
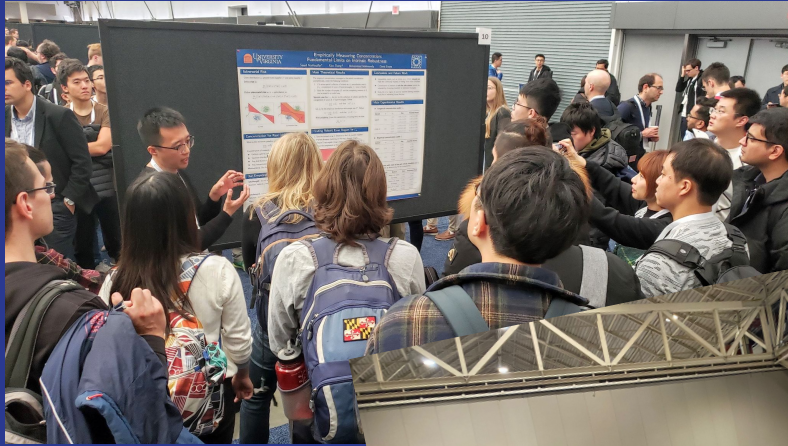


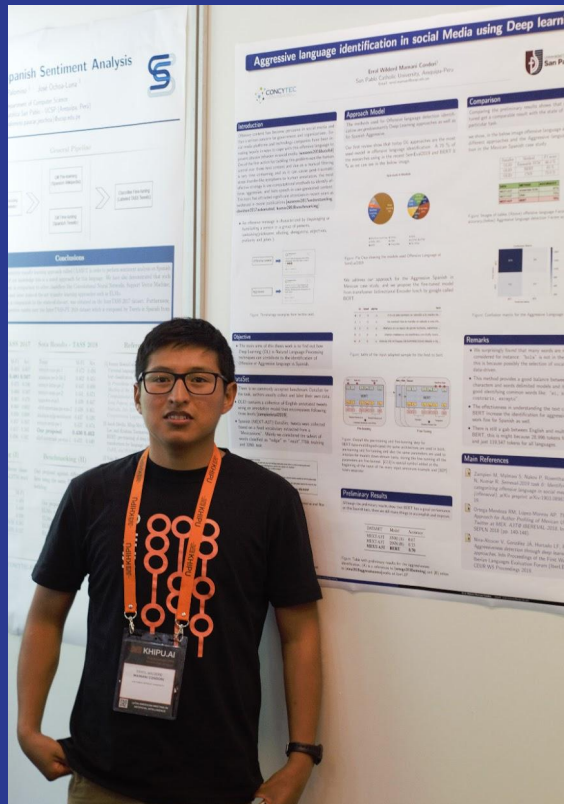
Identificación de lenguaje ofensivo, agresivo usando deep learning

Errol W. Mamani-Condori

outline

- Introducción
 - Motivación y contexto.
 - Descripción de la tarea y Agresividad en redes sociales.
 - Trabajos relacionados
 - transfer learning con BERT, GCN, GAT
 - Metodología y Conjunto de Datos
 - MEX-A3T, OLID, VGCN-BERT
 - Experimentos
 - Conclusiones
- 







Motivación



Donald J. Trump

@realDonaldTrump

Sadly, the overwhelming amount of violent crime in our major cities is committed by blacks and hispanics-a tough subject-must be discussed.

RETWEETS
3,217

LIKES
3,214



Donald J. Trump

@realDonaldTrump

1:05 AM - 5 Jun 2013

Mexico's court system corrupt.I want nothing to do with Mexico other than to build an impenetrable WALL and stop them from ripping off U.S.

RETWEETS
574

LIKES
690



Cinthya Figueroa

@sonrevela

Que esperan que no la deportan a la "srta" #JulietaRodriguez ? Si para ella somos indios marginales q se regrese a Argentina pfff



Miklos Lukacs

@mlukacs · Dec 31, 2017

El criminal proceder de la corrupta organizacion antifujimorista @IDL_R no acabara con lamentos sino con acciones legislativas concretas e inmediatas. El sicariato politico-mediatico que el secuestrado Gorriti y sus secuaces han institucionalizado debe ser duramente sancionado



37

257

416



Donald J. Trump

@realDonaldTrump

Follow

Sorry losers and haters, but my I.Q. is one of the highest -and you all know it! Please don't feel so stupid or insecure,it's not your fault

Reply Retweet Favorite More

1,105
RETWEETS

563
FAVORITES



Jayda Fransen

@Jayda8F

Follow

VIDEO: Muslim migrant beats up Dutch boy on crutches!



