Procesamiento de Lenguaje Natural / TP

Detección de discurso de odio (*hate speech*) en medios sociales

Medios sociales & hate speech

Se considera discurso de odio cualquier forma de comunicación de palabra, por escrito o a través del comportamiento, que sea **un ataque o utilice lenguaje peyorativo o discriminatorio** en relación con una persona o un grupo **sobre la base de quiénes son** o, en otras palabras, en razón de su religión, origen étnico, nacionalidad, raza, color, ascendencia, género u otro factor de identidad.

Medios sociales & hate speech

Se considera discurso de odio cualquier forma de comunicación de palabra, por escrito o a través del comportamiento, que sea **un ataque o utilice lenguaje peyorativo o discriminatorio** en relación con una persona o un grupo **sobre la base de quiénes son** o, en otras palabras, en razón de su religión, origen étnico, nacionalidad, raza, color, ascendencia, género u otro factor de identidad.

- Esta es solo una possible definición.
  - Cada red social tiene su propia definición (y sus propias políticas de acción frente al hate speech).
  - El INADI hizo un documento sobre discurso de odio

- Características principales:
  - Incitación al odio o la violencia.
  - Ataca, subestima o menosprecia.
  - Tiene targets específicos.
  - No puede ser considerado humor.

Se pueden encontrar definiciones más específicas de toxicicy, bullying, misogyny…

https://www.argentina.gob.ar/sites/default/files/12_01_2021_informe_discurso_de_odio.pdf

Medios sociales & hate speech

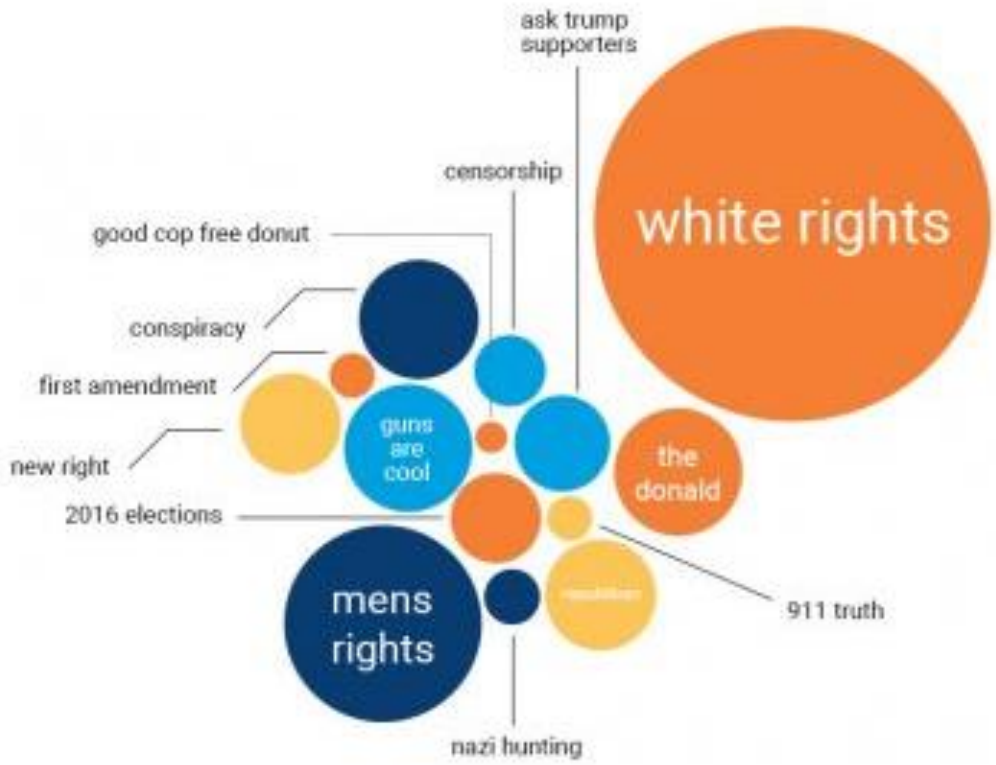
Se considera discurso de odio cualquier forma de comunicación de palabra, por escrito o a través del comportamiento, que sea **un ataque o utilice lenguaje peyorativo o discriminatorio** en relación con una persona o un grupo **sobre la base de quiénes son** o, en otras palabras, en razón de su religión, origen étnico, nacionalidad, raza, color, ascendencia, género u otro factor de identidad.

AVISO

Vamos a ver algunos ejemplos que pueden no ser apropiados y que representan ejemplos de hate speech.

Los ejemplos están a modo ilustrativo, de ninguna forma representan un endorsement a dicho comportamiento.

