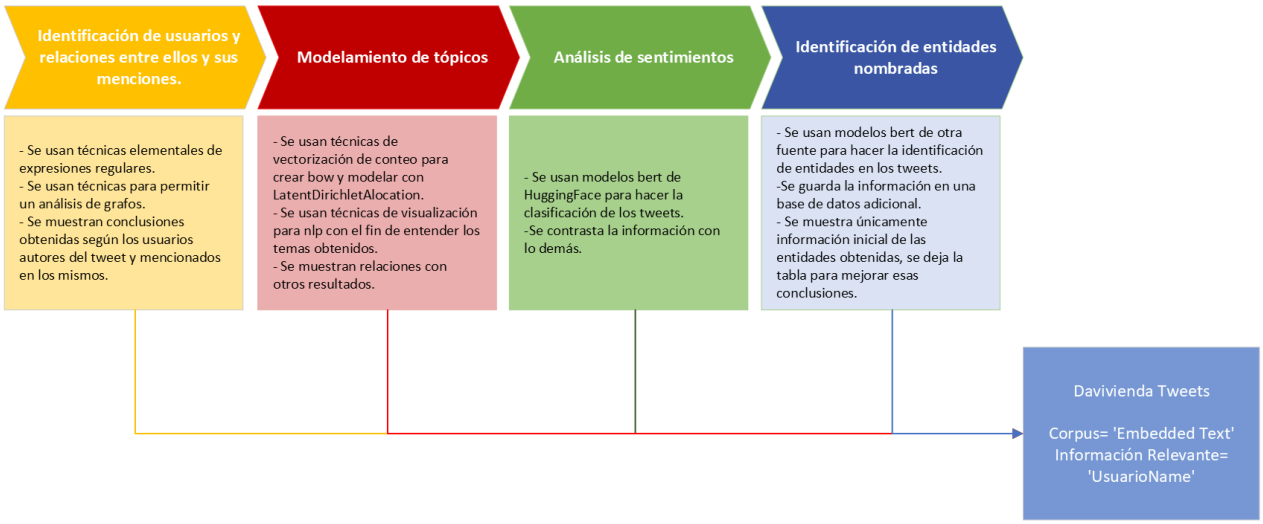
Análisis de tweets

Introducción

En este documento se resume lo obtenido en la Prueba de Conocimiento ViVi para mostrar los resultados de la base davivienda_tweets. Para este ejercicio se aplicaron diferentes modelos para ciertos problemas específicos con el fin de mostrar la variedad de temas que desde el procesamiento de lenguaje natural ayudarían a extraer información relevante del corpus analizado.

Figura 1. Modelos aplicados



Toda la documentación del ejercicio se guardó en el repositorio abierto: https://github.com/Izainea/pruebas_vivi/tree/master/prueba_NLP_cientifico_vivi, utilizando una estructura documental similar a la presentada en la metodología de proyectos de ciencia de datos para equipos de Microsoft. Dicha estructura identifica tres carpetas principales del proyecto:

- **Código** donde encontrarán dos cuadernos:
    - *Análisis de entidades.ipynb* que contiene la implementación de un modelo de deeppavlov para identificar entidades nombradas del texto y que guardo todas esas entidades (Personas, Organizaciones, Locaciones, Obras de arte, entre otras) en la tabla Tweets_entidades.csv (Que guarda texto, entidad e id del tweet donde se encuentra); y el cuaderno,
    - *Exploración.ipynb* que contiene la aplicación técnicas elementales de expresiones regulares que permitieron la identificación de relaciones de usuarios y sus menciones. También la implementación del modelo nlptown/bert-base-multilingual-uncased-sentiment descargado de HuggingFace para hacer un análisis de sentimientos, así como la aplicación de limpieza y preprocesamiento de texto para poder implementar

un modelo de tópicos (LatentDirichletAlocation) que permitio segmentar los tweets en temas relacionados con el texto descrito. Así mismo, contiene una sección con la que se destaca la información contenida en este resumen.

- **Datos** Se guarda toda la información que se obtuvo de este ejercicio, si bien, por la sencillez del proyecto no se generó una carpeta para datos crudos, intermedios y procesados. Se mantuvo los nombres acordes con lo obtenido en cada fase del proyecto. Se destacan las siguientes bases:
    - *Tweets_Entidades.csv* Contiene todas las entidades nombradas en los tweets.
    - *Corpus.mn* Corpus generado con gensim.
    - *Lda_model.model* Modelo LDA con gensim.
    - *Usuarios_mencionados.csv y Relacion_mencionados.csv* un resumen de la interacción de los usuarios que se nombran en los m,ismos tweets.
    - *Tweets_procesados.csv* Información completa que se utilizó para los posteriores análisis.

