Descripción del paper 4

Objetivo general

Analizar cómo distintos grupos sociales en España reaccionan discursivamente ante el conflicto en Gaza.

Títulos tentativos:

- ***"Mapeando el Ecosistema del Debate sobre el conflicto en Gaza en España: Una Aplicación de Clustering de Datos Mixtos en Reddit"***
- ***"Clustering de Datos Mixtos para el Análisis Social: Un Estudio de Caso sobre el Debate de Gaza en Reddit-España"***
- ***"Cuantificando la Polarización: Un Análisis de Clustering Mixto sobre la Conversación de Gaza en la Sociedad Española"***
- ***"Más allá de la Polaridad: Caracterización de Grupos de Opinión sobre Gaza mediante Clustering de Datos Mixtos en Reddit"***

Procedimiento

Fuentes de datos

Se utilizará la red social **Reddit** como fuente de datos primaria, extrayendo información relativa a posts relacionados con el topic objeto de análisis.

La elección de Reddit como plataforma principal se fundamenta en varias ventajas metodológicas clave para esta investigación:

- Viabilidad de Extracción: A diferencia de X (Twitter) y sus actuales restricciones de pago, Reddit mantiene una API robusta, gratuita y accesible (utilizada mediante la librería PRAW en Python), lo que permite una recolección de datos a gran escala viable para la investigación académica.
- Riqueza Nativa para Datos Mixtos: La plataforma es ideal para nuestro objetivo de crear una matriz de datos mixtos. Cada post genera de forma nativa un ecosistema coherente de variables que alimentarán nuestros modelos de clustering.
- Estructura Comunitaria Ideal para el Análisis Social: La arquitectura de Reddit está organizada en subreddits (p.ej., r/es, r/podemos, r/SpainLibre). Estos foros actúan como comunidades temáticas e ideológicas pre-definidas. Esto nos proporciona la variable categórica (subreddit) de altísimo valor que permite anclar el análisis de los clusters discursivos en un contexto social ya establecido, permitiendo comparar grupos de opinión entre diferentes "cámaras de eco".
- Calidad del Discurso: El relativo anonimato de la plataforma fomenta discusiones más extensas y candidas en comparación con otras redes sociales, proporcionando datos textuales más ricos y profundos para el procesamiento y la posterior interpretación.

El Argumento a favor de considerar una sola red social

El objetivo principal es demostrar la aplicabilidad de nuestros métodos de clustering para datos mixtos. Para hacer eso, necesitamos un dataset lo más "limpio" y metodológicamente sólido posible.

1. El Problema Metodológico: El "Efecto Plataforma"

Si mezclamos datos de Reddit, Bluesky y Telegram en una sola matriz para hacer clustering, introducimos un factor de confusión (confounding variable) masivo: la propia plataforma.

¿Qué significa un score de 10 en Reddit? ¿Es comparable a 10 likes en Bluesky? No.

¿Qué significa un flair de "Política" en Reddit? ¿Es comparable a un custom feed de "Política" en Bluesky? No.

Los usuarios de cada plataforma tienen perfiles demográficos y de comportamiento radicalmente distintos.

El Resultado Más Probable de tu Clustering: Tu algoritmo de clustering (que es bueno) encontraría los clusters más obvios:

Cluster 1: "Los posts de Reddit"

Cluster 2: "Los posts de Bluesky"

Cluster 3: "Los mensajes de Telegram"

El "efecto plataforma" dominaría por nuestro análisis. Anularía tu capacidad de encontrar los clusters sociales que realmente buscamoss, porque estos grupos se verían eclipsados por la diferencia estructural entre las fuentes de datos.

---

2. La Fortaleza Metodológica (Publicable en Revistas de Estadística)

Al centrarte solo en Reddit, creas un ecosistema coherente:

Comparabilidad: Todas tus variables cuantitativas (puntuacion, karma_autor_post, num_comentarios) se miden en la misma escala y contexto. Son directamente comparables.

