DCCD-INFOTEC at MeOffendEs@IberLEF21 Subtask 3: A Transfer Learning Approach Based on EvoMSA's Stacked Generalization

José J. Calderón[1,2], Eric S. Tellez[1], and Mario Graff[1]

[1] INFOTEC Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación, México
{juan.calderono,eric.tellez,mario.graff}@infotec.mx
[2] CIMAV Centro de Investigación en Materiales Avanzados, México
juan.calderon@cimav.edu.mx

Abstract A feasible approach to tackle the problem of offensive identification is to treat it as a classification problem. In this contribution, an ensemble of models from domains such as misogyny, aggressiveness identification, and humorous identification are used to tackle the offensive identification task of Non-contextual Binary Classification for Mexican Spanish subtask 3 in the MeOffendEs@IberLEF21. In addition, we also enrich this set of models with a straightforward model based on text reversion which demonstrates a sustained improvement to the final prediction capabilities of the ensemble as it is observed in the results. Our approach is open-source and available through the EvoMSA classification system. Finally, we provide an experimental study of our approach using a brief ablation study with the ensembled models.

Keywords: Offensive Language Detection in Spanish Variants · Text Categorization · Model's Performance Analysis.

1 Introduction

Social media, like Facebook and Twitter, are an almost unlimited information flow without restrictions, playing a vital role in our lifestyle whereby people connect, exchange ideas, points of view, and knowledge [3]. These information exchanges have positively impacted our society; however, a number of problems also arose with these new ways to communicate with people around the world. Paradoxically, given its ease, breadth, and apparent anonymity, social networks can generate problems like social isolation, low self-esteem, fraud, identity theft, grooming, and many kinds of offensive content.

IberLEF 2021, September 2021, Málaga, Spain.

From personal and expontaneous attacks to well-orchestrated actions by groups, the aggressions in social networks can lead to long-term harm in victims [19]. Therefore, understanding the variables and processes that predict the perpetration of aggressions is essential to reduce them.

Fortunately, the international scientific community participates in the search for solutions, addressing the detection of aggressive language in social networks through open forums and workshops. Such is the case of MeOffendEs@IberLEF21 [14, 17], which focuses on detecting and analyzing offensive content in Mexican Spanish using NLP techniques.

Participants must test their proposal solutions for the task by using a training corpus provided by the organizers. However, detection and classification of offensive language are both complex tasks. On the one hand, methods have to deal with ambiguity and subjective statements. On the other hand, tweets are very short texts and often full of typos, grammatical errors, and emoticons.

1.1 Related work

In the context of the classification of short texts such as Twitter, the trend is towards proposals with a particular semantic degree, co-occurrence of terms such as word embedding, use of deep neural networks like transformers, among others. However, lexical and syntactic-based models such as BoW, n-grams, tf-idf, and classifiers such as Support Vector Machines (SVM) and Rule-Based Naïve Bayes (RNB) are still valid given their proven performance as shown by recent surveys [12] [8]. This effectivity is proven in the MEX-A3T competition [4], here the UACH team proposes to use character n-grams together with word embedding and an SVM classifier. In contrast, the CIMAT team proposes an ensemble of BERT models. Teams like Intensos and ITCG-SD keep using BoW with tf-idf text representations. The ITCG-SD team uses a simple method based on detecting the capital letters ratio in the text showing how a simple approach can improve the overall performance.

Hate Speech [7, 11, 13] is found when the offensive language targets a particular group of people. From political, sexual orientation, religion, nationality, skin color, and gender, hate speech spreads rapidly around social media [22, 23]. Along with the inherent complexity of the informal language we found in social media, the task imposes additional difficulties like the message's intention, resources in the specific language, the precise target identification, and the social and political reality of the people involved in the message.

