

GRU with Author Profiling Information to Detect Aggressiveness

María Guadalupe Garrido-Espinosa^a, Alejandro Rosales-Pérez^a and Adrián Pastor López-Monroy^b

^aMathematics Research Center (CIMAT) Monterrey, Alianza Centro 502, 66629, Nuevo León

^bMathematics Research Center (CIMAT), Jalisco s/n Valenciana, 36023, Guanajuato

Abstract

This paper describes our participation for the Aggressiveness Identification Track in the third edition of MEX-A3T. The task focuses on the detection of aggressive tweets in Mexican Spanish. Our approach consists in the use of a Bidirectional Gated Recurrent Unit merged with author profiling derived features. The challenge results indicate that our proposal exceeds a Support Vector Machine baseline.

Keywords

Aggressiveness Detection, Bidirectional GRU, Author profiling

1. Introduction

The social media enables users to be in contact with others they care about. It also offers a way to discuss, and disseminate information as well as share opinions with the particularity that the people can decide to show or hide their identity; this makes easier for the users to express themselves freely, but also removes the face to face incentives to avoid being offensive.

Given the huge amount of shared data, it is difficult to manually catch all aggressive messages. So, there is a need to construct mechanisms that help to detect them automatically to avoid harassment on social media and prevent physical assaults derived from aggressive comments.

The Aggressiveness Identification Track in MEX-A3T [1] encouraged the development of methods to determine whether a tweet written in Mexican Spanish is aggressive or not. Based on the results obtained by [2] to tackle the aggressiveness identification problem, we evaluated the usage of author profiling derived characteristics along with a Gated Recurrent Unit (GRU) network. The challenge results showed that our proposal exceeds a Support Vector Machine (SVM) baseline.

This article is organized as follows, Section 2 details the proposed method and the way that author profiling characteristics were predicted. Section 3 describes the corpus and the results obtained with the training set. Subsequently, in Section 4 the results of the competition are presented and finally, the conclusions and future work are presented in Section 5.


Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)

EMAIL: maria.garrido@cimat.mx (M.G. Garrido-Espinosa); alejandro.rosales@cimat.mx (A. Rosales-Pérez); pastor.lopez@cimat.mx (A.P. López-Monroy)

ORCID:



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

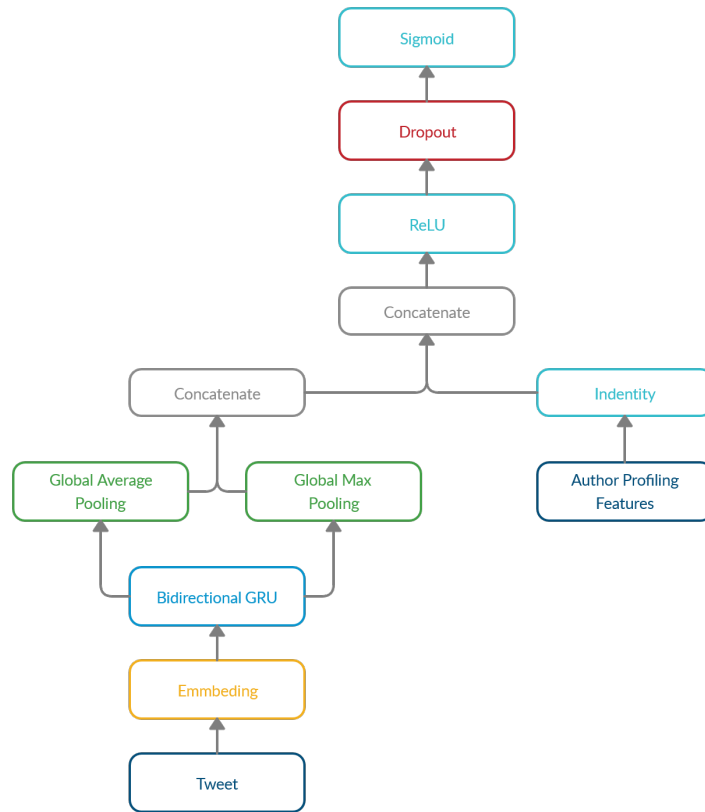


Figure 1: Diagram of architecture proposed to detect aggressive tweets.

2. System

We preserved all the content words in the tweets. To tokenize, all punctuation marks were removed, converting the text into space separated sequences of words. These sequences were split into a list of tokens to form a vocabulary. Each word in the vocabulary is represented as a vector with a pretrained word embeddings. We used FastText embeddings from Spanish Billion Word Corpus [3] of size 300.

