UACh at MEX-A3T 2019: Preliminary Results on Detecting Aggressive Tweets by Adding Author Information Via an Unsupervised Strategy

Marco Casavantes, Roberto López, and Luis Carlos González

Universidad Autónoma de Chihuahua. Facultad de Ingeniería. Chihuahua, Chih., México
{p271673,jrlopez,lcgonzalez}@uach.mx

Abstract. In this paper we describe our participation for the Aggressiveness Detection Track in the second edition of MEX-A3T. We evaluate different strategies for text classification, including classifiers such as Support Vector Machines and a Multilayer Perceptron trained on n-grams (words and characters) and word embeddings. We also study the inclusion of features to try to give context to the text messages and explore if people verbally attack differently depending on their traits and overall environment. Preliminary results show that our strategy is competitive to detect aggression in tweets, ranking in 2nd place with respect to the participants of 2018 and 2019.

Keywords: Spanish text classification · Aggressiveness Detection · Multilayer Perceptron.

1 Introduction

Technology has changed the way in which people communicate with each other, giving rise to new services such as social networks, where a style of informal communication is used. Such social networks, though, present several challenges to maintain communication channels open to the free sharing of ideas. The intolerance and aggressiveness of certain users affects the experience of other consumers or people interested in being part of the communities and their conversations. The fact of not being face to face in the communication channel and even preserve anonymity, encourages these individuals to express themselves offensively. However, the volume of messages that are sent daily, the growth of online communities, and the respective ease of access to these social networks, make the moderation of communication channels a difficult task to be dealt with by conventional means, and as people increasingly communicate online, the need for

high quality automated abusive language classifiers becomes much more profound[1].

One of the goals of the second edition of MEX-A3T[2] is to tackle this problem and further improve the research of this important NLP task, the detection of aggressive tweets in Mexican Spanish. In this work we evaluate strategies proposed before, such as the use of lexical features through TF-IDF representations, and different approaches to add features in order to try to give context to each text. Surprisingly, even tackling the task with such a basic approach our proposal is able to offer competitive results, just slightly behind the top performer of this competition in 2018 and 2019, INGEOTEC. Furthermore, we also investigate how to incorporate author's traits by using unsupervised methods and attempting to include this information as possible features, based on the hypothesis that there are different ways of aggression depending on the author's context.

2 Proposed Method

2.1 Data Pre-processing

After loading the train and test sets, we strip the tweets from non-alphanumeric characters and only keep some relevant Spanish characters (á,é,í,ó,ú,ñ,and ü), all words are then made lowercase and subsequently we noticed that in both sets there exists many different terms to express laughter (mainly due to how many times "ja" is repeated when the word "jaja" appears and because of typos) so that led us to replace every word containing "jaja" to "risa" (laugh), with the purpose of decreasing the number of terms that represent this emotion.
It is worth mentioning that we also created and conducted experiments on a version of the datasets where emojis were converted to text and hashtags were separated by words (e.g., ":)" would turn into "smiling face", and "#FelizMiércoles" would be "feliz miércoles"), however most hashtags were wrongly separated and the performance of the classifiers decreased by incorporating these steps and were therefore discarded.

2.2 Features

We conducted our research using the following features:
Lexical: We use word n-grams (n=1, 2) and char n-grams (n=3, 4) as features, this collection of terms is weighted with its term frequency-inverse document frequency (TF-IDF).
Document Embeddings: The objective was to represent the tweets through Word Embeddings[3] and try different classifiers with these new features, each text message was converted to a vector of size = 300 (mean of the vectors of each word). The model of words in Spanish was computed with fastText[4] and downloaded from [5] .
User Occupation and Location predictions: Although we attempted several strategies to obtain unsupervised author profiles for each document [6], we

ended up using the output of the system developed by [7] as predictions of occupation and location values to explore the possibility of differences in vocabulary that exists according to the profile of the author of the message.
Grouping tweets by theme: An implementation of Self Organizing Maps (SOM) as a clustering strategy called MiniSom[8] was used with aims to find groups in the collection of texts based on underlying or non-explicit features, the clustering was done including all words and also ignoring swear words (to reduce the noise and focus on thematic terms), after training the network we were able to compute the coordinates assigned to a tweet on the map and use these as new features.
Perspicuity score / Inflesz scale: Based on [9], we adapted the idea of capturing the quality of each tweet by using a modified Flesch Reading Ease score (since this test only applies to text written in English), called Perspicuity score and its equivalence to the Inflesz scale, following the equation described in [10] where the number of sentences is also fixed at one.