Significado Consistente: Todas tus variables categóricas (subreddit, flair) pertenecen al mismo universo y tienen un significado compartido por los usuarios de esa plataforma.

Interpretación Válida: Cuando tu método encuentre un cluster, podrás interpretarlo con confianza.

Ej: "El Cluster 2 se caracteriza por un karma_autor_post bajo, flair de 'Debate' y un ratio_upvotes cercano a 0.50".

Traducción: "Es un grupo de usuarios nuevos o polémicos que generan mucha controversia".

Esta interpretación es metodológicamente válida porque todas las variables operan bajo las mismas reglas (las de Reddit).

---

3. Riqueza de Datos Suficiente

Reddit por sí solo te da una matriz de datos mixtos increíblemente rica. Ya tienes:

Texto: titulo, cuerpo_post (y, como siguiente paso, los comentarios).

Quant: puntuacion, ratio_upvotes, num_comentarios, karma_autor_post, antiguedad_cuenta_dias.

Cat: subreddit, flair, es_stickied, es_over18, id_autor (o el estado del autor: 'Suspendido', 'Eliminado').

Procedimiento de extracción de datos

La recolección de datos se llevó a cabo durante [Insertar Rango de Fechas, p.ej., "la tercera semana de octubre de 2025"] utilizando la API oficial de Reddit a través de la librería PRAW (vX.X.X) para Python. Para garantizar una recolección exhaustiva y mitigar los sesgos de muestreo inherentes a un único método de consulta, se diseñó e implementó una estrategia de búsqueda sistemática y multifacética.

Esta estrategia se basó en la combinación de tres dimensiones de búsqueda:

- Contexto (Subreddits): Se definió una lista de 7 subreddits clave que representan una muestra transversal del ecosistema hispanohablante en la plataforma, abarcando desde foros generales (r/es, r/spain, r/AskSpain, r/OpinionesPolemicas) hasta comunidades con un claro sesgo político (r/podemos, r/psoe, r/SpainPolitics).

- Contenido (Queries): Se diseñaron cuatro consultas (queries) booleanas para capturar diferentes "marcos" (frames) del discurso. Estas consultas, insensibles a mayúsculas pero sensibles a tildes (incluyendo variantes con y sin acento), fueron:

    - Marco Político-Doméstico: (gaza OR palestina) AND (sánchez OR sanchez OR feijóo OR feijoo OR gobierno OR pp OR psoe OR podemos)
    - Marco Humanitario-Legal: (gaza OR palestina) AND (genocidio OR humanitaria OR onu OR víctimas)
    - Marco de Seguridad/Conflicto: (gaza OR palestina) AND (hamas OR terrorismo OR ataque OR rehenes OR netanyahu)
    - Consulta General (Catch-all): (gaza OR palestina OR israel)

- Muestreo (Criterios de Ordenación): Para contrarrestar el sesgo algorítmico de la API, cada búsqueda se ejecutó cuatro veces utilizando los diferentes criterios de ordenación (sort) que ofrece Reddit: relevance (relevancia), top (más votados), new (más recientes) y comments (más comentados).

Se llevaron a cabo un total de 112 búsquedas únicas (4 queries x 7 subreddits x 4 sorts). Cada búsqueda se configuró con un filtro temporal de un año (time_filter='year'), es decir se considera la actividad de los últimos 12 meses, desde la fecha de consulta, y un límite máximo de 1.000 posts por consulta, conforme al límite de la API.

Dado que esta estrategia (especialmente al variar los criterios de ordenación) inevitablemente captura los mismos posts múltiples veces, no se consideraron estas duplicidades, garantizando la unicidad de cada registro.

Generación de datos de tipo mixto

Datos extraidos en crudo de Reddit

- **Variables numéricas**
    - Métricas de engagement:
        - *puntuacion* (el score neto del post, proxy de apoyo).
        - *num_crossposts* (Nº de veces que se ha 'crossposteado', mide viralidad interna).
        - *total_premios* (Nº total de premios, mide engagement muy positivo).
    - Métrica de Controversia:
        - *num_comentarios* (proxy de debate o controversia),
        - *ratio_upvotes* (proporción de votos positivos sobre el total de votos (positivos + negativos))
            - Ratio cercano a 1.0 (p.ej., 0.95):

              Significado: Consenso positivo total.