Serpa Olivos

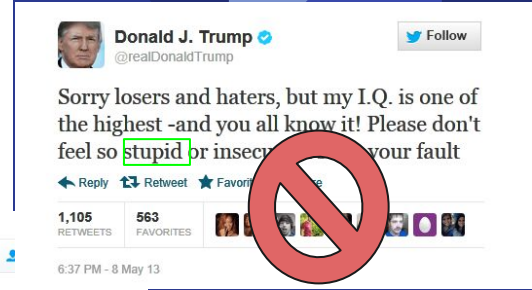
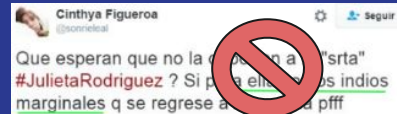
@ElenaOlivos

Seguir

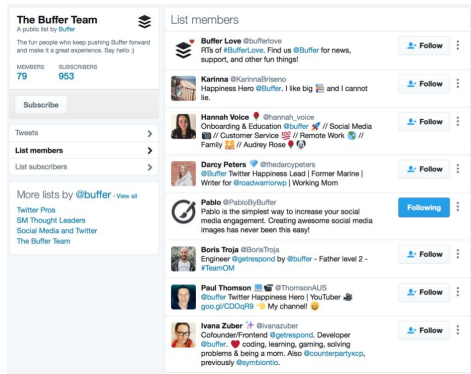
@contab36 No tengo NADA contra San Marcos.Hablo de serranos como tú,que tienen twitter.En mi universidad no entra gente como tú,color puerta

