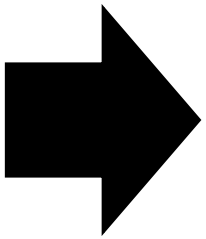
Minería de texto para noticias de U.S

Jairo Fernando Gudiño-Rosero

WHALE & JAGUAR

Nuestro objetivo es descubrir qué luces podemos obtener de titulares y descripciones de noticias de Estados Unidos utilizando herramientas de minería de texto.

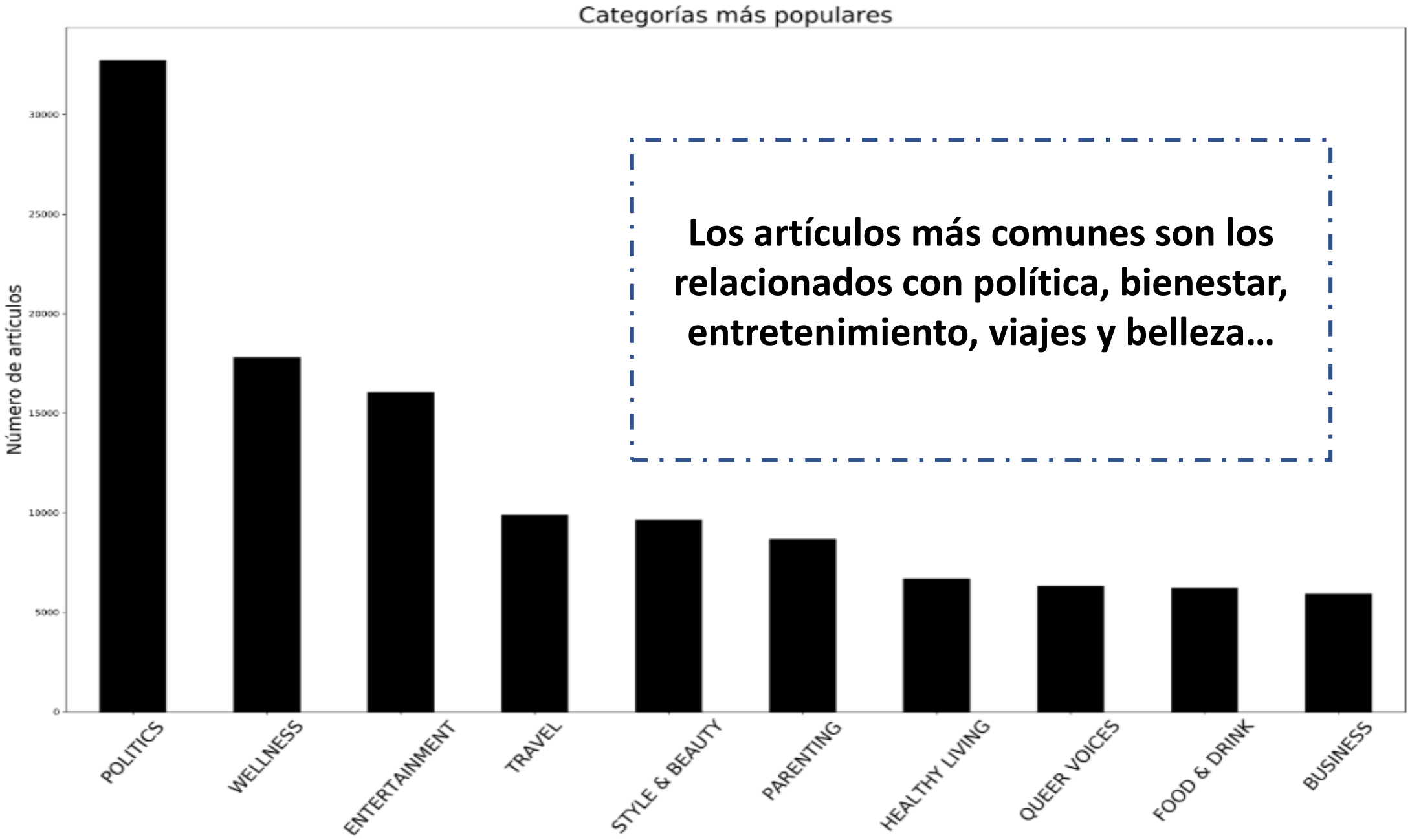


200.850 noticias del portal The Huffington Post recolectadas en una base de Kaggle.

¿Para qué sirven nuestros resultados?

1. Crear campañas de marketing con base en las temáticas encontradas (marketing político, información de interés para agencias de viajes, etc.).
2. Diseñar contenido de noticias con base en las tendencias.
3. Identificar oportunidades de mercado con base en la información pública de la competencia.
4. Desarrollar algoritmos de recomendación de artículos con base en los comentarios o visitas de los lectores.

¿Qué luces podemos obtener de los textos?



Los artículos más comunes son los relacionados con política, bienestar, entretenimiento, viajes y belleza…

