution of the labels across the different split is shown in Table 2. We can observe that the relation among *non-offensive* and *offensive* documents is near 2.6 which can be considered a strong imbalance.

Table 2. Datasets distribution for subtasks 3 and 4

label	total	train	validation	test
non-offensive	3714	2971	743	-
offensive	1422	1137	285	-
Total	5136	4108	1028	2193

For subtask 4, the organisers provided a large variety of contextual data including the date of publication, its number of retweets, the number of times the tweet has been marked as liked by users, and whether the tweet is a reply or an original comment, among other contextual features.

4 Methodology

To accomplish all subtasks, our proposal is grounded on the combination of different feature sets by means of ensembles. Particularly, we focus on two types of features sets. On the one hand, we employ linguistic and interpretable features, compiled by UMUTextStats (LF) [9, 10] and negation features (NE) [11–14]. On the other hand, we study different types of embeddings from word embeddings (WE) and sentence embeddings compiled from fastText (SE), to contextualised word embeddings from Spanish BERT (BE), also known as BETO, and compiled by fine-tuning the Spanish version of BERT (BF) [4], extracting the embeddings from the [CLS] token and applying a mean pooling, as suggested in [23]. Moreover, for subtasks 2 and 4 we compile contextual features (CF) from the datasets provided.

Regarding the linguistic features, the UMUTextStats tool is capable to categorise a total of 365 features regarding semantics, pragmatics, lexical process, social media, figurative language, or correction and style, among others. Concerning the negation features we extract the list of negation cues appearing in each text (simple cues (e.g., "no"/ *not*), continuous cues (e.g. "en mi vida"/ *in my life*) and discontinuous cues (e.g. "ni...ni"/ *nor...nor*) and compute their total. With regard to the contextual features (CF), as commented during corpus analysis (see Section 3), the datasets were very different regarding complexity and number of features. For the generic Spanish dataset we encode the features regarding media and gender with one-hot encoding. In case of Mexican Spanish, we keep the features provided by the organisers of the task but we include extra features from the date of the posts, to know whether a tweet was posted during weekend or working day, and we divide the day into several time slots to distinguish the tweets written in the morning, afternoon, evening and night.

For all feature sets we carry out a process of preprocessing and normalisation of the features. First, for the linguistic features we scale each feature independently in a range [0, 1] with a MinMax scaler. For the negation features, however, we apply a Robust Scaler as we found heavy outliers regarding one of the tweets composed by repeating a negation multiple times. The same technique (Robust Scaler) is applied to the contextual features, regardless the subtask. Next, we apply a feature selection technique based on Mutual Information. According to our evaluation with the development set, we keep this feature selection over LF, SE and BE but keeping all the features for BF, WE, NE and CF as we achieved better results without feature selection for those feature sets.

To combine the features we evaluate different forms of ensembles of the best model per feature set. Specifically, we evaluate three types of ensembles: (1) based on majority voting (mode), (2) based on a modified version of the majority voting ensemble, weighting each vote with the weighted F1-score achieved on the validation set, and (3) training a logistic regression model from the predictions of each neural network model.

Prior to our participation we evaluated different neural networks models and traditional machine learning models in order to get some insights for the reliability of each feature set used in combination or separately, as well as which hyper-parameters of the networks worked well for each subtask. We, therefore, performed an hyperparameter optimisation stage evaluating 110 neural network per feature set in isolation and in combination, both for generic Spanish (subtasks 1 and 2) and for Mexican Spanish (subtasks 3 and 4). We ranked each model based on the micro-averaged F1-score (we used this measure because in CodaLab the organisers indicated that submission would be evaluated with it, but finally for subtasks 3 and 4 they used macro-averaged F1-score). The feature sets evaluated during this stage were LF, NE, SE, BE, BF, and CF. During this stage we also evaluated convolutional and recurrent neural networks with WE from Spanish pre-trained word embeddings from fastText, gloVe, and word2vec.

