ISSN 2007-9737

Verbal Aggressions Detection in Mexican Tweets

Daniel Abraham Huerta-Velasco, Hiram Calvo

Instituto Politécnico Nacional,
Centro de Investigación en Computación,
Mexico

dhuertav2019@cic.ipn.mx, hcalvo@cic.ipn.mx

Abstract. Verbal aggressions are a struggle that a great number of social media users have to face daily. Some users take advantage of the anonymity that social media give them and offend a person, a group of people, or a concept. The majority of proposals which pretend to detect aggressive comments on social media handle it as a classification problem. Although there are a lot of techniques to face this problem in English, there is a lack of proposals in Spanish. In this work, we propose using several Spanish lexicons which have a collection of words that have been weighted according to different criteria like affective, dimensional, and emotional values. In addition to them, structural values, word embeddings and one-hot codification were taken into account.

Keywords. Spanish lexical resources, sentiment analysis, Mexican Spanish tweets, text classification.

1 Introduction

There is no doubt that social media have changed the life of the whole mankind due to the big amount of information that can be shared in them. Social media has a bunch of positive aspects such as the speed in which information can be shared and read in comparison to traditional form of media. Nevertheless, social media have a dark side that can be identified in three main categories: first, social media foster a false sense of online "connections" and superficial friendships leading to emotional and psychological problems. The second harm of social media is that it can become easily addictive taking away family and personal time as well as diminish interpersonal skills, leading to antisocial behavior.

Lastly, social media have become a tool for criminals, predators and terrorists enabling them to commit illegal acts or to offend/harass a person or even a group [1]. Fortunately, some social media have started to work against hateful speech and they have already set policies which determine what kind of features a comment has to have in order to be considered as aggressive and what the consequences would be.

Agressiveness has been a topic studied by various disciplines. Computational linguistics has studied it as a binary classification problem and good results are being obtained by using some machine learning techniques which include classic classifiers (Support Vector Machines, Logistic Regression, Random Forests) and neural networks. Some organizations focus their investigation on this topic and organize competitions where, mainly, ask for new proposals that can classify as good as possible whether a tweet is aggressive or not, among other labels, such as if a tweet is vulgar but not offensive, no vulgar and offensive, if the agression of the tweet is targeted to a person or a group of people, etc.

In English, for instance, in the International Workshop on Semantic Evaluation (SemEval) was organized a task (OffensEval) where three substasks were featured: In sub-task A, the goal was to discriminate between offensive and non-offensive posts. In sub-task B, the focus was on the type of offensive content in the post. Finally, in sub-task C, systems had to detect the target of the offensive posts. This task was part of SemEval in 2019 [25] and 2020 [26] wh¡ere 115 and 87 teams participated, respectively.

In Spanish, in the Iberian Languages Evaluation Forum (IberLEF) a task similar to the one launched at SemEval was presented: its name is **MEX-A3T** and its objective is to encourage research on the analysis of social media content in Mexican Spanish and it only focuses on aggressiveness detection in tweets (like sub-task A). This task was part of IberLEF in 2019 [2] and 2020 [3] where 9 teams participated in each edition of the competition.

Despite there is both incentive and research to detect hate speech on social media, the truth is that research is mainly focused in English rather than any other language. As we can see in the two last years, there were more proposals in English tasks than in Spanish ones. The majority of proposals in Spanish used architectures and data representation based on deep learning models. None of them took importance about what emojis and hashtags can say about the polarity of a post on social media, besides the polarity that a set of contiguous words can have by their own or even together.

This paper will explain the proposed features which compose the model inspired by Spanish lexicons among other Spanish resources. The dataset used for experimenting with this proposed features is the one used in the edition 2020 of the competition MEX-A3T.

2 Related Work

There were 9 different teams which worked on the edition 2020 of MEX-A3T dataset and they were ranked by F_1 score over aggressive class. The system presented in [10] performed the best score with both of their strategies. The first strategy consisted of an ensemble of different BETO models (BERT models trained in Spanish) with majority and weighted voting schemes.

With majority voting scheme the model predicts the most voted class among the classifiers of the ensemble, In case of tie, a random prediction is performed among the classes. The weighted version consists in aggregate the confidence prediction for classes in each model of the ensemble to build a final weighted vote.