- **Documentación** Aloja únicamente este documento.

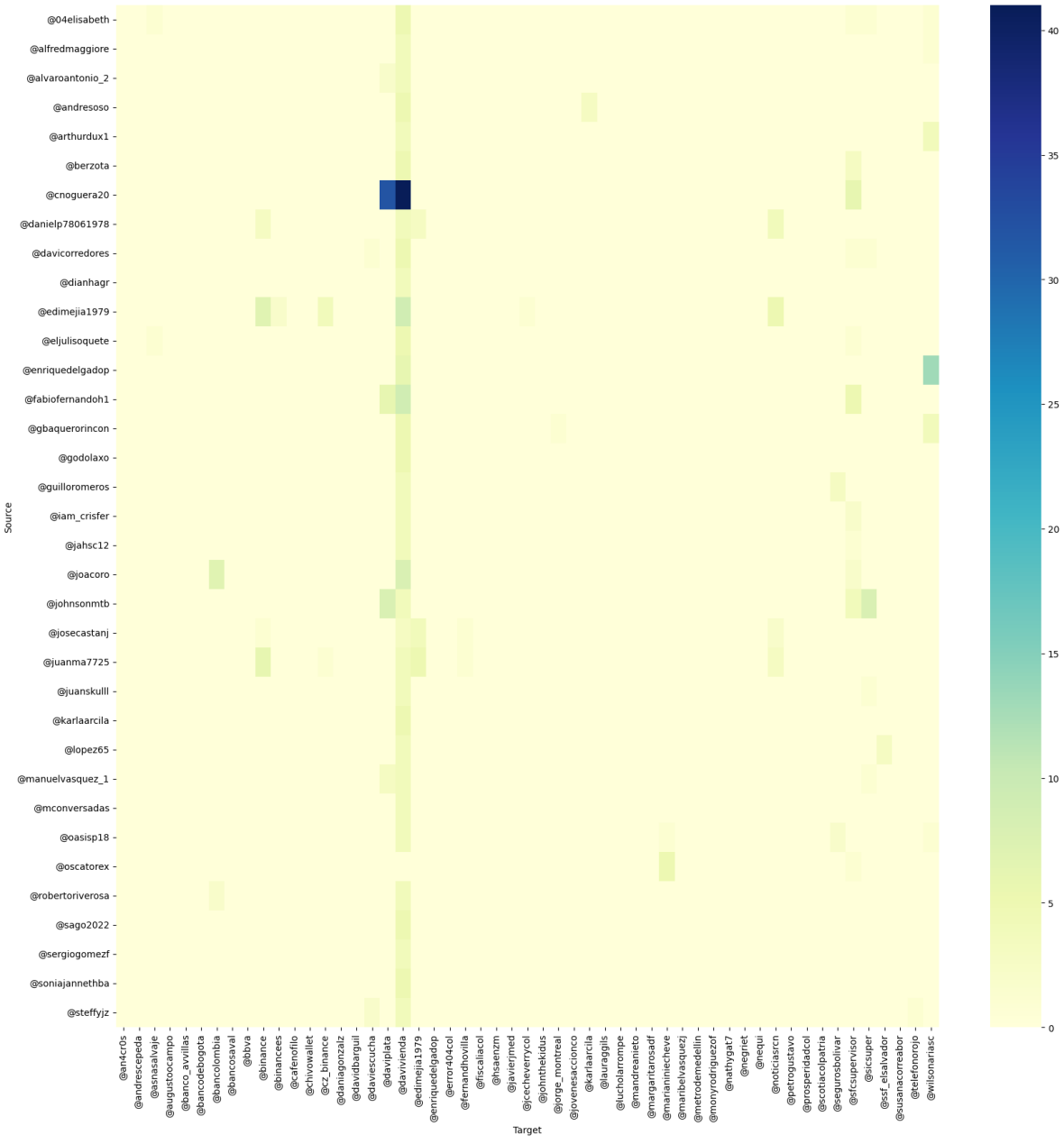
Resultados relevantes

En resumen, se encontró lo siguiente:

Relación entre usuarios

Frente a la relación de usuarios se hizo un esfuerzo por extraer los usuarios mencionados en cada tweet, esta lista de usuarios por tweet se utilizó para hacer conteos entre los usuarios autores del tweet y los usuarios mencionados. Inicialmente se determinó un mapa de calor que muestra lo siguiente:

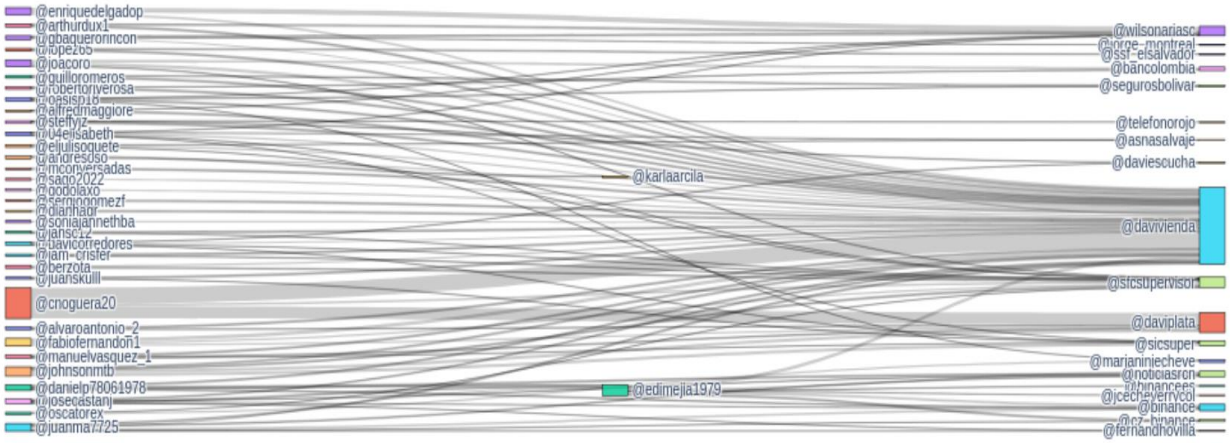
Figura 2. Diagrama de calor menciones



Efectivamente, la gran mayoría de usuarios menciona a @davivienda y @daviplata pero hay uno en particular que ha interactuado con la cuenta en al menos en 41 ocasiones, se trata de @cnoguera.

El siguiente gráfico muestra cómo se configuran las menciones entre usuarios:

Figura 3. Diagrama sankey de interacción de usuarios



Vemos la interrelación entre los usuarios que se mencionan más de tres veces. La mayoría de tweets que tenemos muestran la comunicación constante entre los usuarios y el twitter de @davivienda, @daviplata y algunas necesidades de expresar situaciones ante los medios, pueden ser denuncias o campañas publicitarias. En efecto, esto se reconoce al apreciar una gran cantidad de tweets que hacen referencia a @wilsonariasc (senador de la república), @sfcsupervisor (superfinanciera) y @noticiasrcn.

Revisando con más detalle tenemos:

Tabla1. Las 20 cuentas que más menciones hacen

Source	Usuarios mencionados	Total de menciones
@davivienda	208	248
@juanma7725	11	27
@santini_es	10	10
@davicorredores	10	22
@guillolomeros	9	22
@johnsonmtb	8	40
@damarismarino	8	8
@edimejia1979	8	32
@enriquedelgadop	7	24
@julioorozco29	7	10
@gleniiaaa	6	8
@blan_charjavier	6	6
@valcinate	6	6
@eamdelquindio	6	6
@aposada151	6	6
@jesualmar1	6	6

@luisfer45684347	6	9
@bornacellijimmy	6	10
@andresoso	6	12
@daviescucha	6	7

Tabla2. Las 20 cuentas más mencionadas

Target	Usuarios que lo mencionaron	Total de menciones
@davivienda	746	1103
@wilsonariasc	80	106
@sfcsupervisor	63	88
@daviplata	57	121
@marianiniecheve	53	60
@asnasalvaje	45	47
@bancolombia	44	54
@segurosbolivar	32	39
@sicsuper	20	27
@karlaarcila	20	25
@andrescepeda	13	17
@nequi	13	14
@edimejia1979	12	25
@bancodebogota	11	11
@noticiasrcn	10	22
@petrogustavo	10	10
@fiscaliacol	9	10
@daviescucha	9	11
@johnthekidus	8	8
@binance	8	22