Schmidt and Wiegand [18] survey the field of automatic hate speech detection, concluding that while the set of features examined by different approaches varies greatly, the classification methods mainly focus on traditional supervised learning, like SVM. However, more recent methods are based on deep learning. Finally, Aggrawal [1] surveys methods and systems for detecting aggressive tweets focusing on those using stylistic and content features like message's length, URL embedded, content, retweet pattern, tweet sentiment, author, among others. The author found that the Naïve Bayes classifier performs remarkably well when using this kind of feature to classify hatred, sexual and offensive content.

Humor is expressed with figurative and subjective language. A human learns to interpret this kind of language and expressions from its culture and its environment. Therefore, the automatic identification of humoristic messages in social media has deserved a lot of literature work. For instance, *Humor Analysis based on Human Annotation (HAHA)* [6]. Here, a set of human-labeled messages from Twitter is introduced to train and test humor identification and ranking algorithms. Humor messages have a plethora of topics and complex linguistic structures like induced ambiguity, absurdity, irony, and sarcasm. These same characteristics could explain why humor detectors can work very well for detecting offensive language.

In brief, related works show that the task of offensive language detection can be satisfactorily tackled with classifiers and techniques very well known as SVM and n-grams, even with simple text transformations and transferring knowledge from related domain models.

EvoMSA. Graff et al. [9] introduce EvoMSA as a stacking-based classifier for solving sentiment analysis tasks. A stacking-based classifier has two levels, see Figure 1; the first level comprises several models that solve the main task independently. The second level is then used to improve the classification using the predictions made by each of its composing sub-models. In particular, EvoMSA uses Genetic programming as the second level and B4MSA [20] text classifier as the default first-level modeling tool. The current version of EvoMSA supports different first-level and second-level algorithms.[3]

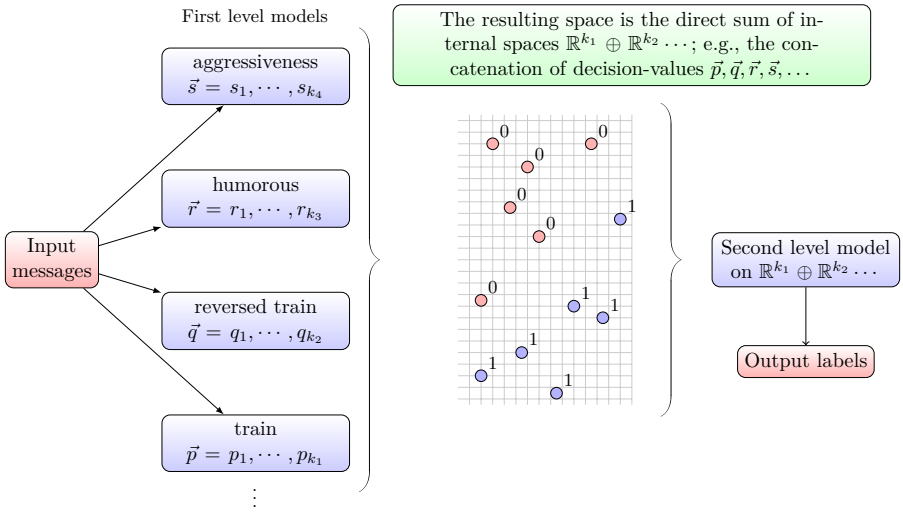

Figure 1. EvoMSA transfer learning scheme via ensembling different domain models.

[3] `https://github.com/INGEOTEC/EvoMSA`

1.2 Our contribution

Our proposal is based on ensembling different pre-trained models to identify related content like misogyny, aggressiveness and a humorous detector. To the best of our knowledge, this is the first approach using this kind of knowledge ensemble to solve the Offensive language identification task. In addition, as explained in §3.2, we use direct and reverse models trained to detect offensive language with the dataset provided for the MeOffendES@IberLEF21 for the mexican Spanish variant.

1.3 Roadmap

The current section introduces and contextualizes our approach to identifying Offensive language as part of the MeOffendES@IberLEF21 challenge. The task is described in Section 2 and our approach to cope with it is detailed in Section 3. The experimental results are presented and discussed in Section 4. Finally, Section 5 summarizes and concludes this study.

