

A vertical rainbow flag with six stripes of equal width: red, orange, yellow, green, blue, and purple.

**Discurso de odio
en tweets.**

Outline

Motivación



NLP Feature Extraction



ML / DL (Models)



State of the art



Discurso de odio (Hate Speech)

Se entiende como toda clase de comunicación verbal, **escrita** o comportamiento que violenta a otra persona (o personas) en **base a quiénes son**, es decir, basado en su religión, raza, etnicidad, nacionalidad, género, orientación sexual, etc.

No existe una definición universal.

A Saudi Arabian woman runs for local office in the drama 'The Perfect Candidate'

WILL COVIELLO | Jun 7, 2021 - 7:00 am

Luai Qubain Gay Pride Month is a time for joy, but also to reflect on those still being persecuted

Let's take a moment from our revelry to think about all those around the world for whom Pride is an inaccessible luxury – and a potentially fatal risk.

El racismo que México no quiere ver

La evidencia estadística sobre el aumento de la discriminación por el color de piel y sus efectos en la vida de los mexicanos es abrumadora, sin embargo en el país latinoamericano apenas se empieza a hablar de este problema



ELÍAS CAMHAJI | SONIA CORONA | GLADYS SERRANO

México - 30 NOV 2019 - 12:04 EST

A Genocide Incited on Facebook, With Posts From Myanmar's Military



A border police officer at a repatriation center for Rohingya returning to Myanmar. Human rights groups blame anti-Rohingya propaganda online for fueling violence and displacement. Adam Dean for The New York Times

East Asia Pacific

US Welcomes Pledge by Myanmar Shadow Government to Help Rohingya

By VOA News
Updated June 07, 2021 04:22 PM

Ethnic cleansing

Facebook Admits It Was Used to Incite Violence in Myanmar



Rohingya refugees after crossing the Naf River, which separates Myanmar and Bangladesh, in 2017. A report commissioned by Facebook found the company failed to keep its platform from being used to “foment division and incite offline violence” in Myanmar. Adam Dean for The New York Times

Human Monitors

THE TRAUMA FLOOR

The secret lives of Facebook moderators in America

By Casey Newton | @CaseyNewton | Feb 25, 2019, 8:00am EST

Illustrations by Corey Brickley | Photography by Jessica Chou



'The Cleaners' Who Scrub Social Media

63K views • 3 years ago

CBC News

Social media platforms say they want to scrub fake news and some ...



'It's the worst job and no-one cares' - BBC Stories

1M views • 3 years ago

BBC News

The release of Facebook's content moderation guidelines has drawn attention ...

**Mexican NLP
Summer School
2021**

**Ethical and social
considerations when
doing NLP**
Panelists: Luciana Benotti,
Ted Pedersen, Daisuke
Kawahara

**2021 Annual Conference of the
North American Chapter of the
Association for Computational
Linguistics**

Online
June 6-11, 2021



4th Workshop Abuse and Harms 2020

**Harassment on
Twitter**

**Discursos de odio
hacia personas
asiáticas**

**Sesgo en las
bases de datos.**

Hateful memes (2021)

**Discurso de odio
referente a
COVID19**



Backgrounder

Hate Speech on Social Media: Global Comparisons

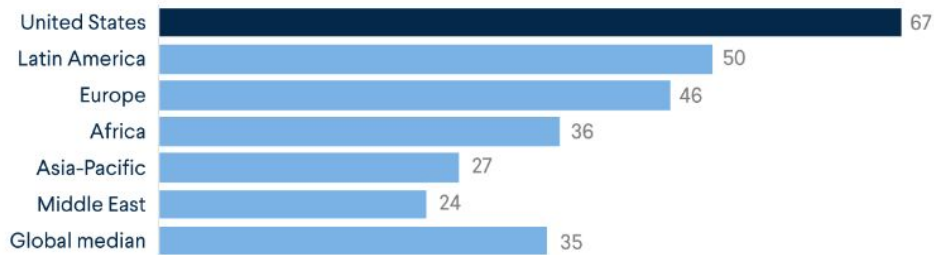
Violence attributed to online hate speech has increased worldwide. Societies confronting the trend must deal with questions of free speech and censorship on widely used tech platforms.



A memorial outside Al Noor mosque in Christchurch, New Zealand. Kai

Hate Speech vs Offensive Language.

