

# Hito 3

“Análisis de comentarios sarcásticos en Reddit”

Integrantes:

- Nicolás García
- Pablo Gutiérrez
- Javier Lavados
- Sebastián Salinas
- José Triviño



# reddit

↑  
20.7k  
↓

r/science · Posted by u/SideBarParty 9 hours ago

🗨️ 11 🗳️ 7 💰 6 🐾 6

A study of cancer patients found that using medical cannabis “improved sleep initiation and continuity, resulted in decreased use of sleep medications, and that improved sleep led to better health.”



[pubmed.ncbi.nlm.nih.gov/338475...](https://pubmed.ncbi.nlm.nih.gov/338475...)

Medicine

565 Comments

Share

Save


Hide

Report

↑  
1.5k  
↓

PH r/popheads · Posted by u/mime454 1 day ago

Apple reveals Apple Music Now Pays Artists \$.01 per stream. Double Spotify.



[wsj.com/articl...](https://wsj.com/articl...)

Report


98% Upvoted

↑  
48.6k  
↓

r/politics · Posted by u/SheepCantFly 6 hours ago

🗨️ 13 🗳️ 9 💰 18 🐾 12 🌟 2 🏠

There was Trump-Russia collusion — and Trump pardoned the colluder



[thehill.com/opinio...](https://thehill.com/opinio...)

2.6k Comments

Share

Save

Hide

Report

79% Upvoted



vammek 6 months ago

Go to Twitter to be heard. Most platforms are shallow & full of jokes. Don't get me wrong jokes are fine, but not when credible info is being shared.



6



Share

Report

Save



C-n0te 6 months ago



For me It was less about being heard and more about refusing to believe that seemingly good people I've known my whole life are just outright dumb and hateful. That and a penchant for calling out propaganda and fake news.

But they sure owned this lefty. /s



15



Share

Report

Save

# Objetivos Principales

- ▶ Descubrir si existen patrones que permitan caracterizar comentarios sarcásticos y predecir la naturaleza de estos últimos.
- ▶ Verificar si el sarcasmo es dependiente del contexto en el que está siendo empleado o los patrones pueden ser aplicados de forma global.
- ▶ Abordar uno de los principales problemas del procesamiento de lenguaje natural.

# Preguntas y Problemas

The background of the slide is composed of several overlapping, semi-transparent triangles in various shades of orange and brown. These triangles are arranged in a way that creates a sense of depth and movement. A thin, light-colored line runs diagonally across the lower portion of the image, intersecting the triangular shapes.

- ▶ ¿Es posible predecir un comentario sarcástico a partir de su contenido?  
¿Cuál es la mejor forma de predecir esto?

- ▶ ¿El patrón de sarcasmo es constante entre un dominio y otro? ¿Influye el tópico en discusión al momento de detectar el sarcasmo?






- ▶ ¿Difieren de gran manera los patrones entre un dominio y otro?






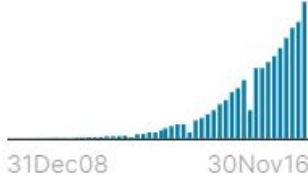
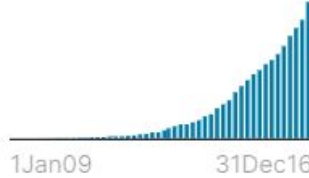
# Exploración de datos



# Exploración inicial

# label 	▲ comment 	▲ author 	▲ subreddit 
Target: Sarcasm	Text of comment	Comment author	Origin subreddit
 01	<b>962295</b> unique values	<b>256561</b> unique values	AskReddit 6% politics 4% Other (905653) 90%
0	NC and NH.	Trumpbart	politics
0	You do know west teams play against west teams more than east teams right?	Shbshb906	nba
0	They were underdogs earlier today, but since Gronk's announcement this afternoon, the Vegas line has...	Creepeth	nfl

# Exploración inicial

# score Comment score	# ups Upvotes	# downs Downvotes	date Created date	created_utc Created datetime	parent_comment If comment in a thread: Parent in thread
					<b>984286</b> unique values
2	-1	-1	2016-10	2016-10-16 23:55:23	Yeah, I get that argument. At this point, I'd prefer is she lived in NC as well.
-4	-1	-1	2016-11	2016-11-01 00:24:10	The blazers and Mavericks (The wests 5 and 6 seed) did not even carry a good enough record to make t...
3	3	0	2016-09	2016-09-22 21:45:37	They're favored to win.

