Aggressive Spanish language identification in social Media using BERT

Errol W. Mamani Condori, Jose E.Ochoa Luna

San Pablo Catholic University, RICS-UCSP, Arequipa - Perú
email: errol.mamani@ucsp.edu.pe, jeochoa@ucsp.edu.pe

1 Introduction

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. Waseem and Hovy (2016) One of the first action for tackling this problem is use computational methods to identify aggression and hate speech in user-generated content. This topic has attracted significant attention in recent years as evidenced in recent publications (Waseem et al. 2017; Davidson et al. 2017; Kumar et al. 2018)

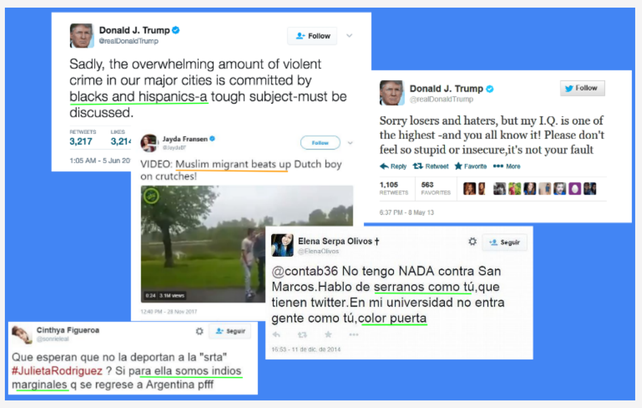

Figure 1: Aggressive tweets in Social Media

- An offensive message is characterized by disparaging or humiliating a person or a group of persons, containing(nickname, alluding, derogatory, adjectives, profanity and jokes).

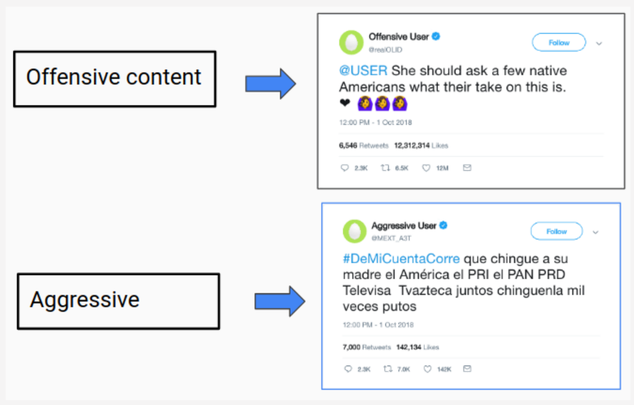

Figure 2: Terminology examples from twitter post

2 Objective

- The main aims of this thesis work is to find out how Deep Learning (DL) in Natural Language Processing techniques can contribute to the identification of Offensive or Aggressive language in Spanish.

3 DataSet

- There is no commonly accepted benchmark DataSet for the task, authors usually collect and label their own data.
- OLID contains a collection of English annotated tweets using an annotation model that encompasses following three levels (Zampieri et al. 2019).
- Spanish (MEXT-A3T) DataSet: tweets were collected based on a fixed vocabulary extracted from a "Mexicanisms". Mainly we considered the subset of words classified as "vulgar" or "insult",7700, training and 3260, test.
- This Spanish data was collected considered a dictionary of "mexicanism". tweets were manually labeled by two persons as aggressive or non aggressive. Taggers were provided with a labeling manual based on the premise that an offensive message.

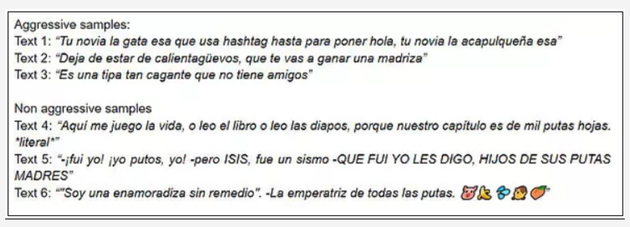

Figure 3: Text sample content of the NEXT-A3T

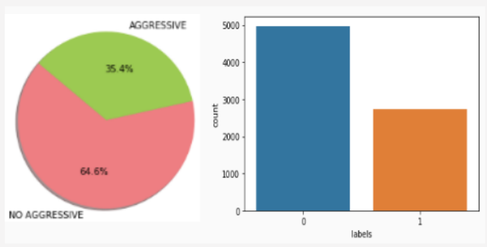

Figure 4: Pie and bar diagram of the imbalanced Dataset

4 Approach Model

Our first review show that today DL approaches are the most used model in offensive language identification. A 70 % of the researches using in the recent SemEval2019 and BERT 8 % as we can see in the below image.

