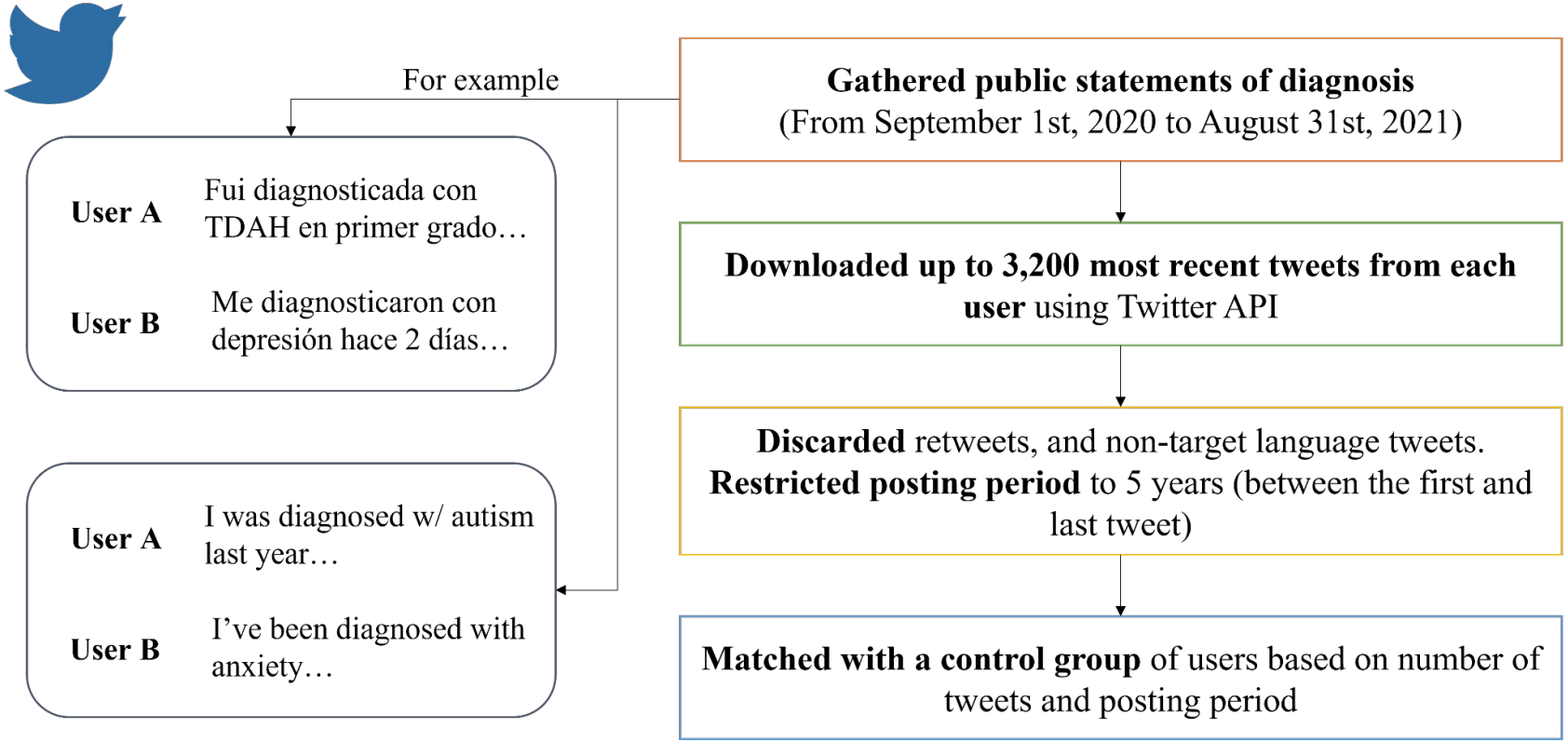
Junta de hoy - 18/abr/2023

(UC Davis)

Recolección de datos ⛏️

Mi metodología de recolección ✏️



Adaptación 

Incendios + Salud mental

1

¿Lugar en donde sucedió?

Recolectar tweets relacionados con X evento utilizando las palabras clave entre las fechas en las que sucedió el evento.

Tweets = 1 millón de tweets Usuarios = 800 mil

Verificación manual de la autenticidad de las declaraciones de diagnóstico y filtrado de spam/anuncios, etc.

Muestra aleatoria y etiquetar por anotadores humanos los tweets.

X meses y máximo de ~200 tweets por usuario

Filtrar usuarios con un máximo X tweets, o X número de palabras.

Grupo Depresión

Buscar usuarios con señales de depresión tanto por el texto como por la descripción del perfil.