16:53 · 11 de dic. de 2014

Motivación



Descripción de la tarea



Hate
speech

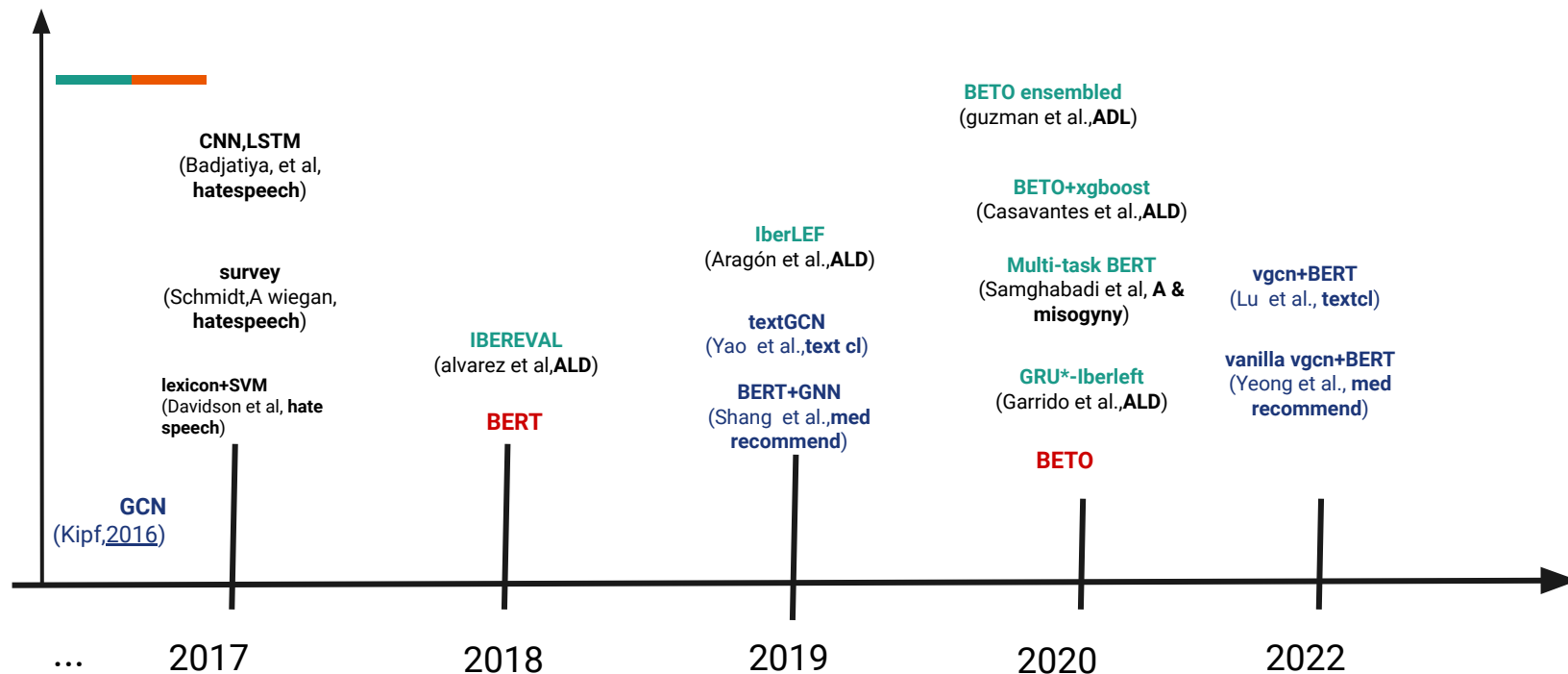
offensive

Aggressive

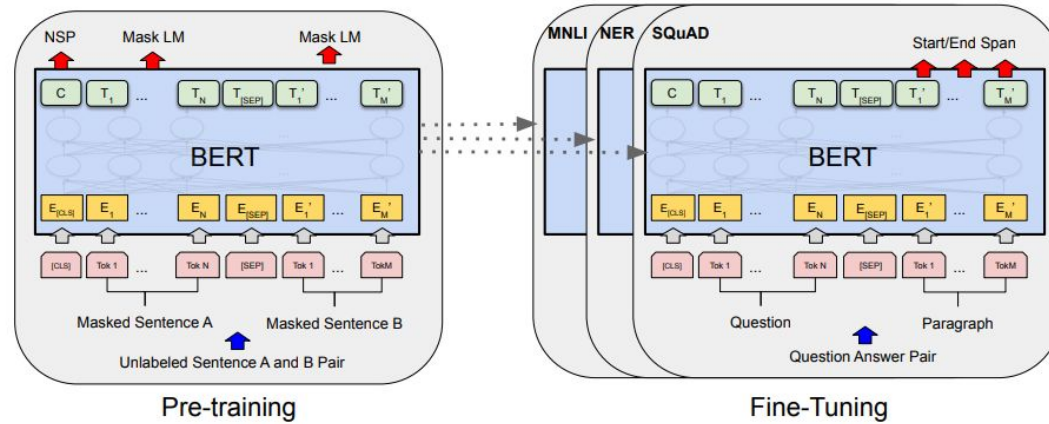


“AGGRESSIVE”

Literatura



Transfer Learning con BERT



BERT
(Devlin et al., [2018](#))

