Aggressive Language Identification in Social Media using Deep Learning

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

The increasing influence from users in social media has made that Aggressive content propagates over the internet. In a way to control and tackle this problem, recent advances in Aggressive and offensive language detection have found out that Deep Learning techniques get good performance as well as the novel Bidirectional Encoder Representations from Transformer called BERT. This work presents an overview of Offensive language detection in English and the Aggressive content detection using this novel approach from Transformer for the case study of Mexican Spanish. Our preliminary results show that pre-trained multilingual model BERT also gets good performance compared with the recent approaches in Aggressive detection track at MEX-A3T.

1 Introduction

The exponential growth of social media such as Twitter and community forum has revolutionized the communication and content publishing, but it also increased explosively the propagation of the hate speech [1, 2, 3]. thus nowadays offensive language is pervasive in social media, this content which has profanity, abusive, aggressive or any kind of words that disparages person or a group is considered hate speech.

Social media platforms and technology companies have been investing heavily in ways to cope with this offensive language to prevent abusive behavior in social media [4] One of the first action for tackling this problem was the human control over those text content and due as a manual filtering is very time consuming and as it can cause post-traumatic stress disorder-like symptoms to human annotators, the most effective strategy is use computational methods to identify offense, aggression, and hate speech in user-generated content. This topic has attracted significant attention in recent years as evidenced in recent publications [5, 6, 7] and in order to improve the research efforts in Spanish Language, we propose to find out how deep learning in NLP techniques can contribute to improve to the identification of offensive and aggressive in Spanish.

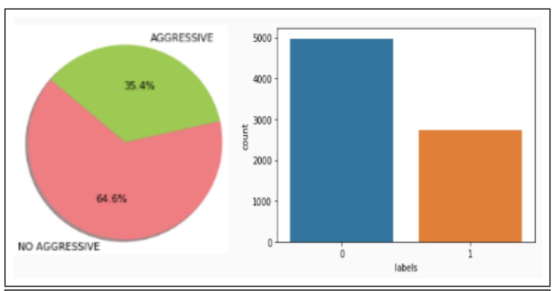
2 Related work

the research of Offensive Language have been increasing in the last years [6, 8, 9]. the scientist have proposed various methods to get features, because on of the most interesting aspect to distinguish approaches is which features are used.Thus, one of the features most used with deep learning is the simple surface features such as *unigrams* and a larger *n-grams* [1, 10, 11] and find out that that character n-grams has better perform than tokens.

In contrast to features extractions, the classification methods for Offensive Language detection are predominantly supervised learning approaches [12]. The first scopes focus on manual features engineering that are then consumed for a Machine learning algorithm such as SVM [2, 6, 11], Naive Bayes [6], Logistic Regression [13, 4], On the other side, recent researches [10, 14, 8] works show up that use deep learning paradigms which employs neural networks to automatically learn abstract features representations has better performance. However, recently Word Embedding trained in neural network have been show applied successfully [1, 7], while another approach appear this year using Bidirectional Encoder Representation from Transformer called BERT [15], which give significant improvements not only in this task if not in others. Although all of those techniques are applied to the English language, recently IberEval and IberLEF for Iberian Languages Evaluation workshops released the task with Aggressive identification task in 2017.[1] In order to develop this task, so far in Spanish the main classifier used is SVM and recently approach in deep learning use CNN [16].

3 Preliminary Approach

In order to identify the Spanish Aggressive language in social media, we decided first re-implemented the current work which achieved good performance in English Offensive Language as it shows in our related work the Deep Learning classification methods CNN, SVM, BERT standing out. At first we decided to apply those Deep Learning classification models in Mexican Spanish DataSet(MEX-A3T) (see image 1), as we found that BERT classifier is highly effective in identifying offensive content in English, then we implement multilingual BERT for Aggressive detection in Mexican Spanish. Although the preliminary results show that *bert-base-multilingual-cased* has a good performance on this Spanish task, there are still many things to accomplish and improve this model. We surprisingly found that many words are not considered for instance: "`hola`" is not in the vocabulary, this is because possibly the selection of vocabulary is data-driven, on the other hand, this method provides a good balance between the characters and words delimited models and it is really good identifying common words like: "`si, no, contrario, excepto`", showing its effectiveness in understanding the text context better than the previous pre-trained such as ELMo. Our preliminary accuracy is shown in the table 1 below.



	id	label	alpha	text
0	0	0	a	!! En el sitio también se atendió a la madre de...
1	1	0	a	!es verdad! !luis le manda un saludo a una chi...
2	2	0	a	#México es un #país de gente luchona, sabemos ...
3	3	0	a	#AlAire Hablamos vía telefónica con Ruffy Gonz...
4	4	0	a	#Alerta #Rt #Chiapas DESAPARECIDAS Madre e hij...

Figure 1: Left: MEX-A3T DataSet distribution 35.4% (green) AGGRESSIVE, 64.6 % NO AGGRESSIVE (pink) ,Right: Data labeled distribution. Below is the sample to feed in BERT

Table 1: Preliminary results for the aggressiveness identification

DATASET	Model	Accuracy
MEXT-A3T	SVM [17]	0.67
MEXT-A3T	DNN [18]	0.73
MEXT-A3T	**BERT**	**0.70**

[1] MEX-A3T: Authorship and aggressiveness analysis in Twitter case study in Mexican Spanish

References

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[2] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[3] Ziqi Zhang and Lei Luo. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, (Preprint):1–21, 2018.

[4] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[5] Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*, 2017.

[6] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[7] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, 2018.

[8] Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*, 2018.

[9] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*, 2019.

[10] 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*, pages 145–153. International World Wide Web Conferences Steering Committee, 2016.

[11] Yashar Mehdad and Joel Tetreault. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, 2016.

[12] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017.

[13] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM, 2015.

[14] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*, 2017.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Mario Ezra Aragón, Miguel Á Álvarez-Carmona, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor-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*, 2019.

[17] Rosa María Ortega-Mendoza and Adrián Pastor López-Monroy. The winning approach for author profiling of mexican users in twitter at mex. a3t@ ibereval-2018. In *IberEval@ SEPLN*, pages 140–148, 2018.

[18] Victor Nina-Alcocer, José-Ángel González, Lluıs-F Hurtado, and Ferran Pla. Aggressiveness detection through deep learning approaches. In *In Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings*, 2019.