For each neural network tested during the hyperparameter optimisation we evaluated different depths and different number of neurons (8, 16, 48, 64, 128, 256). The layers and the neurons were organised in shapes, including *funnel*, *rhombus*, *long_funnel*, *brick*, *diamond*, and *triangle*. It is worth noting that the best results were achieved by simple models in the majority of cases, with one or two layers and few neurons per layer. The majority of architectures evaluated were Multilayer perceptrons because we rely on sentence fixed embeddings and linguistic features, thus, spatial and temporal data cannot be exploded. However, in case of WE, we also evaluated Convolutional Neural Networks (CNN), Bidirectional Long Short Term Memory (BiLSTM) and Bidirectional Gated Recurrent Unit (BiGRU). In addition, we tried different dropout rates to avoid overfitting (0, 0.1, 0.2, and 0.3) and several activation functions including *relu*, *sigmoid*, *tanh*, *selu*, and *elu*. We also included an early stopping mechanism and a learning rate scheduler.

Table 3 depicts the best parameters for each feature set for the generic Spanish and Mexican Spanish datasets. We can observe that the best results were achieved with multi-layer perceptrons, even for the WE, and with shallow neural networks with one or two hidden layers. Deep neural networks achieved the best result for NE in both datasets, with 5 and 7 hidden layers respectively, and WE, both with 6 hidden layers. CF required a very simpler network for generic Spanish but a complex one for Mexican Spanish. This fact can be explained due to the simplicity of the CF for generic Spanish, including only the gender and media type of the documents. It draws attention, when comparing generic and Mexican Spanish, that the models for generic Spanish are generally simpler than those for Mexican Spanish, despite having more instances and being multiclass.

5 Results

This section is divided into the two main subtasks regarding offensive content detection in generic Spanish (see Subsection 5.1) and Mexican Spanish (see Subsection 5.2).

5.1 Subtasks 1 and 2. Non-contextual and contextual multiclass classification for generic Spanish

Subtask 1 and subtask 2 were evaluated by the following metrics: micro-averaged, macro-averaged and weighted-averaged versions of precision (P), recall (R) and F1-score (F1), and Mean Squared Error (MSE). The measure selected by the organisers for ranking the systems was the micro-averaged F1-score.

Table 3. Results of the hyperparameter evaluation stage, including the network architecture (network), the composition of the network (shape, number of hidden layers, and number of neurons), the dropout ratio (dropout), the learning rate (LR), the batch-size (BS), and the activation function (AF)

Feature set	network	shape	layers	neurons	dropout	LR	BS	AF
Generic Spanish (Subtasks 1 and 2)								
LF	MLP	brick	1	256	None	0.001	128	relu
SE	MLP	brick	1	16	0.1	0.1	256	linear
BE	MLP	brick	2	128	None	0.001	512	tanh
NE	MLP	funnel	5	512	None	0.001	512	elu
BF	MLP	brick	2	128	0.3	0.01	256	tanh
CF	MLP	brick	1	4	None	0.01	128	linear
WE	MLP	funnel	6	1024	0.1	0.001	256	elu
Mexican Spanish (Subtasks 3 and 4)								
LF	MLP	brick	2	1024	None	0.001	512	relu
SE	MLP	brick	2	512	0.3	0.1	128	tanh
BE	MLP	brick	1	1024	0.1	0.01	256	linear
NE	MLP	funnel	7	16	0.1	0.001	512	selu
BF	MLP	brick	1	3841	0.1	0.1	256	tanh
CF	MLP	funnel	8	86	0.2	0.01	128	tanh
WE	MLP	brick	6	64	0.2	0.001	512	sigmoid

Each team could participate with up to three submissions and select the best of them as official result. Table 4 contains the results of each of our runs for **subtask 1**, non-contextual multiclass classification for generic Spanish. The first run consists of an ensemble of neural networks models trained with BE, LF, NE, SE, and BF. The ensemble was built using a logistic regression from the individual probabilities of each model. The second run, which is our official result, consists of the same type of ensemble but only with BF, SE, and BE; that is, by removing the linguistic features. We can observe than in this case, we achieved slightly better micro-averaged F1-score but worse macro-averaged precision and F1-score, which suggest that the usage of linguistic features are beneficial for the classes NOM, OFG, and OFP. The third run consists of another type of ensemble, based on the weighted mode of the individual predictions of each model. The weights for this model were calculated based on the weighted F1-score achieved with our custom validation set. We can observe than this run achieved worst results than the ensembles based on logistic regression except for MSE.