This confidence prediction is the output of the last Softmax layer. The second strategy consists in 3 data augmentation techniques: to use **Easy Data Augmentation (EDA)** [22], **Unsupervised Data Augmentation (UDA)** [23], and **Adversarial Data Augmentation (ADA)** [11]. The EDA technique consists in generating new instances by modifying 20% of the original data applying four basic operations in randomly selected words: *Replace* (word is changed by a random synonym), *Insert* (randomly choose a synonym of the chosen word and insert it randomly), *Swap* (select two words and swap their positions), and *Delete* (remove the word of the tweet). UDA implies the use of semisupervised learning, by augmenting each sentence of the original training set and using the Kullback-Leibler divergence to penalize the difference in the distributions of the logits. And finally, ADA consists in using an adversarial method at each epoch of the training to create a new input for the misclassified sentences. Using 20 ensemble models (first strategy) this proposal obtained a F_1 result on the aggressive class of **0.7998**, and on the non-agressive class of **0.9195**. Using the same 20 ensemble models and EDA as the data augmentation strategy, the team obtained **0.7971** and **0.9205** of F_1 score, respectively.

The best result obtained by a team whose model was not based on any Transformer pre-trained model was [24]. Their proposal was divided in 3 phases: data preprocessing, word weightening, and classification. Data preprocessing consisted in converting to lowercase the text of the *tweet*, performing text tokenization, removing stopwords and words which frequency usage was lower than 5 times in the whole dataset, and stemming common words. In the second phase, the *TF-IDF* technique was used. Finally, these features were the input of a SVM. Their obtained results were **0.6619** and **0.8752**, respectively.

3 Model's Description

Some features of this model were inspired in the ones presented in [9]. This research proposed a novel model for handling irony detection in English tweets that explores the use of affective features based on a wide range of lexical resources

available for English, reflecting different facets of affect.

3.1 Data Preprocessing

First of all, a data pre-processing step is performed. In this, four operations are applied:

— *Mentions cleaning*: In the social media slang, a mention means that an user is tagged in a post. In this operation, all mentions are removed in the post but the frequency of them is saved because it will be considered as a feature.

— *Hashtag treatment*: Hashtag is a term associated with topics of discussions that users choose to be indexed in social networks, inserting the hash symbol (#) before the word, phrase or expression with no whitespaces, allowing only the underscore symbol (_) to "separate" the words if wanted. In this pre processing, word segmentation is used in order to have the words as if the user had not used a hashtag. The corpus used by word segment model to learn how to split Spanish words was Spanish Billion Words Corpus [5]. The frequency of hashtags is used as a feature.

— *Emojis cleaning*: All emotional polarity values of emojis which are present in the post are averaged according to values in [12]. It should be said that not all emojis[1] are present in the work of Kralj and her team. That is why four features are extracted: the averaged polarity of the post, number of total emojis which are in the post, and the number of emojis which are both in the work of Kralj and not. Finally, all emojis are removed from the post.

— *URLs cleaning*: URLs are counted and then removed from the post.

[1] https://unicode.org/Public/UNIDATA/emoji/emoji-data.txt

Table 1. Structural features

Features	Description
exclam_marks quest_marks	The frequency of each in a tweet
singulars plurals	The frequency of each inflectional feature
words chars	The total amount of words and characters in a tweet
upper	The total amount of upper-case characters in a tweet
verbs adv adj nouns	The frequency of each POS-tag in a tweet
hashtags mentions urls	The frequency of each specific marker in a tweet
emojis polar_emojis non_polar_emojis	The frequency of emojis in a tweet and a counter of emojis that appear in [12] or not, respectively

3.2 Features' Extraction

Features in this model are extracted using some Spanish lexical resources, also known as lexicons. They are divided in the following categories: Structural features, affective features, dimensional features, and emotional features.

3.2.1 Structural Features

They consist in the quantification of features that can be obtained based on Part-Of-Speech classification. Table 1 shows the features which fall under this description.

3.2.2 Affective Features

They consist in both positive and negative polarity values that a post has according to the sum of the words' polarity present in it. To do that, lexicons with *positive* and *negative* values (categorical

or numerical) were taken in count. The used lexicons were Hate Speech Spanish Lexicons [16], NRC Word-Emotion Association Lexicon (a.k.a. EMOLEX) [13], iSOL [14], Mexican Slang Lexicon [6] (we added 1,373 words and phrases from our own knowledge to this lexicon), ML_Sentincon [7], Multilingual Sentiment Lexicon[2], Sentiment Lexicons in Spanish [15], ElhPolar Lexicon [21], and SenticNet [4].

