# A Machine Learning Approach for Detecting Aggressive tweets in Spanish

Helena Gómez-Adorno, Gemma Bel-Enguix
Gerardo Sierra, Octavio Sánchez, and
Daniela Quezada

Universidad Nacional Autónoma de México (UNAM),
Engineering Institute (II), Mexico City, Mexico
{hgomeza,gbele,gsierram}@iingen.unam.mx
oct_sanc@unam.mx,danielaqu9@gmail.com

**Abstract.** This paper presents our approach to the aggressive detection track at MEX-A3T 2018. The track consists in identifying whether a tweet is aggressive or not. To solve this task we follow a machine learning approach, we trained the logistic regression algorithm on linguistically motivated features, and several types of $n$-grams. We applied several pre-processing steps for standardizing tweets in order to capture relevant information. Our best run achieved 42.85% of f-score on the aggressiveness class, which is between 30% and 40% less than our best cross-validation result on the training set.

**Keywords:** Aggressiveness detection· Machine learning · Logistic regression

## 1 Introduction

Due to the increase of cyberbullying against social media users, the automatic detection of aggressive behavior in these platforms is gaining a lot of attention. Aggressive text detection is the first step towards cyberbullying automatic identification.

The MEX-A3T 2018 [3] workshop launched this year the aggressive detection track in Mexican Spanish tweets. The aim is to increase the research flow in such a critical topic. The organizers of the evaluation lab define an aggressive tweet as a message that tends to disparage or humiliate a person or a group of persons.

From a machine-learning perspective, the task can be seen as a binary classification problem. We experimented with various machine learning algorithms: Support Vector Machines (SVM), multinomial naive Bayes, and logistic regression. As features, we extracted linguistically motivated patterns, and several types of $n$-grams (character, words, POS, and aggressive words). Bellow we explain the pre-processing steps, and the experiments carried out to solve this task.

## 2 Corpus

The organizers collected tweets with a distribution of 75% of non-aggressive and 25% of aggressive messages. The corpus was splitted into training (70%) and testing (30%) partitions. In the training corpus, the distribution is 4973 (64.58%) non-aggressive tweets and 2727 (35.42%) aggressive tweets. This shows that the quantity of the aggressive tweets (the class of interest) was the half of the non-aggressive tweets. This is usual in many text classification problems. According with the data provided by the organizers, the test corpus distribution would be severely unbalanced, with just 23 aggressive tweets. This represent a 0.70% of the test corpus, which mean that the non-aggressive would be 99.30%.

The data set for this track was collected between August and November 2017 [3]. All tweets should contain at least one word considered *vulgar* or *insult* by [1]. Based on a manual labeling, which stated that an offensive message would be humiliating a person or a group of persons or disparaging them, the tweets were labeled by human annotators.

Some pre-processing was performed over the corpus: 1) all the user handles were anonymized and set to @USUARIO and 2) urls were reduced to a <URL> tag. However, some urls were not stripped as some user handles were not substituted. Tweets and labels were kept on separate files.

## 3 Methodology

### 3.1 Pre-processing

Previous research show that pre-processing is useful for several natural language processing (NLP) task, specially when the corpus is composed of social media data [7, 12]. Before the extraction of features, we apply the following pre-processing steps aiming to enhance $n$-grams representation and to reduce part-of-speech (POS) tagging errors:

1. **Lowercase:** This allowed us to improve the POS tagging process.
2. **Digits:** Since the numbers do not carry semantic information, we replace them by a single symbol ( e.g., 1,599 $\rightarrow$ 0,000).
3. **Mentions:** We only remove the @ symbol from the *@USUARIO* label because keeping the mention (without the @) improves the POS tags features.
4. **Pics:** The picture links are also replaced by a single symbol (link $\rightarrow$ 1).
5. **Slangs:** Following the process presented in [7], we replaced slang words by their standardized version from the Spanish social media lexicon.

### 3.2 Features

**Language Patterns.** We have performed a linguistic analysis of the training corpus to identify the language patterns that can help to distinguish aggressivity, detecting two types of them: a) morphological structures; b) some recurrent lexical items, usually combined with some morphological category.

The patterns which are at least the 50% more represented in the aggressive samples have been considered. The morphological combinations are the following:
- `Vb 2p + prep`
- `NC + Adj calif`
- `^NC + anything`
- `NCM + NCM`
- `Fz`
- `possessiv 'tu'+NC`

The lexical patterns are the following:
- `pinche+NC`
- `puto(s)+NC`
- `puto(s)+AQ`

In both cases, the set of patterns was treated as an only feature, and the values for each one were added to the total.

**Character $n$-grams** are language-independent features that have proved to be highly predictive for several natural language processing tasks [13], hence we examined them as single features and in combination with others. The submitted approaches include the variation of $n$ from 3 to 7.

**Word $n$-grams.** In our experiments we found that the combination of word $n$-grams with $n$ varying from 2 to 5 helped to improve our cross-validation results.

**POS tags $n$-grams** are sequences of continuous part-of-speech (POS) tags. They capture syntactic information and are useful, for example, for identifying user's intentions on tweets [8]. In this work we use bigrams of POS tags.

**Aggressive words $n$-grams** are also commonly used features for aggressiveness detection. In our submission we used bigrams of aggressive words, because of the nature of the corpus the use of aggressive words in isolation does not help for discriminating aggressive tweets.

### 3.3 The SMOTE technique

As described in section 2, the training corpus was not balanced. Bayesian methods take this imbalance into account as priors, however, it has been shown [10, 2, 4] that for some text classification tasks, a balanced corpus performs better.