Política

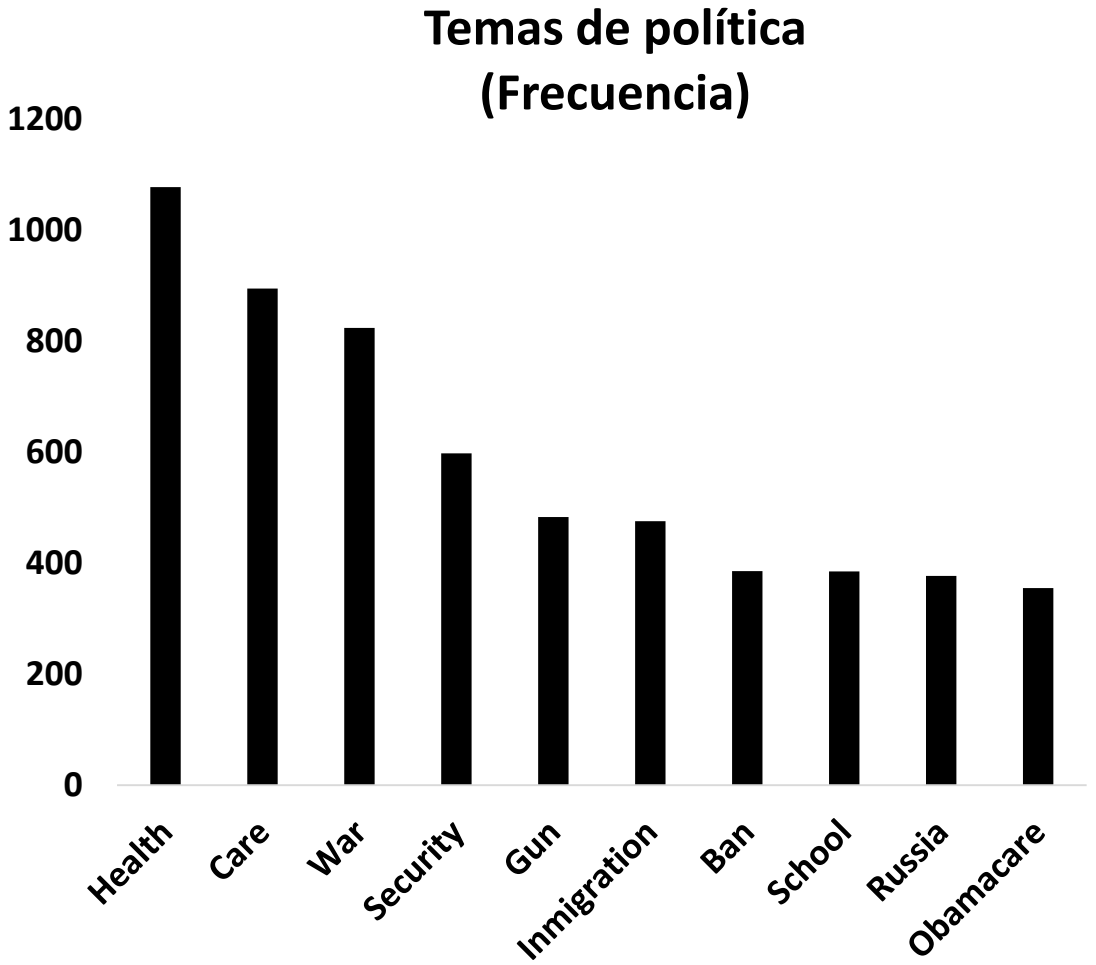
Bienestar

Viajes

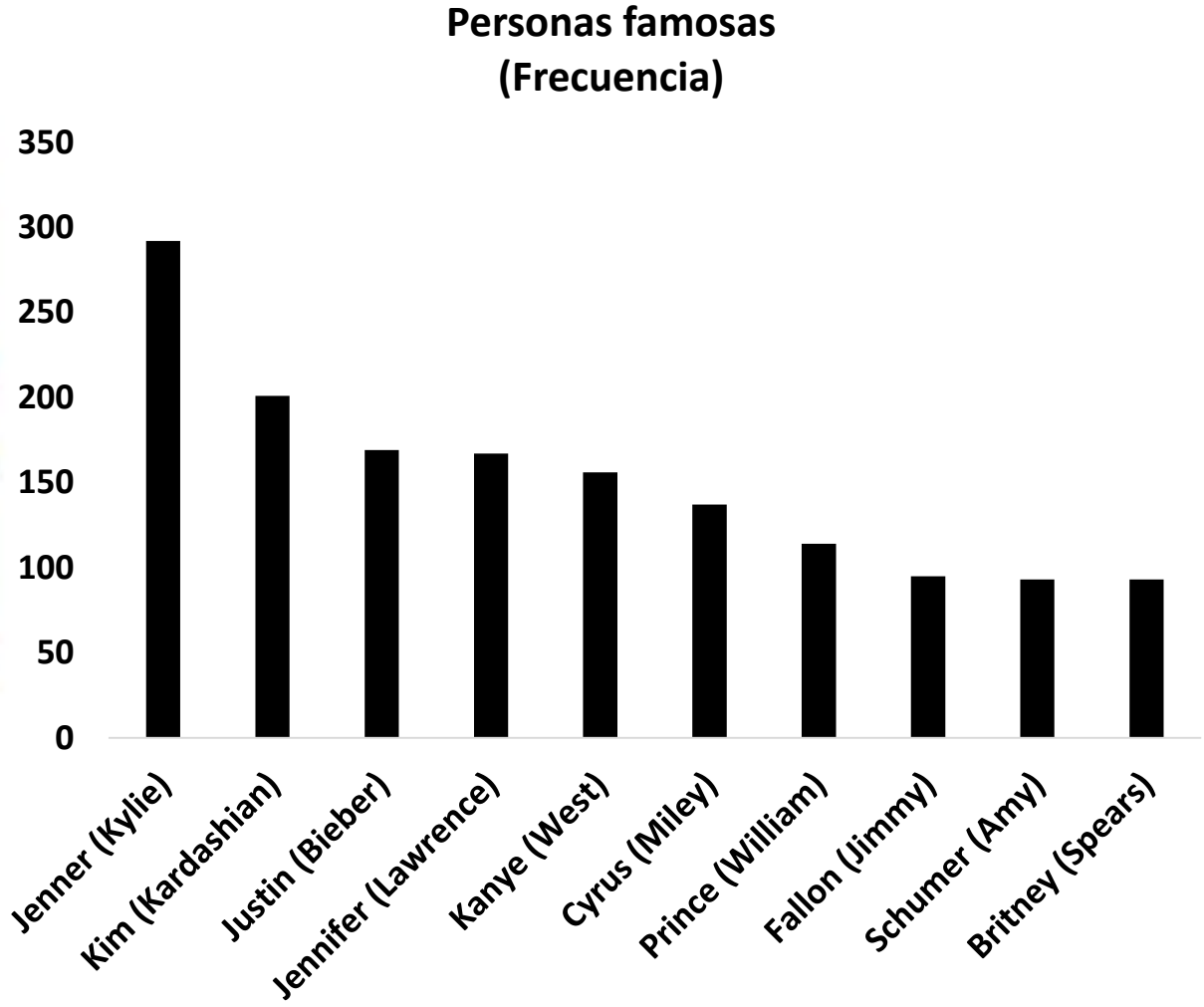
Entretenimiento

1. Trump y el partido republicano dominan la escena política.
2. Los temas relacionados con salud, comida y estudio son los más comunes cuando se habla de bienestar.
3. Los viajes en avión a sitios con playa son los más comunes en la sección de viajes.
4. Trump, Taylor Swift y Beyonce tienen alta importancia en el mundo de la farándula. Los temas de familia, amor, muerte y guerra son los más populares.

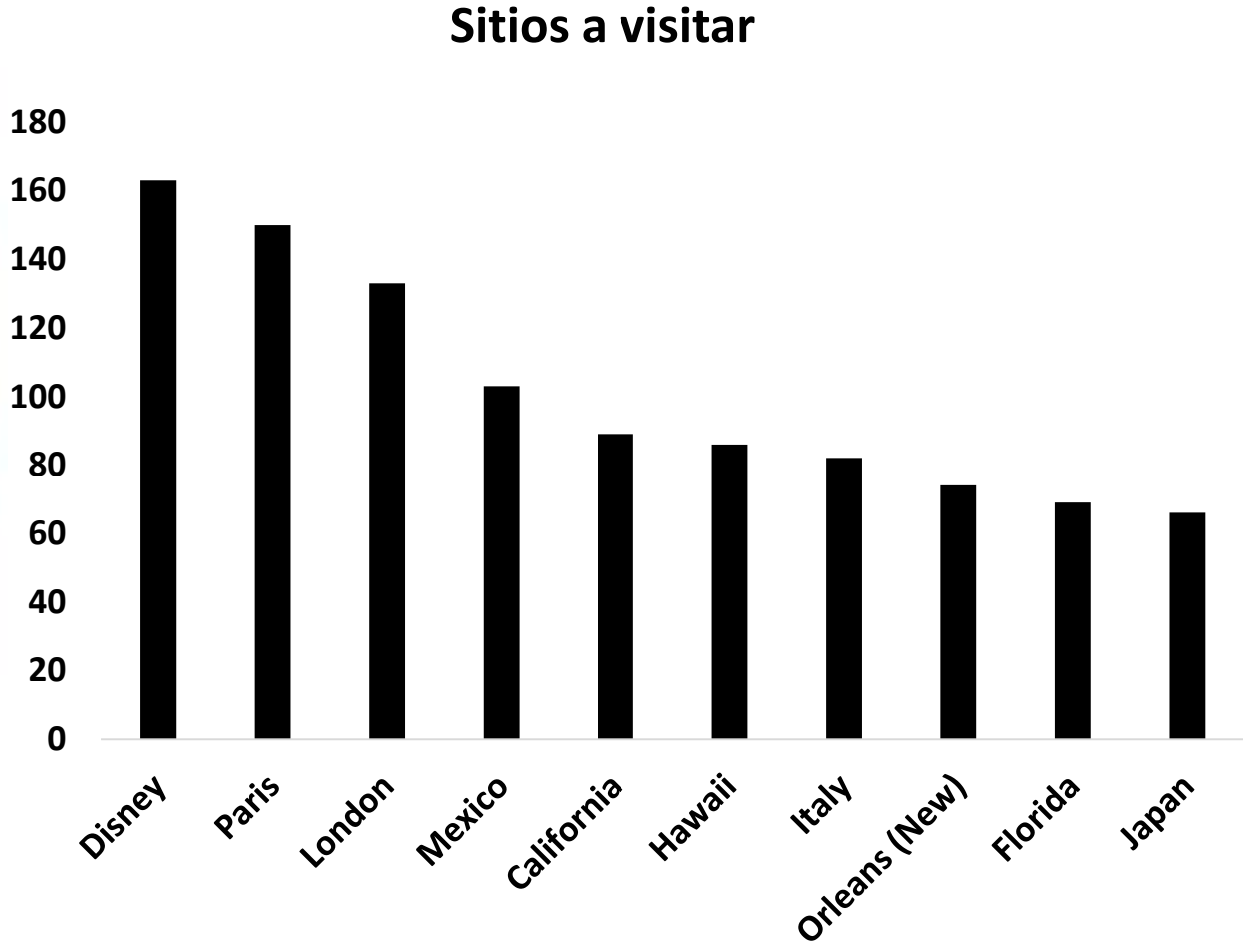
¿Cuáles son los temas de política más discutidos?