Transfer Learning con BERT

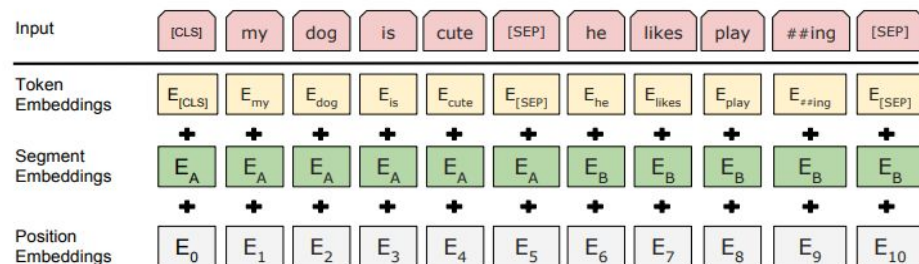
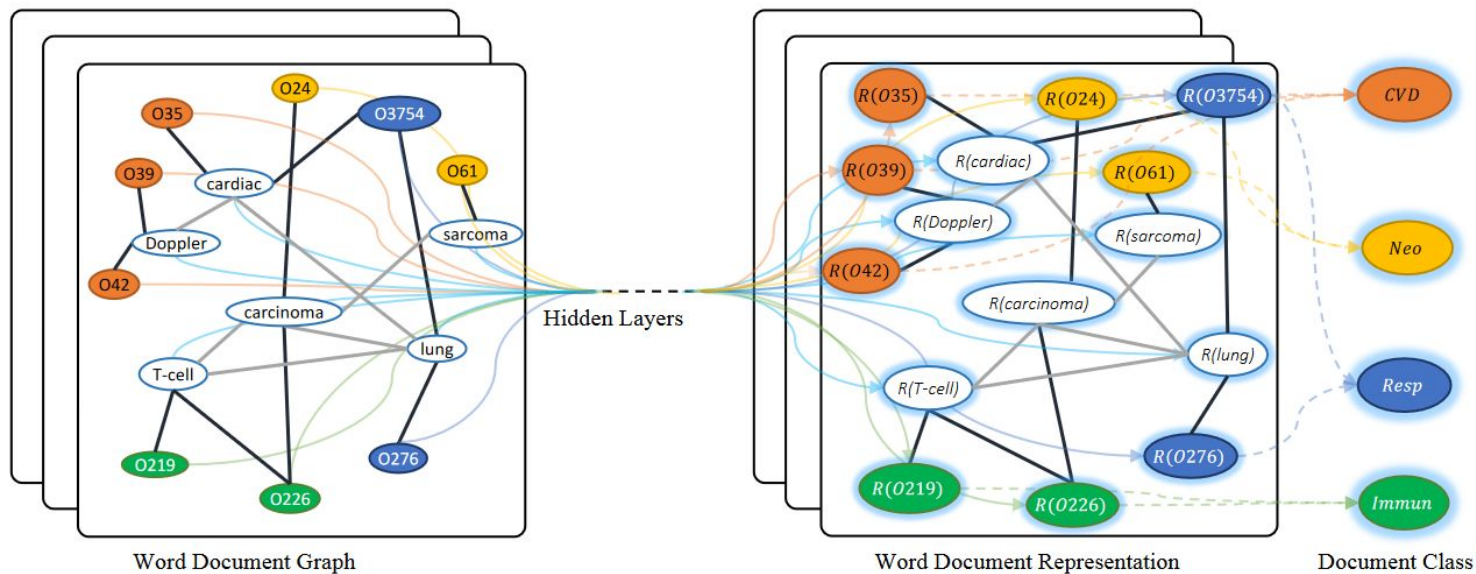


Figure 2: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

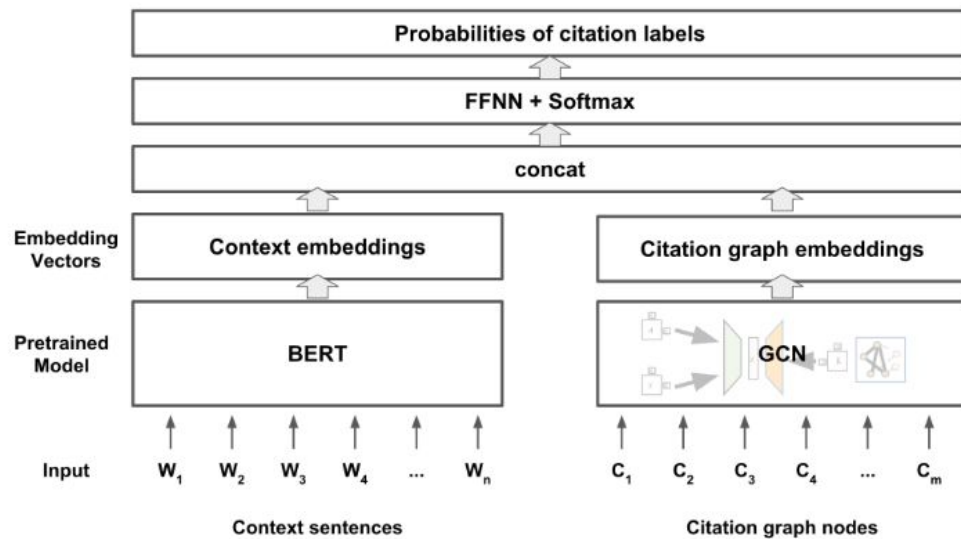
BERT
(Devlin et al., [2018](#))