As it has been previously mentioned, we selected our second run to participate in **subtask 1**.The official leader board is depicted in Table 5. We reached the second best result with a micro-averaged F1-score of 87.8289%. with a difference of 0.3307% with the best result, achieved by *saroyehun* (88.1596% of micro-averaged F1-score), and followed by *xjywing* (87.3291% of micro-averaged F1-score) with a difference of 0.4998%. Due to the high number of non-offensive instances, the results of the micro-averaged F1-score among all participants is simi-

Table 4. Benchmark of our three runs for subtask 1

		Macro			Weighted			
Run	Micro-F1	P	R	F1	P	R	F1	MSE
run1	87.800	**78.779**	69.145	**73.037**	87.438	87.800	87.107	0.041
run2	**87.829**	78.605	**69.186**	73.006	**87.473**	**87.829**	**87.137**	0.041
run3	84.029	74.308	60.287	64.680	84.033	84.029	82.947	**0.031**

lar. By looking the macro-averaged F1-score, we can observe that *Marta_NG_BD* and *Timen* achieved lower results, which can indicate that some minority labels were not classified. This subtask also includes a regression metric based on the probabilities assigned for each class, that is measured using the Mean Squared Error (MSE). As it can be observed, we got an MSE of 0.41134 and the results achieved by the rest of the participants (lower is better) matches with the official rank.

Table 5. Official results of the subtask 1. Non-contextual multiclass classification for generic Spanish

			Macro			Weighted			
#	Team	Micro-F1	P	R	F1	P	R	F1	MSE
1	saroyehun	**88.160**	76.786	**70.930**	**73.238**	**88.036**	**88.160**	**88.038**	**0.023**
2	**UMUTeam**	87.829	**78.605**	69.186	73.006	87.473	87.829	87.137	0.041
3	xjywing	87.329	75.648	70.019	72.386	86.924	87.329	86.767	0.067
4	Marta_NG_BD	84.169	57.819	54.514	55.950	82.161	84.169	82.995	0.070
5	Timen	81.685	61.167	46.638	50.366	80.693	81.685	78.456	0.393

Regarding **substask 2**, contextual multiclass classification for generic Spanish, we employed the same techniques than for subtask 1. Table 6 contains the results of our three runs. The results for the first run, which consisted of the ensemble based on logistic regression over BE, LF, NE, SE, and BF but including contextual features (CF) in the ensemble, improves the micro-averaged F1-score achieved in subtask 1 in 0.0294%. This behaviour was expected as the contextual features of Spanish were few. However, the results improve the macro-averaged metrics, which indicates that the contextual features contribute to the texts with offensive or vulgar content. For the second run, however, we evaluate a different approach from subtask 1. We retrain the model including the validation set but the results were not good. Specially, we can observe a significant drop in the macro-averaged recall. Finally, for the third run we adopted a similar approach for the second run of the first subtask (see Table 4) removing the linguistic features but keeping the ensemble model based on logistic regression from the training dataset. As it can be observed, this run achieves worst micro-averaged F1-score and macro-averaged precision, recall and F1-score. In this case, the first run is the one we selected for the official leader board of **subtask 2**. As it can be observed from Table 7, only two participants sent runs and we achieved the first position in this subtask with a micro-average F1-score of 87.8289%.

Table 6. Benchmark of our three runs for subtask 2

		Macro			Weighted			
Run	Micro-F1	P	R	F1	P	R	F1	MSE
run1	**87.829**	**78.791**	**69.210**	**73.084**	**87.466**	**87.829**	**87.142**	0.041
run2	86.146	77.445	60.113	65.950	85.505	86.146	84.998	**0.032**
run3	87.800	78.484	69.052	72.875	87.436	87.800	87.108	0.041

Table 7. Official results of the subtask 2. Contextual multiclass classification for generic Spanish

			Macro			Weighted			
#	Team	Micro-F1	P	R	F1	P	R	F1	MSE
1	**UMUTeam**	**87.8289**	**78.791**	**69.210**	**73.084**	**87.466**	**87.829**	**87.142**	**0.041**
2	Timen	73.262	53.067	29.271	28.822	71.476	73.262	64.382	0.422

Figure 1 contains the ten top-ranked linguistic features according to mutual information. We can observe that general and negative psycho-linguistic processes are strong features to discern among non-offensive documents with the rest of the classes. However, they are not indicators to discern among offensive (OFP, OFG) or vulgar language (NOM). We also found informal speech and swear language are the strongest correlation with the label NOM. Offensive speech is also a strong marker related to when the offensive message is towards a person (OFP) and when the usage of the language is vulgar or not adequate (NOM). As it can be observed, the number of words have greater influence on NO and OFG classes. We can find a correlation of this feature with the number of syllables, the overall length of the documents and readability scores.