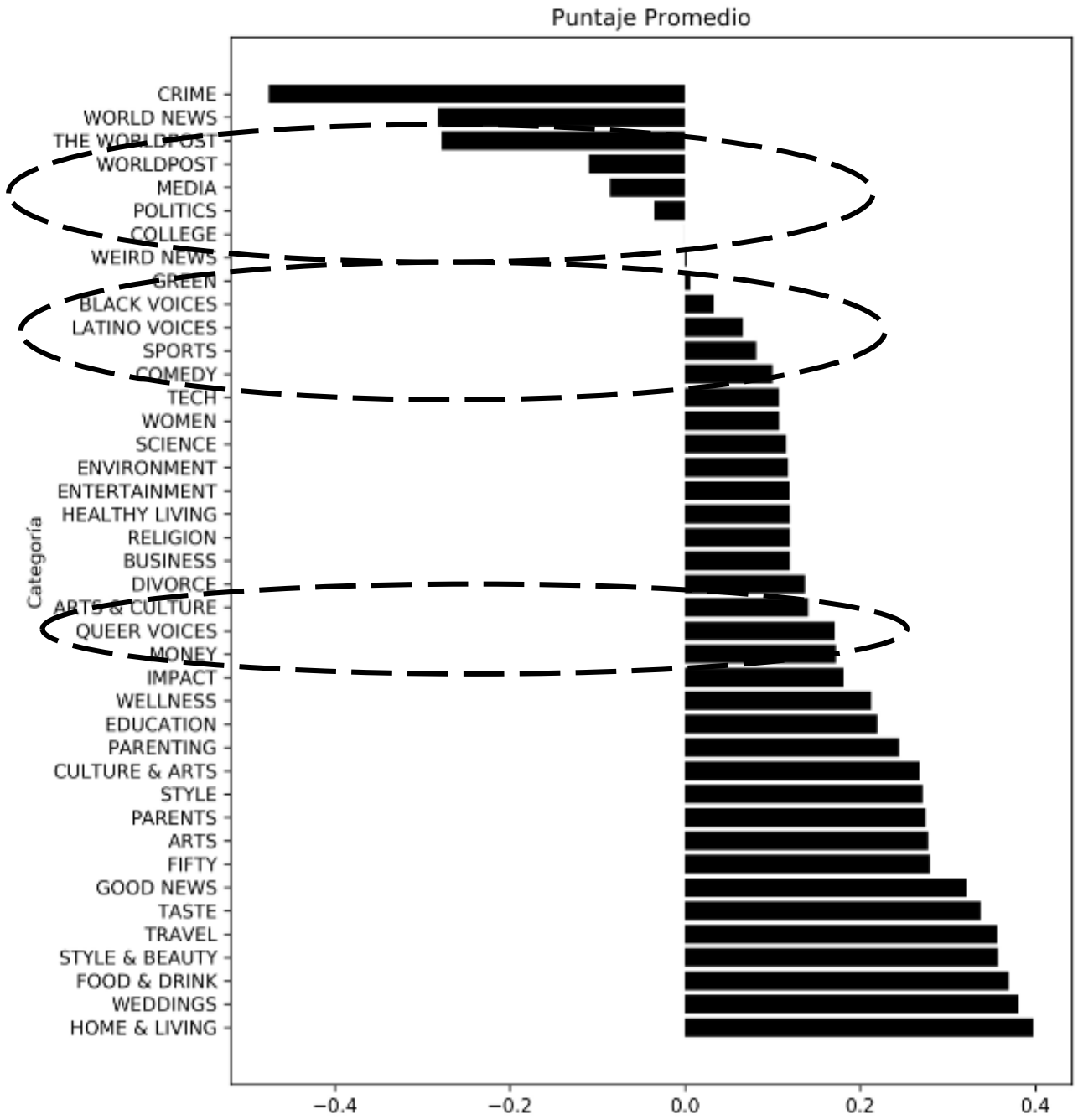


¿Cuáles son los más famosos (aparte de Trump, Beyonce y T. Swift)?



¿Qué sitios se recomienda visitar?



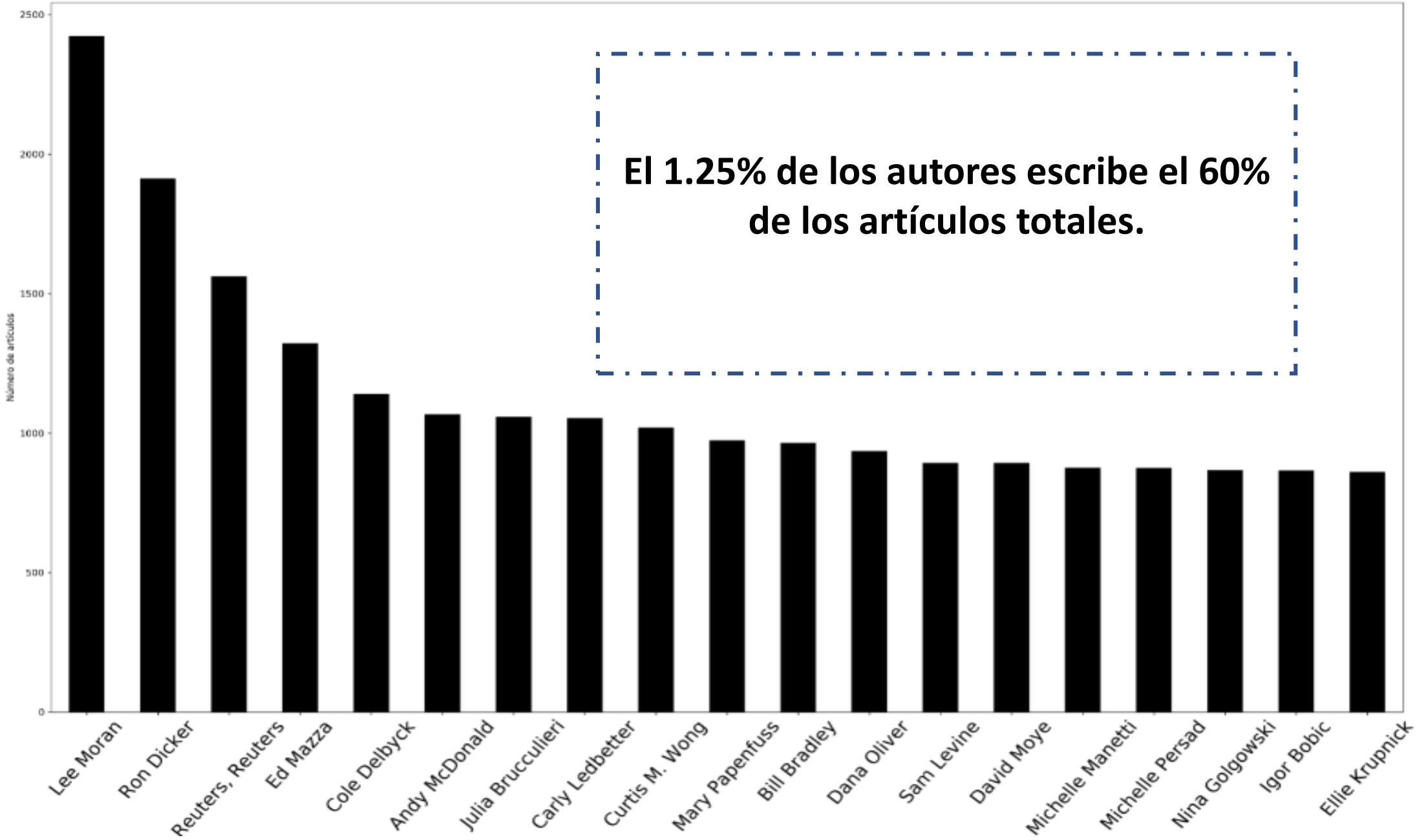


Mayor puntaje, más positivos son los textos..

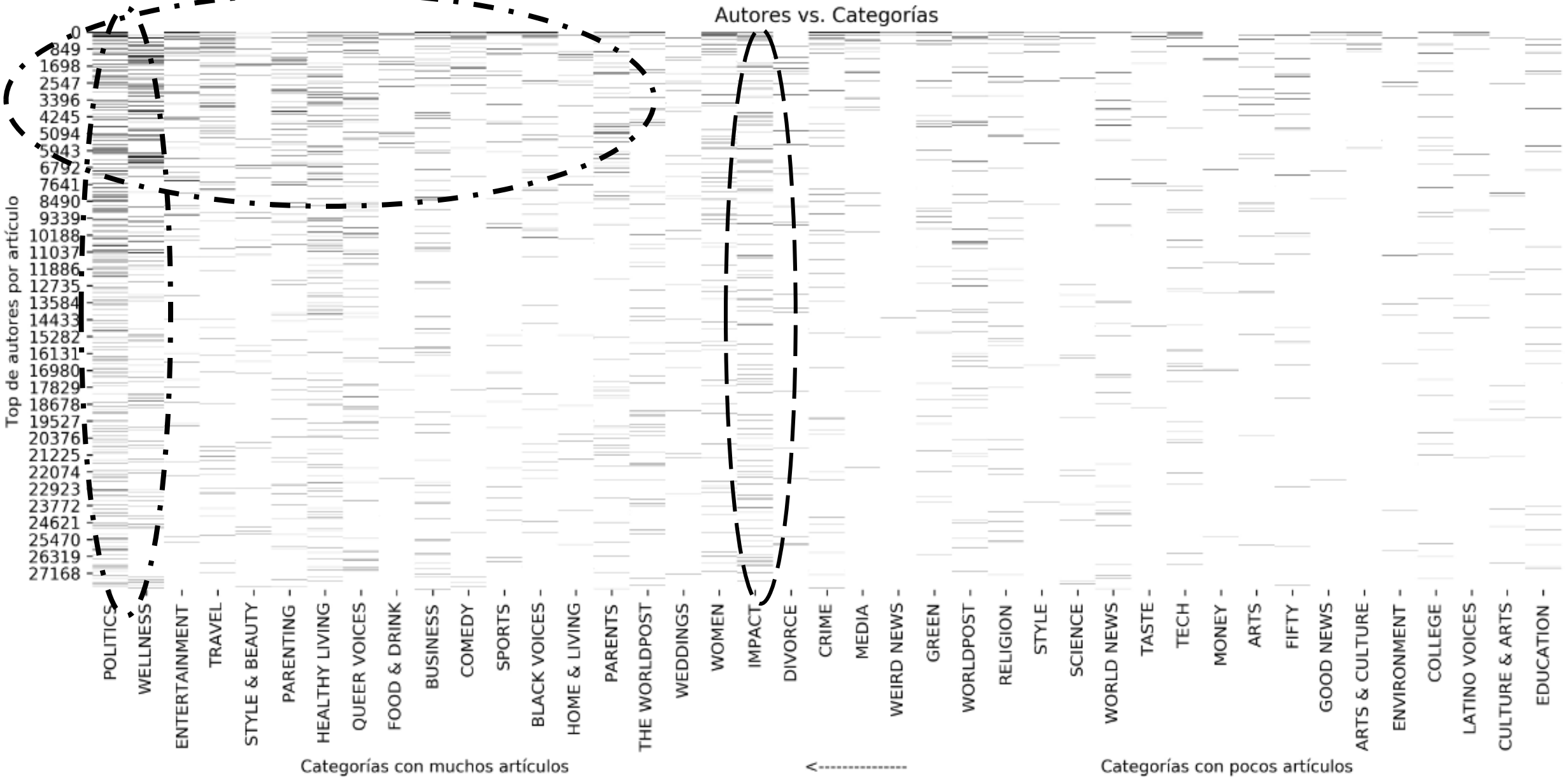
Los temas de política, medios, latinos y afro-americanos tienen un tono promedio negativo.