From each lexicon, except the Hate Speech Spanish Lexicon, two features were extracted: the positive and negative sum of the words in the tweet, separately. For Hate Speech Spanish Lexicon four features were extracted: the frequency of words in the tweet that were labeled in the lexicon as insults, xenophobic, misogynistics, and words that refer to the nationality of an immigrant.

It should be noted that, due to some lexicons (EMOLEX, Mexican Slang Lexicon, ML_Senticon, and Elhpolar) have not only words but also phrases formed by one, two, three and four words, 6 more features were extracted: positive and negative values per *n-gram*.

Besides these features, three more are part of this group. They are the positive and negative sums of polarity value of the *emojis* (according to the values in [12]), and the difference between them.

3.2.3 Dimensional Features

They consist in those which are inspired in some theories which propose that the nature of an emotional state is determined by a word's position in a space of independent dimensions. According to a dimensional approach, emotions can be defined as a coincidence of values on a number of different strategic dimensions. The used lexicons, i.e. those which were labeled according to the values of some of these theories, were: SENTICNET [4], Spanish ANEW (S-ANEW) [18], and Spanish DAL (S-DAL) [19]. Table 2 shows the dimensions' name used in each of these three lexicons.

[2] https://sites.google.com/site/datascienceslab/projects/multilingualsentiment

Table 2. Features extracted per each Dimensional lexicon in our model

SenticNet	S-ANEW	S-DAL
aptitude	valence	pleasantness
attention	arousal	activation
pleasantness	dominance	imagery
sensitivity		

3.2.4 Emotional Features

They consist in those which are inspired in the work of [17, 8] who defined 8 and 6 basic emotions, respectively: anger, disgust, fear, joy, sadness, surprise, anticipation, and trust. The Spanish lexicons that are labeled according these emotions, and therefore those that were used, were EMOLEX [13] and Spanish Emotion Lexicon (SEL) [20]. EMOLEX has not only words but also phrases formed by 4 words as maximum, so the very same features per emotion were extracted per *n*-gram.

In general our model is composed by 114 mainly features (in the future refered as CVAD features): 17 structural features, 49 affective, 10 dimensional, and 38 emotional. In addition to them, 300 word embeddings and features extracted by the Boolean weighting technique (word's presence representation in a vector where elements are turned from 0 to 1 if the word is present in the text) are added. The way in which these word embeddings were trained is described in [5]. For Boolean weighting features, the number of these *n*-features depens on how many words are used at least *m*-times in the whole training dataset. The results of the experimentation to know the value of *m* that obtained the best result are explained in the next section.

4 Experimentation and Results

4.1 Dataset Description

MEX-A3T 2020 corpus is a set of *tweets* written by Mexican users. The way in which they were extracted was using Twitter's API and a set of terms that were used as seeds. These terms were words identified by the *Diccionario de Mexicanismos de*

Table 3. Distribution of the classes in the corpus used in aggresiveness detection task at MEX-A3T 2020

Class	Training set	Test set
Non-aggressive	5,222	2,238
Aggressive	2,110	905
Total	7,332	3,143

Table 4. Length of each vector representation for each tested value of *m*

Value of *m*	Number of words
1	14,917
2	5,472
3	3,394

la Academia Mexicana de la Lengua as vulgar words and Mexican slang words. Table 3 shows the distribution of the recollected *tweets* into the corpus of the competition.

An important note in this corpus is that the *tweets*' labels in the Test set are not released. Despite this situation, the corpus' owners have shown a willingness to receive predictions of the *tweet*'s labels from anyone and respond with the corresponding prediction's results.

4.2 Experimentation

A SVM classifier with linear kernel[3] was used in order to train a model to detect aggressiveness in the MEX-A3T 2020 corpus. SVM hyperparameters' tuning and Cross Validation over training dataset were performed in order to find out which configuration of both features and hyperparameters yielded the best theoretical results and then, predict the labels of test dataset. We used scikit-learn GridSearchCV[4] and cross_validate[5] methods to obtain these results taking F_1 score over aggressive class as the metric to be optimized. One thing to be noted is that the experiments are not only focused on CVAD features + 300 word embeddings + Boolean Weighting codification but also the combination of them, being 300 Word Embeddings + Boolean Weighting, together and by their own, the baselines for our proposal due to these representations have been furtherly studied before.

