Detection of Aggressive Tweets in Mexican Spanish Using Multiple Features with Parameter Optimization

Germán Ortiz[1], Helena Gómez-Adorno[2][0000−0002−6966−9912], Jorge Reyes-Magaña[1,3][0000−0002−8296−1344], Gemma Bel-Enguix[1][0000−0002−1411−5736], and Gerardo Sierra[1][0000−0002−6724−1090]

[1] Instituto de Ingeniería, Universidad Nacional Autónoma de México, México
{jortizb,gbele,gsierram}@iingen.unam.mx
[2] Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, México
helena.gomez@iimas.unam.mx
[3] Universidad Autónoma de Yucatán, Mérida, Yucatán, México
jorge.reyes@correo.uady.mx

CEUR-WS.org/Vol-2421/MEX-A3T_paper_5.pdf

Abstract. This paper explains our approach to Aggressiveness Identification in the MEX-A3T shared task, whose aim is the detection of aggressive tweets. The task proposes a binary classification for every tweet: aggressive and non-aggressive. We approached the problem using linguistically motivated features and several types of n-grams (words, characters, functional words, punctuation symbols, among others). We trained a Support Vector Machine using a combinatorial framework that optimizes the results of the classifier. Our best run achieved a F1-score of 0,4549, which is the 5$^{\text{th}}$ best among the twenty-six runs.

Keywords: Aggressiveness detection · Support Vector Machine · Machine learning.

1 Introduction

Aggressiveness is an emotional state that consists of hate feelings and desires to physically or psychologically hurt a person or group of people. Nowadays, communication through social networks plays a crucial role in society life. Social Networking Services open a whole world of possibilities, but they also represent a significant threat, since users are exposed to many risks and attacks; among them aggressive comments, which can cause short-term and long-term damage to victims.

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). IberLEF 2019, 24 September 2019, Bilbao, Spain.

For the second year in a raw, the MEX-A3T 2019 workshop [2] launched the aggressiveness detection track in Mexican Spanish tweets with the aim of promoting research on the analysis of the content of social networks in this language. For this task, the organizers define an aggressive tweet as follows: it *contains messages that despise or humiliate a person or group of people, using the following elements: nicknames, jokes or derogatory adjectives.* Our approach uses a Machine Learning perspective in which the problem results in a binary classification, between aggressive or not. To do this, we use the Support Vector Machine (SVM) algorithm as a classifier. For feature extraction, different types of n-grams were used (n-grams of words, n-grams of characters, skipgrams, among others).

2 Related work

In recent years, the automatic detection of aggressive behavior in social media is gaining a lot of attention.

Our approach is based on previous work on hate speech detection in twitter [3] and aggressive detection of tweets in Mexican Spanish [6], which were presented in the MEX-A3T 2018 Workshop [1], and the SemEval-2019 Workshop, respectively. The former follows a classical machine learning approach, in which a logistic regression algorithm is trained on linguistically motivated characteristics and various types of n-grams. The latter uses a Support Vector machine as classifier with a combinatorial framework for parameter optimization.

Concerning to aggressiveness detection related work, [8] classifies Facebook comments using three deep learning architectures, Convolutional Neural Networks, Long Short Term Memory networks, and Bi-directional Long Short Term Memory networks and a majority voting-based ensemble method to combine them.

Djuric et al. [5] used the generated list to annotate a publicly available corpus of more than 16k tweets. They analyzed the impact of various extra-linguistic features along with character n-grams for the detection of hate speech. In turn, they elaborated a dictionary based on the most indicative words in their data.

Chatzakou et. al [4] studied the properties of bullies and aggressors, and found that stalkers post with less frequency, participate in fewer online communities and are less popular than users with standard models of behaviour. Their research shows that machine learning classification algorithms can accurately detect users who exhibit bullying and aggressive behavior, with more than 90% of accuracy.

3 Corpus

The corpus was collected between August and November 2017. The training dataset has 7700 tweets, with a distribution of 35% of aggressive messages and 65% non-aggressive messages, keeping the texts and labels on separate files.

Aggressive tweets contain at least one word considered vulgar or insulting based on a Mexicanisms dictionary. The dataset was manually labeled by two taggers following the premise that an aggressive message pretends to humiliate a person or people with jokes or derogative adjectives.

In the corpus, all user handlers were replaced by @USUARIO and all URL's were replaced by <URL>.

4 Methodology

This section shows in detail the processing that was carried out in the corpus to subsequently perform the classification task. This is a very important stage to maximize the classifier performance, as well as allow to manipulate the data in a simplified way. In addition, text representation features are described.

4.1 Pre-processing

- **Diacritic symbols:** These were removed to avoid composed symbols, that are an errors source in informal texts.
- **Text normalization:** Tweets were standardized to lowercase to avoid multiple copies of the same words along the corpus.
- **Abbreviations replacement:** Abbreviations, contractions and slangs were replaced by the original text using a social networks-based dictionary [7].
- **Emojis** were removed.