Los relacionados con el movimiento LGBT tienen un tono más positivo.

¿Qué luces podemos obtener de los autores?

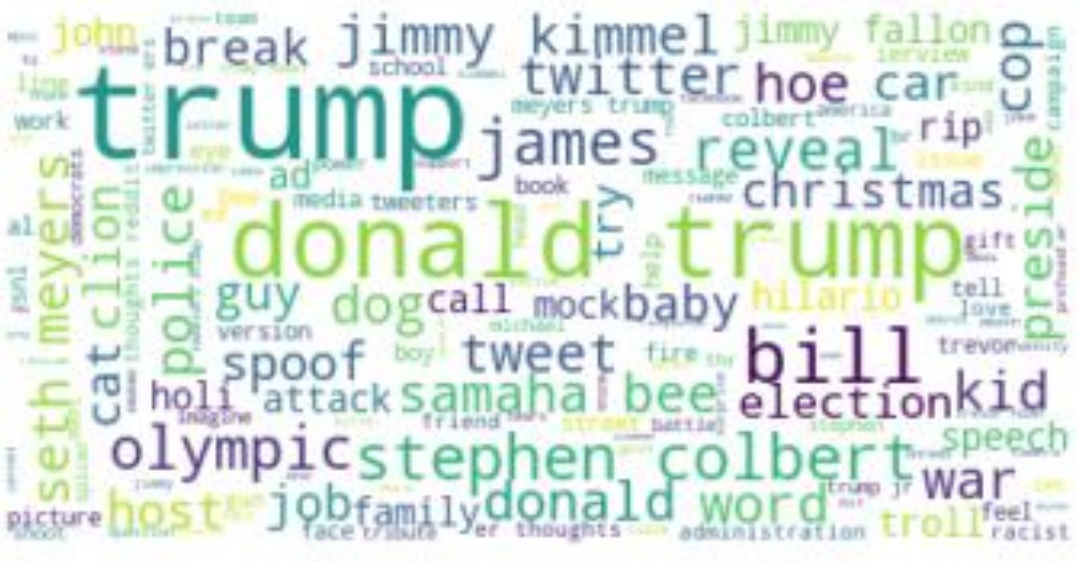

El 1.25% de los autores escribe el 60% de los artículos totales.
Número de artículos
2500
2000
1500
1000
500
0
Lee Moran
Ron Dicker
Reuters, Reuters
Ed Mazza
Cole Delbyck
Andy McDonald
Julia Brucculieri
Carly Ledbetter
Curtis M. Wong
Mary Papenfuss
Bill Bradley
Dana Oliver
Sam Levine
David Moye
Michelle Manetti
Michelle Persad
Nina Golgowski
Igor Bobic
Ellie Krupnick

Los autores que escriben mucho son los que más participan en la escritura de casi todas las categorías…


Autores vs. Categorías
Top de autores por artículo
0
849
1698
2547
3396
4245
5094
5943
6792
7641
8490
9339
10188
11037
11886
12735
13584
14433
15282
16131
16980
17829
18678
19527
20376
21225
22074
22923
23772
24621
25470
26319
27168
POLITICS
WELLNESS
ENTERTAINMENT
TRAVEL
STYLE & BEAUTY
PARENTING
HEALTHY LIVING
QUEER VOICES
FOOD & DRINK
BUSINESS
COMEDY
SPORTS
BLACK VOICES
HOME & LIVING
PARENTS
THE WORLDPOST
WEDDINGS
WOMEN
IMPACT
DIVORCE
CRIME
MEDIA
WEIRD NEWS
GREEN
WORLDPOST
RELIGION
STYLE
SCIENCE
WORLD NEWS
TASTE
TECH
MONEY
ARTS
FIFTY
GOOD NEWS
ARTS & CULTURE
ENVIRONMENT
COLLEGE
LATINO VOICES
CULTURE & ARTS
EDUCATION
Categorías con muchos artículos
<----------
Categorías con pocos artículos

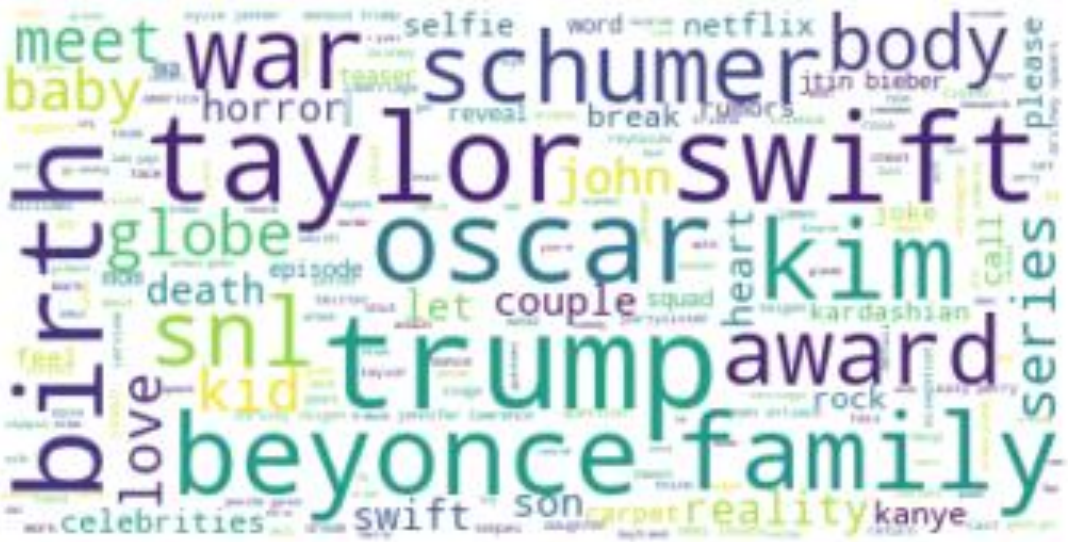
Las temáticas de los autores top tienden a ser muy parecidas… Trump, Taylor Swift, fútbol y entretenimiento…