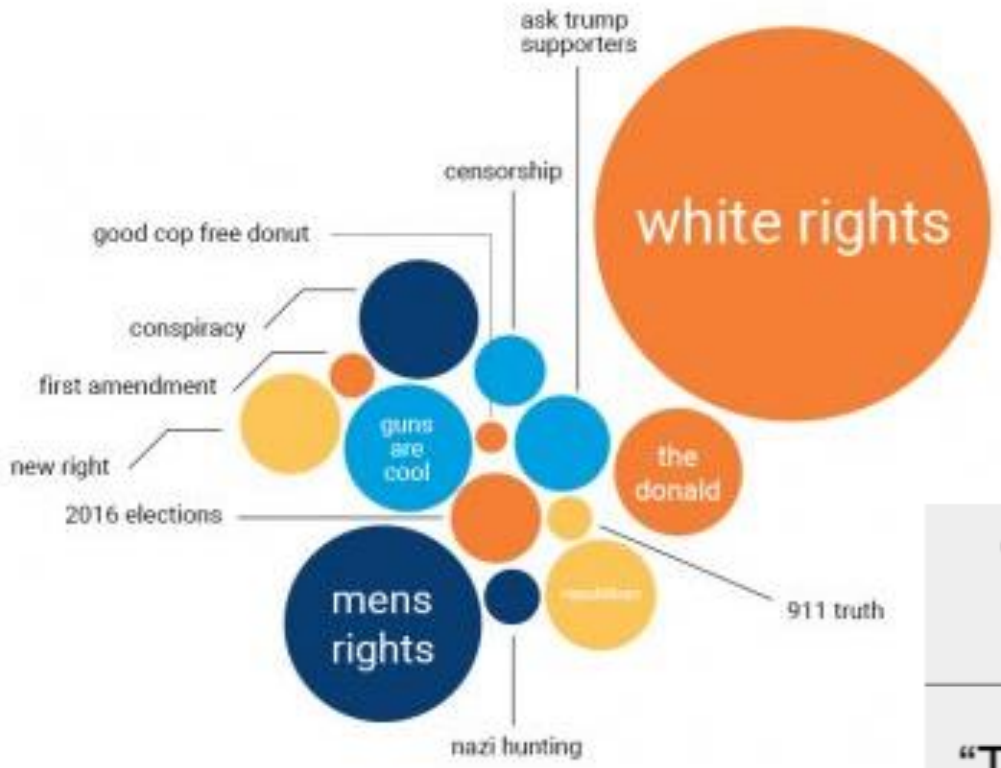
Medios sociales & hate speech



Most common words found in Hate Comments

Most common words found in Non-hate Comments

Medios sociales & hate speech

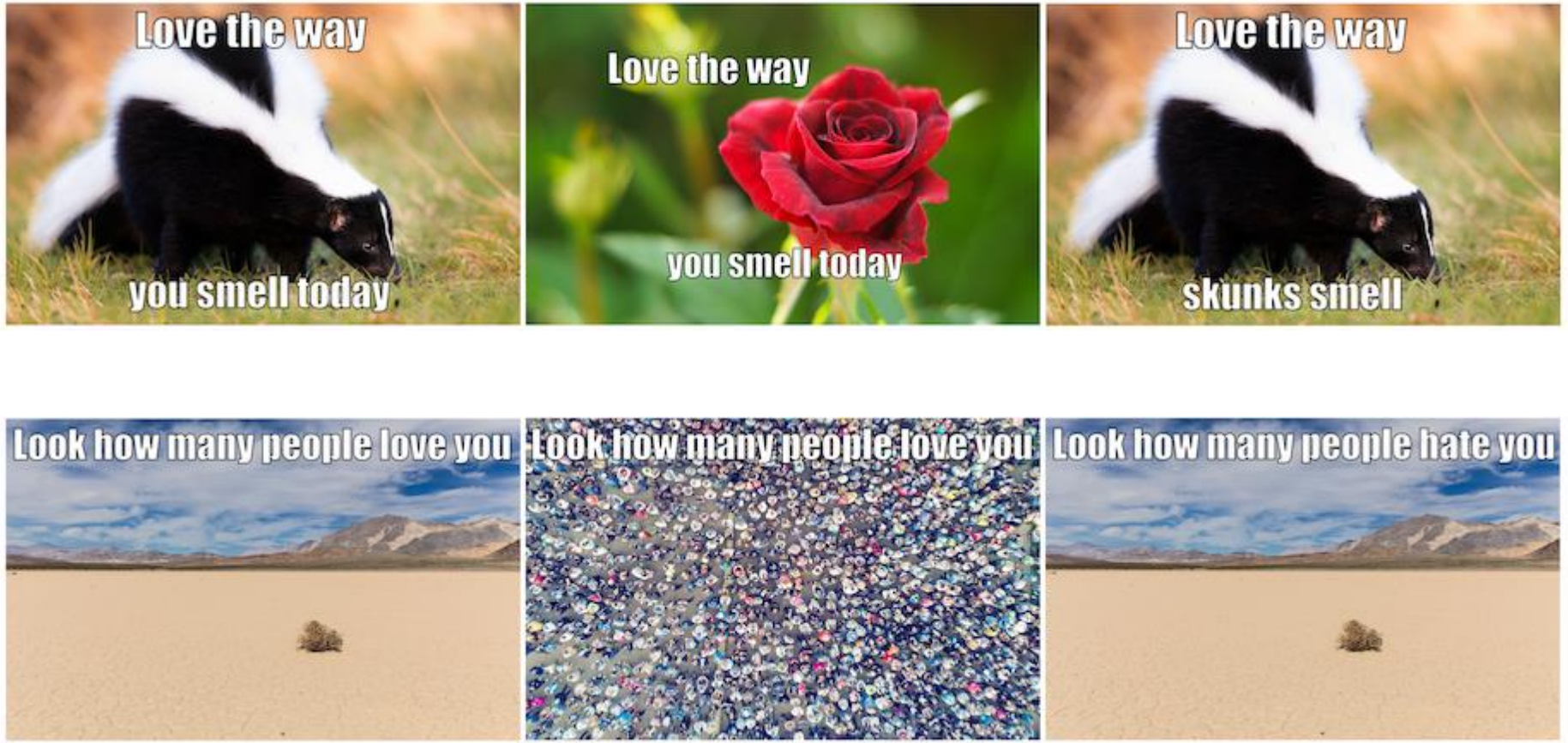


"[F]or many **Africans**, the most threatening kind of ethnic hatred is **black** against **black**." - *New York Times*