2 Task description

In the MeOffendEs@IberLEF21, Subtask 3: *Non-contextual binary classification for Mexican Spanish*, the goal is to create a model that identifies tweet messages as offensive or non-offensive. For this matter, a corpus of labeled messages is given as the training set; the organizers also provide a set of unlabeled messages as a test. The subtask of non-contextual binary classification has no more information than the text (tweets) and its associated labels. These messages are written in the Mexican variation of the Spanish language.

The training Dataset provided for this task (OffendMEX) consists of 5060 tweets; 73% are classified as no-offensives and 27% as offensives. Most of them with short texts of between 8 and 40 words, and although they contain typos, spelling errors, and a variety of symbols, it can be said that most are legible enough to apply classification techniques. Table 2 shows some examples of the tweets.

Table 1. Examples of Tweets in the OffendMEX DataSet

Offensive:	Actualmente ya pasó de moda la pucha joto, ahora sólo quedamos los verdaderos seguidores, los que amamos este estilo de vida
Non-offensive:	Soy el Clint Eastwood de los Puentes de Madison en todas las putas historias de amor que me 🌹 han tocado
Typos:	Mmmmta, pos neh veo las de wismichu
Emojis:	😄 ☹ 😃 tu que putas hablas de futbol
Others:	@USUARIO, #SiLUP, #MasterChefMx espero y ya saquen a Rogelia por joto, aparte no sabe cocinar ˆ<URL>

3 Our Offensive-language detection approach

Our starting point was the evaluation of the strategies proposed by INGEOTEC. Their results using EvoMSA, described in the Section 1.1 as a multilingual sentiment analysis system based on genetic programming to detect aggressive text, presents the top performance in similar tasks in 2018 [5] and 2019 [6].

3.1 Text transformations

As it has been described in the state-of-the-art (e.g. [10]), one crucial point of text classification is to transform the text into tokens to produce a vector. EvoMSA has already integrated some lexicon-based models that can be configured to achieve better results in this step according to the dataset's characteristics.

In addition, the lexicon models achieve a better performance to predict aggressiveness using unigrams, bigrams, trigrams, q-grams of 1 and 5 characters length, and skip-grams, according to the description given in [9]. Therefore, a simple variation on these text transformations could produce different vectors when the models are trained, and consequently, improve the final predictions.

3.2 Reverse-order text

Starting from the idea that some transformations of the input text could impact the performance of the training process, we experiment with a model where the text is input to the training algorithm in reverse order; that is to say, to invert the tweets character by character, as in the example below:

Normal: *Soy el Clint Eastwood de los Puentes de Madison en todas las putas historias de amor que me han tocado*

Inverted: *odacot nah em euq roma ed sairotsih satup sal sadot ne nosidam ed setneup sol ed dowtsae tnilc le yos*

This pre-processing of each tweet allows generating an entirely different token list for the same tweet producing a new sparse vector space, impacting the training process.

To find the best tokenization scheme, we performed some tests using the task the dataset and μTC [21], a minimalist tool that generates text classifiers. As a result, we found that the best performance for inverted-order text transformation is achieved with the following tokenizing scheme (in combination): unigrams, bigrams, skip-grams, and q-grams of 3 characters.

TextModelInv. Once the best tokenization scheme was achieved, we proceeded to develop the model with inverted text to be added to the EvoMSA set models. We extended B4MSA's TextModel class, and override the method used for tokenization. The new extended class was called TextModelInv. We take advantage of EvoMSA which takes a set of different models allowing the combination of different text transformations and tokenizers, and secondly, it allows us to add a new model with our own text transformations.

4 Experiments and results

As noted in §1.1, EvoMSA can combine the predictions of pre-trained models (see Table 2) with diverse meanings (counting of positive negative words, emoji prediction, hatred, misogyny, humor, among others). On the other hand, it also allows the construction of new models based on B4MSA, a text classifier using a bunch of text tokenization, text transformations, weighting methods that use an SVM with a linear kernel as the classifier. We use B4MSA to produce TextModelInv and ensemble other models found in EvoMSA to produce a highly competitive model for offensive language identification.