5.2 Subtasks 3 and 4. Non-contextual and contextual binary classification for Mexican Spanish

The subtasks of contextual and non-contextual binary classification for Mexican-Spanish were ranked by macro-averaged F1-score of the *offensive* class. Similar to subtasks 1 and 2, the number of participants was lower in the contextual subtask, with only three participants in subtask 4.

Table 8 depicts the results achieved individually for each of our runs in **subtask 3**, non-contextual binary classification for Mexican Spanish. Similar to subtask 1 and subtask 2, the first run consisted of an ensemble based on logistic regression from the probabilities of the neural networks models trained with

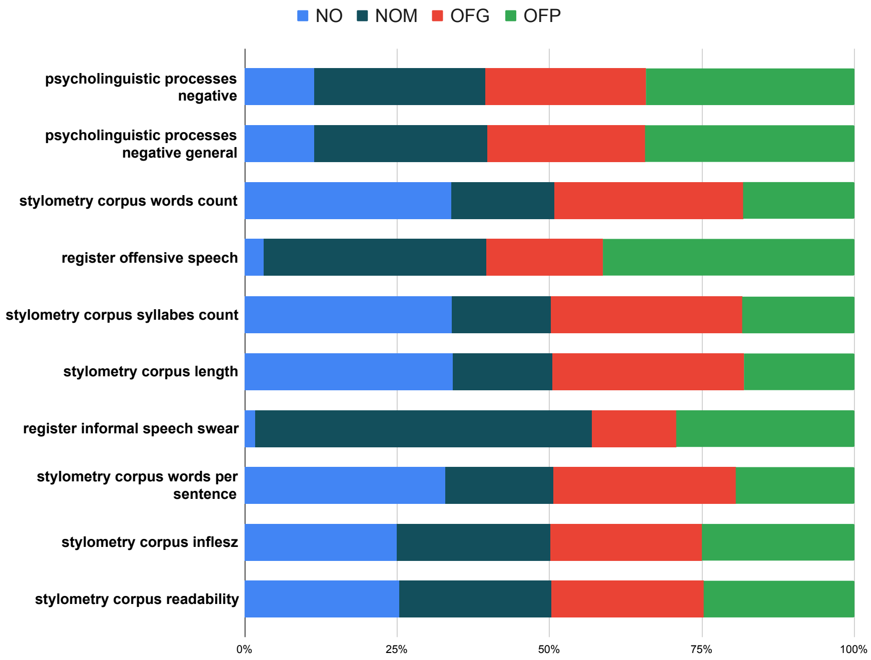

Fig. 1. Mutual Information of the ten top-ranked linguistic features averaged by label for the generic Spanish dataset

only one feature set (BE, LF, NE, SE, and BF). We achieved a macro-averaged F1-score of 66.7791%. Our second run is the one that is on the leader board and it consisted of removing the linguistic features (LF and NE). We can observe than in this case, better results are achieved for macro-averaged metrics. One explanation for this fact is that the linguistic and negation features could provide contradictory results for Mexican Spanish as they were designed for generic Spanish. Third run also consisted of an ensemble based on regression but using BE, LF, NE, and BF. We achieved better results than in the first run by excluding the SE features.

Table 8. Benchmark of our three runs for subtask 3

Run	Macro-P	Macro-R	Macro-F1
run1	66.000	67.577	66.779
run2	66.500	**67.627**	**67.059**
run3	**66.833**	67.057	66.833

For **subtask 3** we selected our second run as official result. Table 9 depicts the results of the leader board for this subtask. We reached position 5 in the rank with a macro-averaged F1-score of the offensive class of 67.0588%. The best result was achieved by *vic_gomez* with a macro-averaged F1-score of 70.2619%. Compared with the best results, we achieved a similar macro-averaged recall score, even greater than the two best overall submits, but their macro-averaged precision was higher (76% and 75% vs. our 66.50%). Other teams achieved even greater macro-averaged precision (xjywing with 88.8333%, aomar with 87.5000%, and 91.8333%) with smaller macro-averaged recall.

