Aggressive Language Detection using VGCN-BERT for Spanish Texts

E. Wilderd Mamani Condori
errol.mamani@ucsp.edu.pe
Advisor: Dr. Jose E. Ochoa Luna

Department of Computer Science
in partial fulfillment of the requirements for the
degree of Master in Computer Science

Universidad Católica San Pablo

CONCYTEC
CONSEJO NACIONAL DE CIENCIA, TECNOLOGÍA E INNOVACIÓN TECNOLÓGICA

Outline

- Introduction
  - Aggressive, Offensive and Hate speech in social Medias,
- Related work
  - transfer learning with BERT, GCN
- Methodology and Data
  - MEX-A3T, VGCN, BERT
- Experimental Results

Social Media

The exponential growth of Social Media has revolutionized, but this increasingly exploited the propagation of the Aggressive language over the internet.

Online communities and social media platforms are heavily investing to cope this behaviour and prevent abusive on social media.

Related work

Offensive content has become pervasive in social media and thus a serious concern for government and organizations.

Terminology

It has many names… a little bit confusing?

🤔

No at all...!

Offensive

Aggressive

hate speech

- Offensive language (Xiang -2012,zampieri -2016)
- Hostil message or flames, Abusive messages (Spertus -1997)
- Cyberbullying (Dinakar, Warner & Xu et ...2012, zhong 2015, dadvar 2013, Burnap-2015)

Insult, profanity(Razavi 2010)

Transfer Learning with BERT

- BERT Transformer (Devlin et al., 2018)
- BERT, tokenization lowercase, WordPiece token,
- Batch = 24, epochs= 2, adams = 2e-5, dropout= 0.1
- BERT-multilingual uncased

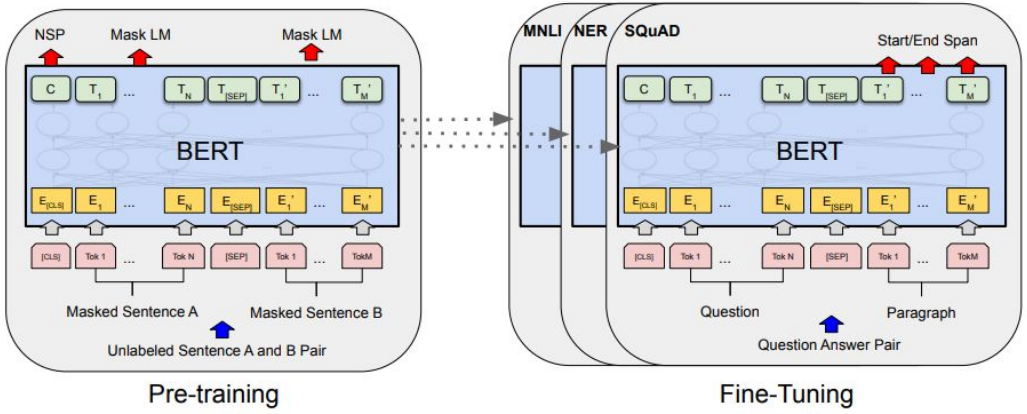


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

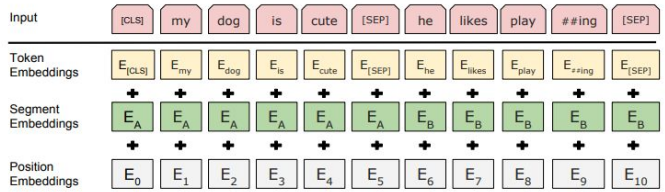
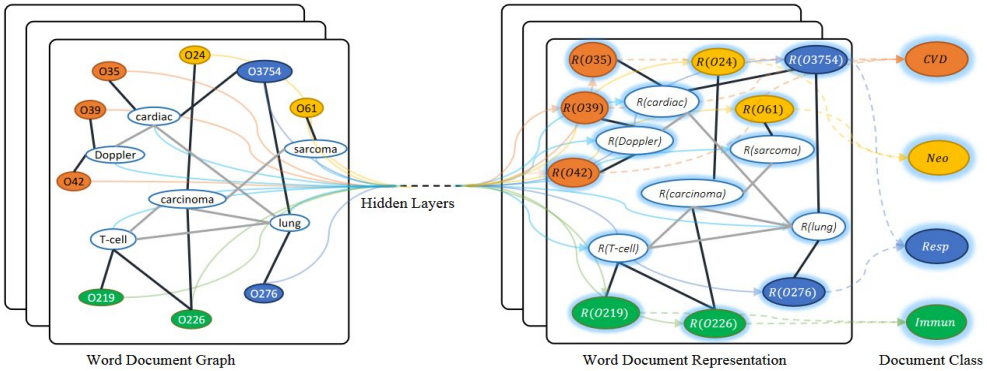
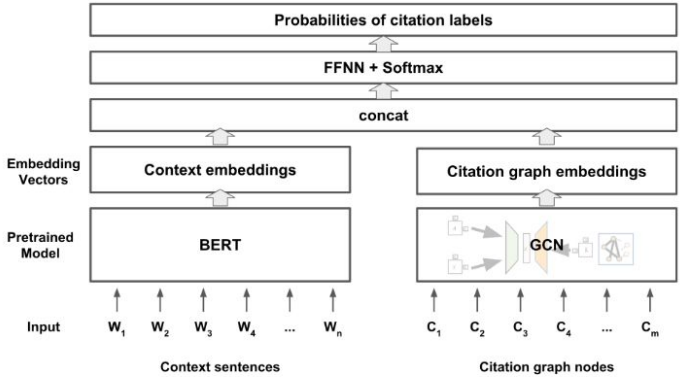