All the extra categorical features mentioned above were concatenated following a One Hot Encoding scheme.

3 Experiments and Results

The datasets were provided by MEX-A3T Team. Table 1 shows the distribution of training and test partitions for Spanish tweets.

Table 1. Data distribution for Spanish tweets corpus

Class	Training	Test
Aggressive	2727	N/A
Non-aggressive	4973	N/A
Total	7700	3156

We separated the training set in 67% for training and 33% for validation to evaluate our experiments with different combinations of features discussed in section 2.2. We started our research by recreating the baselines described in the overview of the first edition of MEX-A3T[11], particularly focusing on the character trigrams baseline, as it holds the best performance in comparison to the BoW baseline.
We trained Linear Support Vector Machines and a Multilayer Perceptron as classifiers for this task, and we decided to use the perceptron as the final system to submit our predictions since it exhibited the best results in the validation stage, as shown in Table 2 where we obtained the F1-score macro and the F-Measure over the aggressive class. We performed all modeling regarding the creation of tf-idf feature matrices and SVM classifiers using scikit-learn[12], and for the Multilayer Perceptron, we used the implementation described in [13],

there was only an instance were this Perceptron couldn't be trained with Word Embeddings, so we tried another configuration on the MLPClassifier from scikit-learn getting low scores similar to the ones obtained using LinearSVM, and therefore casting aside this approach.

3.1 Results

As stated before, the Multilayer Perceptron was chosen as final system, however, because of time and memory constraints we had to train this model using only character n-grams of range [3,4] for this task even though later results shows better performance by using n-grams of range [3,5]. Table 3 list the top five final rankings for the aggressiveness detection task for 2019, more details of all results of the contest are shown at [2]. It is interesting to observe that even when our system relied on such a basic approach, it is able to compete face-to-face against INGEOTEC, a model based on an ensemble of classifiers, which specially tailors discriminative features for aggressive detection via a Genetic Programming strategy.

3.2 Analysis

To breakdown our results, we started by getting the 10 most valuable n-grams at character level separated by length, as shown in Table 4. With respect to the aggressive class, our final configuration had more false positives than false negatives, meaning that it was easier for an aggressive tweet to be missclassified as non-aggressive than the other way around. Despite running several experiments and adding new features trying to give context to the tweets, in hopes of improving classification in this task, unfortunately these strategies showed, at best, almost unnoticeable changes in the results, and hinder of classification at worst. After manual inspection, we observed that this could have happened because:

- Occupation and Location predictions did not group the messages in a balanced way, in fact, most tweets would fall under only one out of eight available categories for occupation and six categories for location.
- SOM Coordinates would not enhance the classification scores before as the clusters were capturing word repetition instead of thematic aspects for each tweet. Later experiments (after submission of results) showed that this behaviour was caused because the clustering was made with n-grams; training the SOM with word embeddings created with the train set of this task (without external resources) solved this issue and did a better job at grouping the tweets by subjects.
- There was no relevant pattern by applying a perspicuity score to each tweet, as there were multiple cases of similar scores assigned to both aggressive and non-aggressive messages.

Table 2. Detailed classification with F1-scores in the validation stage.

Added features	**Classifier**	**Char n-gram range**	**F1-score macro**	**F1-score** (aggressive_class)
None	LinearSVM	[3,3]	0.76	0.68
None	MLP	[3,3]	0.77	0.66
None	LinearSVM	[3,4]	0.77	0.69
None	MLP	[3,4]	0.79	0.69
None	LinearSVM	[3,5]	0.77	0.70
None	MLP	[3,5]	0.79	0.70
Word Embeddings	LinearSVM	N/A	0.59	0.39
Word Embeddings	MLPClassifier	N/A	0.56	0.34
Occupation (O)	LinearSVM	[3,3]	0.76	0.69
Occupation (O)	MLP	[3,3]	0.78	0.67
Occupation (O)	LinearSVM	[3,4]	0.77	0.69
Occupation (O)	MLP	[3,4]	0.76	0.68
Location (L)	LinearSVM	[3,3]	0.77	0.68
Location (L)	MLP	[3,3]	0.77	0.65
Location (L)	LinearSVM	[3,4]	0.77	0.70
Location (L)	MLP	[3,4]	0.76	0.67
Perspicuity (P)	LinearSVM	[3,3]	0.76	0.68
Perspicuity (P)	MLP	[3,3]	0.77	0.66
Perspicuity (P)	LinearSVM	[3,4]	0.77	0.70
Perspicuity (P)	MLP	[3,4]	0.76	0.67
SOM Coordinates (S)	LinearSVM	[3,3]	0.76	0.68
SOM Coordinates (S)	MLP	[3,3]	0.78	0.66
SOM Coordinates (S)	LinearSVM	[3,4]	0.76	0.69
SOM Coordinates (S)	MLP	[3,4]	0.79	0.69
O + L + P + S	LinearSVM	[3,3]	0.76	0.69
O + L + P + S	MLP	[3,3]	0.78	0.67
O + L + P + S	LinearSVM	[3,4]	0.77	0.69
O + L + P + S	MLP	[3,4]	0.77	0.69