Lee Moran

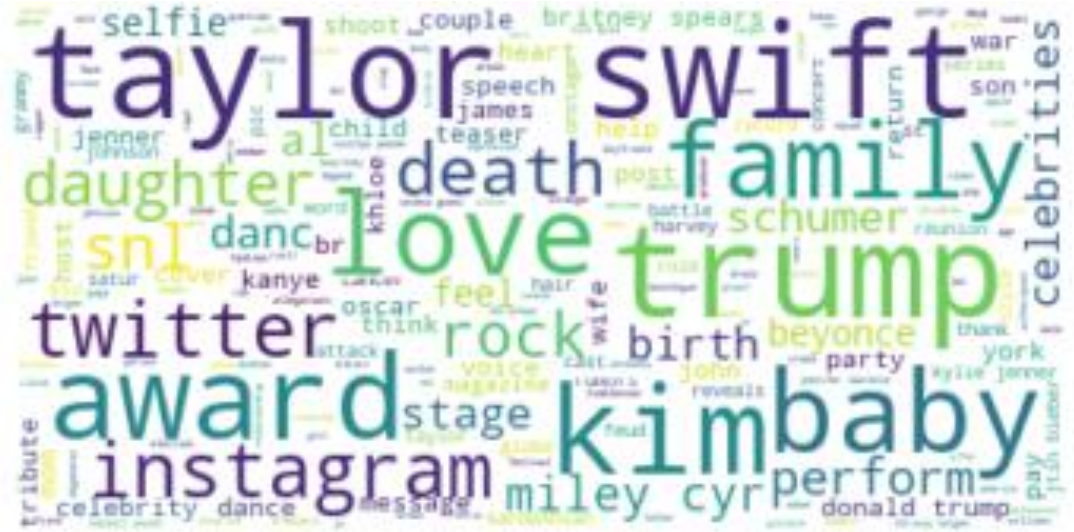


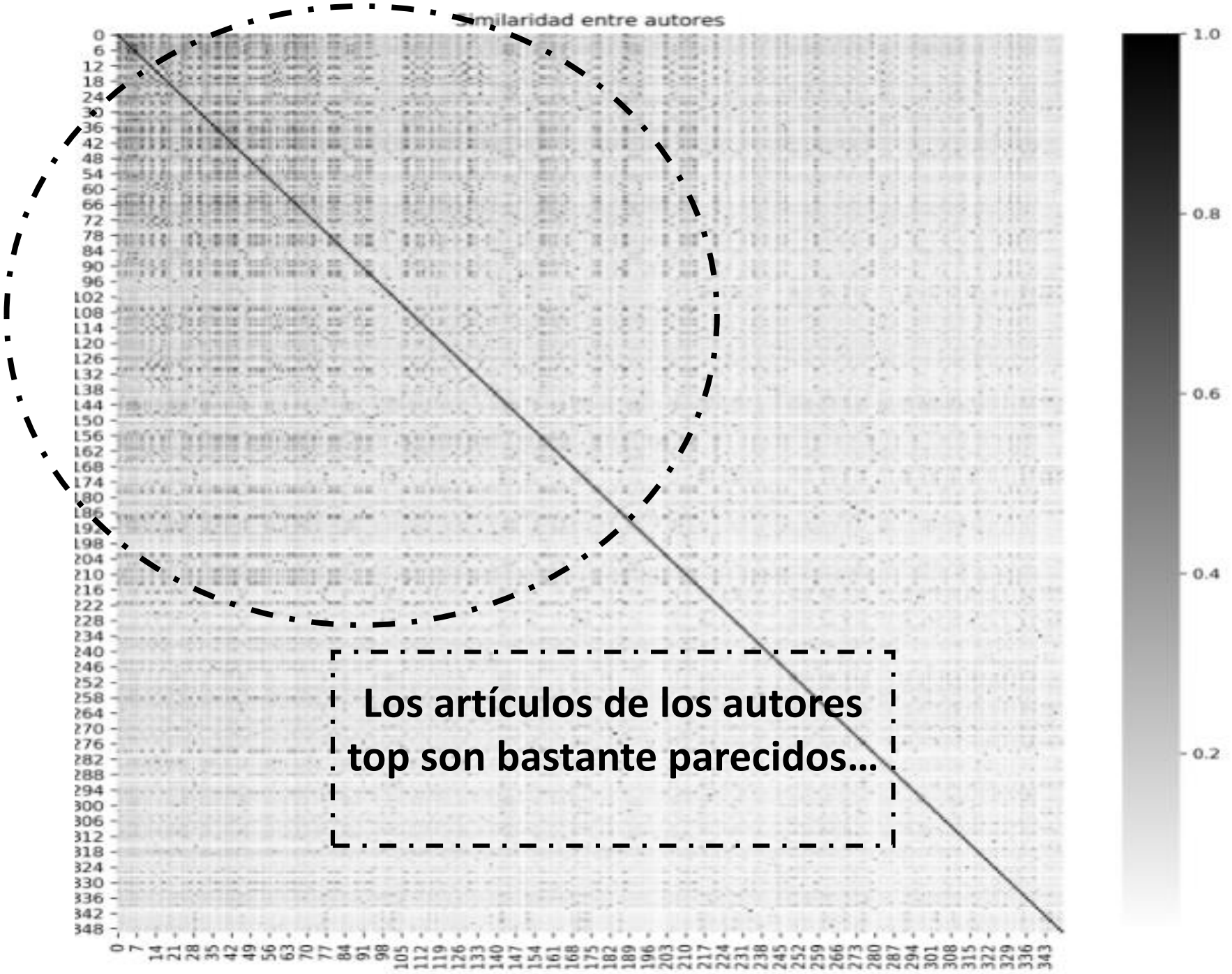
Ron Dicker

Cole Delbyck



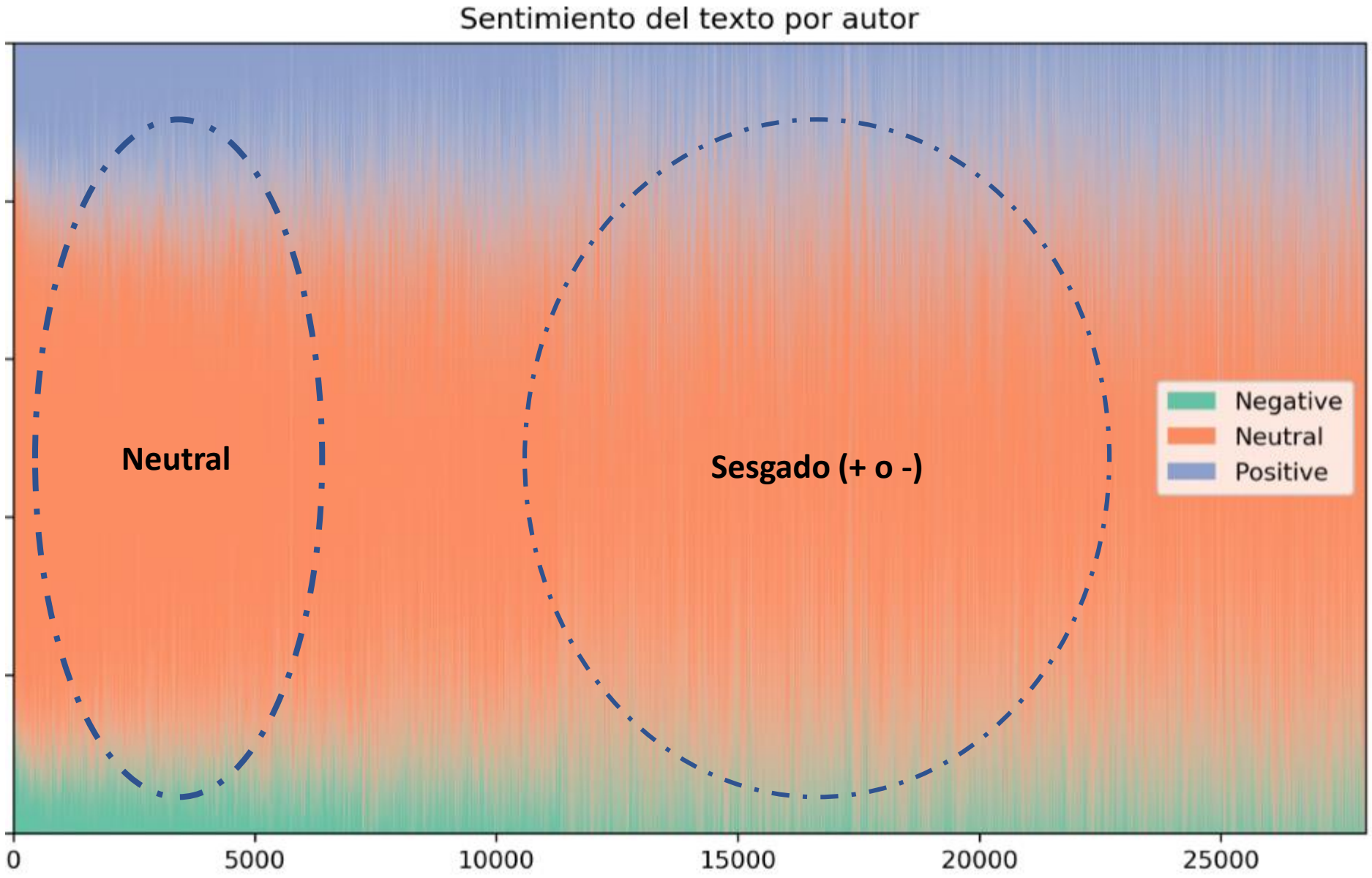
Julia Brucculieri




Similaridad entre autores
Los artículos de los autores top son bastante parecidos…