Cross validation was performed using *k*-Fold technique, where the value of *k* used in the experiments was 5. Table 5 shows the scores and the hyperparameters for the experiments where WE stands for 300 word embeddings, and *m*-BW stands for *m*-Boolean Weighting. The value of *m* indicates the minimum number of a word's occurrences in the whole training dataset to be taken in count in the vector representation. Table 4 shows the number of features extracted for each value of *m*.

4.3 Results

The best configuration of the SVM's model (the features and hyperparameters), i.e. the ones in grey in Table 5, were used to predict the labels in Test dataset. Table 6 shows the obtained scores per each configuration, and Table 7 shows the updated ranking of the results of agressiveness detection task at MEX-A3T 2020 including our best two results (refered as *CICDanHv-1*, and *CICDanHv-2*).

Despite the scores of using only CVAD features did not obtain competitive results (actually it is the worst ranked), the top three best configurations in experimentation phase included these features.

In order to know a bit about the usefulness of them in the aggressiveness detection in Mexican Spanish *tweets* task, a feature selection process was performed using scikit-learn feature_selection[6] method using as the predictor the same configuration of our tested linear SVM. Table 8 shows the percentages of usefulness per group of CVAD features and per configuration.

[3] We used the one described in `https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html#sklearn.svm.LinearSVC`

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`.

[5] `https://scikit-learn.org/stable/modules/cross_validation.html`.

[6] `https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html`.

Table 5. Ranking of experiments' scores in training dataset

Used features	Hyperparameters		F_1 agg	F_1 non-agg	F_1 macro	Accuracy
	C	penalty				
CVAD + 2-BW + WE	0.22	l1	**0.7047**	**0.8924**	**0.7985**	**0.8423**
CVAD + 3-BW + WE	0.22	l1	0.7042	**0.8924**	0.7983	0.8422
CVAD + 1-BW + WE	0.23	l1	0.7032	0.8912	0.7972	0.8408
baseline (1-BW + WE)	0.25	l1	0.7001	0.8914	0.7957	0.8406
CVAD + 1-BW	0.11	l2	0.6993	0.8920	0.7957	0.8411
CVAD + 3-BW	0.29	l1	0.6991	0.8929	0.7945	0.8417
baseline (3-BW + WE)	0.25	l1	0.6950	0.8901	0.7926	0.8385
baseline (2-BW + WE)	0.31	l1	0.6949	0.8891	0.7920	0.8374
CVAD + 2-BW	0.36	l1	0.6943	0.8911	0.7927	0.8395
baseline (1-BW)	0.11	l2	0.6884	0.8915	0.7900	0.8391
baseline (3-BW)	0.33	l1	0.6806	0.8897	0.7852	0.8361
baseline (2-BW)	0.32	l1	0.6803	0.8905	0.7854	0.8370
CVAD + WE	0.13	l1	0.6512	0.8777	0.7644	0.8189
baseline (WE)	0.15	l1	0.5905	0.8644	0.7275	0.7964
CVAD	1	l2	0.5234	0.8580	0.6907	0.7812

Table 6. Obtained scores of our model in test dataset

Features	F_1 agg	F_1 non-agg	F_1 macro	Accuracy
CVAD + 2-BW + WE	**0.6952**	0.8886	0.7919	0.8368
CVAD + 1-BW	0.6946	**0.8895**	**0.7921**	**0.8377**
1-BW + WE	0.6938	0.8890	0.7914	0.8371
1-BW	0.6840	0.8874	0.7857	0.8339
CVAD + WE	0.6406	0.8735	0.7571	0.8129
WE	0.5913	0.8624	0.7269	0.7941
CVAD	0.5152	0.8549	0.6850	0.7766

It should be said that each configuration includes CVAD and 300 Word Embeddings features.