Graph Convolutional Network



Text GCN
(Yao et al., [2019](#))

Vanilla VGCN-BERT



Text Med citation
(Jeong et al., [2019](#))

DataSet Español

(MEXT-A₃T)

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”

DataSet Español

(MEXT-A₃T)

*Las fans de odiseo se ven bien bonitas en sus fotos de twitter y **están bien feas en persona.***

*(Odisseo fans look really pretty in their twitter photos and **they are pretty ugly in person.**)*

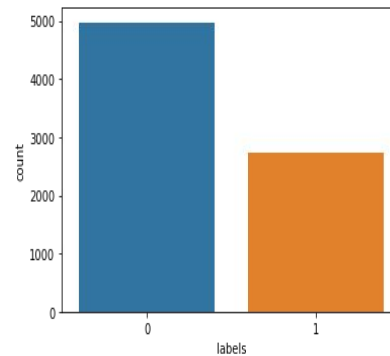
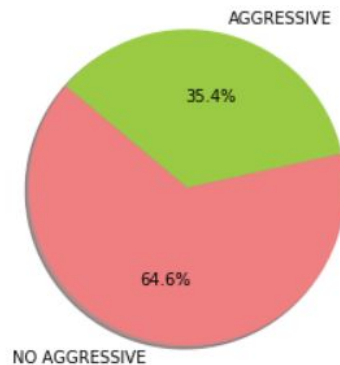
Odisseo fans look really pretty...