Sentimiento de textos por autor

Los autores que escriben mucho tienen un estilo más neutral que los que escriben poco…

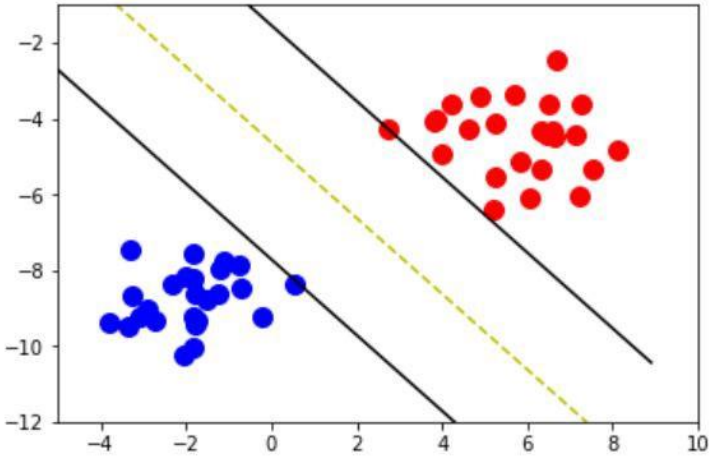


Top de autores por número de artículos: 1 (Escribe mucho) – 25.000 (Escribe poco)

¿Podemos predecir satisfactoriamente la categoría de las noticias a partir de las descripciones que tienen?

Relativamente sí (60.33% de aciertos analizando 40 categorías de texto).

¿Cómo lo hacemos?

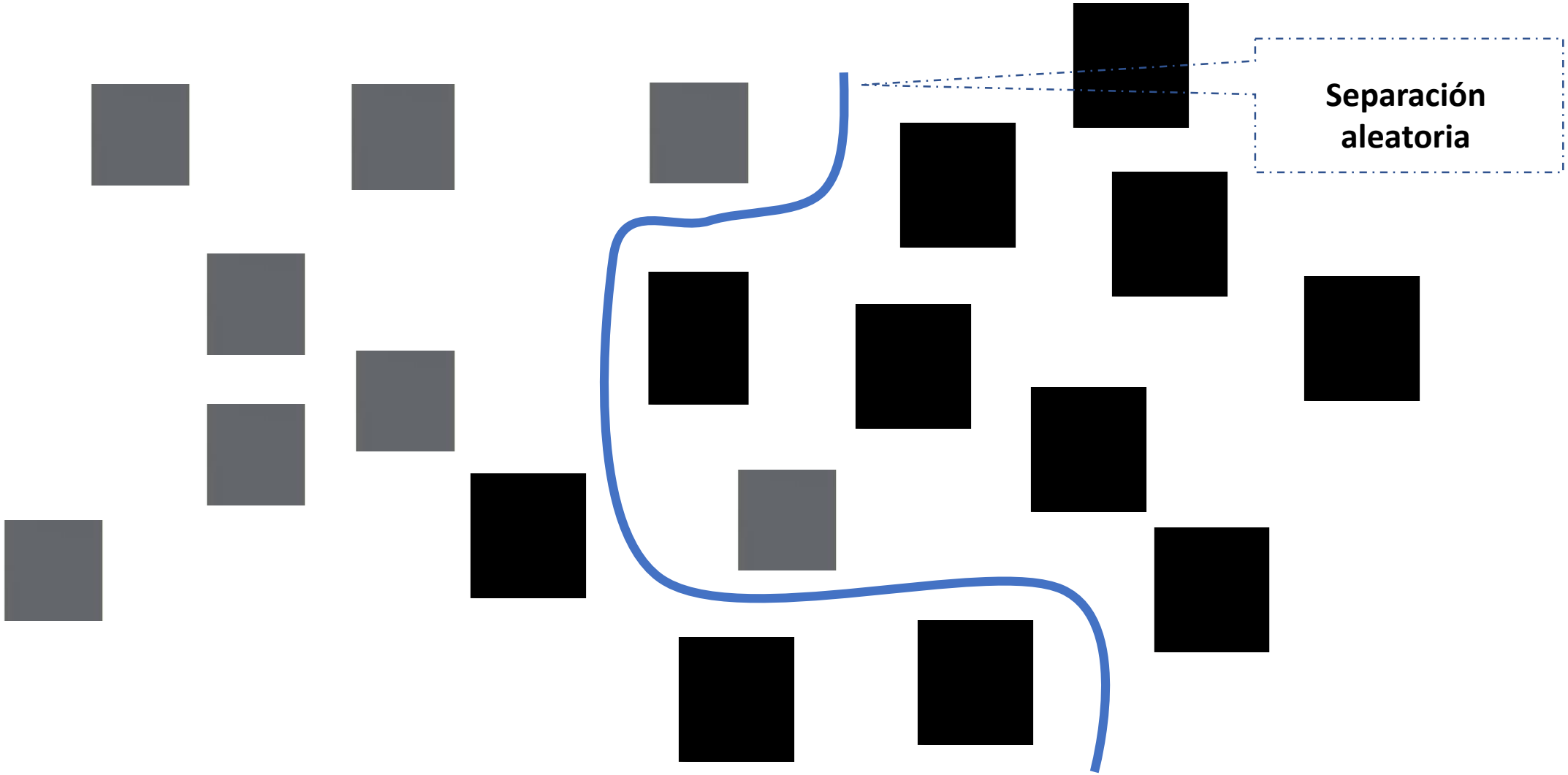


Utilizamos un algoritmo de aprendizaje de máquinas para este fin (SVM con función kernel lineal).

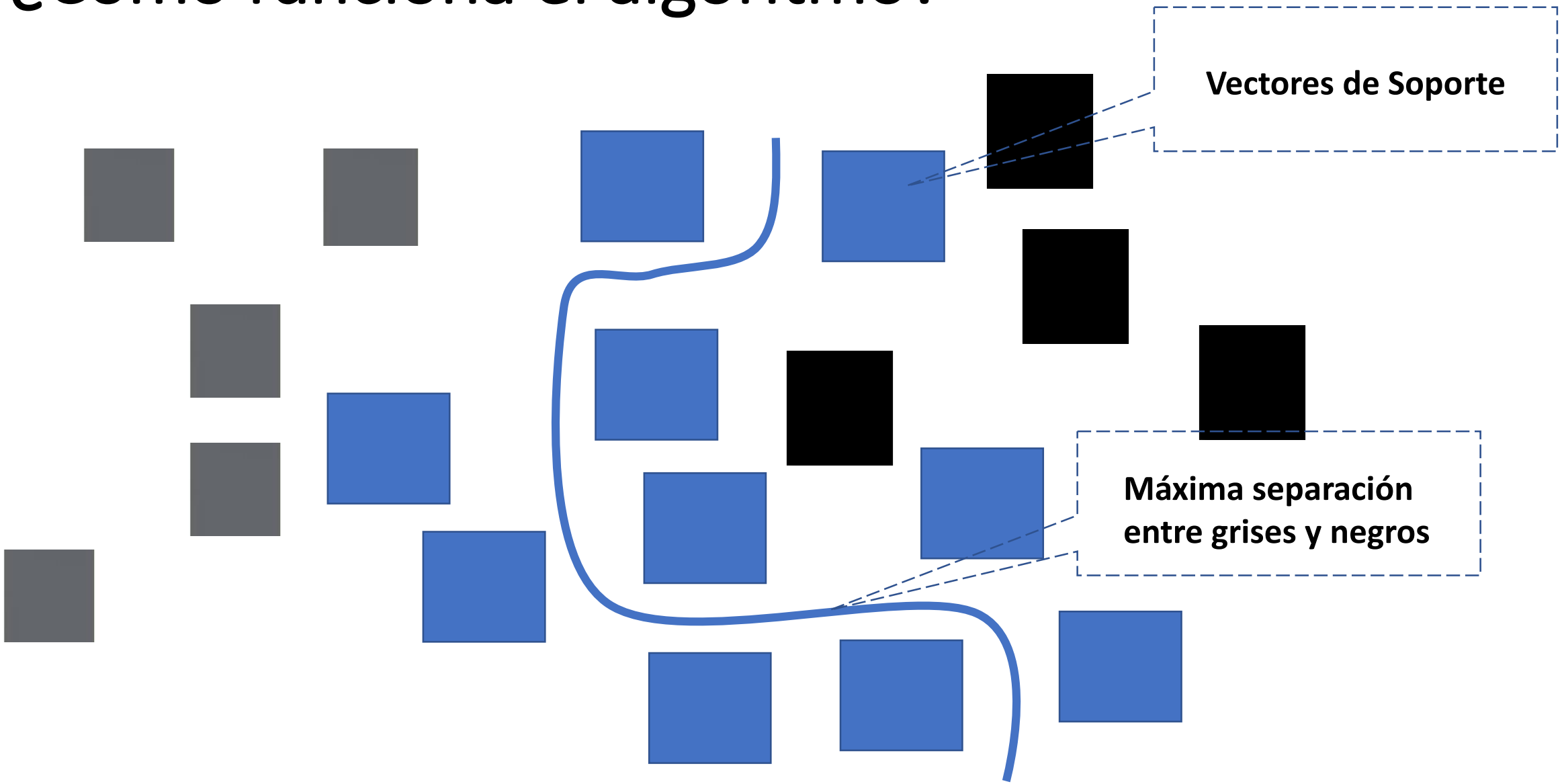
Tomamos el 70% de los textos para construir el algoritmo y el 30% restante lo usamos para analizar su poder predictivo.