One well known oversampling technique is the Synthetic Minority Oversampling Technique (SMOTE) [5, 6]. This technique shows an improvement on performance of the classifier (measured in ROC space) than just undersampling or modifying Bayes priors.

In this technique, the minority class is over-sampled by creating synthetic examples. These new samples are created by taking each minority class sample and pointing to new samples that occur along the line segments between the $k$ nearest data points of the minority class. The new samples are created by the difference between feature vector of the real sample and its nearest neighbor. This difference is then multiplied by a random number between 0 and 1. The result of this operation is then added to the feature vector of the real sample. This creates a new data point between the two considered samples.

### 3.4 Classification

We used the logistic regression algorithm[1], which showed better performance than other machine-learning algorithms we examined: SVM and multinomial naive Bayes. We performed 10-fold cross-validation experiments for selecting the best features, weighting scheme, and frequency threshold. The final configuration of our system implements a binary weighting scheme, and considered only those features that occur at least 10 times in the entire corpus and that occur in at least 2 documents in the corpus.

## 4 Results

The performance measure of the aggressive detection track is the F1-score on the aggressive class. Table 1 shows the 10-fold cross-validation results on the training corpus, as well as the official results on the testing corpus. The table also shows a baseline when all instances are predicted as the aggressive class and the overall results of the best performing teams in the shared task. We achieved the $5^{\text{th.}}$ best result with the *run 2*.

It can be observed that *run 2* (without SMOTE) showed higher results on the test corpus, 42.85%. However, in the training corpus, *run 1* achieved results up to 10% higher than the *run 2*. The obtained results of both runs are much higher than the baseline.

**Table 1.** Results under 10-fold cross-validation on the training corpus (Train) and the official results on the testing corpus (Test). Both in terms of F1-score (%).

| **Run** | **Train** | **Test** |
|---|---|---|
| Shared task $1^{\text{st.}}$ (INGEOTEC) | – | 48.83 |
| Shared task $2^{\text{nd.}}$ ($\text{CGP}_{Team}$) | – | 45.00 |
| Shared task $3^{\text{rd.}}$ (GeoInt-b4msa) | – | 43.40 |
| Shared task $4^{\text{th.}}$ (aragon-lopez) | – | 43.12 |
| run 1 (with SMOTE) | **85.53** | 40.20 |
| run 2 (without SMOTE) | 74.32 | **42.85** |
| baseline (aggressive class) | 52.31 | 01.38 |

## 5 Conclusions

We presented our approach for detecting aggressive tweets in Mexican Spanish. We trained a logistic regression classifier on a combination of linguistic patterns, aggressive words lexicon, and different types of $n$-grams (character, words, and POS tags). Our best run achieved 42.85% F1-score on the aggressive class. We

[1] The scikit-learn [11] implementation

used an oversampling technique (SMOTE) to overcome the problem of unbalanced data which allowed us to achieve better results in the training corpus, but it did not generalized well on the testing corpus. We achieved the $5^{\text{th.}}$ place out of 12 participating systems.

One of the directions for future work is to tackle the unbalance problem with a deeper analysis of the SMOTE process. We will also examine the use of our linguistic patterns in an statistical-based approach [9], which achieved outstanding results in another NLP problem.

## Acknowledgments

This work was possible thanks to the funding of CONACYT fellowship 387405, CONACYT project number 002225, DGAPA-PAPIIT projects IN403016 and IA400117.

## References

1. Academia Mexicana de la Lengua: Diccionario de Mexicanismos. Academia Mexicana de la Lengua (2016)
2. Aggarwal, C.C.: Outlier analysis. In: Data mining. pp. 237–263. Springer (2015)
3. Álvarez-Carmona, M.Á., Guzmán-Falcón, E., Montes-y Gómez, M., Escalante, H.J., Villaseñor-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 $3^{\text{rd.}}$ SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September (2018)
4. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A.: Mean birds: Detecting aggression and bullying on twitter. In: Proceedings of the 2017 ACM on Web Science Conference. pp. 13–22. ACM (2017)
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)
6. Fernández, A., Garcia, S., Herrera, F., Chawla, N.V.: Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. Journal of Artificial Intelligence Research **61**, 863–905 (2018)
7. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Sanchez-Perez, M.A., Chanona-Hernandez, L.: Improving feature representation based on a neural network for author profiling in social media texts. Computational Intelligence and Neuroscience **2016**, 13 pages (October 2016)
8. Gómez-Adorno, H., Pinto, D., Montes, M., Sidorov, G., Alfaro, R.: Content and style features for automatic detection of users' intentions in tweets. In: Ibero-American Conference on Artificial Intelligence. pp. 120–128. Springer (2014)
9. Markov, I., Gómez-Adorno, H., Sidorov, G., Gelbukh, A.: The winning approach to cross-genre gender identification in Russian at RUSProfiling 2017. In: FIRE 2017 Working Notes. FIRE'17, vol. 2036, pp. 20–24. CEUR-WS.org, Bangalore, India (December 2017)
10. Mladenic, D., Grobelnik, M.: Feature selection for unbalanced class distribution and naive bayes. In: ICML. vol. 99, pp. 258–267 (1999)

11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
12. Pinto, D., Vilarino, D., Alemán, Y., Gómez, H., Loya, N.: The soundex phonetic algorithm revisited for sms-based information retrieval. In: II Spanish Conference on Information Retrieval CERI (2012)
13. Sanchez-Perez, M.A., Markov, I., Gómez-Adorno, H., Sidorov, G.: Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 145–151. Springer (2017)