# Preprocesamiento

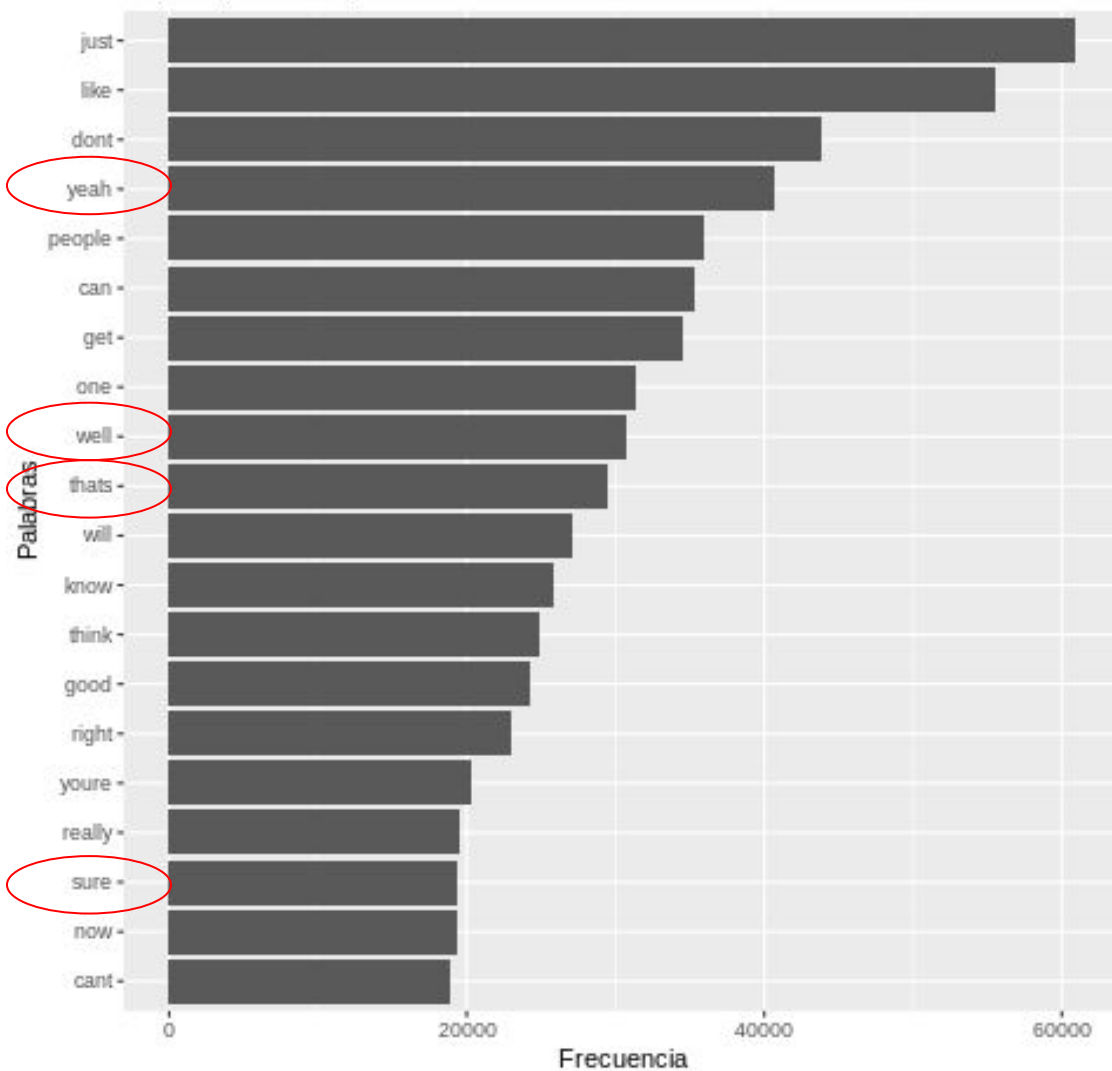
```
# A tibble: 59 x 10
  label comment author subreddit score ups downs date created_utc
  <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <dbl> <chr>   <dtm>
1     0 0      Anxietyf... pokemongo     6    -1    -1 2016... 2016-12-26 00:27:13
2    NA <NA>    <NA>    <NA>    NA     NA     NA <NA>   NA
3    NA <NA>    <NA>    <NA>    NA     NA     NA <NA>   NA
4    NA <NA>    <NA>    <NA>    NA     NA     NA <NA>   NA
5    NA <NA>    <NA>    <NA>    NA     NA     NA <NA>   NA
6    NA <NA>    <NA>    <NA>    NA     NA     NA <NA>   NA
7    NA <NA>    <NA>    <NA>    NA     NA     NA <NA>   NA
8     0 0      sleeplyb... pcmaster...     1     1     0 2016... 2016-05-22 01:30:49
9    NA <NA>    <NA>    <NA>    NA     NA     NA <NA>   NA
10   NA <NA>    <NA>    <NA>    NA     NA     NA <NA>   NA
# ... with 49 more rows, and 1 more variable: parent_comment <chr>
```