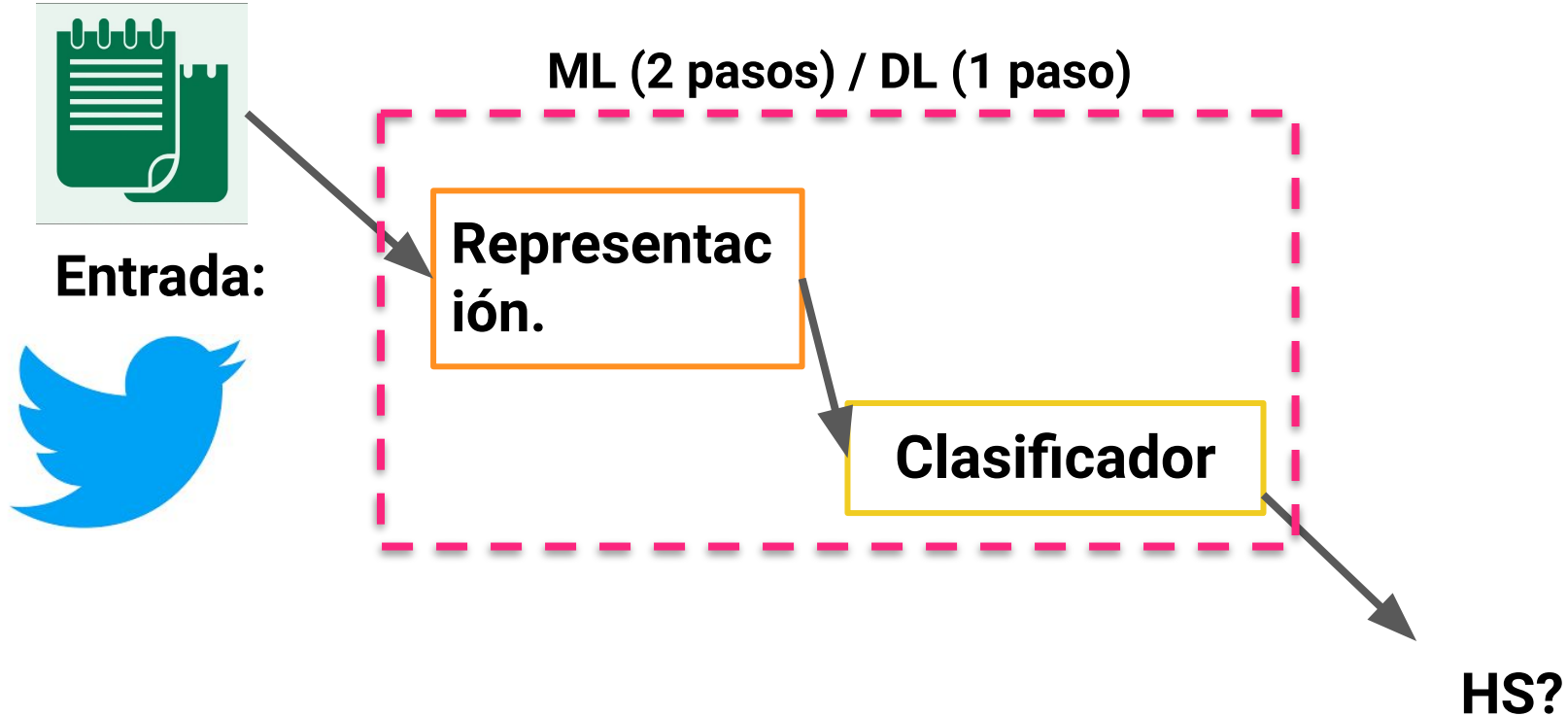
Percent that agree “People should be able to make statements that are offensive to minority groups publicly” (2015)



Note: Displays the median among countries included in the survey.

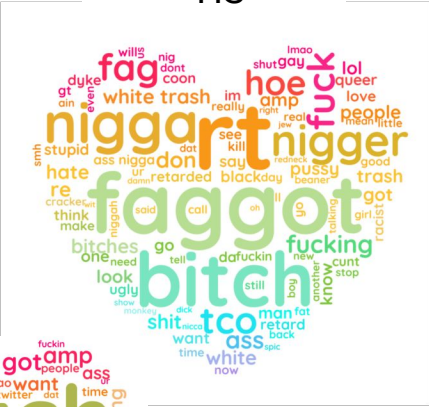
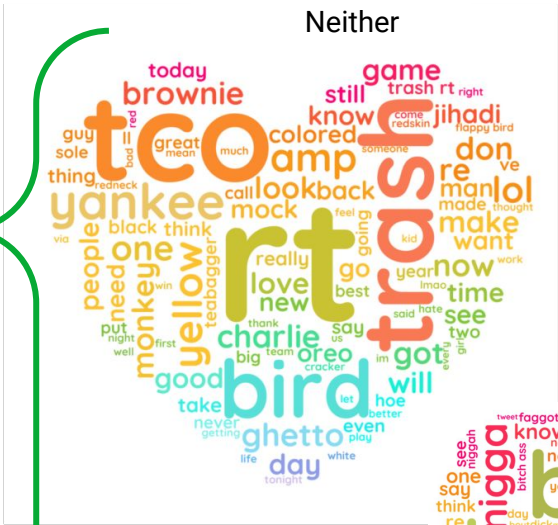
Source: Pew Research Center.