Table 2. Examples of EvoMSA pre-trained models used in similar competitions with similar datasets.

Model	**Competition**	**Task**
Humor	IberLEF'2019 [15]	Humor Analysis based on Human Annotation
Aggressiveness	MEX-A3T [2]	Author profiling and aggressiveness
Misogyny	SemEval-2019 [16]	Multilingual detection of hate speech against immigrants and women in Twitter
		Identification and categorization of offensive language in social media

The tests were performed using the OffendMEX training Dataset of 5060 tweets, split at 80% / 20%, 4048 train set, and 1012 test set. EvoMSA was trained: i) one model at a time, ii) combining pairs the TextModelInv with each pre-trained model, and finally, iii) combining all. Also, TextModelInv was tested using a simple and a complex selection of lexicon parameters.

The following non-decisive points could be considered to improve the reliability of the tests:

1. A larger number of instances could improve the training.
2. Classes balance is labeled 73% for non-offensive and 27% for offensive, which could bias predictions.
3. We can expect a certain degree of randomness from different train instances due to random splits and the stochastic models being used.

With respect to the test dataset provided by the organization, it consists of 2183 tweets, a balance of classes and a style similar to the test dataset. Noting this similitud and our result shown in Table 3 with respect to the training dataset, we decided to participate in the task using the five best results under the same parameters.

According the official results of MeOffendEs@IberLEF21, computed on the official ground truth, our system proposed reached the third place overall (just 1.7% below rank #1) and the first place in *recall*.

Table 3. Performance of EvoMSA with different models and OffendMEX. Best values per column are in bold font; entries are ranked by *Macro F1* score. The top five models combinations were used with the test dataset.

Models							Macro	Macro
Straight	Reverse	Aggress.	Emoji	Humor	Misogyny	Acc.	F1	Recall
	yes		yes			0.8158	**0.7689**	0.7622
	yes	yes	yes	yes	yes	0.8182	0.7683	0.7748
	yes	yes		yes	yes	0.8073	0.7679	**0.7793**
				yes		**0.8291**	0.7668	0.7572
yes						0.8152	0.7658	0.7575
		yes				0.8113	0.7591	0.7639
	yes					0.8162	0.7549	0.7477
	yes	yes				0.8024	0.7527	0.7518
					yes	0.8182	0.7480	0.7380
			yes			0.8134	0.7479	0.7424
yes$^-$						0.7885	0.7470	0.7485
		yes		yes	yes	0.7984	0.7453	0.7521
	yes$^+$					0.7984	0.7442	0.7348
			yes	yes		0.8016	0.7404	0.7349
	yes	yes			yes	0.8014	0.7369	0.7286
	yes$^-$					0.7806	0.7323	0.7353
	yes			yes		0.7992	0.7250	0.7086
	yes				yes	0.7976	0.7150	0.6964

yes $\rightarrow$ text model uses unigrams, bigrams, and skip-grams.

yes$^+$ $\rightarrow$ text model uses unigrams, bigrams, skip-grams, and q-grams of 3 characters.

yes$^-$ $\rightarrow$ text model uses only unigrams.

4.1 Analysis

Table 3 shows the results of our model selection process using the OffendMEX dataset. We can observed that the reverse-order input text model –TextModelInv– remains at the top of our internal procedure; this is notorious when it is combined with all the pre-trained base models. Regarding Macro-F1 we observe that the best model is the ensemble of Reverse and Humor models, followed by the ensemble of all models (excepting for the Straight model). With respect to Macro-Recall the best model corresponds to the ensemble of four models: Reverse, Aggressiveness, Humor, and Mysogyny. It is remarkable to note that the best accuracy is found by the Humor singleton model; it is also noticeable that Humor is part of the best performing models in the table. Singleton models of Straight and Reverse are competitive but far from being at the table's top.