# Vectorización

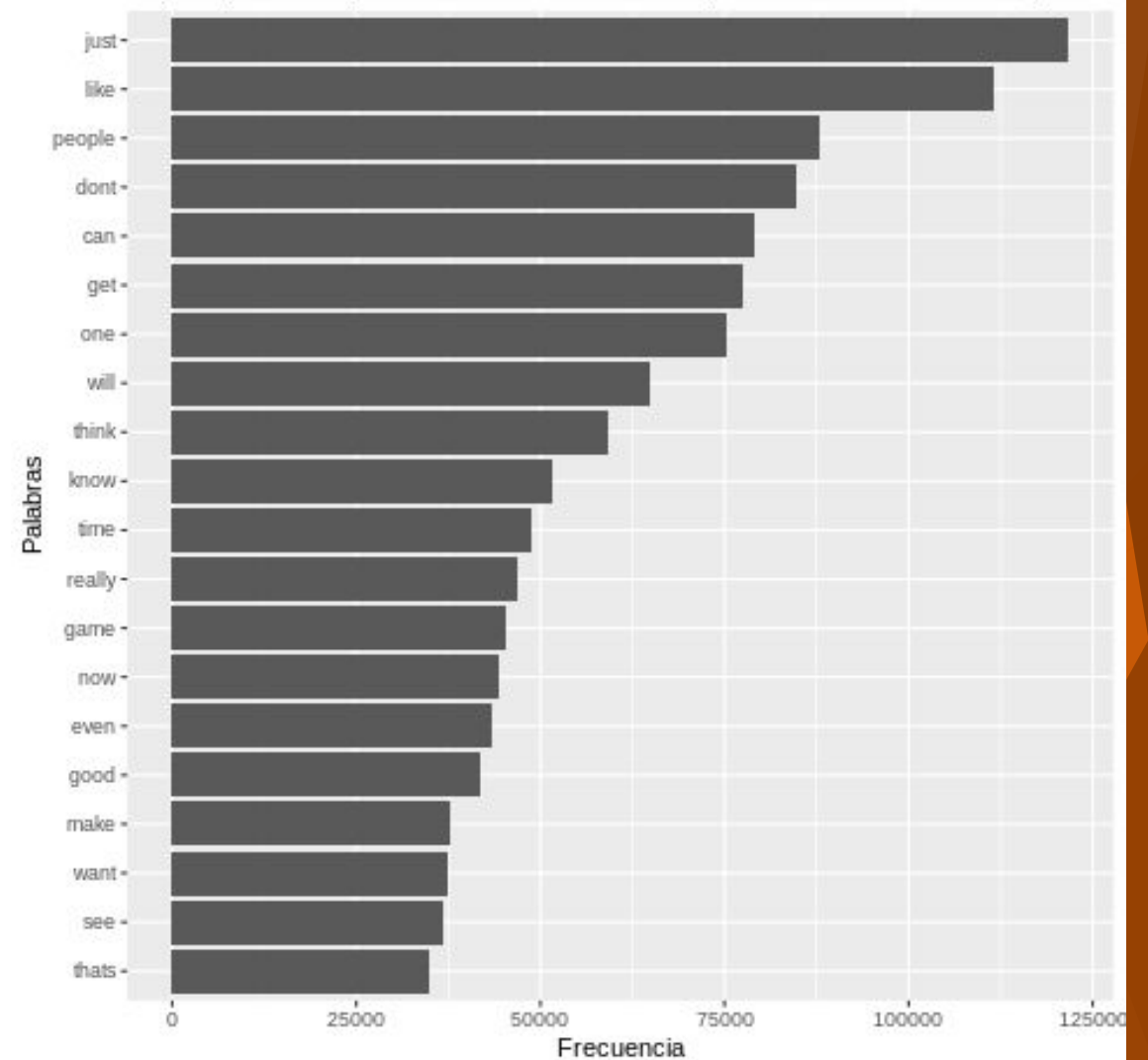
	bolsa	cancha	comer	comprarme	delicioso	ensuciar	fútbol	ganar	jugar	juro	noche	plato	plástico	suadero	taco	vida
Document 1	0	0	1	0	1	0	0	0	0	1	0	0	0	3	1	1
Document 2	1	0	0	0	1	1	0	0	0	0	0	1	1	1	1	0
Document 3	0	1	0	1	0	0	1	1	1	0	1	0	0	0	1	0