Procesamiento del **Lenguaje** Natural



25K

T. Davidson et al (2017)



“Bitch Plz
Whatever”

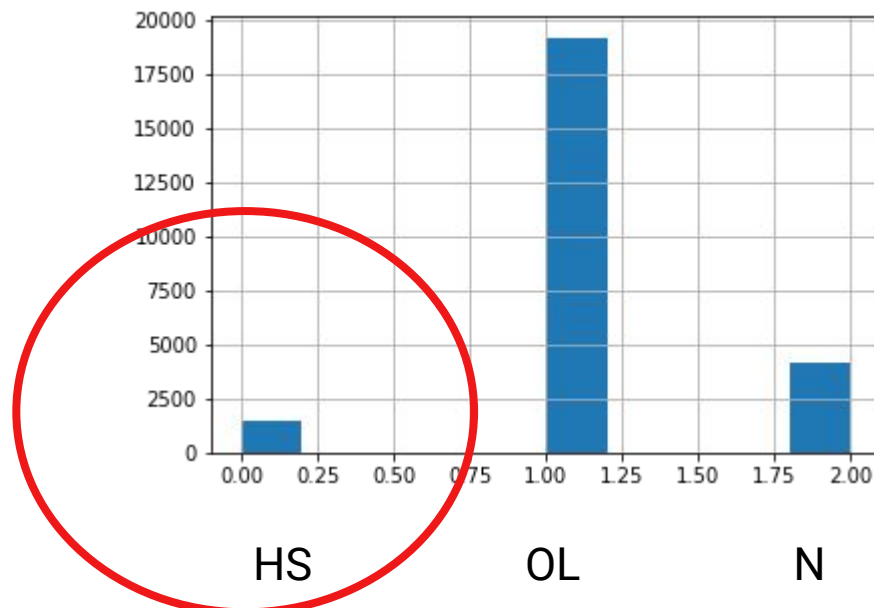


class

Hate speech (0),
Offensive L. (1) or
Neither (2)

0

T. Davidson et al (2017). **Automated hate speech detection and the problem of offensive language.** In *Proceedings of the International AAAI Conference on Web and Social Media*





Los anotadores consideran términos **homofóbicos o racistas** como lenguaje de odio. Pero, tweets **sexistas o derogatorios** hacía mujeres **sólo ofensivos**,



Sesgo

El clasificador asigna **más términos ofensivos** **O de discurso de odio** que los mismos anotadores.



Discurso de odio

Si no contiene groserías, tienden a no ser clasificados como discursos de odio.



Fenómeno

¿Qué sucede cuando alguien contesta un tweet racista con uno homofóbico?

Representación del texto (ML POV)

Feature extraction

Bolsa de palabras

¿Qué tan “fácil” es leerlo?



Wicked Jen
@wickedsga

Follow

I don't think it's a coincidence that diet has the word die in it.

2:07 PM - 30 Nov 2015

410 Retweets 560 Likes



4



410



560



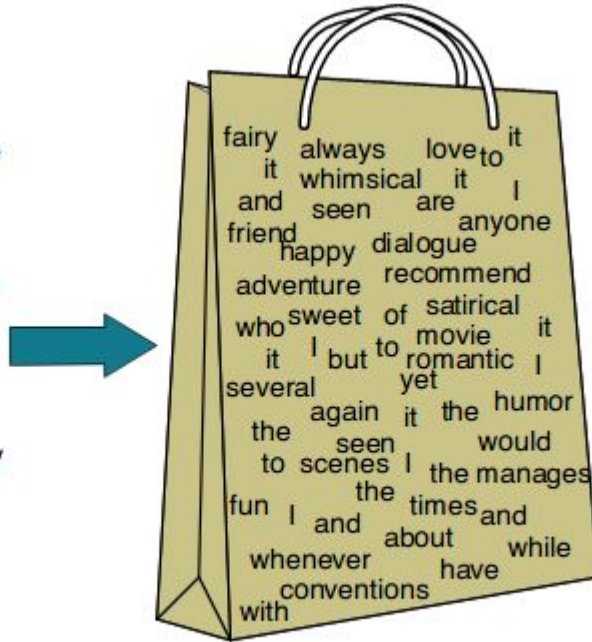
Sentimiento
Pos, neg, neutro

Part of speech (
adverbios,
proposiciones,
adjetivos)

#hashtags,
#RT,
#menciones

Bolsa de palabras

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	5
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

BOW + pesado

Binario.

	it	is	puppy	cat	pen	a	this
it is a puppy	1	1	1	0	0	1	0
it is a kitten	1	1	0	0	0	1	0
it is a cat	1	1	0	1	0	1	0
that is a dog and this is a pen	0	2	0	0	1	2	1
it is a matrix	1	1	0	0	0	1	0

TF-IDF

	and	antagonistic	are	cats	dogs	four	hate	have	he	legs
vector1	0.000000	0.000000	0.000000	0.402040	0.000000	0.528635	0.000000	0.528635	0.000000	0.528635
vector2	0.490479	0.490479	0.490479	0.373022	0.373022	0.000000	0.000000	0.000000	0.000000	0.000000
vector3	0.000000	0.000000	0.000000	0.000000	0.473630	0.000000	0.622766	0.000000	0.622766	0.000000