Table 9. Official results of the subtask 3. Non-contextual binary classification for Mexican Spanish

#	User	Macro-P	Macro-R	Macro-F1
1	vic_gomez	76.000	65.330	**70.262**
2	saroyehun	75.500	64.074	69.319
3	JuanCalderon	67.333	**69.655**	68.475
4	cimatgto	66.333	69.580	67.918
5	**UMUTeam**	66.500	67.627	67.059
6	Timen	60.000	60.811	60.403
7	DanHv94	53.500	68.737	60.169
8	xjywing	88.833	34.167	49.352
9	aomar	87.500	32.387	47.276
10	Sreelakshmi	**91.833**	31.432	46.834

Regarding **subtask 4**, contextual binary classification for Mexican Spanish, we sent three runs whose results are shown in Table 10. The first one, consisted of an ensemble model based on logistic regression of neural networks models for the following feature set: LF, NE, BE, BF, SE, and CF. In the second run, we sent the same model but we adjusted the CF features by using Robust Scaler, as we found heavy outliers in the data. However, as we can observe, the results were the same that we achieved with our first run. Finally, our third run consisted of an ensemble with the following feature set: LF, NE, BF, and CF. Compared with the first run, we removed SE. This run was submitted because we achieved good results with our custom validation set.

Table 10. Benchmark of our three runs for subtask 4

Run	Macro-P	Macro-R	Macro-F1
run1	66.000	**67.577**	66.779
run2	66.000	**67.577**	66.779
run3	**66.833**	67.057	**66.945**

Table 11. Official results of the subtask 4. Contextual binary classification for Mexican Spanish

#	User	Macro-P	Macro-R	Macro-F1
1	**UMUTeam**	**66.833**	67.057	**66.945**
2	DanHv94	53.833	**68.432**	60.261
3	Timen	42.333	44.561	43.419

In **subtask 4**, as previously stated, we submit the third run to the official leader board, that achieve slightly better macro-averaged precision and slightly worse macro-averaged recall than the other two runs. Table 11 depicts the official ranking for subtask 4. Only three teams sent runs for this subtask. We achieved the first position, with a macro-averaged F1-score for the offensive class of 66.9449%, improving slightly our results from the previous subtask. Compared with the second best result, we achieved slightly worse macro-averaged recall but higher macro-averaged precision.

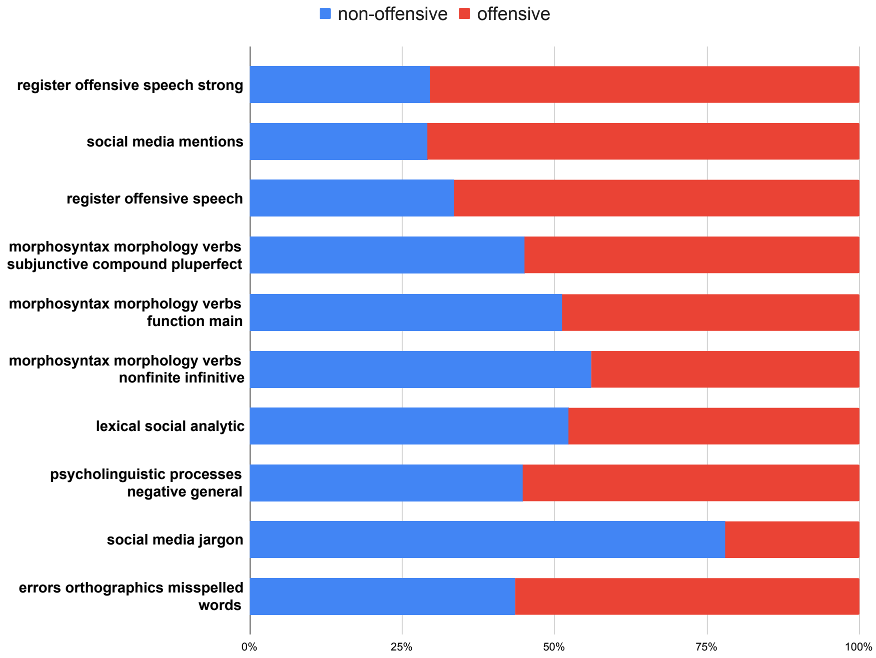

Fig. 2. Mutual Information of the ten top-ranked linguistic features averaged by label for the Mexican-Spanish dataset