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

Graph Convolutional Network

- GNN for text classification
- Text GCN (***Yao 2019***)
- BERT-GCN (***jeaon 2020***)



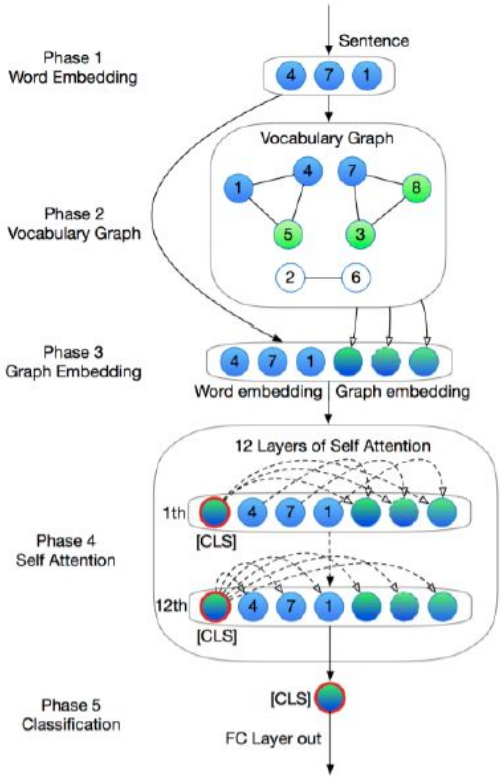


Methodology
VGCN-BERT

- ***VGCN***

  we use graph embedding from vocabulary graph (built) with pmi-word relation

- **BERT** : for our task we chose BETO



DataSet Spanish

(MEXT-A3T)

- Data collected in November 2017 https://mexa3t.wixsite.com/home/aggressive-detection-track
- tweets were collected based on a fixed vocabulary extracted from a “Mexicanisms”. Mainly we considered the subset of words classified as “vulgar” or “insult”
- an offensive message is characterized by disparaging or humiliating a person or a group of persons, containing(nickname, alluding, derogatory, adjectives, profanity and jokes)
- 7700, training
- 3260, test

Aggresssive Samples:

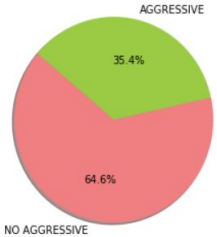
“Sólo a las Pu#..aS MOCOS#S GORDAS Y feas les gusta ese al%man xd”

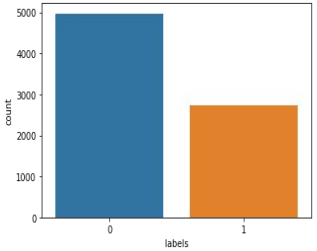
“Profe hijo de las mil pu#..as 6 de calificación como es posible. ”

Non Aggressive Samples:

“Put#s Madres ahora comprendo todo, tu tan lava y yo tan frio ”

“ Segunda vez que me pasa. Estoy hasta la madre”



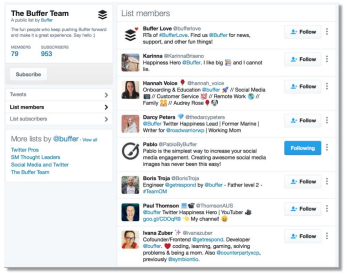


Category	Percentage
Aggressive	25%
Non Aggressive	75%

Table 1. MEX-A3T tweets with Aggressive data set distribution of classes.

Class	Train Corpus	Test Corpus
Non Aggressive	5222	2238
Aggressive	2110	905
Total	**7332**	**3143**

Experimental Results



“AGGRESSIVE”

Table 2. Results of the Aggressive Detection using F1-score in test set.

Model	F1 Aggressive	F1 non-agressive	F1 macro
BoW-SVM	0.6760	0.8780	0.7770
BI-GRU	0.7124	0.8841	0.7983
BETO+msg	0.7720	0.9042	0.8381
bert (**multi***)	0.7809	0.9094	0.8452
bert (**1 BETO**)	0.7998	0.9195	0.8596
bert (20 BETOs)	0.7994	0.92.23	0.8608
VGCN-BERT *	**0.8124**	**0.9169**	**0.8642**

* meas our developed system model.

thanks :)