Regarding the official results, we have observed that our top five models proposed in our system kept the same order in the rank with respect to training and

test dataset, being able to assume the same observations. Although the metrics were a bit lower, maybe explained by differences in the metrics computation.

5 Conclusions

We used EvoMSA's stacked generalization machinery to borrow knowledge from aggressive, misogyny, and humor classification models. The combination of these models and a pair of classifiers created on the official dataset of MeOffend-MEX@IberLEF21 produce a robust framework to solve the offensive language task.

The ensemble combines several related models to solve the offense language identification and two more models working with the current target. These two models trained to identify offensive language are based on B4MSA text classifiers; perhaps the unique significant difference between these two B4MSA models is the direction of the input text. With these facts in mind, we observed that offensive language models work consistently better jointly than separated. We also observed a remarkable impact of the versatility of the humor model, which could be a good predictor by itself of the task, perhaps due to the connection between the offensive language and sarcasm and other kinds of rude jokes.

Our future research focuses on studying how each model help to identify offensive language and other related tasks. We also plan to add support for other models that can help to unravel how these kinds of rude language models affect each other.

References

1. Aggrawal, N.: Detection of offensive tweets: A comparative study. Computer Reviews Journal **1**(1), 75–89 (2018)
2. Álvarez-Carmona, M.Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H.J., Villasenor-Pineda, L., Reyes-Meza, V., Rico-Sulayes, A.: Overview of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook papers of 3rd sepln workshop on evaluation of human language technologies for iberian languages (ibereval), seville, spain. vol. 6 (2018)
3. Amedie, J.: The impact of social media on society (2015)
4. Aragón, M., Jarquín, H., Gómez, M.M.y., Escalante, H., Villaseñor-Pineda, L., Gómez-Adorno, H., Bel-Enguix, G., Posadas-Durán, J.: Overview of mex-a3t at iberlef 2020: Fake news and aggressiveness analysis in mexican spanish. In: Notebook Papers of 2nd SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Malaga, Spain (2020)
5. Castro, S., Chiruzzo, L., Rosá, A., Garat, D., Moncecchi, G.: A crowd-annotated spanish corpus for humor analysis. In: Proceedings of SocialNLP 2018, The 6th International Workshop on Natural Language Processing for Social Media (2018)
6. Chiruzzo, L., Castro, S., Etcheverry, M., Garat, D., Prada, J.J., Rosá, A.: Overview of HAHA at IberLEF 2019: Humor Analysis based on Human Annotation. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings, CEUR-WS, Bilbao, Spain (9 2019)