              Interpretación: A casi todos los que lo vieron y votaron, les gustó. Es un contenido que genera mucho acuerdo (p.ej., un meme muy gracioso, una noticia con la que todos están de acuerdo).
            - Ratio cercano a 0.50 (p.ej., 0.55):

              Significado: Controversia máxima.

              Interpretación: El post divide a la audiencia por la mitad. Por cada persona que le da upvote, otra le da downvote. Este es el verdadero "post polémico".
            - Ratio cercano a 0.0 (p.ej., 0.18):

              Significado: Consenso negativo total.

              Interpretación: A la inmensa mayoría de los que votaron, no les gustó.
    - Métricas de Influencia del Autor:
        - *karma_autor_post* (Experiencia/Influencia del autor)
        - *karma_autor_comment* (Experiencia/Influencia del autor como debatidor).
        - *antiguedad_cuenta_dias* (Veteranía del autor).
- **Variables categóricas**
    - Contexto Ideológico:
        - *subreddit* (La comunidad de origen, p.ej., 'spain' vs 'podemos' vs 'SpainLibre')
        - *flair* (La etiqueta que el usuario/mod le dio al post, p.ej., 'Política', 'Debate').
    - Metadatos del Autor:
        - *estado_autor*: (Procesando id_autor y nombre_autor). (Cat: 'Activo', 'Suspendido', 'Eliminado').
    - Metadatos del Post:
        - *es_over18*: post con contenido para mayores de edad (True/False).

        - *es_stickied*: te dice si un post ha sido "pineado" o "fijado" por los moderadores de ese subreddit (True/False).
        - *esta_bloqueado* (Post con comentarios bloqueados por mods, señal de toxicidad: True/False).
    - Tipo de Contenido/Fuente:
        - *es_self_post* (Booleano: True si es un post de texto, False si es un link/imagen).
        - *dominio* (El dominio del link, p.ej., 'eldiario.es', 'self.spain', 'youtube.com'. ¡Variable clave!).

Datos generados a partir de datos textuales extraidos de Reddit

- **Variables numéricas**
    - *Embeddings:* genera K variables cuantitativas que capturan el significado semanatico del texto. Como K suele ser grande, se pueden usar técnicas de reducción de la dimensión como PCA.
    - *Score de sentimientos:** una puntuación continua que indica la polaridad del texto (negativo, neutro, positvo)
    - *Scores emocionales:* puntuaciones continuas que miden emociones especificas como tristeza, ira, miedo.
    - *Tamaño del post:* medido a través de variables como número de palabras o frases.
- **Variables categoricas**
    - *Posición en el conflicto:* pro palestina, pro israel, critico con ambos, indefinido, etc.
    - *Tema principal:* 'Tópico 1: Crítica Política ES', 'Tópico 2: Ayuda Humanitaria', 'Tópico 3: Debate Seguridad/Terrorismo', 'Tópico 4: Crítica a Medios', etc.
    - *Intención del discurso:* 'Opinión', 'Noticia', 'Pregunta', 'Queja', 'Meme/Sátira'.
    - *Menciones:* menciona_politico_es, menciona_hamas, menciona_onu, etc