Sparsity



feature

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	1.291631	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
1	2.583261	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	1.78288	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	6.898406	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
2	2.583261	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	1.78288	0.0	0.0	0.0	0.0	0.0	1.829224	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	2.490969
3	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
4	5.166523	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	1.78288	0.0	0.0	0.0	0.0	0.0	1.829224	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	4.981938
...
24778	2.583261	3.907834	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.958858	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
24779	3.874892	3.907834	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	6.473522	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
24780	1.291631	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
24781	1.291631	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	1.78288	0.0	0.0	0.0	0.0	0.0	1.829224	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.000000
24782	2.583261	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	2.490969

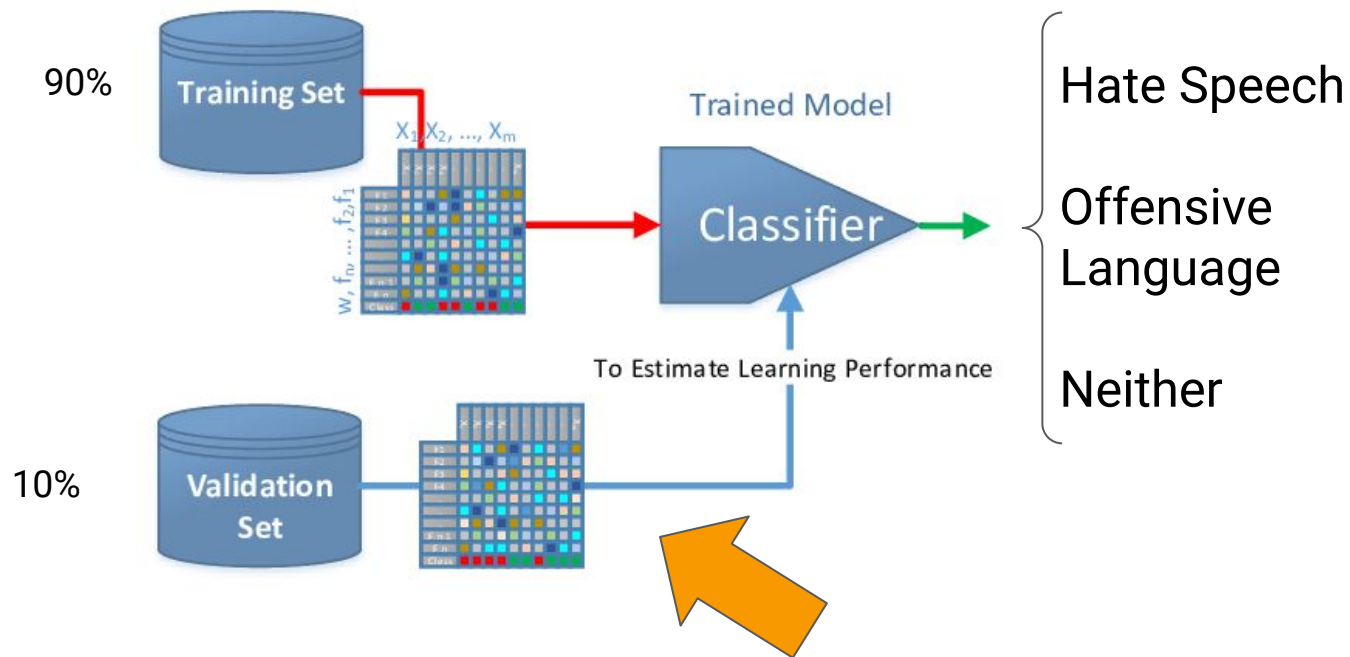
24783 rows × 3437 columns



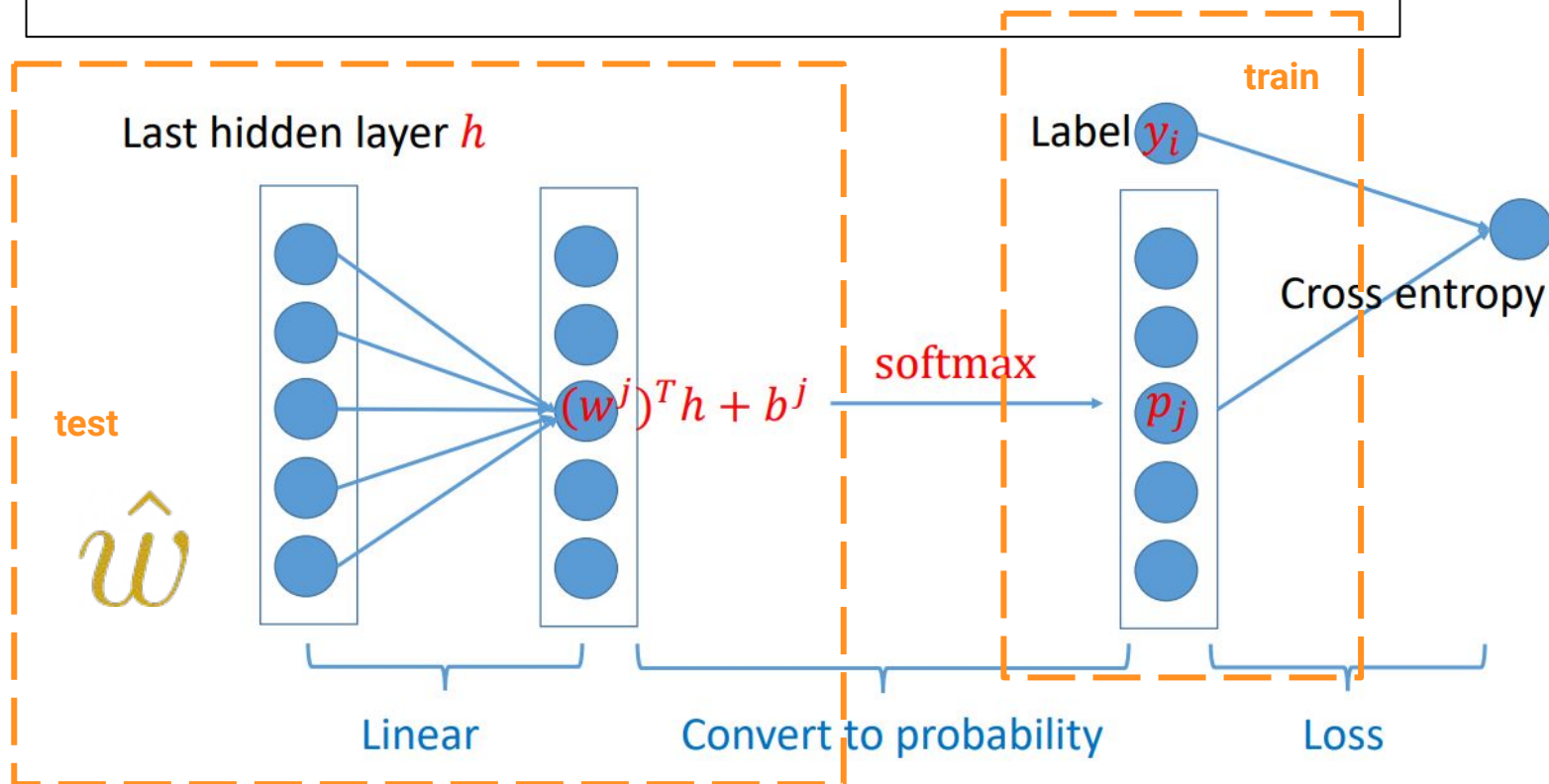
Tweet

(24783, 4023)

Clasificador



Clasificador: MLR



Regresión Multinomial

$$\mathbb{P}(Y = k|x), \quad k \in \{\text{HS}, \text{OL}, \text{N}\}$$

$$\begin{aligned} \min_{w_k, b_k} L_{\text{CE}}(\hat{y}, y) &= - \sum_{k=1}^K \mathbb{1}\{y = k\} \log \hat{p}(y = k|x) \\ x \in \text{Test Set} \quad &= - \sum_{k=1}^K \mathbb{1}\{y = k\} \log \underbrace{\frac{\exp(w_k \cdot x + b_k)}{\sum_{j=1}^K \exp(w_j \cdot x + b_j)}}_{\mathbb{P}(Y = k|x)}, \end{aligned}$$

Regresión con penalización (Feature Selection)

$$\hat{\beta}_{\text{PLS}} = \arg \min_{\beta} \left[L_{\text{CE}}(\hat{y}, y) + \lambda \cdot \text{pen}(\beta) \right]$$

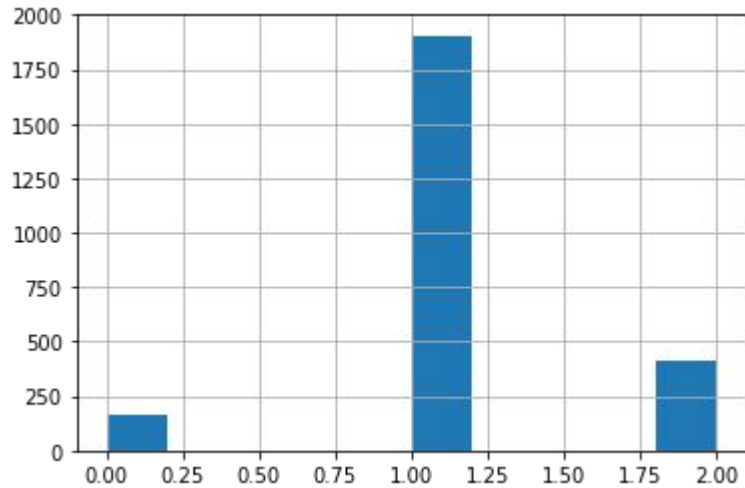
$$\text{pen}(\beta) = \sum_{j=1}^k |\beta_j|$$

Least Absolute Shrinkage and Selection Operator (LASSO)

- ❑ Coeficientes pequeños serán más susceptibles a ser cero.
- ❑ **Se puede usar como un método para reducir dimensionalidad.**
- ❑ Alimentar al clasificador con las variables sobrantes.