Usuarios = 3,500

1) Descargar timeline (3,200 más recientes) sin retweets.
2) Eliminar tweets que no estén en inglés.
3) Quedarse con los tweets de los últimos [1,2,3] meses desde el momento de la publicación del tweet relacionado con la depresión.
4) Si el usuario fue identificado por la descripción entre A y B fechas.

Usuarios = 2,500

Grupo Control

Seleccionar aleatoriamente 3,500 usuarios distintos en los que no aparecen términos relacionados con la depresión.

Age-gender match

Usuarios = 3,500

1) Descargar timeline (3,200 más recientes) sin retweets.
2) Eliminar tweets que no estén en inglés.
3) Verificar que sus tweets no aparezcan términos relacionados con la depresión.
4) Quedarse con los tweets entre A y B fechas.

Usuarios = 3,100

2

Grupo Depresión

Recolectar tweets con señales de depresión entre las fechas en las que sucedió el evento.

¿Lugar en donde sucedió?

Expresiones regulares para filtrar usuarios con señales verdaderas de depresión.

Verificación manual de la autenticidad de las declaraciones de diagnóstico y filtrado de spam/anuncios, etc.

Muestra aleatoria y etiquetar por anotadores humanos los tweets.

1) Descargar timeline (3,200 más recientes) sin retweets.
2) Eliminar tweets que no estén en inglés.
3) Quedarse con los tweets de los últimos [1,2,3] meses desde el momento de la publicación del tweet relacionado con la depresión y con un máximo de ~200 tweets por usuario.

X meses y máximo de ~200 tweets por usuario

Filtrar usuarios con un máximo X tweets, o X número de palabras.

Grupo Control

Seleccionamos aleatoriamente X usuarios distintos que hayan posteado entre las fechas en las que sucedió el evento y que no aparecen términos relacionados con la depresión.

Para cada usuario del grupo depresión hacemos match con un usuario de control basados en el número de tweets que tienen (?)

Age-gender match

1) Descargar timeline (3,200 más recientes) sin retweets.
2) Eliminar tweets que no estén en inglés.
3) Verificar que sus tweets no aparezcan términos relacionados con la depresión.
4) Quedarse con los tweets entre A y B fechas.

⚠️ Comentarios ⚠️

(Coppersmith et al., 2015)[1] (Guntuku et al., 2015)[2]

En general, los grupos de control se selección aleatoria de usuarios de Twitter. Sin embargo, las afecciones físicas y mentales tienen diferentes tasas de prevalencia en función de la edad y el sexo.

Dos Reis y Culotta (2015) demostraron que, si no se tienen en cuenta estos factores puede dar lugar a grupos de control sesgados que distorsionen los resultados, por lo que nuestro objetivo es formar grupos de control de edad y género similares.

Existe abundante bibliografía que investiga la influencia de la edad y el sexo en el lenguaje (Pennebaker, 2011).

(Cohan et al., 2018)[3]

Los usuarios de control se eligieron de un grupo de usuarios de control candidatos en función de su similitud con los usuarios diagnosticados, medida por su número de publicaciones y los subreddits en los que publicaban. Esto se hace para evitar sesgos entre los grupos de control y diagnosticados en el conjunto de datos y evitar hacer la tarea de identificar a dichos usuarios artificialmente fácil.

⚠️ Comentarios ⚠️

Tweet ancla ⚓ → tweet relacionado con la depresión

Puede pasar que no existan tweets de los últimos X meses desde el tweet ancla.

- Por ejemplo: los más recientes serán del 2023 para abajo y el tweet ancla es del 2018…

Expresiones o frases para buscar 🔍

Tweets:

I was diagnosed with (major/severe/clinical) depression

I've been diagnosed with (major/severe/clinical) depression

I am diagnosed with (major/severe/clinical) depression

I have (major/severe/clinical) depression

I have developed (major/severe/clinical) depression

I suffer(ed) from (major/severe/clinical) depression

My (major/severe/clinical) depression

I’m healing from (major/severe/clinical) depression

Descripción:

depression fighter/sufferer/survivor

No practitioner/counselor

Twitter API 

Academic Research Access 📕

Key benefits	Access Twitter's real-time and historical public data with additional features and functionality that support collecting more precise, complete, and unbiased datasets. **More details on included endpoints**
Tweet cap	10 million Tweets / month
Query rules	1024 characters, 1000 streaming rules

[How to apply for academic research access?]

Search Tweets

GET /2/tweets/search/all

This endpoint is only available to those users who have been approved for Academic Research access.