"There is a great discrepancy between **whites** and **blacks** in SA. It is … [because] **blacks** will always be the most backward race in the world." Anonymous user, *Gab.com*

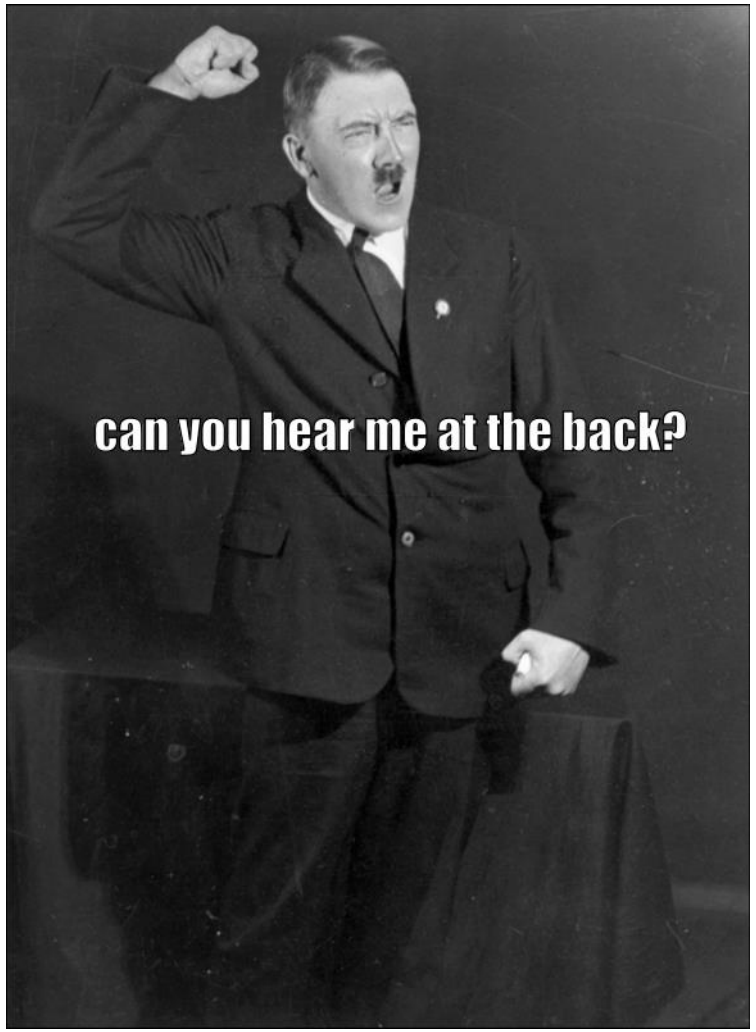
Medios sociales & hate speech

No es solo texto!



Medios sociales & hate speech

No es solo texto!



[redacted] ...