7. ElSherief, M., Kulkarni, V., Nguyen, D., Wang, W.Y., Belding, E.: Hate lingo: A target-based linguistic analysis of hate speech in social media. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 12 (2018)
8. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR) **51**(4), 1–30 (2018)
9. Graff, M., Miranda-Jiménez, S., Tellez, E.S., Moctezuma, D.: Evomsa: A multilingual evolutionary approach for sentiment analysis. Computational Intelligence Magazine **15**, 76 – 88 (Feb 2020), https://ieeexplore.ieee.org/document/8956106
10. Liu, N., Zhang, B., Yan, J., Chen, Z., Liu, W., Bai, F., Chien, L.: Text representation: from vector to tensor. In: Fifth IEEE International Conference on Data Mining (ICDM'05). pp. 4 pp.– (2005). https://doi.org/10.1109/ICDM.2005.144
11. Matamoros-Fernández, A., Farkas, J.: Racism, hate speech, and social media: A systematic review and critique. Television & New Media **22**(2), 205–224 (2021)
12. Mladenović, M., Ošmjanski, V., Stanković, S.V.: Cyber-aggression, cyberbullying, and cyber-grooming: A survey and research challenges. ACM Computing Surveys (CSUR) **54**(1), 1–42 (2021)
13. Mondal, M., Silva, L.A., Benevenuto, F.: A measurement study of hate speech in social media. In: Proceedings of the 28th ACM Conference on Hypertext and Social Media. p. 85–94. HT '17, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3078714.3078723, https://doi.org/10.1145/3078714.3078723
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. Ortiz-Bejar, J., Tellez, E.S., Graff, M., Moctezuma, D., Miranda-Jiménez, S.: Ingeotec at iberlef 2019 task haha. In: IberLEF@ SEPLN. pp. 203–211 (2019)
16. i Orts, Ò.G.: Multilingual detection of hate speech against immigrants and women in twitter at semeval-2019 task 5: Frequency analysis interpolation for hate in speech detection. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 460–463 (2019)
17. 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)
18. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media. pp. 1–10 (2017)
19. Siddiqui, S., Singh, T., et al.: Social media its impact with positive and negative aspects. International Journal of Computer Applications Technology and Research **5**(2), 71–75 (2016)
20. Tellez, E.S., Miranda-Jiménez, S., Graff, M., Moctezuma, D., Suárez, R.R., Siordia, O.S.: A Simple Approach to Multilingual Polarity Classification in Twitter. Pattern Recognition Letters (2017). https://doi.org/10.1016/j.patrec.2017.05.024, http://www.sciencedirect.com/science/article/pii/S0167865517301721
21. Tellez, E.S., Moctezuma, D., Miranda-Jiménez, S., Graff, M.: An automated text categorization framework based on hyperparameter optimization. Knowledge-Based Systems **149**, 110–123 (2018), https://github.com/INGEOTEC/microtc

22. Vidgen, B., Yasseri, T.: Detecting weak and strong islamophobic hate speech on social media. Journal of Information Technology & Politics **17**(1), 66–78 (2020). https://doi.org/10.1080/19331681.2019.1702607, `https://doi.org/10.1080/19331681.2019.1702607`
23. Ziems, C., He, B., Soni, S., Kumar, S.: Racism is a virus: Anti-asian hate and counterhatein social media during the covid-19 crisis. arXiv preprint arXiv:2005.12423 (2020)

A Source code of our model

```
### Installing EvoMSA
# Full instructions in https://github.com/INGEOTEC/EvoMSA

# EvoMSA can be easly install using anaconda
conda install -c ingeotec EvoMSA

# or can be install using pip, it depends on numpy, scipy, scikit-learn and b4msa
pip install cython
pip install sparsearray
pip install evodag
pip install EvoMSA

### Usage

import pandas as pd
import sklearn.model_selection as model_selection

# load datasets
tweets_mx = pd.read_csv('mx-train-data-non-contextual.csv', names=['texto'])
out_mx = pd.read_csv('mx-train-outputs.sol', names=["clase"])

# Split 80/20
X_train, X_test, y_train, y_test =
    model_selection.train_test_split(tweets_mx, out_mx, train_size=0.80)

# Load the enhanced models
from EvoMSA.utils import download
from EvoMSA.base import EvoMSA

# pre-trained models
haha = download('haha2018_Es.evomsa')
mexa3t = download('mexa3t2018_aggress_Es.evomsa')
misoginia = download('misoginia_Es.evomsa')
```

```
# reverse text model
from EvoMSA.model import TextModelInv

# create EvoMSA model with enhaced models; including Emoji space.
# uses sklearn.naive_bayes.GaussianNB as stacked classifier
evo = EvoMSA(TR=True, B4MSA=False, lang='es', Emo=True,
        stacked_method='sklearn.naive_bayes.GaussianNB',
        models=[
            [TextModelInv,"sklearn.svm.LinearSVC"],
            [misoginia, "sklearn.svm.LinearSVC"],
            [haha, "sklearn.svm.LinearSVC"],
            [mexa3t, "sklearn.svm.LinearSVC"]
        ])

# train EvoMSA model
evo.fit(X_train, y_train)

# Make prediction with the test dataset
pred = evo.predict(X_test)

# report resulting metrics
from sklearn import metrics
from sklearn.metrics import classification_report

classification_report(y_test, pred, digits=4)
```