5 Conclusion and Future Work

One of the objectives of this work was to propose a novel model to tackle aggressiveness on social media posts and test it in a Spanish Corpus of a social media. The corpus which was used in this paper is the one used in the edition 2020 of MEX-A3T competition. Our top two obtained results overcame the simplest baseline (BoW and SVM) developed for this task. The best result obtained by a team whose model was not based on a deep learning approach, and some other teams' proposals which used deep learning approaches. This suggests that the

Table 7. Updated ranking of results at MEX-A3T 2020 with our scores

Teams	F_1 agg	F_1 non-agg	F_1 macro	Accuracy
CIMAT-1	**0.7998**	0.9195	**0.8596**	0.8851
CIMAT-2	0.7971	**0.9205**	0.8588	**0.8858**
UPB-2	0.7969	0.9107	0.8538	0.8759
UACh-2	0.772	0.9042	0.8381	0.8651
baseline (INGEOTEC)	0.7468	0.8933	0.82	0.8498
Idiap-UAM-1	0.7255	0.8886	0.8071	0.8416
baseline (Bi-GRU)	0.7124	0.8841	0.7983	0.8348
Idiap-UAM-2	0.7066	0.8953	0.801	0.8451
UACh-1	0.7062	0.8861	0.7961	0.8358
DeepMath-1	0.7001	0.8544	0.7773	0.804
DeepMath-2	0.6957	0.8537	0.7747	0.8024
CICDanHv-1	0.6952	0.8886	0.7919	0.8368
CICDanHv-2	0.6946	0.8895	0.7921	0.8377
baseline (BoW-SVM)	0.676	0.878	0.777	0.8228
UMUTeam-2	0.6727	0.8706	0.7716	0.8145
Intensos-1	0.6619	0.8752	0.7686	0.8177
UMUTeam-3	0.6516	0.8771	0.7644	0.8183
UGalileo-2	0.6388	0.8208	0.7298	0.7604
UGalileo-1	0.6387	0.843	0.7408	0.7811
ITCG-SD	0.608	0.882	0.745	0.8186
UMUTeam-1	0.5892	0.843	0.7161	0.7728
UPB-1	0.3437	0.8463	0.595	0.7509
Intensos-2	0.2515	0.7664	0.509	0.644

usage of our features in a multilayer perceptron or their combination with a deep neural network architecture can outperform the obtained results and more conclusions about the usage of lexical resources for facing aggressiveness detection on social media can be done.

Analyzing the usefulness of our CVAD features in the configuration which gave us the best result, we can observe that more than the half of them were found useful. Individually, not every group of CVAD features were helpful to fulfill the main objective: Structural and dimensional features were the most useful ones (they overpassed the 90% of usefulness) meanwhile affective and emotional features' percentages did not obtain good results.

Trying to find out the reason of these phenomenon, we observed that a great number of

Table 8. Usefulness percentages per CVAD's groups and configuration

Features	Configuration (CVAD + WE +		
	1-BW)	2-BW)	3-BW)
Structural	94.12%	94.12%	94.12%
Affective	73.46%	73.46%	73.46%
Dimensional	90.00%	100.0%	100.0%
Emotional	28.94%	31.58%	28.94%
All CVAD	63.15%	64.91%	64.03%

phrases present in the affective and emotional lexicons are not frequently used by Mexicans. This indicates us that there is a necessity of more Spanish lexicons labeled in affective, dimensional, and emotional contexts with not only words but also phrases that includes those words and phrases which are actually used by Spanish speakers in the internet environment. As future work, we plan to publish our list of 1,373 words and phrases (which were added to Mexican Slang Lexicon) in these three contexts.

Acknowledgments

This work was supported in part by the Mexican Government through Instituto Politécnico Nacional (IPN) under SIP Multidisciplinary Project 2083, Project SIP 20210189, EDI, COFAA-SIBE, BEIFI-IPN; and CONACYT.