Replying to [redacted]

When you see the nigga who gave you the gun in court

3:49 pm · 22 Mar 2019 · Twitter for iPhone

Medios sociales & hate speech

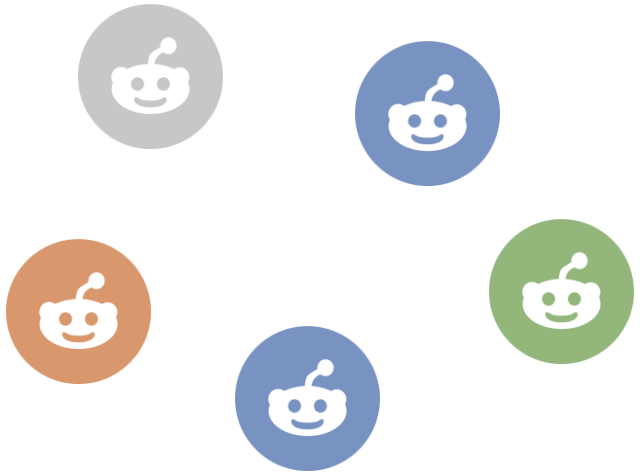
Cuál es el objetivo?

Detectar aquellos comentarios de Reddit que contienen hate speech distinguiéndolos de aquellos que no!

Medios sociales & hate speech

Cuál es el objetivo?

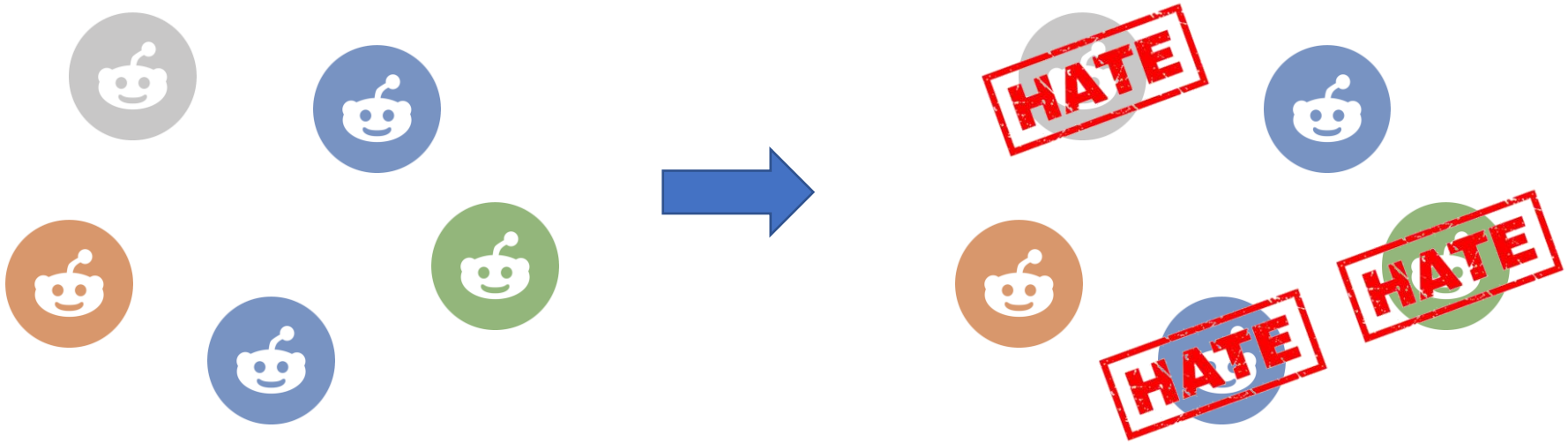
Detectar aquellos comentarios de Reddit que contienen hate speech distinguiéndolos de aquellos que no!



Medios sociales & hate speech

Cuál es el objetivo?

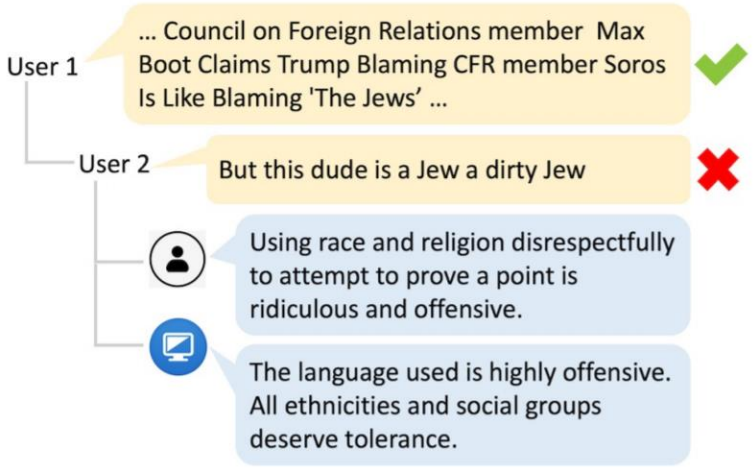
Detectar aquellos comentarios de Reddit que contienen hate speech distinguiéndolos de aquellos que no!