Acc 83%, F1-score: (0) 27%, (1) 90, (2) 70%

Real

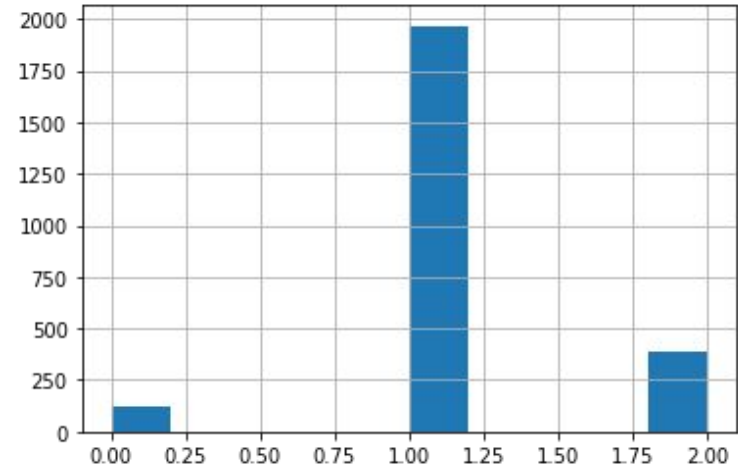


HS

OL

N

Estimados

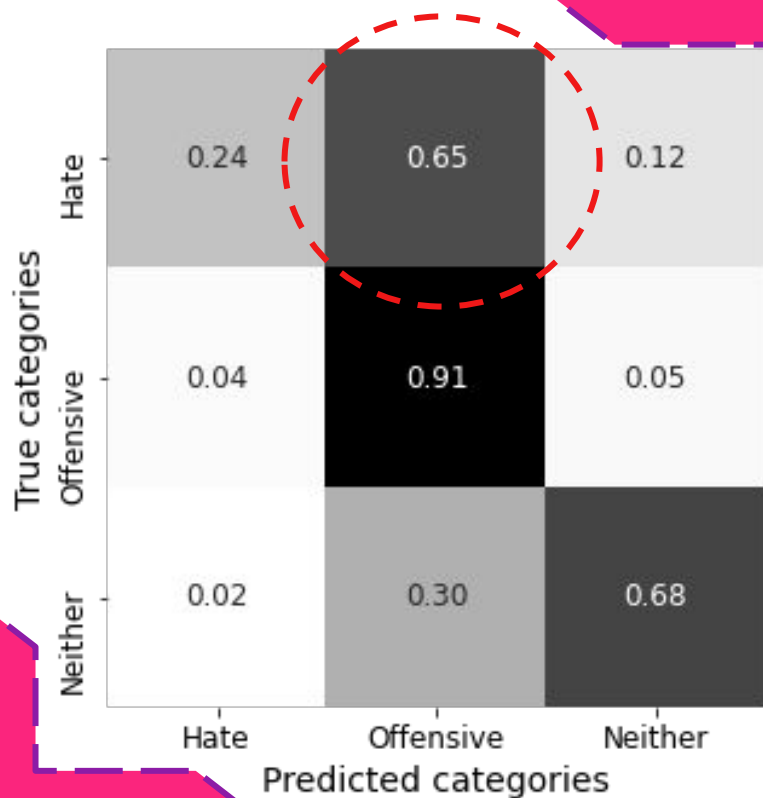


HS

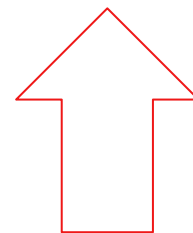
OL

N

Resultados



No se diferencia el discurso de odio, respecto al lenguaje ofensivo



Podemos considerar erróneamente un discurso de odio.

tweet: fucking queer

True: 1

Pred: 0

tweet: @The__Sweetest no shit you are, what euros would call a typical American, fat as fuck and goes to eat fast food all the time, dumb bitch.

True: 1

Pred: 0

tweet: I hate when white trash try to act like they're my equal. It only makes it that much clearer how white trash they are.

True: 0

Pred: 2

tweet: Benzino is a bitch point blank period

True: 1

Pred: 0

tweet: Teacher: You had all weekend to do you homework! Me: Uhm, sorry bitch but I have a life..

True: 1

Pred: 0

- Los tweets con **groserías** tienden más a ser considerados **discurso de odio**.

tweet: Their #1 insult. Even if it's self-deprecating. The first thing straights go to is faggot, queer, tranny, gay. Even supposed "allies".

