Using Lexical Resources for Detecting Offensiveness in Mexican Spanish Tweets

Daniel Abraham Huerta-Velasco and Hiram Calvo

Centro de Investigación en Computación, Instituto Politécnico Nacional, Ciudad de México, México
{dhuertav2019, hcalvo}@cic.ipn.mx

Abstract. This work presents a description of our participation in subtasks 3 and 4 at MeOffendEs@IberLEF 2021 which consisted in classifying tweets as offensive or non-offensive in the OffendMEX corpus. For both subtasks, we proposed to use 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. The scores of recall metric obtained in both subtasks was competitive comparing to both the baseline of the competition's and the other teams'.

Keywords: Lexical Resources · Sentiment Analysis · Mexican Spanish Tweets · Text Classification.

1 Introduction

Social media have had a great impact in the history of humanity. Nowadays it is very easy to share information, thoughts, images, videos, etc, only with a click. Despite there are positive aspects associated with social media usage, there are negative ones that many social media users have to face daily. One of the most dangerous for most people is that many users take advantage of the anonymity that social media gives them and insult, harass, provoke and threat to an individual or a group of people.

Offensiveness 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 the investigation on this topic and organize competitions where, mainly, ask for new proposals that can classify as good as possible whether a tweet is offensive or not, among other labels, such as if a tweet is vulgar but not offensive, not vulgar and offensive, if the aggression of the tweet is targeted to a person or a group of people, etc.

IberLEF 2021, September 2021, Málaga, Spain.

This year (2021) MeOffendEs competition [16] at the Iberian Languages Evalutation Forum (IberLEF) [14] was organized. The aim of this competition was to boost research on a sensitive topic for the Spanish language. 4 subtasks were part of this competition. This work presents our solution for the last two, which consisted in classifying tweets as offensive or non-offensive in the OffendMEX corpus. It should be said that metadata about each tweet were provided in Task 4.

As further detailed next, our proposed features derived from several lexicons which have a collection of Spanish words that have been weighted according to different criteria like affective, dimensional, and emotional value among others derived from POS-tagging analysis of the tweets and other models which have been already proved such as word-embeddings and one-hot codification. This data representation was the input of a Support Vector Machine and obtained competitive scores of recall metric in the subtasks, and the usefulness percentage of the lexical features overcame the 50% in each subtask.

2 Model's Description

For these two tasks, the OffendMEX corpus was used. It is divided into 2 sets: The *Training* set is formed by 5,060 tweets where 3,679 of them were labeled as non-offensive, and the rest (1,381) as offensive. The *Test* set is formed by 2,183 tweets. In addition to these sets, another one was released named as *trial* set and was formed by 76 tweets (35 non-offensive, 41 offensive).

Figure 1 shows the flow process of how we faced these two subtasks. In a nutshell, the process consists in the extraction of the features similar to [8]. In this work, the authors extracted features from lexical resources called lexicons, which are lists of words weighted according a value, in this case, the polarity value of a word or a phrase in English to detect irony in English tweets. Then, they used these features as inputs of some machine learning algorithms such as Support Vector Machines, Decision Trees, and Naive-Bayes. The same strategy is followed here, but we used different lexicons and proposed other kind of features which include an special treatment for emojis and hashtags. These steps will be explained in detail in the following sections.

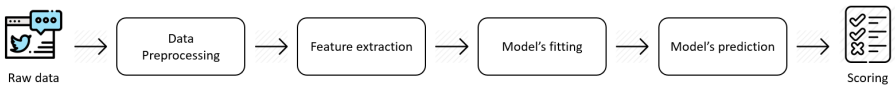

Fig. 1. Block diagram of our proposal

2.1 Data preprocessing

Before the feature extraction process, a data preprocessing is performed. In this step, 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 preprocessing, 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 [4]. The frequency of hashtags is used as a feature.
- *Emojis cleaning*: All emotional polarity values of emojis which are present in the post are summed both positive and negative values individually, and the combination of them according to values in [9]. It should be said that not all emojis[1] are present in the work of Kralj and her team. That is why six features are extracted: the sum of the polarity of positive and negative emojis in the post, the sum of polarity of positive and negative emojis separately, the 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.