# Palabras frecuentes

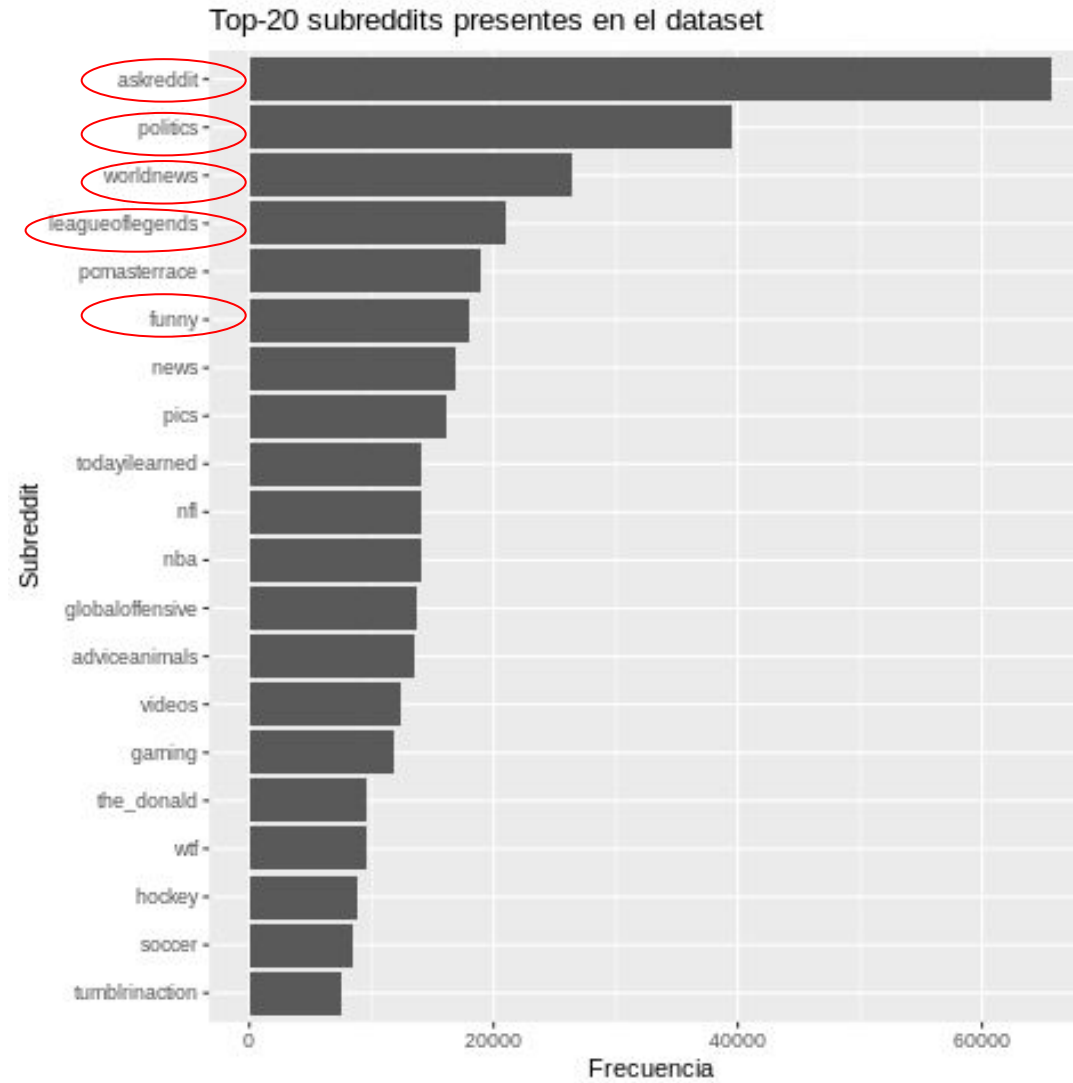
Top-20 palabras presentes en comentarios sarcásticos sin considerar stop



Top-20 palabras presentes en comentarios padre sin considerar stopwords



# Subreddits frecuentes





# Dataset final

comment	subreddit	score	sarcasm
I don't pay attention to her, but as long as she's legal I wouldn't kick her out of bed (before she took a load)	AskReddit	0	yes
Trick or treating in general is just weird...	AskReddit	1	yes
what the fuck	AskReddit	22	yes
This would make me cry.	AskReddit	1	yes
My stuffed animal I've had since I was born.	AskReddit	1	yes
How I'm considered an asshole because I let people know that most of the shit they do or have is because of a trend that'll die out i	AskReddit	1	yes
For me, it's suce... success that give me the most trouble.	AskReddit	1	yes
Oh, I never realized it was so easy, why had I, and every other lonely person on earth never thought of that before?	AskReddit	1	yes
About tree fiddy	AskReddit	2	yes
BMW drivers	AskReddit	1	yes
Imagine doing that shit in a forest or park at night Fuck	AskReddit	2	yes
In archery, if at full draw their elbow is below their heart.	AskReddit	11	yes
Your bf sounds like he might have an extra chromosome or two	AskReddit	1	yes
Mr Garrison is the next President.	AskReddit	3	yes
Wtf lmaoo	AskReddit	1	yes
And even wars such as Egyptian Unification war, a whopping 5316 years ago (lsh)	AskReddit	2	yes
I imagine they are small marks and if you weren't looking for them wouldn't see them.	AskReddit	1	yes
Smoke weed	AskReddit	32	yes
talk about my penis	AskReddit	1	yes
~~I think, therefore~~ I am	AskReddit	2	yes
Love and support.	AskReddit	2	yes
dairy	AskReddit	1	yes
The issue many people have with it is that it targets a lot of other things too - by the phrasing of the law, caffeine, alcohol and nicot	AskReddit	6	yes
Isn't that like illegal?	AskReddit	4	yes
Or is this just fantasy	AskReddit	1	yes