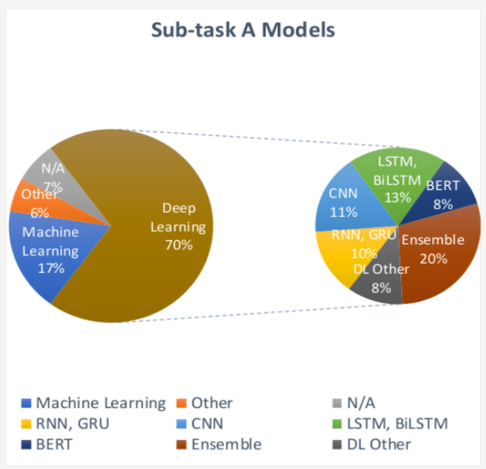

Figure 5: Pie of models used Offensive Language at SemEval2019

We address our approach for the Aggressive Spanish in Mexican case study, and we propose the fine-tuned model from transformer bidirectional Encoder lunch by google called BERT.

	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 6: table of the input adapted sample for the feed to bert.

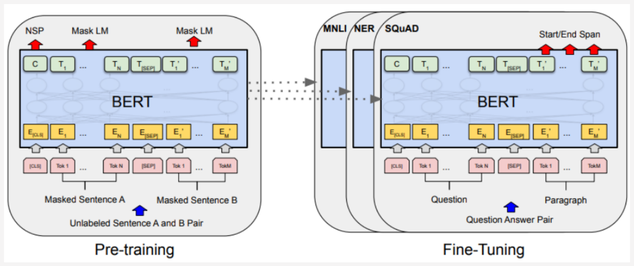

Figure 7: Overall the pre-training and fine-tunning step for BERT-base-multilingual-cased,the same architecture are used in both, pre-training and fin-tunning and also the same parameters are used to initialize for models down-stream tasks, during the fine tunning all the parameters are fine-tunned. [*CLS*] is special symbol added at the beginning of the input of the every input sentences example and [*SEP*] token separator.

5 Preliminary Results

Although the preliminary results show that BERT has a good performance on this Spanish task, there are still many things to accomplish and improve.

DATASET	MODEL	Accuracy
MEXT-A3T	SVM (a)	0.63
MEXT-A3T	DNN (b)	0.73
MEXT-A3T	**BERT**	**0.70**

Table 1: Table with preliminary results for the aggressiveness identification, (a) is a references to Ortega-Mendoza and López-Monroy 2018 and (b) refers to Nina-Alcocer et al. 2019works at iberLEF.

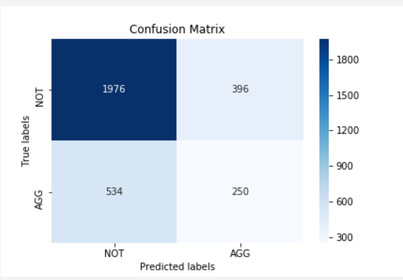

Figure 8: Confusion matrix for the Aggressive Language detection.

6 Additional

Comparing the preliminary results shows that our Bert fine-tuned get a comparable result with the state of the art in this particular task. we show, in the below table offensive English language accuracy using different approaches as a references resource.

DATASET	MODEL	Accuracy
OLID	Ensamble SVM	0.66
OLID	CNN	0.73
OLID	**BERT**	**0.83**

Table 2: Table with preliminary results for the similar task of offensive language identification in English.

7 Remarks

- 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.
- 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`".
- The effectiveness in understanding the text context of BERT increase the identification for aggressiveness and work fine for Spanish as well.
- There is still a gap between English and multi language BERT, this is might because 28,996 tokens for English and just 119,547 tokens for all languages.
- there is recently additional model based on the transformers architectures such as RoBERTa using the BERT-base architecture to say one of them.