2.2 Features' Extraction

After all tweets have been preprocessed, the next step is to extract the features of the text. As it is widely known, most machine learning algorithms require a numeric representation of text as the input, so it has to be casted to a vectorial representation where each element represents a feature. They are categorized depending on their nature.

Structural features 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.

Affective features consist in both positive and negative polarity values that a tweet has according to the sum of the words' polarity present in it. To do that, several lists of Spanish words (lexicons) classified by an amount (positive amounts means positive emotional polarity, otherwise, negative) or a label (positive, negative, neutral) are used. Table 2 shows the features which fall in this description and the name of the lexicon which was used for computing each feature.

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

Table 1. Structural features

Features	Description
exclam_marks quest_marks	The frequency of each punctuation mark in a tweet
singulars plurals	The frequency of each inflectional feature of nouns, pronouns, adjectives, determiners, numerals, and verbs.
words chars	The total amount of words and characters in a tweet, respectively
upper	The total amount of uppercase 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 [9] or not, respectively

Dimensional Features consist in those which are inspired in some theories which propose that the nature of an emotional state is determined by its 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. Table 3 shows the features inspired by these theories.

Emotional Features consist in those which are inspired in the work of [18] and [7] who defined 8 and 6 basic emotions, respectively: anger, disgust, fear, joy, sadness, surprise, anticipation, and trust. Table 4 shows the features inspired by these emotions.

Contextual Features consist in those which are meta-data of the tweet. These features were only used for subtask 4. Table 5 shows a description of the meta-data given for this subtask and how we used them as features.

In total, a tweet is represented as a vector composed by 114 features for subtask 3, and by 126 for subtask 4. In the future, they will be refered as CVAD features. One thing to note is that the lexicons used in affective, dimensional and emotional features contain words or phrases not in a specific variant of Spanish except the Mexican Slang Lexicon. In addition to them, 300 word-embeddings and a one-hot codification features are added. The way in which these word-embeddings were trained is described in [4]. For one-hot codification, all words in the training dataset are obtained. Then, these n-features (where n depends on how many words are used at least m-times in the whole training dataset) are vectorized as zeros. Finally, if each feature (word) is present in the post, its

Table 2. Affective features

Features	**Description**
emojis_polarity	Sum of tweet's polarity according to the emojis present in the post
pos_emojis neg_emojis	Sum of polarity value of "positive" and "negative" emojis, respectively.
HL_insults HL_xenoph HL_misog HL_inmigrants	Hate speech Spanish lexicons[17] contain 4 lexicons which described general insults, hateful lexicons toward immigrants and women, and words that refer to the nationality of an immigrant in Spanish. Each lexicon contains 279, 44, 183, and 250 words respectively.
EMOLEX_*n*+ EMOLEX_*n*-	NRC Word-Emotion Association Lexicon (aka EMOLEX) [12] is a list of English and Spanish words/phrases and their associations with two sentiments (positive and negative). Each feature is the sum of positive and negative (separately) per n-gram in the lexicon. n goes from 1 to 4
ISOL_1+ ISOL_1-	iSOL[13] is a list of words labeled as positive or negative. Each feature is the sum of positive and negative words in the post.
MXSL_int1+ MXSL_int1- MXSL_phr*n*+ MXSL_phr*n*-	Mexican Slang lexicon [5] consists in lists of interjections and phrases used in mexican slang. Each feature is the sum of positive and negative (separately) per n-gram in the lexicon. n goes from 1 to 4. We added 1,373 Mexican expressions from our own knowledge to this list.
ML_SENTICON_*n*+ ML_SENTICON_*n*-	ML-Senticon [6] is a list of Spanish words/phrases which, for each lemma, provides an estimation of polarity (from very negative -1.0 to very positive +1.0). Each feature is the sum of positive and negative words in the post per n-gram in the lexicon. n goes from 1 to 4
MS_1+ MS_1-	Multilingual Sentiment lexicon [10] is a list of Spanish words labeled as positive or negative. Each feature is the sum of positive and negative words in the post
SSL_1+ SSL_1-	Sentiment Lexicons in Spanish [15] is a list of Spanish words which are labeled as positive and negative according to English and Spanish annotations
ELHPOLAR_*n*+ ELHPOLAR_*n*-	Elhpolar lexicon[22] is a list of Spanish words/phrases labeled as positive and negative. Each feature is the sum of positive and negative words in the post per n-gram. n goes from 1 to 4
SENTICNET_+ SENTICNET_-	SenticNet [2] is a list of words which have an emotional polarity floating value from -1 (negative) to +1 (positive). Each feature is the sum of these values according their polarity