Medios sociales & hate speech

Y los datos?

- **“A Benchmark Dataset for Learning to Intervene in Online Hate Speech”**
  - El objetivo del trabajo no fue detectar hate, sino diseñar estrategias de intervención con respuestas automáticas a conversaciones quqe tienen hate speech.



Conversation	Hate Speech	Human-Written Intervention Responses
1. User 1: United Kingdom: 'Schoolboy, 15, given detention for backing UKIP during classroom debate' 2. User 2: The education system is full of re***ds! Yes, most school teachers are ret***ed lefties! Teach your children to laugh at these ret***ed lefties! 3. User 3: Asking a teacher to not be a leftist is like asking a medieval munk to question the Pope. 4. User 4: The Jews are like Sjws, they infest everything.	2, 4	➢ Use of this language is not tolerated and it is uncalled for. ➢ Use of the slurs and insults here is unacceptable in our discourse as it demeans and insults and alienates others. ➢ I recommend that you research the holocaust, you might change your opinion.

https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech

Medios sociales & hate speech

Y los datos?

- **“A Benchmark Dataset for Learning to Intervene in Online Hate Speech”**
    - El objetivo del trabajo no fue detectar hate, sino diseñar estrategias de intervención con respuestas automáticas a conversaciones quqe tienen hate speech.

- Aproximadamente 5k conversaciones, con 22k comentarios.
    - Actualmente se encuentran disponibles alrededor de 4k con 18k comentarios.

- Comentarios pertenecientes a 10 subreddits:
    - `r/DankMemes`
    - `r/Imgoingtohellforthis`
    - `r/KotakuInAction`
    - `r/MensRights`
    - `r/MetaCanada`
    - `r/MGTOW`
    - `r/PussyPass`
    - `r/PussyPassDenied`
    - `r/The Donald`
    - `r/TumblrInAction`

- Para cada subreddit, recolectarons los 200 posts más “hot”.
- Buscaron posts con keywords de hate.
- Recolectaron las conversaciones.
- No hay repetidos.

https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech

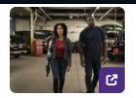
Medios sociales & hate speech

Y los datos?

Posted by u/JustOneAmongMany **Knitta, please!** 2 years ago

595

[SocJus] The Mary Sue: "In a Shocking Move, Netflix Cancels Marvel's Luke Cage After Two Seasons" (From the article: "Its cancellation is also a massive blow for representation, as Luke Cage was easily the most inclusive of all of Marvel's Netflix shows.") archive.fo/uSlxX



SOCJUS

275 Comments Share Save Hide Report

This thread is archived
New comments cannot be posted and votes cannot be cast

SORT BY BEST

View all comments

pasta4u · 2y

wouldn't the defenders or whatever they are as a group be the most diverse group ? Since you know it has a blind dude , orphan , black man , woman and all of their supporting cast ?

I also believe all the Marvel stuff is going to get canceled on Netflix because next year Disney is launching their own streaming platform and will want as many characters as it can get for shows on that

406 Share Report Save

MosDaf · 2y

'inclusive' = not white

472 Share Report Save

Continue this thread →

Posted by u/Traxorbomber 2 years ago

312

444 (Hungarian news site) on Gab: "Nazis can freely spew their idiocity on the hate speech of Facrbook"

[Journalism ethics] +2 [Company under attack from media] +1 My first post here, so apology in advance if it breaks the guidlines.

Link to the newspiece (it's in hungarian) http://archive.is/OLB3T

Sone highlights (translated):

-(Regarding the Pitsburg shooter): "The shooter openly admitted his anti-simitic views, which he spread on a social-media site designed by people who think when they can't wish openly for the death of muslims and jews, then their free speech rights are violated."

- "He spread his views mainly on the social media platfor Gab, where his neighbours naturally didn't follow him, because the site mainly provides a safe haven for those far-right users,who where alredy banned because of hates speech from the sites reserved to the clear-thinking part of the world." -(On Gabs history): "Their goal is to make hatespeech and far-right ideologies appear revolutionary opposed to the liberal "opinionterror" forced on them".

46 Comments Share Save Hide Report 98% Upvoted

This thread is archived
New comments cannot be posted and votes cannot be cast

SORT BY BEST

CrankyDClown · 2y
Groomy Beardman

What car did the shooter drive? We must burn the factory to the ground. I'm sure a lot more nazis drive the same brand.

176 Share Report Save

Medios sociales & hate speech

Y los datos?