# Experimentos y resultados relevantes



# Experimento 1

- ¿Es posible predecir un comentario sarcástico a partir de su contenido?  
¿Cuál es la mejor forma de predecir esto?

# Metodología

- ▶ Se trabajará con los datos de cada comentario con el fin de crear un **clasificador**.
- ▶ Se segmentará el dataset en los distintos subreddits
- ▶ Se vectorizará el dataset
- ▶ Se entrenarán los 5 clasificadores con los distintos subreddits

## Clasificadores utilizados

- ▶ Decision Tree
- ▶ Support Vector Machine (SVC)
- ▶ Linear Support Vector Machine (LinearSVC)
- ▶ K-Nearest Neighbors (KNN)
- ▶ Dummy Classifier

# Subreddits estudiados

## r/AskReddit

- ▶ Subreddit con mayor cantidad de comentarios (90658)
- ▶ Proporción del 72% de comentarios sarcásticos

## r/worldnews

- ▶ 3er subreddit con más comentarios (51288)
- ▶ Proporción del 51% de comentarios sarcásticos

# Tabla comparativa de Desempeños

r/AskReddit	Decision Tree	SVC	LinearSVC	KNN	Dummy
Precision	0.81	0.81	0.83	0.74	0.73
Recall	0.88	0.98	0.91	1.00	0.72
F1-score	0.84	0.89	0.87	0.85	0.72
Acurracy	0.76	<u>0.82</u>	<u>0.80</u>	0.74	0.60

r/worldnews	Decision Tree	SVC	LinearSVC	KNN	Dummy
Precision	0.71	0.76	0.78	0.55	0.51
Recall	0.76	0.91	0.84	0.99	0.51
F1-score	0.73	0.83	0.81	0.71	0.51
Acurracy	0.71	<u>0.81</u>	<u>0.79</u>	0.58	0.50

# Tabla comparativa de Desempeños

r/AskReddit	Decision Tree	SVC	LinearSVC	KNN	Dummy
Precision	0.81	0.81	0.83	0.74	0.73
Recall	0.88	0.98	0.91	1.00	0.72
F1-score	0.84	0.89	0.87	0.85	0.72
Acurracy	0.76	<u>0.82</u>	0.80	0.74	<u>0.60</u>

r/worldnews	Decision Tree	SVC	LinearSVC	KNN	Dummy
Precision	0.71	0.76	0.78	0.55	0.51
Recall	0.76	0.91	0.84	0.99	0.51
F1-score	0.73	0.83	0.81	0.71	0.51
Acurracy	0.71	<u>0.81</u>	0.79	<u>0.58</u>	<u>0.50</u>

# Experimento 1:

## Conclusiones

- ▶ Es posible entrenar un clasificador que reconozca comentarios sarcásticos, obteniendo resultados razonables.
- ▶ El mejor desempeño se obtiene con los clasificadores SVC y LinearSVC
- ▶ Nos da una idea de la distribución que toman los comentarios sarcásticos
- ▶ La proporción de comentarios sarcásticos solo afecta a algunos clasificadores.

# Experimento 2

- ¿El patrón de sarcasmo es constante entre un dominio y otro? ¿Influye el tópico en discusión al momento de detectar el sarcasmo?



# Metodología

- ▶ Se realizará **validación cruzada**, entrenando clasificadores con datos de un subreddit en específico y probándolos con datos de los otros subreddits.
- ▶ El clasificador que se escogerá será el que obtuvo el mejor desempeño en un tiempo razonable en el experimento anterior.
- ▶ Se evaluará el desempeño de los 5 clasificadores y se graficarán los resultados en una **matriz de accuracy**.