Table 3. Dimensional features

Features	Description
SENTICNET_aptitude SENTICNET_attention SENTICNET_pleasantness SENTICNET_sensitivity	SenticNet [2] is a list is a list of Spanish words which are associated with the four dimensions of the Cambria Hourglass of Emotions model [3]
S-ANEW_val S-ANEW_aro S-ANEW_dom	Spanish ANEW [20] is a list of words which is inspired by Affective Norms for English Words (ANEW) [1]. Words are associated with emotional ratings in terms of the Valence-Arousal-Dominance model
SDAL_pleasantness SDAL_activation SDAL_imagery	Spanish DAL (SDAL) [21] is a list of Spanish words which are manually annotated with regard to this three dimensions. SDAL is inspired by [23]

Table 4. Emotional features

Features	Description
EMOLEX_n_anger EMOLEX_n_disgust EMOLEX_n_fear EMOLEX_n_joy EMOLEX_n_sadness EMOLEX_n_surprise EMOLEX_n_anticipation EMOLEX_n_trust	EMOLEX [12] is a list of English and Spanish words or phrases and their associations with the 8 basic emotions identified by Plutchik. Each feature is the sum of each emotion per n-gram in the lexicon. n goes from 1 to 4
SEL_1_anger SEL_1_disgust SEL_1_fear SEL_1_joy SEL_1_sadness SEL_1_surprise	Spanish Emotion Lexicon (SEL) [11][19] is a list of Spanish words that are associated with the measure of Probability Factor of Affective use (PFA) with respect to the 6 basic emotions identified by Ekman

Table 5. Contextual features

Features	Description
acc_verified acc_followers acc_followings acc_listed acc_favs acc_tweets acc_theme acc_default_image	These features describe the data of the user who twitted: whether his/her account is verified, how many followers he or she has, how many users he or she is following, how many public lists that he or she is a member of, how many tweets he or she has published, if he or she has altered the theme or background of his/her profile, and if he or she has his/her own profile image
tweet_rt tweet_favs tweet_isrt tweet_isquote	These are the information about the tweet itself: how many retweets it has, how many times it has been marked as favorite, if it is a reply of another tweet, and if it is a quote of a tweet.

representation in the vector is changed to 1. It should be noted that tweets in the trial dataset were included into training dataset.

2.3 Model's training

These features were the inputs of a Support Vector Machine (SVM). SVM hyperparameters' tuning and cross validation over training dataset were performed to know which configuration of both features and hyperparameters yielded the best theoretical results and then, predict the labels of testing dataset using them. We used scikit-learn GridSearchCV[2] and cross_validate[3] methods to perform this step. The metric used for optimizing the hyperparameters was F_1 macro. Cross validation was performed using the K-Fold technique which consists in dividing all samples in k groups (k-folds). The prediction function is learned using $k-1$ folds, and the fold left out is used for testing. The value of k used in the experiments was 5. Finally, to obtain one-hot codification, tested word frequencies were from bigger or equal than 1 to 5, separately.

Tables 6 and 7 show the ranked results of the experimentation for subtask 3 and 4, respectively. All experiments include CVAD features, 300 word-embeddings and n-one hot codification. Tables show the experimentation among the different number of features derived of the number of words which frequencies are bigger or equal to n.

There are 11,544 different words in the training dataset of which 4,102 are used at least twice, 2,462 at least thrice, 1,721 at least four times, and 1,333 at least five times.

Table 6. Experimental scores in the training dataset for subtask 3

n	Hyperparameters	F_1 macro
1	C = 0.14, penalty = l1	0.7239
2	C = 0.18, penalty = l1	0.7228
4	C = 0.18, penalty = l1	0.7223
3	C = 0.15, penalty = l1	0.7218
5	C = 0.18, penalty = l1	0.7214

2.4 Model's prediction

Using the configuration of the best experimental results, labels from the test dataset are obtained and the results of these are shown in Table 8.