4.2 Classifier

We used a combinatorial framework (μTC) developed by [9]. The framework approaches any text classification task as a combinatorial optimization problem; where there is a search space containing all possible combinations of different text pre-processing methods, text features and weighting schemes with their respective parameters, and, on this search space, a local search-based metaheuristic is used to search for a configuration that produces a highly effective text classifier. Considering all the combinations established in the implementation of (μTC), we optimized the features described in Section 4.3. Once the best configuration was selected, we trained an SVM with a linear kernel.

Different from previous work [3] where the features added to (μTC) are static, that is, the feature sets that are not considered in the (μTC) framework were selected based on their individual classification performance. Once the best configuration space was found, all n-grams types with all n variations are added to the final vector for each text. In our approach, all features were included and optimized in the (μTC) framework, which selects only those feature sets that are likely to offer the best classifier performance.

4.3 Features

Beside the features already considered in the μTC framework we took into account other features such as the one mentioned bellow:

- **Character n-grams:** They are language-independent powerful features for many natural language processing tasks where many words are likely to be poor written. For our approach a variation of n from 3 to 5 is used.
- **Word n-grams:** These features capture the identification of a word and its possible neighbors. We use a variation of n from 3 to 5.
- **Aggressive words n-grams:** In our approach we manually collected an aggressive words lexicon obtained from the web and some word extracted from the training corpus. Variation of n from 2 to 3 is used.
- **Skipgrams:** For our approach we capture 2-words groups with skips from 2 to 4 words.
- **Stopwords n-grams:** We use the stopwords list from NLTK library to build them, with a variation of n from 2 to 4. Stopwords frequencies are one of the best features to detect aggressiveness messages.
- **Punctuation-symbols n-grams:** These n-grams helps to detect patterns in aggressiveness analysis. We use a variation of n from 2 to 5 to build them.

5 Results

The system performance in the aggressiveness detection track was measured using F1-score on aggressive class. Table 5 shows results for the best run on the training corpus with 10-fold cross-validation using static and optimized features (as we describe above), along with the evaluation phase official results on the test corpus.

Position	Team	Training	Eval
1	INGEOTEC	-	0.4796
2	Casavantes	-	0.4790
3	GLP	-	0.4749
4	**mineriaUNAM (optimized)**	0.7438	0.4549
6	**mineriaUNAM (static)**	0.7433	0,4516
7	LyR	-	0.4288
8	Victor	-	0.4081

Table 1. Results of Aggressiveness detection task of training phase and the evaluation phase (Eval) official results

In the final configuration space, besides the features already considered in (μTC), from our additional feature sets just punctuation symbols n-grams were used with $n = 5$, while the other features are ignored.

The results we obtained in the 2019 edition were clearly better than the ones from 2018. We improved our results from 0.4285 to 0.4549. The main difference

was the use of (μTC) [9]. This means that this combinatorial framework is a good complement that helps to optimize the feature set for the classification process.

6 Conclusions

We presented an approach for aggressiveness detection in Mexican Spanish tweets.

We trained a Support Vector Machine using a combinatorial framework (μTC), to which we added different types of n-grams such as punctuation symbols n-grams, stop-words n-grams, and aggressive words n-grams to be optimized. The results we obtained are better than the ones obtained last year, achieving an improvement from 0.4285 to 0.4549 on the F1-score on the aggressive class.

In addition, the obtained results in this task were improved by the optimization of extra features added in previous work [3].

References

1. 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 3rd. SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL), Seville, Spain, September (2018)
2. Aragón, M.E., Álvarez-Carmona, M.Á., Montes-y Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Moctezuma, D.: Overview of mex-a3t at iberlef 2019: authorship and aggressiveness analysis in mexican spanish tweets. In: Notebook Papers of 1st SEPLN Workshop on Iberian Languages Evaluation Forum (IberLEF), Bilbao, Spain, September (2019)
3. Argota, L., Reyes-Magaa, J., Gómez-Adorno, H., Bel-Enguix, G.: MineriaUNAM at SemEval-2019 Task 5: Detecting Hate Speech in Twitter using Multiple Features in a Combinatorial Framework. In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Association for Computational Linguistics (2019)
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. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th international conference on world wide web. pp. 29–30. ACM (2015)
6. Gómez-Adorno, H., Bel-Enguix, G., Sierra, G., Sánchez, O., Quezada, D.: A machine learning approach for detecting aggressive tweets in spanish. In: Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceedings (2018)
7. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J.P., 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**, 2 (2016)

8. Madisetty, S., Sankar-Desarkar, M.: Aggression detection in social media using deep neural networks. In: Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (2018)
9. 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://doi.org/10.1016/j.knosys.2018.03.003