References

1. **Amedie, J. (2015).** The impact of social media on society. Pop Culture Intersections, Vol. 2.
2. **Aragón, M. E., Carmona, M. A. A., Montes-y Gómez, M., Escalante, H. J., Pineda, L. V., Moctezuma, D. (2019).** Overview of MEX-A3T at IberLEF 2019: Authorship and aggressiveness analysis in Mexican Spanish tweets. IberLEF@ SEPLN, pp. 478–494.
3. **Aragón, M. E., Jarquín-Vásquez, H. J., Montes-Y-Gómez, M., Escalante, H. J., Pineda, L. V., Gómez-Adorno, H., Posadas-Durán, J. P., Bel-Enguix, G. (2020).** Overview of MEX-A3T at IberLEF 2020: Fake news and aggressiveness analysis in Mexican Spanish. IberLEF@ SEPLN, pp. 222–235.
4. **Cambria, E., Li, Y., Xing, F. Z., Poria, S., Kwok, K. (2020).** SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 105–114.
5. **Cardellino, C. (2019).** Spanish billion words corpus and embeddings.
6. **Castro-Sánchez, N. A., Baca-Gómez, Y. R., Martínez, A. (2015).** Development of affective lexicon for Spanish with Mexican slang expressions. Res. Comput. Sci., Vol. 100, pp. 9–18.
7. **Cruz, F. L., Troyano, J. A., Pontes, B., Ortega, F. J. (2014).** Building layered, multilingual sentiment lexicons at synset and lemma levels. Expert Systems with Applications, Vol. 41, No. 13, pp. 5984–5994.
8. **Ekman, P. (1992).** An argument for basic emotions. Cognition & Emotion, Vol. 6, No. 3-4, pp. 169–200.
9. **Farías Hernández, D. I., Patti, V., Rosso, P. (2016).** Irony detection in twitter: The role of affective content. ACM Transactions on Internet Technology (TOIT), Vol. 16, No. 3, pp. 1–24.
10. **Guzman-Silverio, M., Balderas-Paredes, Á., López-Monroy, A. P. (2020).** Transformers and data augmentation for aggressiveness detection in Mexican Spanish. IberLEF@ SEPLN, pp. 293–302.
11. **Jin, D., Jin, Z., Zhou, J. T., Szolovits, P. (2020).** Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. Proceedings of the AAAI conference on artificial intelligence, volume 34, pp. 8018–8025.
12. **Kralj Novak, P., Smailović, J., Sluban, B., Mozetič, I. (2015).** Sentiment of emojis. PloS one, Vol. 10, No. 12, pp. e0144296.
13. **Mohammad, S. M., Turney, P. D. (2013).** Crowdsourcing a word-emotion association lexicon. Vol. 29, No. 3, pp. 436–465.
14. **Molina-González, M. D., Martínez-Cámara, E., Martín-Valdivia, M.-T., Perea-Ortega, J. M. (2013).** Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications, Vol. 40, No. 18, pp. 7250–7257.
15. **Perez-Rosas, V., Banea, C., Mihalcea, R. (2012).** Learning sentiment lexicons in Spanish. LREC, volume 12, Citeseer, pp. 73.

16. **Plaza-Del-Arco, F.-M., Molina-González, M. D., Ureña-López, L. A., Martín-Valdivia, M. T. (2020).** Detecting misogyny and xenophobia in Spanish tweets using language technologies. ACM Transactions on Internet Technology (TOIT), Vol. 20, No. 2, pp. 1–19.
17. **Plutchik, R. (2001).** The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. American scientist, Vol. 89, No. 4, pp. 344–350.
18. **Redondo, J., Fraga, I., Padrón, I., Comesaña, M. (2007).** The Spanish adaptation of ANEW (affective norms for English words). Behavior research methods, Vol. 39, No. 3, pp. 600–605.
19. **Ríos, M. D. A., Gravano, A. (2013).** Spanish dal: a spanish dictionary of affect in language. Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 21–28.
20. **Sidorov, G., Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Suárez-Guerra, S., Treviño, A., Gordon, J. (2012).** Empirical study of opinion mining in Spanish tweets. LNAI, Vol. 7629, pp. 1–14.
21. **Urizar, X. S., Roncal, I. S. V. (2013).** Elhuyar at TASS 2013. Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS 2013), pp. 143–150.
22. **Wei, J., Zou, K. (2019).** Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196.
23. **Xie, Q., Dai, Z., Hovy, E., Luong, M.-T., Le, Q. V. (2019).** Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848.
24. **Zaizar-Gutiérrez, D., Fajardo-Delgado, D., Carmona, M. Á. Á. (2020).** ITCG's participation at MEX-A3T 2020: Aggressive identification and fake news detection based on textual features for Mexican Spanish. IberLEF@ SEPLN, pp. 258–264.
25. **Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R. (2019).** Semeval-2019 Task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.
26. **Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Çöltekin, Ç. (2020).** Semeval-2020 Task 12: Multilingual offensive language identification in social media (offenseval 2020). arXiv preprint arXiv:2006.07235.

Article received on 31/07/2021; accepted on 30/09/2021.
Corresponding author is Hiram Calvo.