A bi-directional GRU model using words as inputs is proposed, this model is combined with the predictions on gender and occupation of users (using a reference model and using a one-hot-encoding). Then a ReLU activation is applied, followed by a dropout, and a dense layer for making predictions; Fig. 1 shows the architecture diagram. At the end, the model considered only the gender and Sciences-Student occupation categories (the remaining categories were discarded by a χ^2 criterion).

2.1. Bidirectional Gated Recurrent Unit

The bidirectional recurrent neural networks perform better on certain tasks where the order is meaningful and are frequently used on natural language processing [4].

The Bidirectional GRU is formed by two regular GRU, each of which processes the input sequence in one direction, left-to-right and right-to-left, and then it merges their representations. By proceeding in this way, the Bidirectional GRU can capture patterns that might be pass over by a unidirectional GRU.

A regular GRU calculates each hidden state h_t as follows:

$$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1}) && \text{(Update Gate)} \\ r_t &= \sigma(W_r x_t + U_r h_{t-1}) && \text{(Reset Gate)} \\ \tilde{h}_t &= \tanh(W x_t + U(r_t \odot h_{t-1})) && \text{(Candidate)} \\ h_t &= (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t && \text{(Output)} \end{aligned}$$

where W_z , U_z , W_r , U_r , W and U are the parameters to be learned in the training phase. The function σ is the logistic sigmoid function and \odot is the element-wise multiplication [5].

The method proposed in this work uses a Bidirectional GRU network with $\hat{h}_t = \overrightarrow{h}_t + \overleftarrow{h}_t$ as the way of merging the two GRUs.

2.2. Author Profiling features

In order to introduce more information to the model, we used the Mexican corpus for author profiling from MEX-A3T 2019 [6] to predict three labels: gender, place of residence, and occupation, where we considered a different model for each label. The occupation label has eight classes: *arts*, *student*, *social*, *sciences*, *sports*, *administrative*, *health*, and *others*, while the place of residence has six classes: *north*, *northwest*, *northeast*, *center*, *west*, and *southeast*.

We adopted the n-gram ensemble approach proposed by [7] for each one of the attributes to forecast with a little variation in the size of n-grams. The n-gram ensemble approach involves four steps: the first extracts groups of n-grams of size one to three at word level and size three to five at character level. In the second step, for each group, the best n-grams are selected using χ^2 criterion. This process led to choose the best five thousand, two thousand, and thousand n-grams at word level, and the best two thousand, three thousand, and five thousand at character level. All of them are concatenated in the third step and used to classify with a SVM in the fourth step.

Once the prediction is done, the one-hot-encoding is applied to each label, and the resulting features are further filtered with the χ^2 criterion to select the best three features. This process leaved three author profiling features: gender, student, and sciences occupation.

3. Experimental Settings and Preliminary Evaluation

In this section we describe the corpus provided by the organizers, the partitions used to make experiments, the architecture used, and the preliminary results obtained. Table 1 shows the

Table 1
MEX-A3T corpus distribution

Class	Train	Percent.	Test
Non-aggressive	5,222	71.2%	-
Aggressive	2,110	28,8%	-
Total	7,332	100%	3,143

Table 2
F1 scores in the validation stage

Added Features	F1-score (agressive class)
None	0.7256
Gender, Sciences, Student	0.7311
Gender, Sciences	0.7328

tweets distribution in the training and test set

To perform experiments, we made a partition with the 7,332 samples set: 70% was taken to train, 10% to validate, and 20% to test. Fig. 1 shows the architecture of our Bidirectional GRU model. The embedding layer outputs an embedding vector of size 86×300 and feeds a Bidirectional GRU layer with 128 hidden units. Next, a global max pooling layer and a global average pooling layer flatten the Bidirectional GRU output by taking the average and max value, both of them are concatenated into a vector of size 1×256 .

In other channel, the author profiling features feed a dense layer with identity activation and with 16 hidden units. The outcome of this layer is concatenated with the pooling outcome and form a vector of size 1×272 . It is then passed to another layer with ReLU activation and 64 units. Before the final prediction, a dropout layer with a rate of 0.10 is used to regularize the network.

Table 2 shows the results obtained by the method described in Section 2 at the validation stage. The F1 obtained with the fusion of gender, sciences, and bi-GRU features is slightly better than the model that incorporates student variable but is nearly a point better than the method without author profiling features.