Table 3. Final scores of the aggressiveness detection task.

Rank	**Team**	**F1-score** (aggressive_class)	**F1-score** (non-aggressive_class)	**Accuracy**
1	INGEOTEC	0.4796	0.8131	0.7250
2	Casavantes (Our approach)	0.4790	0.8164	0.7285
3	GLP (run 2)	0.4749	0.7949	0.7050
4	GLP (run 4)	0.4635	0.7774	0.6854
5	mineriaUNAM (run 2)	0.4549	0.8016	0.7075

Table 4. Best n-grams at character level in training set

Length	n-gram	Frequency in aggressive class	Frequency in non aggressive class
3 chars	'os '	3074	3207
	' de'	2571	3879
	'as '	2205	3252
	'que'	1991	3540
	' qu'	1965	3667
4 chars	' de '	1860	2798
	'que '	1768	3262
	' que'	1649	2954
	' put'	1589	1517
	' la '	1062	2195

4 Conclusions and Future Work

In this paper, we describe our strategy to classify aggressive and non-aggressive tweets in Mexican Spanish. In our best performing system, we use only lexical features and our results show a better performance than most results of all participants. This outcome, and the fact that the F-measure for the aggressive class is still low compared to the score on the non-aggressive class, motivates the idea of future work focusing on feature analysis for aggressiveness detection and explore which representations are truly relevant, including word embeddings, Bag of Words and Characters of different n-gram ranges, see if these complement each other and if so, how to combine them. We analyzed our clustering strategies, and after changing the way they were trained we could observe slight improvement in classification results, motivating us to keep experimenting on ways to try to add context to the text messages. We also believe in the potential that neural networks display for this task, and that more research on how to build and train them properly will certainly improve the current situation of this task.
As future work, we look forward to develop new strategies based on deep neural networks, such as Recurrent Neural Networks, which are tools aimed to work with sequential data similar in nature to time series.

References

1. Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee.
2. Mario Ezra Aragón, Miguel Á Álvarez-Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, and Daniela Moctezuma. 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. Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1188–II–1196. JMLR.org, 2014.
4. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
5. Github - mquezada/starsconf2018-word-embeddings: Material para el taller "representaciones vectoriales de palabras basadas en redes neuronales" de la starsconf 2018. https://github.com/mquezada/starsconf2018-word-embeddings. (Accessed on 06/02/2019).
6. Roberto López Santillán, L.C. González-Gurrola, and Graciela Ramírez-Alonso. Custom document embeddings via the centroids method: Gender classification in an author profiling task. In Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors, *CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 10-14 September, Avignon, France*. CEUR-WS.org, September 2018.
7. Rosa Marıa Ortega-Mendoza and A Pastor López-Monroy. The winning approach for author profiling of mexican users in twitter at mex. a3t@ ibereval-2018.
8. Github - justglowing/minisom: Minisom is a minimalistic implementation of the self organizing maps. https://github.com/JustGlowing/minisom. (Accessed on 06/03/2019).
9. Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017.
10. Escala inflesz — legible. https://legible.es/blog/escala-inflesz/. (Accessed on 06/02/2019).
11. Miguel Álvarez-Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, Verónica Reyes-Meza, and Antonio Rico-Sulayes. Overview of MEX-A3T at IberEval 2018: Authorship and aggressiveness analysis in Mexican Spanish tweets. *CEUR Workshop Proceedings*, 2150:74–96, 2018.
12. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
13. Github - afshinrahimi/sparsemultilayerperceptron: Lasagne / theano based multilayer perceptron mlp which accepts both sparse and dense matrices and is very easy to use with scikit-learn api similarity. https://github.com/afshinrahimi/sparsemultilayerperceptron. (Accessed on 06/03/2019).