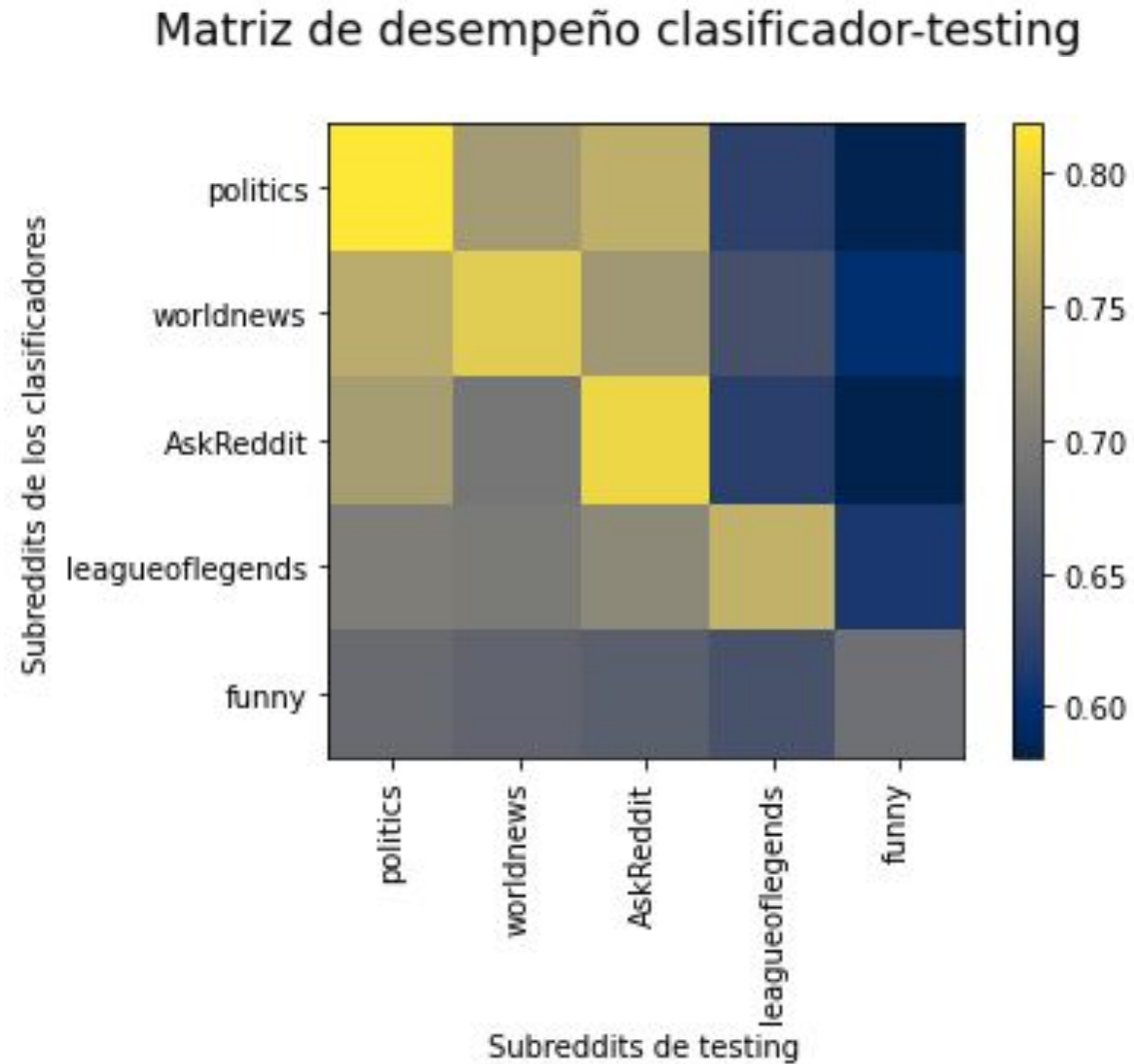
# Matriz de accuracy cruzada

Datos de Testing

Datos de Training

	politics	worldnews	AskReddit	leagueoflegends	funny
politics	81.8%	73.8%	76.2%	62.3%	58.3%
worldnews	75.8%	79.2%	73.4%	64.3%	59.9%
AskReddit	74%	69.4%	80.3%	62.1%	58.1%
leagueoflegends	70.3%	69.8%	71.7%	76.5%	61.1%
funny	67.8%	67%	66.4%	64.6%	68.6%