Comparamos las categorías predichas por el algoritmo para ese 30% vs. las categorías reales…

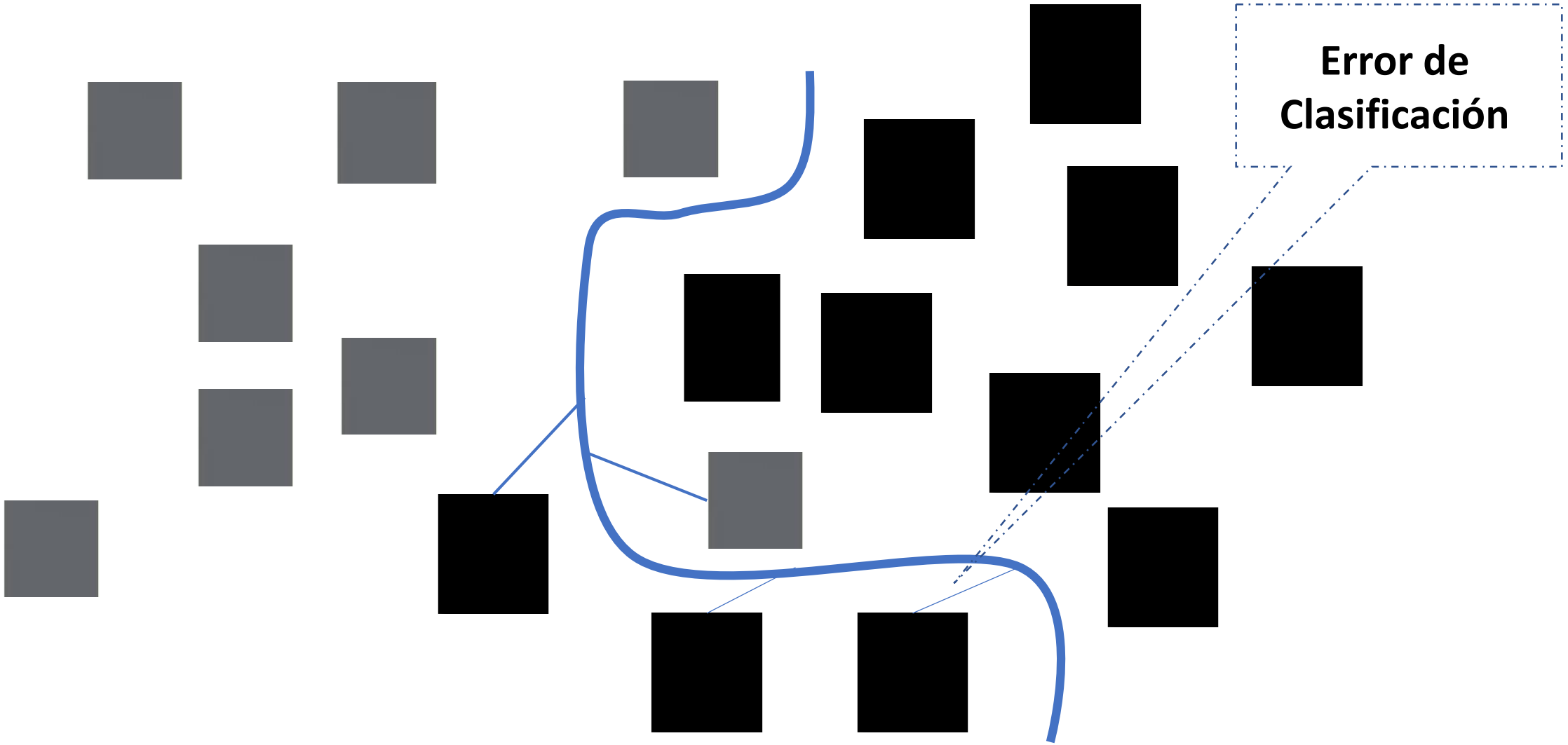
¿Cómo funciona el algoritmo?



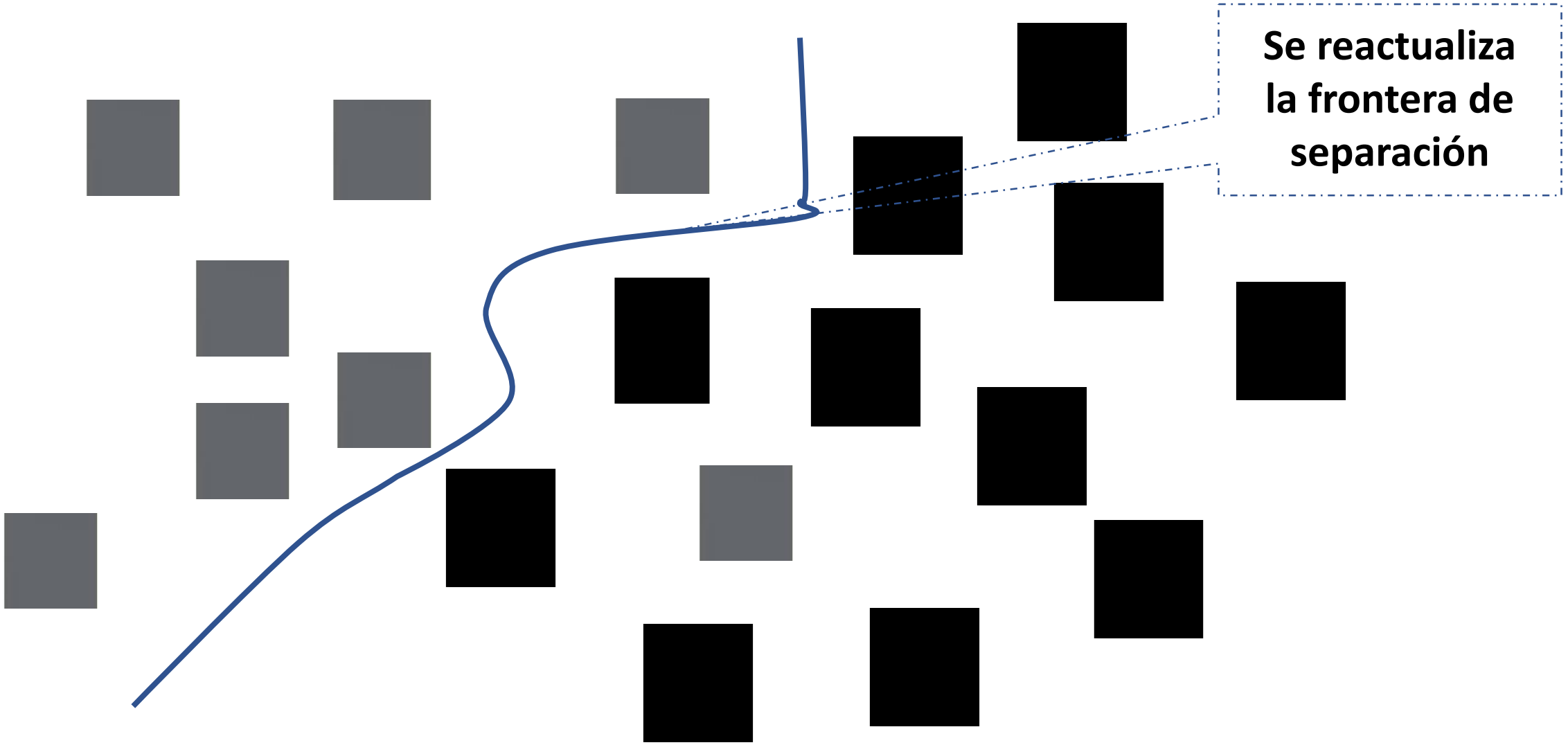
¿Cómo funciona el algoritmo?



¿Cómo funciona el algoritmo?