---

Muestra de datos crudos extraidos de Reddit:

id_post	subreddit	flair	subreddit_subs	id_autor	nombre_autor	karma_autor_post	karma_autor_comment
str	str	str	i64	str	str	i64	i64
"1kgv8ds"	"OpinionesPolemicas"	"Opinión Polémica (Geopolítica)...	77190	"tw5nuqr6"	"Brimiasclasab"	2327	1829
"1gfkehp"	"spain"	null	1040373	"96djz"	"hysbald"	17169	21187
"1ncrij6"	"podemos"	"Canal RED"	32047	"9rejp"	"aritza"	2926	3895
"1o0b0df"	"SpainPolitics"	null	25255	"jzihpgxh"	"Oskarin23-vk-723"	1951	182
"1lsa2u4"	"OpinionesPolemicas"	"Opinión Polémica (General) 🔒"	77190	"Suspendido"	"5mesesintento"	0	0
"1npdc3m"	"SpainPolitics"	null	25255	"124qwber7x"	"TheLastRole"	9051	32995
"1ivd2j7"	"OpinionesPolemicas"	"Opinión Polémica (Política) 🔒"	77190	"5g1nnrgk"	"Valuable_Mirror_6433"	534	2308
"1n8t0qk"	"OpinionesPolemicas"	"Opinión Polémica (Política) 🔒"	77190	"wgk25ifk0"	"Impossible_Neck_7134"	626	1252

antiguedad_cuenta_dias	puntuacion	ratio_upvotes	num_comentarios	es_stickied	es_over18	es_self_post	esta_bloqueado
i64	i64	f64	i64	bool	bool	bool	bool
1084	1160	0.83	1254	false	false	true	false
4768	406	0.9	96	false	false	false	true
4707	12	0.94	0	false	false	false	false
1336	34	0.73	11	false	false	true	false
-1	70	0.73	100	false	false	true	false
498	44	0.95	7	false	false	false	false
1681	94	0.74	140	false	false	true	false

	581	110	0.76	63	false	false	true	false

num_crossposts	total_premios	titulo	cuerpo_post	url	dominio	fecha_creacion
i64	i64	str	str	str	str	datetime[μs, UTC]
2	0	"Estamos a pocas semanas de que...	"Si ninguna potencia hace nada,...	"https://www.reddit.com/r/Opini...	"self.OpinionesPolemicas"	2025-05-07 11:42:34 UTC
0	0	"Portadas de la prensa tras la ...	""	"https://www.reddit.com/gallery...	"reddit.com"	2024-10-30 11:37:24 UTC
0	0	"La Base 6x6 | ¿En qué consiste...	"En el programa de hoy, 9/9/202...	"https://www.youtube.com/watch?...	"youtube.com"	2025-09-09 18:52:23 UTC
0	0	"Hay que romper con Israel y ex...	"A la señora embajadora en Tel ...	"https://www.reddit.com/r/Spain...	"self.SpainPolitics"	2025-10-07 10:29:02 UTC
0	0	"La gente en México va a seguir...	"Con todo lo de la gentrificaci...	"https://www.reddit.com/r/Opini...	"self.OpinionesPolemicas"	2025-07-05 13:37:16 UTC
0	0	"El Rey, ante la ONU: "Nos cues...	""	"https://elpais.com/internacion...	"elpais.com"	2025-09-24 14:16:31 UTC
0	0	""🤡👺 El gObiErnO de TrUmpnO Es...	"[Aquí está el video.](https://...	"https://www.reddit.com/r/Opini...	"self.OpinionesPolemicas"	2025-02-22 07:07:01 UTC
1	0	"La increible similitud entre A...	"La figura de los guetos: Los c...	"https://www.reddit.com/r/Opini...	"self.OpinionesPolemicas"	2025-09-05 02:07:17 UTC

Detección y análisis de comunidades (clustering y análisis de datos)

- Aplicar *nuestros* métodos de clustering para detectar comunidades
- Interpretación de las comunidades basado en:
    - Análisis de datos mixtos y textuales
        - Análisis de datos "clásico" para las variables de tipo mixto "clásicas".
        - Ideas para datos textuales: Ranking de terminos más relevantes basados en frecuencia de palabras, tf-idf, odss ratio. Nubes de palabras.
        - Idea para refinar el análisis: Centrar el análisis de los clusters en el top X posts más relevantes de cada cluster, identificandolos de algún modo, como por ej: el post medoid y los 9 más cercanos a él, para cada cluster.
    - Conocimiento de expertos en el topic social analizado.