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

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

Table 7. Experimental scores in the training dataset for subtask 4

n	Hyperparameters	F_1 macro
1	C = 0.24, penalty = l1	0.7305
3	C = 0.2, penalty = l1	0.7251
2	C = 0.2, penalty = l1	0.7249
4	C = 0.19, penalty = l1	0.7243
5	C = 0.16, penalty = l1	0.7234

Table 8. Scores for both subtasks in the test dataset

Task	Precision	Recall	F_1 macro
3	0.535	0.687	0.602
4	0.538	0.684	0.603

3 Results in the competition

The organizers of MeOffendEs [16] reported a baseline performance per subtask. For Subtask 3 they reported 0.719, 0.41, and 0,522 scores for precision, recall and F_1 score respectively, and for Subtask 4, 0.663, 0.698, and 0.68. As they ranked the participants by using the F_1 macro metric, our solution was better ranked than baseline for Subtask 3, but it was not able to outperform it in Subtask 4.

After analyzing cross validation process to find out what type of tweets in training dataset our proposal was not able to classify correctly in both subtasks, we realized that tweets with sexual connotations or with negative words (not vulgar) but not attacking someone are some of them. Table 9 shows some instances which falls under these descriptions.

Table 9. Tweets in training dataset that were misclassified by our model

Tweet	Actual label
@USUARIO como luchar contra la corrupción de los oficiales no sólo nos enfoquemos en la de los ciudadanos esa moneda tiene dos caras feas	Non-aggressive
Woou rulo invita..yo tambien quiero mamar esa panocha deliciosa y clavarsela	Aggressive

Comparing our results to the rest of competitors, our solution was ranked at 7th place of 10 teams for Subtask 3, and at 2nd place out of 3 participants for Subtask 4. In order to know which CVAD features (i.e. the ones derived by lexical resources) were useful for these problems, a feature selection process was

performed. To do this, we used the SelectFromModel[4] method, which selects features based on importance weights, on our top solutions per subtask.

For Subtask 3, 13 structural features out of 17 (76.47%), 26 affective ones out of 49 (53.06%), 9 dimensional of 10 (90%), and 13 emotional of 38 (34.21%) were found useful. For Subtask 4, 16 (94.12%), 30 (61.22%), 9 (90%), 16 (42.10%), and 8 contextual features out of 12 (66.67%) were selected.

As can be seen, the usage percentage per type of CVAD feature increased when the metadata of the tweet was supplied to detect whether a tweet is offensive or not. This phenomenon can be observed in the obtained scores which showed a slightly better classification in subtask 4 than 3.

Another interesting feature to be observed is that both affective and emotional features were less useful in subtasks 3 and 4 compared to the other features. The reason of this is that phrases with 3 or 4 words (i.e. trigrams and 4-grams) which are present in the used affective and emotional lexicons are not frequently used by Mexican users except for those present in the combination of the Mexican Slang lexicon [5] and our list. If we removed these features, the usage percentage turns into 70.27% affective features, and 59.09% emotional features for Subtask 3. For Subtask 4, the percentages after removing said features are 81.08% and 72.73%, respectively.

4 Conclusions and Future Work

For these subtasks, a relatively simple model was proposed to classify Mexican Spanish tweets as offensive or non-offensive. This model was mainly based on lexical resources as features, as well as other kind of features which have been used previously. This representation allowed our model to learn contextual features which are the meta-data provided for subtask 4.

One thing to be noted is that our recall scores obtained in both subtasks were better than the majority of competitors' whose models were better ranked, but our precision scores were not as good as theirs. This evidence suggests that using lexical resources to detect offensiveness in Mexican Spanish tweets is a good option when there is a high cost associated with False Negatives, i.e. when a model is preferred to detect offensiveness or non-offensiveness in tweets when they actually are.

As a future work, we plan to perform experiments using these features with different Machine Learning algorithms such as the multilayer perceptron; additionally, we plan to update the used lexicons with words or phrases which mexicans actually use both in the real life and on social media according to the criteria adopted to make these lists.