# Visualización matriz de accuracy



# Experimento 2:

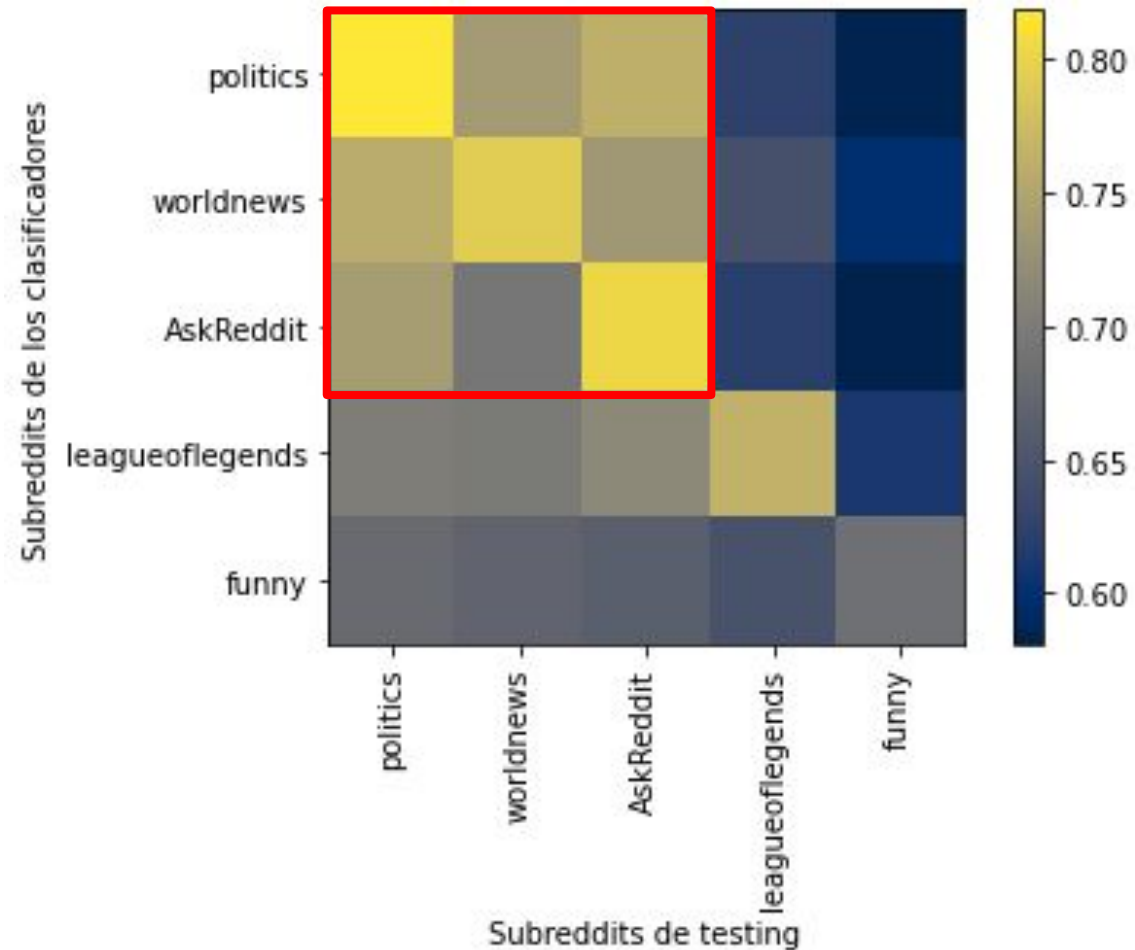
## Conclusiones

- ▶ Los clasificadores entrenados con los datos de un subreddit en específico tienen un mejor desempeño al ser evaluados con datos de este mismo subreddit.
- ▶ El desempeño del clasificador está correlacionado con la seriedad del tema abordado en el subreddit.

### Validación Cruzada:

- ▶ Mejor desempeño: Comentarios de **r/AskReddit** en clasificador entrenado con comentarios de **r/politics**.
- ▶ Accuracy: 76.26%.
- ▶ Peor desempeño: Comentarios de **r/funny** en clasificador entrenado con comentarios de **r/AskReddit**.
- ▶ Accuracy: 58.11%

# Experimento 2: Conclusiones



# Experimento 3

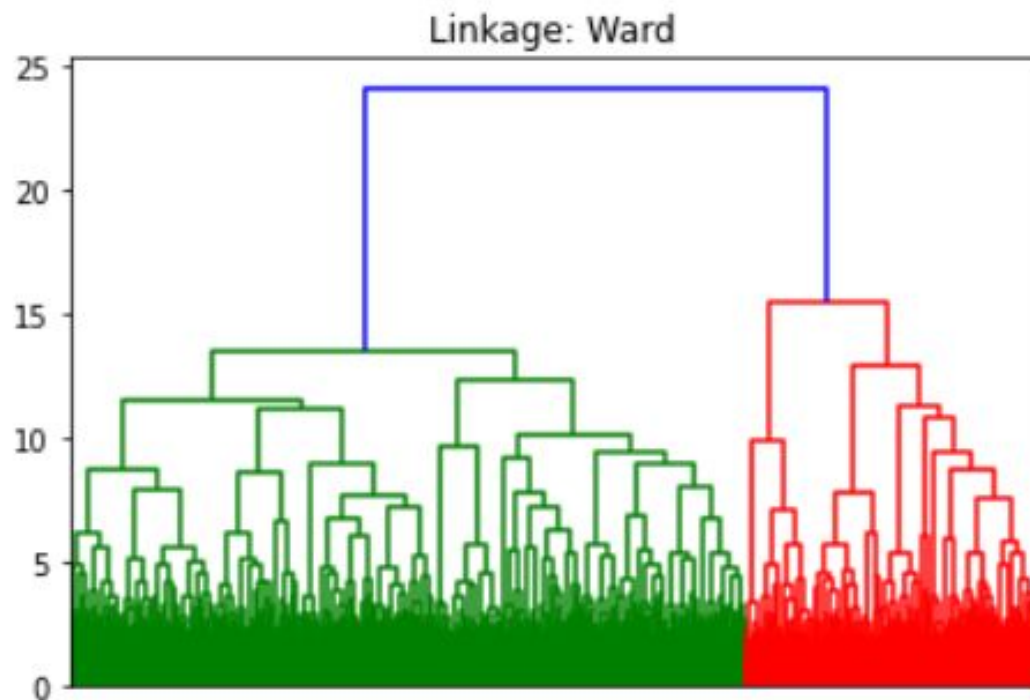
- ▶ ¿Difieren de gran manera los patrones entre un dominio y otro?