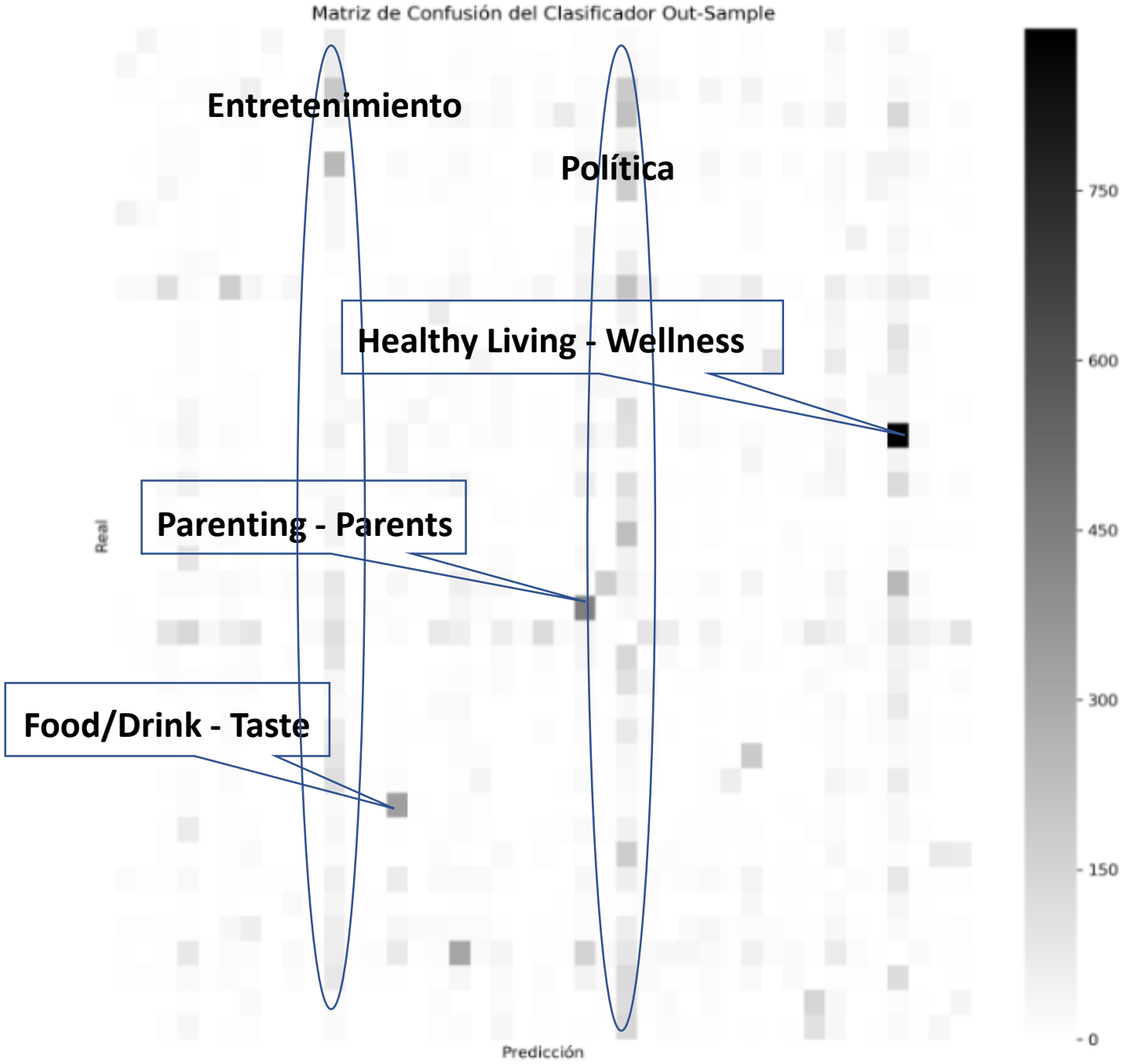


¿Cómo funciona el algoritmo?



El algoritmo no es perfecto en parte porque hay categorías muy parecidas…

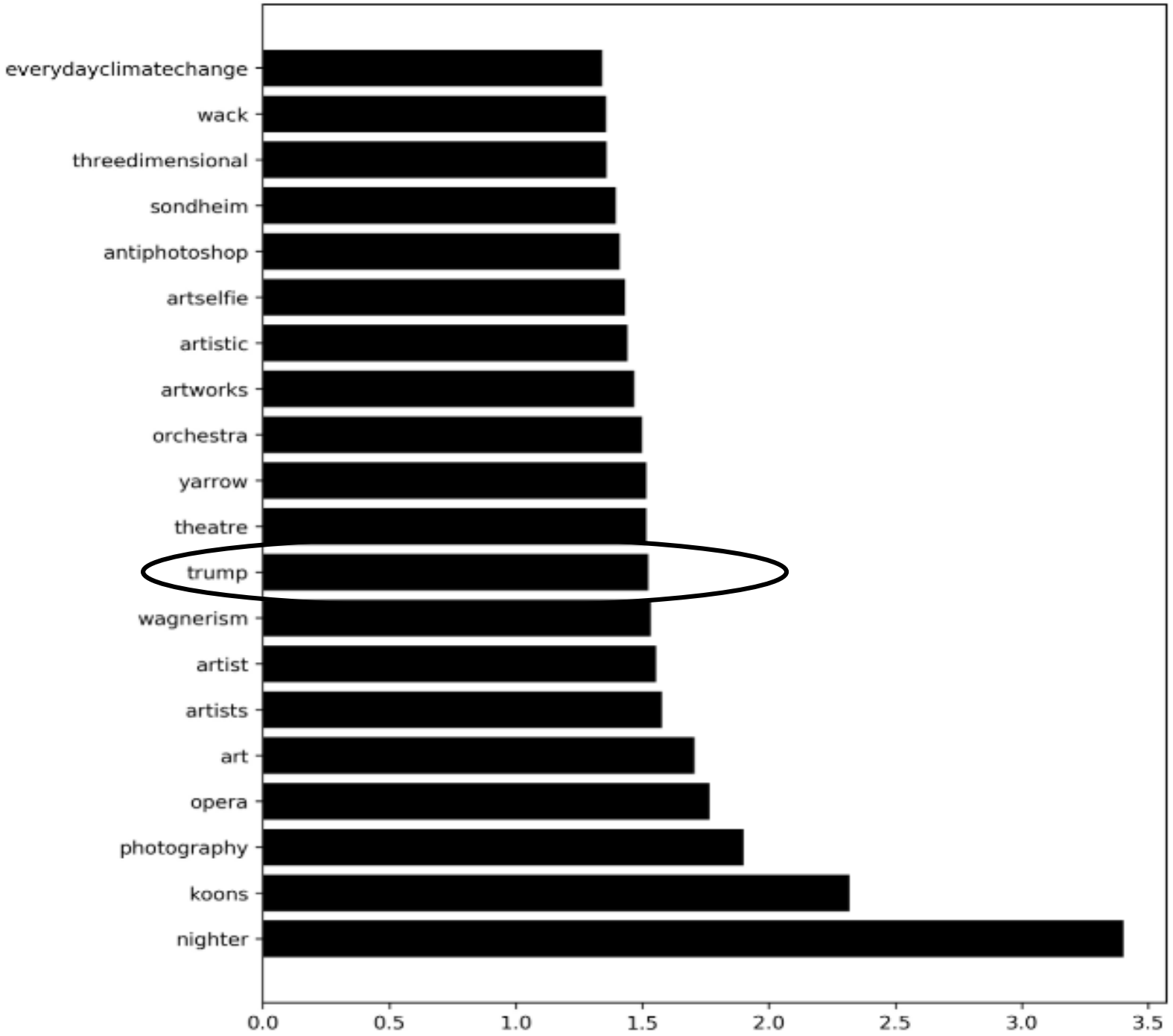
WHALE & JAGUAR



Porcentaje de aciertos: 60.33%

Sensibilidad: 78.67%

Especificidad: 65.09%



Las palabras relacionadas con arte y política son las más importantes en la identificación de categorías.

Conclusiones

1. La mayoría de los textos se concentran en política, vida y bienestar. Mientras en política el sentimiento es negativo, en los demás es positivo.
2. Unos pocos autores escriben muchos textos! Y estos pocos tienen un tono neutral y homogéneo.
3. Identificamos temas de política, ciudades recomendadas y actores famosos.
4. Tenemos éxito relativo en predecir las categorías de noticias a partir de sus descripciones.

Mejoras posibles

- ✓Reducir el número de categorías para mejorar la capacidad predictiva del algoritmo.
- ✓Obtener insights adicionales extrayendo contenido de los links de las noticias (webscrapping) (el código para acceder al contenido de estas noticias se puede solicitar al autor).
- ✓Encontrar más información de los sentimientos implícitos analizando mayúsculas, emojis ó utilizando herramientas más sofisticadas para identificar tonos (Watson Analyzer, IBM).

GRACIAS!

WHALE & JAGUAR