- **“A Benchmark Dataset for Learning to Intervene in Online Hate Speech”**
    - El objetivo del trabajo no fue detectar hate, sino diseñar estrategias de intervención con respuestas automáticas a conversaciones quqe tienen hate speech.

- **`reddit_comments.json`**. Un `jsonarray` donde cada elemento `json` representa un comentario. Para cada comentario se tienen diferentes atributos. El único obligatorio de utilizar es body, que contiene el texto del comentario. Cada comentario tiene un atributo llamado is_hate que indica si el comentario tiene hate speech (`1`) o no (`0`).

- **`conversations.csv`**. Cada línea del archivo representa un hilo conversacional. Se tiene una lista de comentarios separados por comas indicando el orden en el que se sucedieron.

- **`reddit_authors.json`**. Un `jsonarray` donde cada elemento `json` representa un autor. Complementa la información de `reddit_comments.json` con atributos de los autores. Puede suceder que no se encuentre la información de todos los autores de comentarios dado que algunos se encuentran actualmente suspendidos.

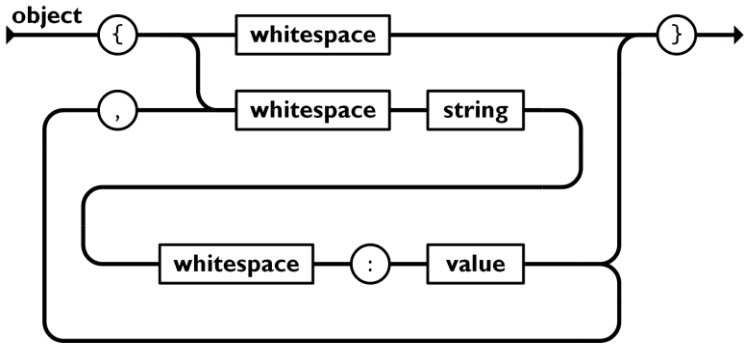
https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech

Medios sociales & hate speech

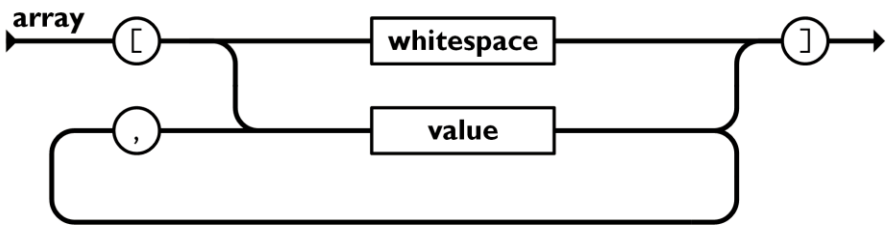
Y los datos?

- **“A Benchmark Dataset for Learning to Intervene in Online Hate Speech”**
  - El objetivo del trabajo no fue detectar hate, sino diseñar estrategias de intervención con respuestas automáticas a conversaciones quqe tienen hate speech.

- Formato JSON → JavaScript Object Notation



Cada **objeto** es una colección de pares <nombre atributo, valor atributo>.



Los **objetos** se pueden combinar en la forma de una secuencia ordenada denominada **array**.

Medios sociales & hate speech

Formato de los datos


```
'{"_replies": [], "id": "e8q18lf", "total_awards_received": 0, "approved_at_utc": null, "comment_type":
null, "edited": false, "mod_reason_by": null, "banned_by": null, "author_flair_type": "text",
"removal_reason": null, "link_id": "t3_9smha5", "author_flair_template_id": null, "likes": null,
"user_reports": [], "saved": false, "banned_at_utc": null, "mod_reason_title": null, "gilded": 0,
"archived": true, "no_follow": false, "author": "PessimisticPaladin", "can_mod_post": false, "created_utc":
1540905081.0, "send_replies": true, "parent_id": "t3_9smha5", "score": 4, "author_fullname": "t2_12qeux",
"approved_by": null, "mod_note": null, "all_awardings": [], "subreddit_id": "t5_33726", "body": "A
subsection of retarded Hungarians? Ohh boy. brace for a livid Bulbasaur coming in here trying to hate a hole
in some of her stupider countrymen.", "awarders": [], "author_flair_css_class": null, "name": "t1_e8q18lf",
"author_patreon_flair": false, "downs": 0, "author_flair_richtext": [], "is_submitter": false, "body_html":
"<div class=\\"md\\"><p>A subsection of retarded Hungarians? Ohh boy. brace for a livid Bulbasaur coming in
here trying to hate a hole in some of her stupider countrymen.</p>\\n</div>", "gildings": {},
"collapsed_reason": null, "distinguished": null, "associated_award": null, "stickied": false,
"author_premium": false, "can_gild": true, "top_awarded_type": null, "author_flair_text_color": "dark",
"score_hidden": false, "permalink":
"/r/KotakuInAction/comments/9smha5/444_hungarian_news_site_on_gab_nazis_can_freely/e8q18lf/", "num_reports":
null, "locked": false, "report_reasons": null, "created": 1540933881.0, "subreddit": "KotakuInAction",
"author_flair_text": "You were thrown into the GG pit. I was born in it, molded by it.", "treatment_tags":
[], "collapsed": false, "subreddit_name_prefixed": "r/KotakuInAction", "controversiality": 0,
"author_flair_background_color": "", "collapsed_because_crowd_control": null, "mod_reports": [],
"subreddit_type": "public", "ups": 4, "is_hate": 1}'
```