Figure 2 shows the ten top-ranked linguistic features for the Spanish Mexican dataset averaged by label. As it was expected, linguistic features related to offensive speech are highly related to the offensive class but it also appears significantly on tweets labelled as non-offensive, which suggests that there are texts considered as non-offensive that contain offensive words. Regarding social media, the usage of mentions is most common on offensive tweets, which suggests that offensive speech is more common to individuals than groups. However, social media jargon that includes words such as *retweets*, *posts*, or *direct messages* are more common on non-offensive tweets. The rest of the features only have slightly variations regarding the label, as happens on the usage of verbs on subjunctive compound pluperfect or the number of misspellings that are slightly more common on offensive tweets. On contrasts, verbs in infinitive, or analytic thinking are more common on non-offensive tweets.

6 Conclusions

In this working notes we describe the participation of the UMUTeam in the task MeOffendES regarding offensive language detection in different variants of Spanish. Our proposal for solving the different subtasks have been grounded on the combination of linguistic and negation features with state-of-the-art transformers, combined as ensembles of neural networks classifiers or combined in the same neural network. Our results have achieved very good results in the official leader board. We reached the first position on subtasks 2 and 4 regarding offensive detection with contextual features on generic and Mexican Spanish respectively, and position 2 and 5 on the non-contextual subtasks. It is worth mentioning, however, that the number of participants who sent runs for contextual tasks was lower than for non-contextual features tasks.

The following insights were obtained during the participation in this task. First, the results achieved indicate that linguistic features and transformers mutually benefit from each other, increasing their reliability. Our results indicate that ensembles learned from training a logistic regression machine-learning classifier from the individual probabilities of each model achieve better results than ensembles based on the mode (the label most voted) or the weighted mode. Second, we observe than sentence-fixed embeddings from the fine-tuned model of BETO, which we called BF, outperform plain BE vectors in all cases, and it is more easier to combine them with other feature sets. Third, during the hyperparameter evaluation stage we observed that simpler neural network models with only a few layers and few neurons behave better than models with more than four hidden layers. Forth, we found that those linguistic features concerning negative processes, such as anger, sadness, anxiety, are discriminatory features regarding offensive language detection. However, our results suggest that these features are reliable to distinguish between non-offensive documents from the ones that include offensive or vulgar language but not to differentiate among different types of offensive language.

Lastly, we are pleased with the opportunity we have been give to participate in this task. This work has been a collaboration between Universidad de Murcia and Universidad de Jaén.

Acknowledgments

This work was supported by the Spanish National Research Agency (AEI) through project LaTe4PSP (PID2019-107652RB-I00/AEI/10.13039/501100011033). In addition, José Antonio García-Díaz has been supported by Banco Santander and University of Murcia through the industrial doctorate programme. Salud María Jiménez-Zafra has been partially supported by a grant from Fondo Social Europeo, Administration of the Junta de Andalucía (DOC_01073), Fondo Europeo de Desarrollo Regional (FEDER) and LIVING-LANG project (RTI2018-094653-B-C21) from the Spanish Government.