*... **They are pretty ugly in person.***

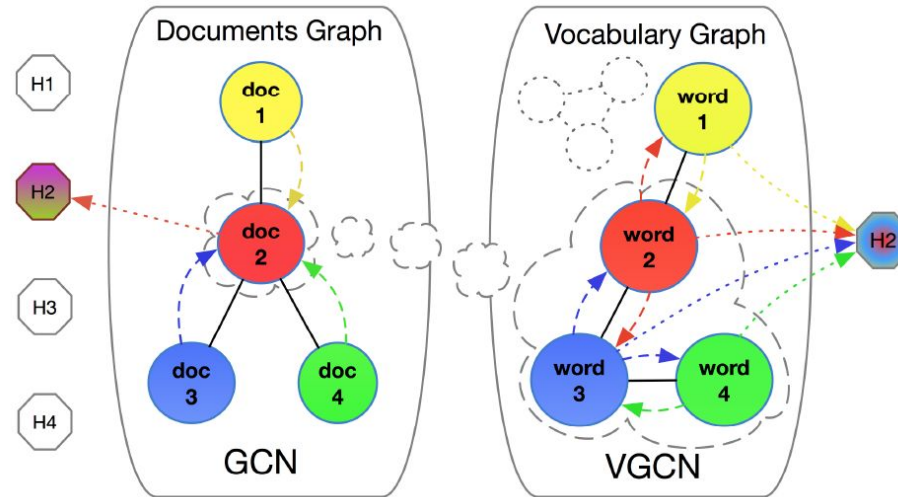
Conjunto de Datos

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

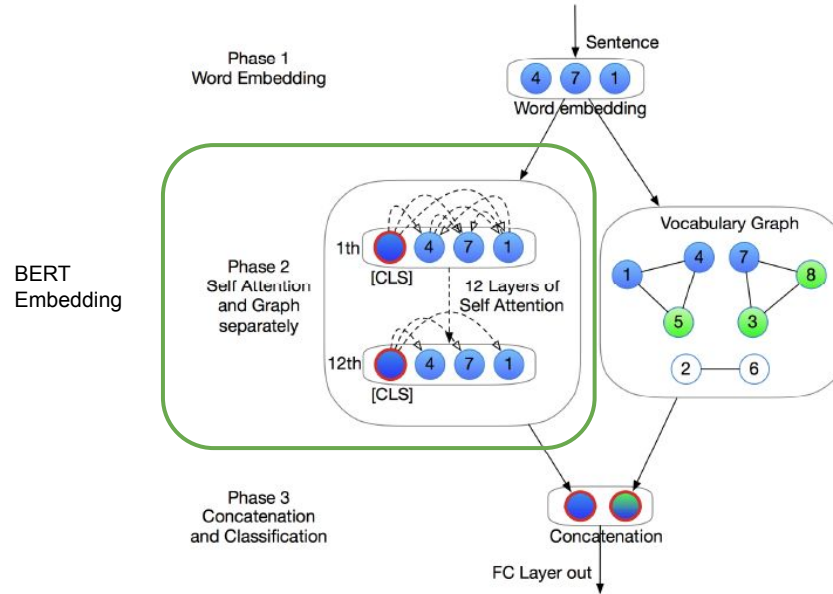


Metodología VGCN-BERT



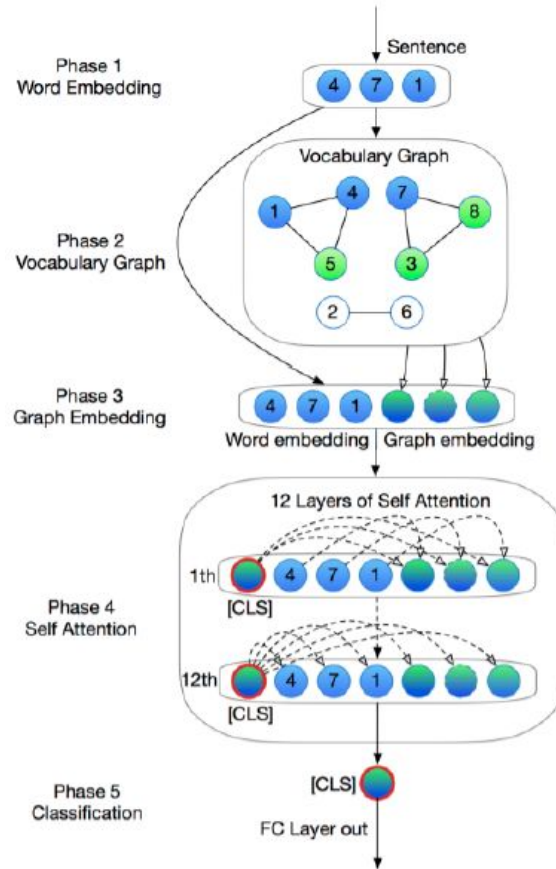
VGCN-BERT
(Lu et al., [2019](#))

Metodología VGCN-BERT



**VGCN-BERT based on
Jeon, 2019**
(Lu et al., [2019](#))

Metodología VGCN-BERT



VGCN-BERT
(Lu et al., [2019](#))

Experiments

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)	<u>0.7994</u>	0.92.23	<u>0.8608</u>
VGCN-BERT *	0.8124	0.9169	0.8642

* meas our developed system model.

conclusion

- Hemos demostrado que agregar un gráfico de vocabulario, incluso en un conjunto pequeño de vocabulario para BERT en español, podría mejorar la información local de contexto.
- Nuestros resultados experimentales nos permiten afirmar que el uso de VGCN-BERT podría ser beneficioso para detectar contenido de agresión, especialmente para el idioma español.
- Un vocabulario e información global de un idioma es un camino para mejorar el modelo BERT como un classier.
- Creemos que una palabra dentro de una oración puede aportar mucho si su relación dentro de un tuit contribuye a detectar agresividad.

Contributions

- Combinando la información local y global utilizando VGCN-BERT sin embeddings de palabras o conocimientos externos lo cual es adecuado y novedoso para detectar las palabras de agresividad con una noción global para el idioma español.
- La combinación de VGCN con un modelo BERT (una versión pre-entrenada en español llamada BETO) permite obtener resultados comparables con respecto a los modelos BERT de conjunto en la detección de agresividad en español.
- Uso de GCN-BERT, GAT-BERT para detección de comentarios ofensivos capturando de contexto global através de “attention” usando symbolic language, uso de “lexicon” para captura de conocimiento