

Aggressive language identification in social Media using Deep learning



Errol Wilderd Mamani Condori^{1,*}, Jose E. Ochoa Luna².
San Pablo Catholic University, Arequipa-Peru
Email: errol.mamani@ucsp.edu.pe



Introduction

Offensive content has become pervasive in social media and thus a serious concern for government and organizations. 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 [waseem2016hateful]. 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 [waseem2017understanding, davidson2017automated, kumar2018benchmarking].

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

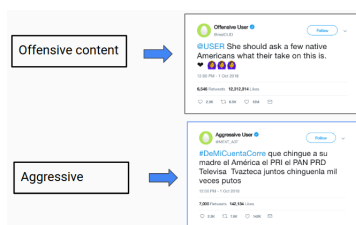


Figure: Terminology examples from twitter post

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.

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 [zampieri2019].
- 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.

Aggressive samples:

Text 1: "Yo novia la gata esa, yo una hashtag hasta poner hola, tu novia la acapulqueña esa"

Text 2: "Deja de estar de calentilleros, que te vas a ganar una madrita"

Text 3: "Es una fije tan cagante que no tiene amigos"

Non aggressive samples

Text 4: "Aquí me juego la vida, o leo el libro o leo las diapos, porque nuestro capítulo es de mil putas hojas."

Text 5: "Niqui yul yio putas, yul-peao ISIS, fue un sísmo - QUE FUI YO LES DIGO, MUJOS DE SUS PUTAS MADRES"

Text 6: "Soy una enomadrada sin remedio". -Le emperatriz de todas las putas. 🇵🇷 🇺🇸 🇩🇪 🇪🇸

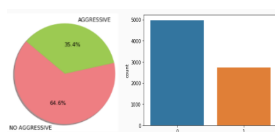


Figure: Pie and bar diagram of the imbalanced Dataset

Approach Model

The methods used for Offensive language detection identification are predominantly Deep Learning approaches as well as for Spanish Aggressive.

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.

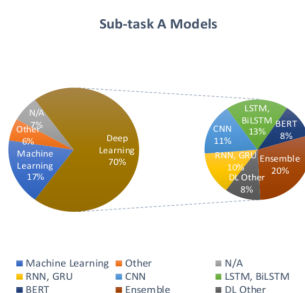


Figure: Pie Char showing the models used Offensive Language at SemEval2019

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	les verdad! fluis le manda un saludo a una chi...
2	2	0	a	#México es un #pais de gente luchona, sabemos ...
3	3	0	a	#AlAire Hablamos vía telefónica con Ruffy Gonzá...
4	4	0	a	#Alerta #Rt #Chiapas DESAPARECIDAS Madre e hi...

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

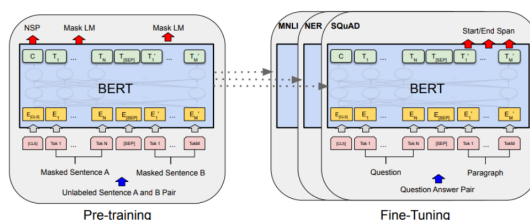


Figure: Overall the pre-training and fine-tuning step for BERT-base-multilingual-cased, the same architecture are used in both, pre-training and fin-tuning and also the same parameters are used to initialize for models down-stream tasks, during the fine tuning all the parameters are fine-tuned. `[CLS]`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

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)	0.67
MEXT-A3T	DNN (B)	0.73
MEXT-A3T	BERT	0.70

Figure: Table with preliminary results for the aggressiveness identification, (A) is a references to [ortega2018winning] and (B) refers to [nina2019aggressiveness]works at iberLEF.

Comparison

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 image offensive language accuracy using different approaches and the Aggressive language identification in the Mexican Spanish case study.

DataSet	Method	F1 score
OLID	Ensamble SVM	66.4 %
OLID	BERT	83 %
OLID	CNN	73.9 %

DATA	METHOD	ACCURACCY
MEXT-A3T	ensemble SVM	67
MEXT-A3T	DNN	73
MEXT-A3T	BERT	70

Figure: Images of tables (Above) offensive language f-score accuracy.(below) Aggressive language detection f-score accuracy

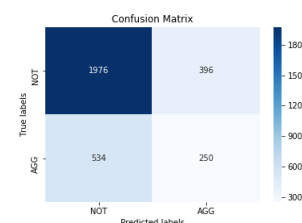





Figure: Confusion matrix for the Aggressive Language detection.P

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.

Main References

-  Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R. *Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval)*. arXiv preprint arXiv:1903.08983. 2019 Mar 19.
-  Ortega-Mendoza RM, López-Monroy AP. *The Winning Approach for Author Profiling of Mexican Users in Twitter at MEX. A3T@ IBEREVAL-2018*. InIberEval@ SEPLN 2018 (pp. 140-148).
-  Nina-Alcocer V, González JÁ, Hurtado LF, Pla F. *Aggressiveness detection through deep learning approaches*. InIn Proceedings of the First Workshop for Iberian Languages Evaluation Forum (IberLEF 2019), CEUR WS Proceedings 2019.