References

1. Aragón, M.E., Carmona, M.A.A., Montes-y Gómez, M., Escalante, H.J., Pineda, L.V., Moctezuma, D.: Overview of mex-a3t at iberlef 2019: Authorship and aggressiveness analysis in mexican spanish tweets. In: IberLEF@ SEPLN. pp. 478–494 (2019)
2. Balakrishnan, V., Khan, S., Arabnia, H.R.: Improving cyberbullying detection using twitter users' psychological features and machine learning. Computers & Security **90**, 101710 (2020). https://doi.org/https://doi.org/10.1016/j.cose.2019.101710
3. Basile, V., Bosco, C., Fersini, E., Debora, N., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M., et al.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics (2019)
4. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish pre-trained bert model and evaluation data. PML4DC at ICLR **2020** (2020)
5. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 11 (2017)
6. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 5 (2011)
7. Fersini, E., Nozza, D., Rosso, P.: Overview of the evalita 2018 task on automatic misogyny identification (ami). EVALITA Evaluation of NLP and Speech Tools for Italian **12**, 59 (2018)
8. Fersini, E., Rosso, P., Anzovino, M.: Overview of the task on automatic misogyny identification at ibereval 2018. IberEval@ SEPLN **2150**, 214–228 (2018)
9. García-Díaz, J.A., Cánovas-García, M., Valencia-García, R.: Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in latin america. Future Generation Computer Systems **112**, 614–657 (2020). https://doi.org/10.1016/j.future.2020.06.019

10. García-Díaz, J.A., Cánovas-García, M., Colomo-Palacios, R., Valencia-García, R.: Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. Future Generation Computer Systems **114**, 506 – 518 (2021). https://doi.org/10.1016/j.future.2020.08.032, http://www.sciencedirect.com/science/article/pii/S0167739X20301928
11. Jiménez-Zafra, S.M.: Negation processing in spanish and its application to sentiment analysis. Procesamiento del Lenguaje Natural **66**, 193–196 (2021)
12. Jiménez-Zafra, S.M., Cruz-Díaz, N.P., Taboada, M., Martín-Valdivia, M.T.: Negation detection for sentiment analysis: A case study in spanish. Natural Language Engineering **27**(2), 225–248 (2021)
13. Jiménez-Zafra, S.M., Morante, R., Blanco, E., Valdivia, M.T.M., Lopez, L.A.U.: Detecting negation cues and scopes in spanish. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 6902–6911 (2020)
14. Jiménez-Zafra, S.M., Taulé, M., Martín-Valdivia, M.T., Urena-López, L.A., Martí, M.A.: Sfu review sp-neg: a spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. Language Resources and Evaluation **52**(2), 533–569 (2018)
15. Kumar, R., Bhanodai, G., Pamula, R., Chennuru, M.R.: Trac-1 shared task on aggression identification: Iit (ism)@ coling'18. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 58–65 (2018)
16. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Benchmarking aggression identification in social media. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018). pp. 1–11 (2018)
17. Kumar, R., Ojha, A.K., Malmasi, S., Zampieri, M.: Evaluating aggression identification in social media. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. pp. 1–5. European Language Resources Association (ELRA), Marseille, France (May 2020), https://www.aclweb.org/anthology/2020.trac-1.1
18. Malmasi, S., Zampieri, M.: Detecting hate speech in social media. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017. pp. 467–472 (2017)
19. Mandl, T., Modha, S., Kumar M, A., Chakravarthi, B.R.: 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. pp. 29–32 (2020)
20. Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., Patel, A.: 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. pp. 14–17 (2019)
21. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Ángel Álvarez Carmona, M., Álvarez Mellado, E., de Albornoz, J.C., Chiruzzo, L., Freitas, L., Adorno, H.G., Gutiérrez, Y., Jiménez-Zafra, S.M., Lima, S., de Arco, F.M.P., (eds.), M.T.: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021). (CEUR Workshop Proceedings) (2021)
22. Plaza-del-Arco, F.M., Casavantes, M., Escalante, H., Martín-Valdivia, M.T., Montejo-Ráez, A., Montes-y-Gómez, M., Jarquín-Vásquez, H., Villaseñor-Pineda, L.: Overview of the MeOffendEs task on offensive text detection at IberLEF 2021. Procesamiento del Lenguaje Natural **67**(0) (2021)
23. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Nat-

ural Language Processing. Association for Computational Linguistics (11 2019), http://arxiv.org/abs/1908.10084

24. Struß, J.M., Siegel, M., Ruppenhofer, J., Wiegand, M., Klenner, M., et al.: Overview of germeval task 2, 2019 shared task on the identification of offensive language (2019)
25. Wiegand, M., Siegel, M., Ruppenhofer, J.: Overview of the germeval 2018 shared task on the identification of offensive language. In: 14th Conference on Natural Language Processing KONVENS 2018 (2018)
26. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 75–86 (2019)
27. Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç.: Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 1425–1447 (2020)