True: 1

Pred: 0

- Los **comentarios sexistas** no se clasificaban como discursos de odio.

tweet: RT @stayfoqued: Don't eat pussy until she show you the Vagfax

True: 2

Pred: 1

tweet: Winking emoji? I'm about to fuck her right in the pussy

True: 2

Pred: 1

f*ggot

tweet: @mjs79 @ChingonAbe well he is at Genos. So he is the ultimate faggot

True: 2

Pred: 1

tweet: I wish everyone I knew wasn't a faggot a try hard or a sketchball

True: 2

Pred: 1

tweet: @alyssawiens faggot

True: 2

Pred: 1

tweet: This nigga Magic Johnson got a Grade A faggot for a son 😂😂.. He had to have done some terrible shit growin up lol..

True: 2

Pred: 1

n*gga

tweet: @kieffer_jason @C_janacek07 nigga stop being a bitch and come to the high school You already said yes

True: 2

Pred: 1

tweet: Lemme find out that bitch nigga talking shit

True: 2

Pred: 1

tweet: A real gangster ass nigga know the play, the real gangster ass niggas gets the flyest of the bitches, ask that gangster ass nigga little J

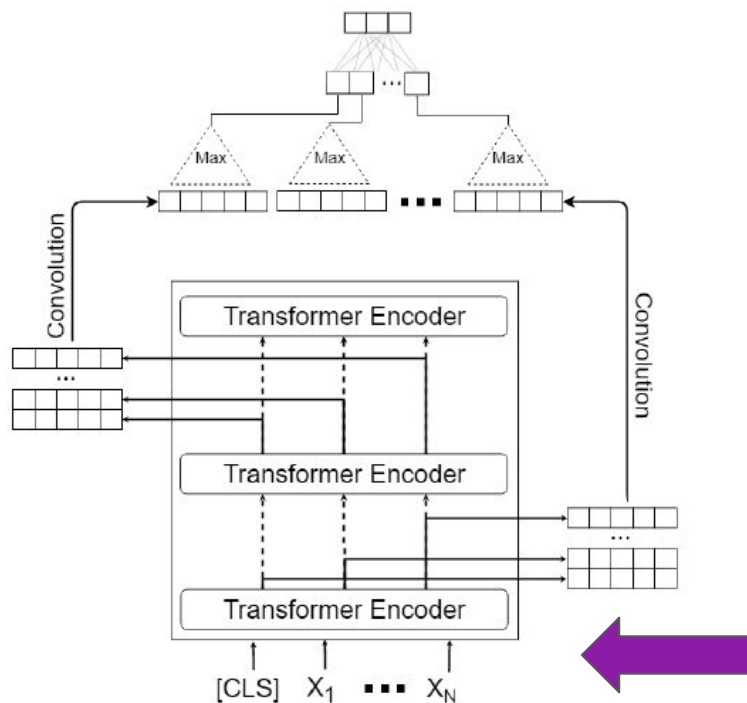
True: 2

Pred: 1



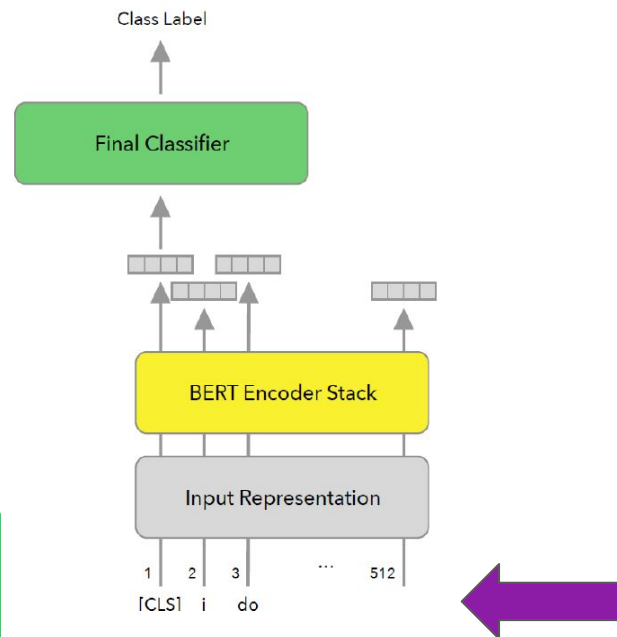
Deep Learning

Deep Learning for Hate Speech Detection in Tweets, (Pinkesh Badjatiya et al **2017**).




A **BERT**-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media. (Noel Crispi et al **2019**)


Using Transfer-based Language Models to Detect Hateful and Offensive Language Online. (Vebjørn Isaksen et al **2020**).





A BERT-Bases Transfer Learning Approach for Hate Speech Detection in Online Social Media. (Noel Crispi et al **2019**)

Mismo fenómeno!!

tweet: @AustinG1135 I do not like talking to you faggot and I did but in a nicely way fag
pred: Offensive Language 
true: Hate Speech

tweet: @justinbieber have fun getting deported you fucking faggot
pred: Hate Speech 
true: Offensive Language

tweet: Accept your flaws and imperfections because that's what makes you, YOU! See I'm a fuck, small dick faggot with parents who don't love me.
pred: Offensive Language 
true: Hate Speech

tweet: Starks being a faggot
pred: Hate Speech 
true: Offensive Language

tweet: @AustinG1135 answer my snapchat faggot
#butthurt
pred: Offensive Language
true: Hate Speech



Al menos muestra el sesgo de las anotaciones.....?

tweet: RT @dril: ah, i can smell it,. its just about ready. *opens the oven up and pulls out a sshitty burnt up ritz cracker* my perfect boy's lu…

True: Offensive Language

Pred: Neither ←

tweet: I hit raw while im listening to papoose

True: Offensive Language

Pred: Neither ←

tweet: When they bitch about "carbon emissions" tell them to stop breathing because exhaling adds carbon dioxide to the atmosphere...