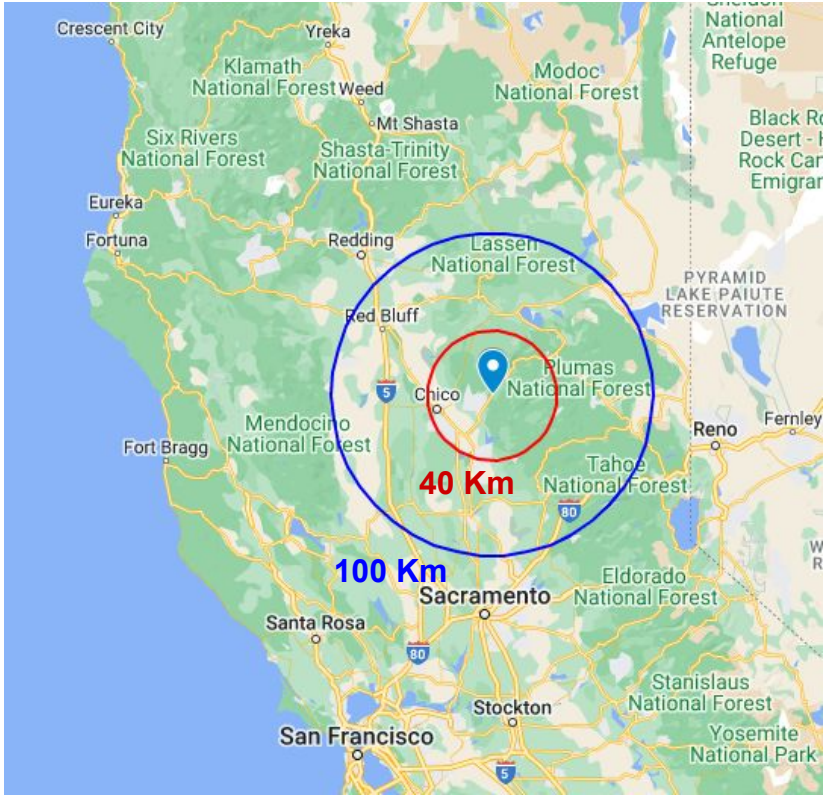
The full-archive search endpoint returns the complete history of public Tweets matching a search query; since the first Tweet was created March 26, 2006.

GET /2/tweets/Search/all

Ejemplo

Incendio Camp Fire

- **Inicio:** Noviembre 8, 2018
- **Fin:** Noviembre 25, 2018
- **Lugar:** Condado de Butte (Norte de California)



Ejemplo

```
bearer_token = 'XXX'

## Point_radius:[longitude latitude radius] - radius must be less than 25mi
query = '#CaliforniaFires OR wildfire OR (Camp Fire) point_radius:[-121.437222 39.810278 40km] lang:en
-is:retweet'
start_time = '2018-11-08T00:00:00Z' # inclusive
end_time = '2018-11-26T00:00:00Z' # exclusive
max_results = 500 # default
limit = 10 # default

client = getAuthentication(bearer_token)
response_lst = searchHistoricalTweets(client, query, start_time, end_time, max_results, limit)

if response_lst[0].meta['result_count'] > 0:
    processTweets(response_lst, 'resultados')
```

Código search_tweets.py

Otros queries

```
query = '-RT wildfire depression place:"Sacramento, CA" OR place:"Los Angeles, CA" lang:en -is:retweet -is:verified'

query = '-RT depression place:"California, USA" lang:en -is:retweet'

query = """-RT ("I was diagnosed" OR "I've been diagnosed" OR "I have been diagnosed" OR "I'm diagnosed") (depression OR "depressive disorder") lang:en -is:retweet"""
```

[Building queries]

Timelines

GET /2/users/:id/tweets

Returns Tweets composed by a single user, specified by the requested user ID. By default, the most recent ten Tweets are returned per request. Using pagination, the most recent 3,200 Tweets can be retrieved.

The Tweets returned by this endpoint count towards the Project-level Tweet cap.

GET /2/users/:id/tweets

Ejemplo

```
bearer_token = 'XXX'
client = getAuthentication(bearer_token)

# Directorio para guardar  los timelines
directory = r'Timeline_users'
LIMIT_TIMELINES = 32

# Archivo csv con los ids de los usuarios a buscar
users_retrieve_list = pd.read_csv(r'ids_search_timelines.csv', dtype={'user_id': str})['user_id'].values.tolist()


# Iterar sobre todos los usuarios a buscar
for i, userID in enumerate(users_retrieve_list, start=1):

    # Buscar timeline
    response_timeline = getUserTimeline(client, userID, LIMIT_TIMELINES)

    # Procesar segun la respuesta
    if response_timeline[0].data:
        processTweets(response_timeline, directory)
```

Código get_user_timeline.py

<> Github </>

Estadísticas Básicas

	Depression	Control
Users	478	1,704
Total Tweets	339,927	1,484,651
# tweets/user		
Mean	711.1	871.3
Median	795.5	844.0
Mean tweet frequency		
Per day	10.8	11.9
Per hour	2.3	2.4