Medios sociales & hate speech

Formato de los datos

'{"_replies": [], "id": "e8q18lf", "total_awards_received": 0, "approved_at_utc": null, "comment_type": null, "edited": false, "mod_reason_by": null, "banned_by": null, "author_flair_type": "text", "removal_reason": null, "link_id": "t3_9smha5", "author_flair_template_id": null, "likes": null, "user_reports": [], "saved": false, "banned_at_utc": null, "mod_reason_title": null, "gilded": 0, "archived": true, "no_follow": false, "author": "PessimisticPaladin", "can_mod_post": false, "created_utc": 1540905081.0, "send_replies": true, "parent_id": "t3_9smha5", "score": 4, "author_fullname": "t2_12qeux", "approved_by": null, "mod_note": null, "all_awardings": [], "subreddit_id": "t5_33726", "body": "A subsection of retarded Hungarians? Ohh boy. brace for a livid Bulbasaur coming in here trying to hate a hole in some of her stupider countrymen.", "awarders": [], "author_flair_css_class": null, "name": "t1_e8q18lf", "author_patreon_flair": false, "downs": 0, "author_flair_richtext": [], "is_submitter": false, "body_html": "<div class=\\"md\\"> A subsection of retarded Hungarians? Ohh boy. brace for a livid Bulbasaur coming in here trying to hate a hole in some of her stupider countrymen. \\n ", "gildings": {}, "collapsed_reason": null, "distinguished": null, "associated_award": null, "stickied": false, "author_premium": false, "can_gild": true, "top_awarded_type": null, "author_flair_text_color": "dark", "score_hidden": false, "permalink": "/r/KotakuInAction/comments/9smha5/444_hungarian_news_site_on_gab_nazis_can_freely/e8q18lf/", "num_reports": null, "locked": false, "report_reasons": null, "created": 1540933881.0, "subreddit": "KotakuInAction", "author_flair_text": "You were thrown into the GG pit. I was born in it, molded by it.", "treatment_tags": [], "collapsed": false, "subreddit_name_prefixed": "r/KotakuInAction", "controversiality": 0, "author_flair_background_color": "", "collapsed_because_crowd_control": null, "mod_reports": [], "subreddit_type": "public", "ups": 4, "is_hate": 1}'