True: Offensive Language

Pred: Neither ←

tweet: RT @MrHoratioSanz: Vin Scully once called me a camel jockey. #liesaboutvinscully

True: Hate Speech

Pred: Neither ←

Ventaja de pre-entrenar no con tweets. **Pre-entrenamiento con wikipedia.**

Conclusiones.

- El problema de diferenciar discurso de odio, de lenguaje ofensivo **sigue siendo complicado.**
- **Existe un sesgo en las las anotaciones también.**

tweet: The most beautiful women be havin' the most retarded captions on their pics! Sh*t be extra dumb...

True: 2

Pred: 1

Entonce ¿Cómo podemos evaluar de manera correcta el clasificador?

tweet: @Im_a_Asshole_ @ImTooMuch @MyAssHoleSoWet So you gon call me a nigger because of that? Fuck wrong with you?

True: 2

Pred: 1

Davidson, T., Bhattacharya, D., & Weber, I. (2019). **Racial bias in hate speech and abusive language detection datasets.**



¿ML o DL? Gröndahl et al. (2018) probaron distintos modelos desde regresión logística hasta modelos de aprendizaje profundo actuales, y no hubo gran diferencia.



Estudiar más los datos.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. *All you need is "love": Evading hate speech detection*. 2018

T Arango, A., Pérez, J., & Poblete, B. (2020). **Hate speech detection is not as easy as you may think: A closer look at model validation.**

Method	Class	Prec.	Rec.	F1
Badjatiya et al. [7] Emb. over all dataset	Neither	95.5	96.8	96.1
	Racist	94.5	93.5	94.0
	Sexist	91.2	87.5	89.3
	Micro avg.	94.6	94.6	94.6
	Macro avg.	93.7	92.6	93.1
~				
Method	Class	Prec.	Rec.	F1
Badjatiya et al. [7] Emb. over train set	Neither	82.3	94.7	88.1
	Racist	78.0	64.0	70.2
	Sexist	84.5	47.8	60.9
	Micro avg.	82.3	82.1	80.7
	Macro avg.	81.6	68.9	73.1

Considerando de dónde provenían los tweets. Sólo dos eran propietarios del casi 80% de HS tweets.

Español? Falta información!!.

Language	Method	Acc
Spanish	LR	.704
	CNN	.650
	RNN	.674
	BERT	.605

Huang, X., Xing, L., Deroncourt, F., & Paul, M. J. (2020).
**Multilingual Twitter corpus and baselines for evaluating
demographic bias in hate speech recognition.**

+ Age, Race, Gender, County.

HaterNet	SVM (Pereira-Kohatsu et al., 2019)	–	48.3	–
	LSTM+MLP (Pereira-Kohatsu et al., 2019)	–	61.1	–
	BETO (Our proposal)	88.7	65.8	77.2
HatEval	multi-channel BERT (Sohn & Lee, 2019)	–	–	76.6
	Ensemble voting classifier (Plaza-del-Arco et al., 2020)	80.0	68.8	74.2
	BERT (Gertner et al., 2019)	73.0	72.7	72.9
	SVM (Argota Vega et al., 2019)	76.1	69.9	73.0
	BiGRU (Paetzold et al., 2019)	77.1	52.1	64.6
	BETO (Our proposal)	79.7	75.5	77.6

Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). **Comparing pre-trained language models for Spanish hate speech detection.**

Un intento.... Mex_a3t **Agresividad** **Acc: 77%**

50 % clasificó como NO
agresividad como agresividad

tweet: venir a pagar el cable y ahorrar para comprarle ropa a mi novia no me hace menos hombre putos.

True: No Agresivo

pred Agresivo

tweet: "llégale a la verga pinche pendejo" es el título de una canción que estoy escribiendo. es de amor.

True: No Agresivo

pred Agresivo



Mismo fenómeno: Existen palabras que hacen que el clasificador se equivoque.

Referencias

- Davidson, T., Warmusley, D., Macy, M., & Weber, I. (2017, May). **Automated hate speech detection and the problem of offensive language**. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1).
- Davidson, T., Bhattacharya, D., & Weber, I. (2019). **Racial bias in hate speech and abusive language detection datasets**. *arXiv preprint arXiv:1905.12516*.
- Isaksen, V., & Gambäck, B. (2020, November). **Using Transfer-based Language Models to Detect Hateful and Offensive Language Online**. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 16-27).