4. Competition Results

In this section we will present our results in the competition. Table 3 lists the final rankings for the challenge in the aggressiveness detection task. DeepMath-1 corresponds to the experiment with gender and sciences while DeepMath-2 also includes the student trait. They ranked ninth and tenth correspondingly.

5. Conclusions and future work

In this paper, we reported our participation in the MEX-A3T 2020 project to classify aggressive and non-aggressive tweets written in Mexican Spanish. We proposed a Bidirectional GRU

Table 3

Final scores of aggressiveness detection task

Rank	Team Name	F1-score (agressive class)
1	CIMAT-1	0.7998
2	CIMAT-2	0.7971
3	UPB-2	0.7969
4	UACH-2	0.7720
5	INGEOTEC	0.7468
6	Idiap-UAM-1	0.7255
	Baseline (Bi-GRU)	0.7124
7	Idiap-UAM-2	0.7066
8	UACH-1	0.7062
9	DeepMath-1	0.7001
10	DeepMath-2	0.6957
	Baseline (BoW-SVM)	0.6760
11	UMUTeam-2	0.6727
12	Intensos-1	0.6619
13	UMUTeam-3	0.6516
14	UGalileo-2	0.6388
15	UGalileo-1	0.6387
16	ITCG-SD	0.6080
17	UMUTeam-1	0.5892
18	UPB-1	0.3437
19	Intensos-2	0.2515

at word level with author profiling information. The results showed that the use of extra information as gender and sciences occupation allows us to get a better performance than those obtained without author profiling features. The competition results also showed that the proposed method was able to outperform the BoW-SVM baseline provided by the organizers as well as several proposed methods by other competitors.

Future work includes conducting experiments with Bidirectional GRU at the character level to capture dependencies in text missed by the one at the word level.

Acknowledgments

First author would like to thank CONACyT for financial support through scholarship number 718246.

References

- [1] M. E. Aragón, H. Jarquín, M. Montes-y Gómez, H. J. Escalante, L. Villaseñor-Pineda, H. Gómez-Adorno, G. Bel-Enguix, J.-P. Posadas-Durán, 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, September, 2020.
- [2] M. Casavantes, R. López, L. C. González, UACH at MEX-A3T 2019: Preliminary Results on Detecting Aggressive Tweets by Adding Author Information Via an Unsupervised Strategy, in: In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings, 2019.
 - [3] C. Cardellino, Spanish Billion Words Corpus and Embeddings, 2019. URL: <https://crscardellino.github.io/SBWCE/>.
 - [4] F. Chollet, Deep Learning with Python, Manning Publications, 2018.
 - [5] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, arXiv preprint arXiv:1412.3555 (2014).
 - [6] M. E. Aragón, M. Á. Álvarez-Carmona, M. Montes-y Gómez, H. J. Escalante, L. Villasenor-Pineda, D. 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, 2019.
 - [7] M. E. Aragón, A. P. López-Monroy, Author Profiling and Aggressiveness Detection in Spanish tweets: MEX-A3T 2018., in: IberEval@ SEPLN, 2018, pp. 134–139.
 - [8] M. Graff, S. Miranda-Jiménez, E. S. Tellez, D. Moctezuma, V. Salgado, J. Ortiz-Bejar, C. N. Sánchez, INGEOTEC at MEX-A3T: Author Profiling and Aggressiveness Analysis in Twitter Using μ tc and EvoMSA, in: IberEval@ SEPLN, 2018, pp. 128–133.
 - [9] Y. Kim, Convolutional Neural Networks for Sentence Classification, arXiv preprint arXiv:1408.5882 (2014).
 - [10] N. Albadi, M. Kurdi, S. Mishra, Investigating the Effect of Combining GRU Neural Networks with Handcrafted Features for Religious Hatred Detection on Arabic Twitter Space, Social Network Analysis and Mining 9 (2019) 41.
 - [11] Z. Zhang, D. Robinson, J. Tepper, Detecting Hate Speech on Twitter Using a Convolution GRU Based Deep Neural Network, in: European semantic web conference, Springer, 2018, pp. 745–760.
 - [12] C. E. M. Cuza, G. L. De la Peña Sarracén, P. Rosso, Attention Mechanism for Aggressive Detection, in: CEUR Workshop Proc., volume 2150, 2018, pp. 114–118.