Medios sociales & hate speech

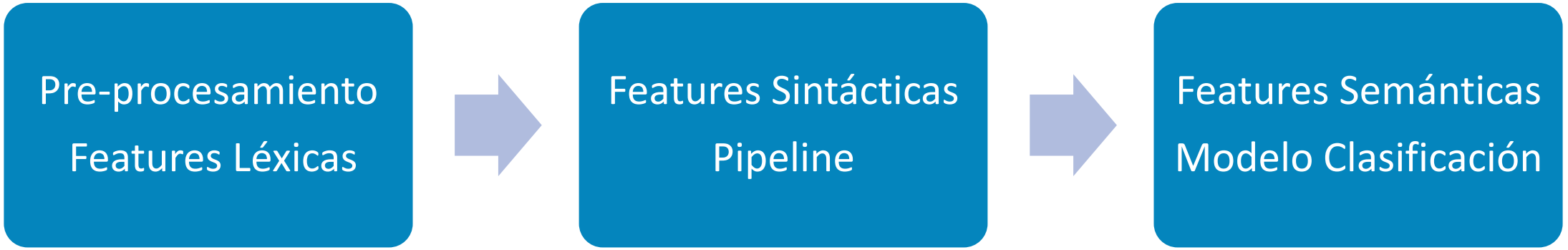
Formato de los datos

'{"_replies": [], "id": "e8q18lf", "total_awards_received": 0, "approved_at_utc": null, "comment_type": null, "edited": false, "mod_reason_by": null, "banned_by": null, "author_flair_type": "text", "removal_reason": null, "link_id": "t3_9smha5", "author_flair_template_id": null, "likes": null, "user_reports": [], "saved": false, "banned_at_utc": null, "mod_reason_title": null, "gilded": 0, "archived": true, "no_follow": false, "**author**": "PessimisticPaladin", "can_mod_post": false, "created_utc": 1540905081.0, "send_replies": true, "parent_id": "t3_9smha5", "score": 4, "author_fullname": "t2_12qeux", "approved_by": null, "mod_note": null, "all_awardings": [], "subreddit_id": "t5_33726", "**body**": "A subsection of retarded Hungarians? Ohh boy. brace for a livid Bulbasaur coming in here trying to hate a hole in some of her stupider countrymen.", "awarders": [], "author_flair_css_class": null, "name": "t1_e8q18lf", "author_patreon_flair": false, "downs": 0, "author_flair_richtext": [], "is_submitter": false, "body_html": "<div class=\\"md\\"> A subsection of retarded Hungarians? Ohh boy. brace for a livid Bulbasaur coming in here trying to hate a hole in some of her stupider countrymen. \\n ", "gildings": {}, "collapsed_reason": null, "distinguished": null, "associated_award": null, "stickied": false, "author_premium": false, "can_gild": true, "top_awarded_type": null, "author_flair_text_color": "dark", "score_hidden": false, "permalink": "/r/KotakuInAction/comments/9smha5/444_hungarian_news_site_on_gab_nazis_can_freely/e8q18lf/", "num_reports": null, "locked": false, "report_reasons": null, "created": 1540933881.0, "**subreddit**": "KotakuInAction", "author_flair_text": "You were thrown into the GG pit. I was born in it, molded by it.", "treatment_tags": [], "collapsed": false, "subreddit_name_prefixed": "r/KotakuInAction", "controversiality": 0, "author_flair_background_color": "", "collapsed_because_crowd_control": null, "mod_reports": [], "subreddit_type": "public", "ups": 4, "is_hate": 1}'

https://praw.readthedocs.io/en/latest/code_overview/models/submission.html

Medios sociales & hate speech

Qué tienen que hacer?



Semana 1

Semana 2

Semana 3

Medios sociales & hate speech

Qué tienen que hacer?

Pre-procesamiento

Features Léxicas

- Procesar el dataset.
- Pre-procesar el texto (remover stopwords, puntuación, stemmer, lemmatization, …)
- Definir carácterísticas léxicas para representar los comentarios.
- Definir una representación para los comentarios (BOW, N-grams, …)
  - La semana que viene vamos a ver más sobre esto.
- Pueden considerar características que no sean textuales.
- Pueden incluir características extraídas de las conversaciones.

- Analizar cómo varía el vocabulario de los comentarios con la aplicación del pre-procesamiento.
  - Por ejemplo, mostrar distribuciones de frecuencias, términos más frecuentes…

Medios sociales & hate speech

Qué tienen que entregar?

Pre-procesamiento

Features Léxicas

- Notebook con:
  - Carga de dataset.
  - Selección de atributos de los comentarios. Mencionar brevemente por qué eligieron cada uno de ellos.
  - Estadísticas del dataset.
    - Tamaño del vocabulario, términos más frecuentes, términos más frecuentes por clase…
  - Aplicación de pipeline de pre-procesamiento. Mencionar brevemente por qué eligieron cada uno de los pasos.
  - Comparación de estadísticas del dataset antes y después del pre-procesamiento.
  - Selección de características léxicas.

Medios sociales & hate speech

Qué tienen que entregar?

Pre-procesamiento

Features Léxicas

Fecha de entrega: **26 de Junio 2021**

- Notebook con:
  - Carga de dataset.
  - Selección de atributos de los comentarios. Mencionar brevemente por qué eligieron cada uno de ellos.
  - Estadísticas del dataset.
    - Tamaño del vocabulario, términos más frecuentes, términos más frecuentes por clase…
  - Aplicación de pipeline de pre-procesamiento. Mencionar brevemente por qué eligieron cada uno de los pasos.
  - Comparación de estadísticas del dataset antes y después del pre-procesamiento.
  - Selección de características léxicas.

Medios sociales & hate speech

Qué tienen que entregar?

Pre-procesamiento

Features Léxicas

Fecha de entrega: **26 de Junio 2021**

La notebook debe poder ejecutarse sin errores y debe incluir los outputs generados!

- Notebook con:
  - Carga de dataset.
  - Selección de atributos de los comentarios. Mencionar brevemente por qué eligieron cada uno de ellos.
  - Estadísticas del dataset.
    - Tamaño del vocabulario, términos más frecuentes, términos más frecuentes por clase…
  - Aplicación de pipeline de pre-procesamiento. Mencionar brevemente por qué eligieron cada uno de los pasos.
  - Comparación de estadísticas del dataset antes y después del pre-procesamiento.
  - Selección de características léxicas.

Procesamiento de Lenguaje Natural / TP

Detección de discurso de odio (hate speech) en medios sociales