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.SelectFromModel.html`.

References

1. Bradley, M.M., Lang, P.J.: Affective norms for english words (ANEW): Instruction manual and affective ratings. Tech. rep., Technical report C-1, the center for research in psychophysiology (1999)
2. Cambria, E., Li, Y., Xing, F.Z., Poria, S., Kwok, K.: Senticnet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. pp. 105–114 (2020)
3. Cambria, E., Livingstone, A., Hussain, A.: The hourglass of emotions. In: Cognitive behavioural systems, pp. 144–157. Springer (2012)
4. Cardellino, C.: Spanish Billion Words Corpus and Embeddings (August 2019), `https://crscardellino.github.io/SBWCE/`
5. Castro-Sánchez, N.A., Baca-Gómez, Y.R., Martínez, A.: Development of affective lexicon for spanish with mexican slang expressions. Res. Comput. Sci. **100**, 9–18 (2015)
6. Cruz, F.L., Troyano, J.A., Pontes, B., Ortega, F.J.: Building layered, multilingual sentiment lexicons at synset and lemma levels. Expert Systems with Applications **41**(13), 5984–5994 (2014)
7. Ekman, P.: An argument for basic emotions. Cognition & emotion **6**(3-4), 169–200 (1992)
8. Farías, D.I.H., Patti, V., Rosso, P.: Irony detection in twitter: The role of affective content. ACM Transactions on Internet Technology (TOIT) **16**(3), 1–24 (2016)
9. Kralj Novak, P., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. PloS one **10**(12), e0144296 (2015)
10. Lab, D.S.: Multilingual sentiment, `https://sites.google.com/site/datascienceslab/projects/multilingualsentiment`
11. Miranda-Jiménez, S., Viveros-Jiménez, F., Gelbukh, A., Castro-Sánchez, N., Velásquez, F., Díaz-Rangel, I., Gordon, J.: Empirical study of opinion mining in Spanish Tweets (2012)
12. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon **29**(3), 436–465 (2013)
13. Molina-González, M.D., Martínez-Cámara, E., Martín-Valdivia, M.T., Perea-Ortega, J.M.: Semantic orientation for polarity classification in Spanish reviews. Expert Systems with Applications **40**(18), 7250–7257 (2013)
14. Montes, M., Rosso, P., Gonzalo, J., Aragón, E., Agerri, R., Álvarez-Carmona, M.Á., Álvarez Mellado, E., Carrillo-de Albornoz, J., Chiruzzo, L., Freitas, L., Gómez Adorno, H., Gutiérrez, Y., Jiménez-Zafra, S.M., Lima, S., Plaza-de Arco, F.M., Taulé, M. (eds.): Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) (2021)
15. Perez-Rosas, V., Banea, C., Mihalcea, R.: Learning sentiment lexicons in Spanish. In: LREC. vol. 12, p. 73. Citeseer (2012)
16. Plaza-del-Arco, F.M., Casavantes, M., Escalante, H., Martin-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)
17. Plaza-Del-Arco, F.M., Molina-González, M.D., Ureña-López, L.A., Martín-Valdivia, M.T.: Detecting misogyny and xenophobia in Spanish tweets using language technologies. ACM Transactions on Internet Technology (TOIT) **20**(2), 1–19 (2020)

18. Plutchik, R.: 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 **89**(4), 344–350 (2001)
19. Rangel, I.D., Sidorov, G., Guerra, S.S.: Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. Onomazein **5**(29), 31–46 (2014)
20. Redondo, J., Fraga, I., Padrón, I., Comesaña, M.: The Spanish adaptation of anew (affective norms for english words). Behavior research methods **39**(3), 600–605 (2007)
21. Ríos, M.D., Gravano, A.: Spanish dal: a spanish dictionary of affect in language. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. pp. 21–28 (2013)
22. Urizar, X.S., Roncal, I.S.V.: Elhuyar at tass 2013. In: Proceedings of the Workshop on Sentiment Analysis at SEPLN (TASS 2013). pp. 143–150 (2013)
23. Whissell, C.: Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. Psychological reports **105**(2), 509–521 (2009)