# Metodología

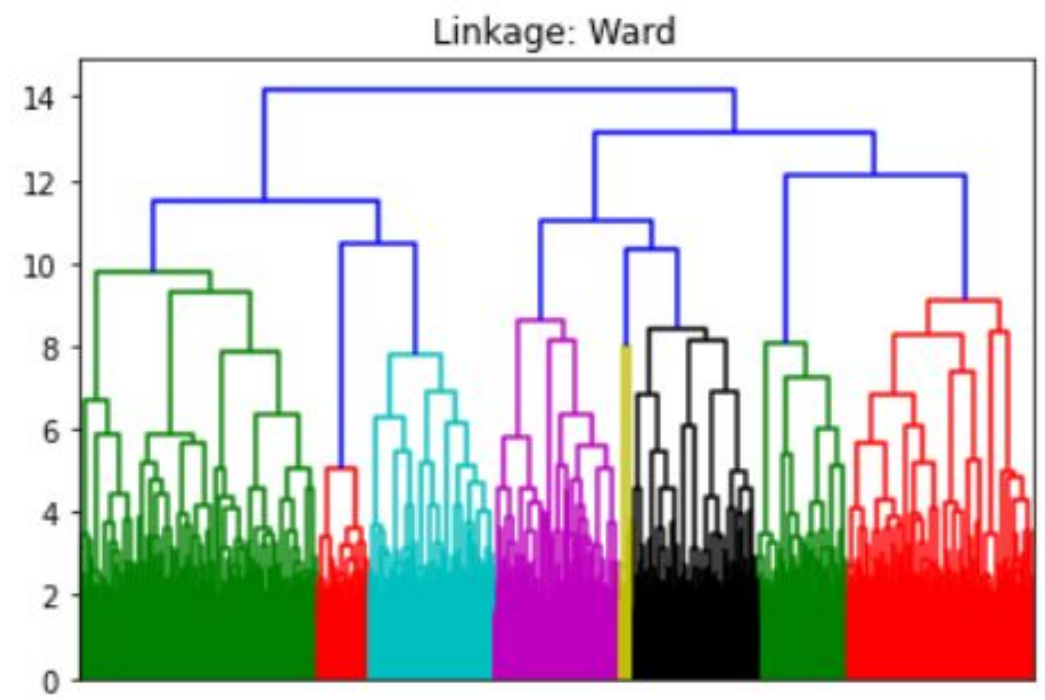
- ▶ Se usarán los pares de subreddits con mejor y peor accuracy de la parte anterior.
- ▶ Para cada par de subreddits se unirán los comentarios en una sola tabla.
- ▶ Utilizando una nueva librería llamada “sentence\_transformers”, se utilizará agglomerative clustering en cada tabla, generando dendogramas que permitan la visualización.

# Resultados

Dendograma para  
r/AskReddit + r/politics



Dendograma para  
r/AskReddit + r/funny





# Experimento 3:

## Conclusiones

- ▶ Mayor cantidad de clusters en subreddits con menor sinergia, menor dispersión con mayor sinergia.
- ▶ Buen desempeño al momento de clasificar en subreddits serios puede deberse a pocos patrones en los comentarios.
- ▶ Mal desempeño al momento de clasificar en subreddits poco serios puede deberse a múltiples patrones en los comentarios.

# Futuras direcciones



# Planificación futura

- ▶ Intentar entrenar el clasificador bajo otros parámetros que no involucren vectorizar el texto.
- ▶ Buscar mejores maneras de entrenar un clasificador basado en texto.
- ▶ Aprovechar la presencia de comentarios padre en la base de datos para extraer información útil a partir de estos.

# Hito 3

“Análisis de comentarios sarcásticos en Reddit”

Integrantes:

- Nicolás García
- Pablo Gutiérrez
- Javier Lavados
- Sebastián Salinas
- José Triviño