Jaccard's index
0.35

IR
3.6

Siguientes pasos… ⏳

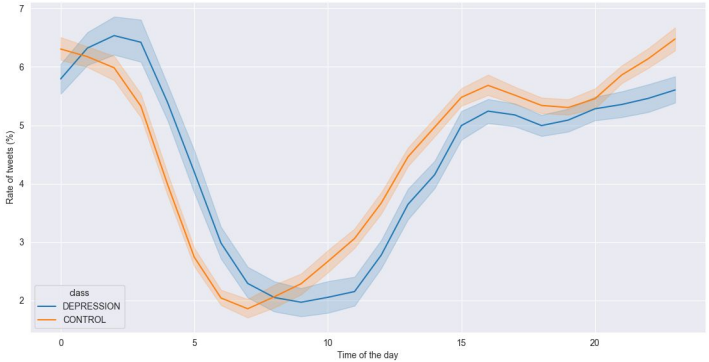
Preprocesamiento 🧼🧽

- @username
- URL
- Hashtag #
- Signos de puntuación
- Minúsculas
- Lemmatizar

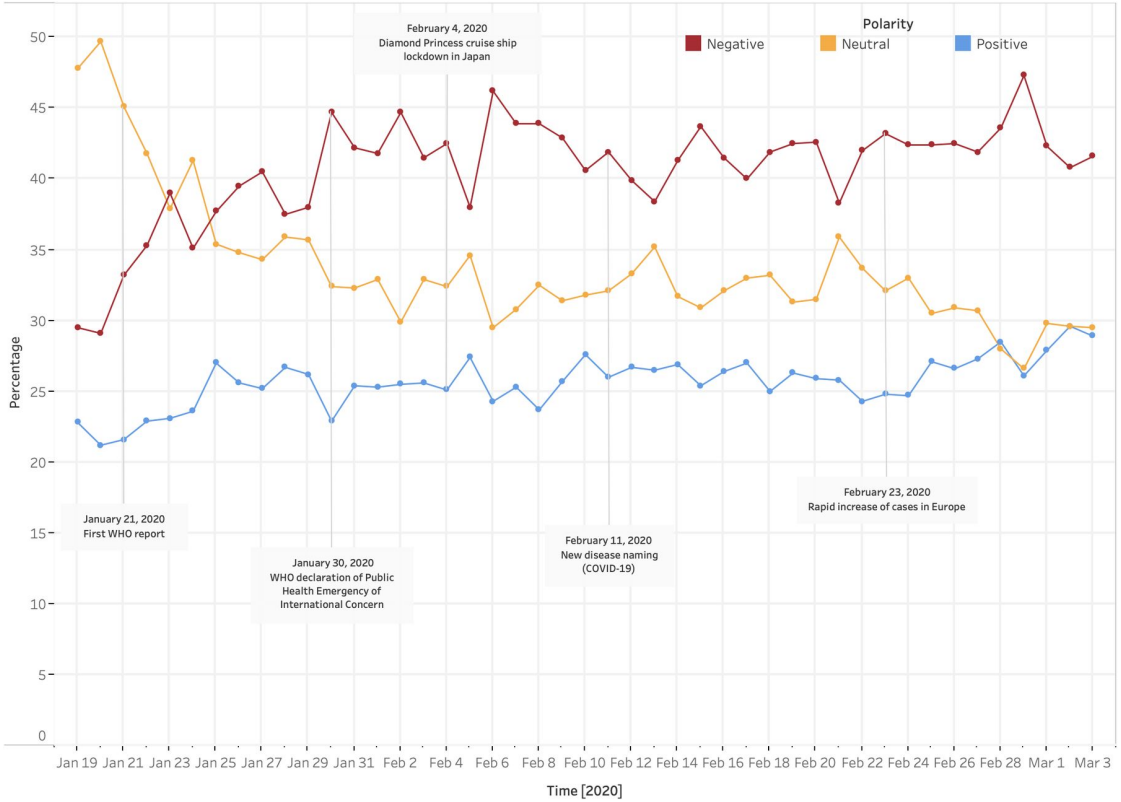
Análisis 🔍📝

- LIWC (I pronoun, swear, anger, family, death…)
- Emolex (sentimiento y emociones)
- Personality (Big5 - ocean)
- Demographics (age, gender)
- Topic Modeling

Rate of tweets per hour



Proportion of tweets by sentiment polarity



Métodos de clasificación 🗂️🎯

- LIWC
- Emolex
- Personality
- Topic
- Unigrams
- *Deep learning*

Referencias

1 Coppersmith, G.A., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. CLPsych@HLT-NAACL.

2 Guntuku, S., Ramsay, J.R., Merchant, R.M., & Ungar, L.H. (2019). Language of ADHD in Adults on Social Media. Journal of Attention Disorders, 23, 1475 - 1485.

3 Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., & Goharian, N. (2018). SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. ArXiv, abs/1806.05258.