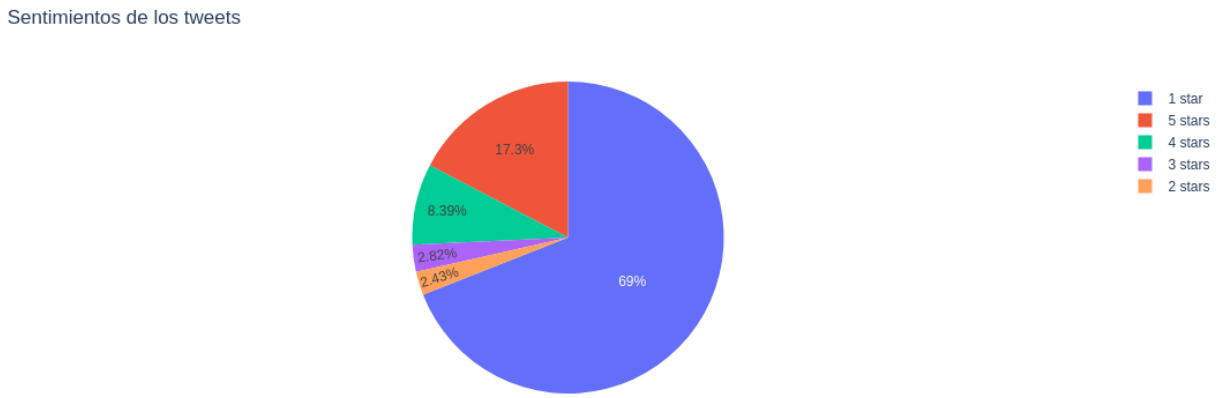
Análisis de sentimientos

Como se mencionó en la introducción se utilizó un modelo de huggingface para identificar los sentimientos de los tweets (no fue reentrenado), se trata del modelo nlptown/bert-base-multilingual-uncased-sentiment. Cuya documentación se puede ver haciendo click al hipervínculo.

Según este modelo, un comentario que tenga una estrella implica que el usuario tiene una opinión negativa respecto a su mención y cinco estrellas simboliza una opinión positiva. En ese sentido, se uso este modelo pensando en que los tweets que hacen referencia a @davivienda estan sugiriendo una opinión a un servicio que presta el banco.

Según este modelo se obtuvo:

Figura 4. Distribución de sentimientos

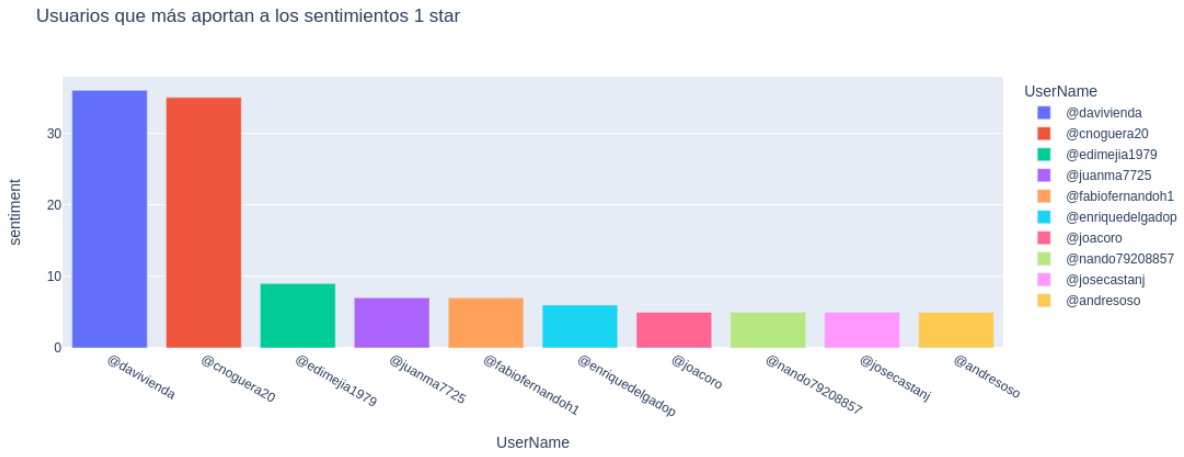


Que se puede resumir en:

Sentimiento	Opiniones
1 star	936
2 stars	43
3 stars	48
4 stars	43
5 stars	175

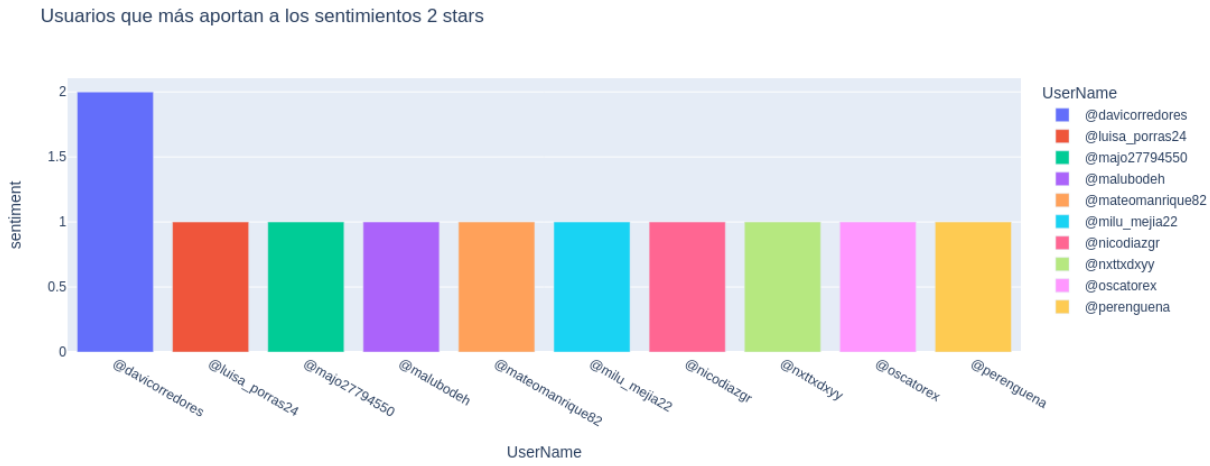
Aprovechando la información que tenemos de usuarios hacemos un reporte de los usuarios que más contribuyen a los diferentes sentimientos de los tweets:

Figura 5. Distribución de usuarios con comentarios de una estrella



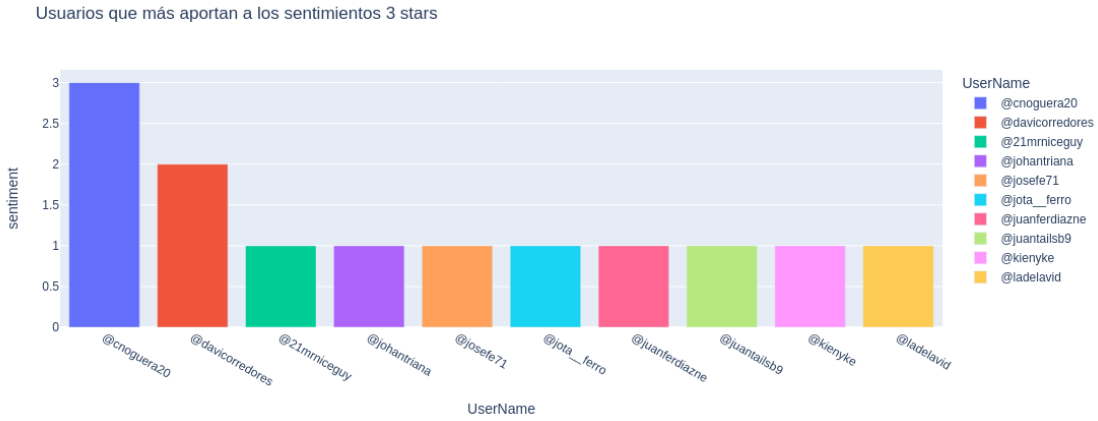
En este caso, estos tweets pueden hacer alusión a problemas de servicio y la respuesta de @davivienda se mantiene con el tono de las conversaciones.

Figura 6. Distribución de usuarios con comentarios de dos estrellas



Son tweets que muestran una inconformidad por parte de los usuarios.

Figura 7. Distribución de usuarios con comentarios de tres estrellas



Aquí aparecen pocos tweets de cada usuario.

Figura 8. Distribución de usuarios con comentarios de cuatro estrellas

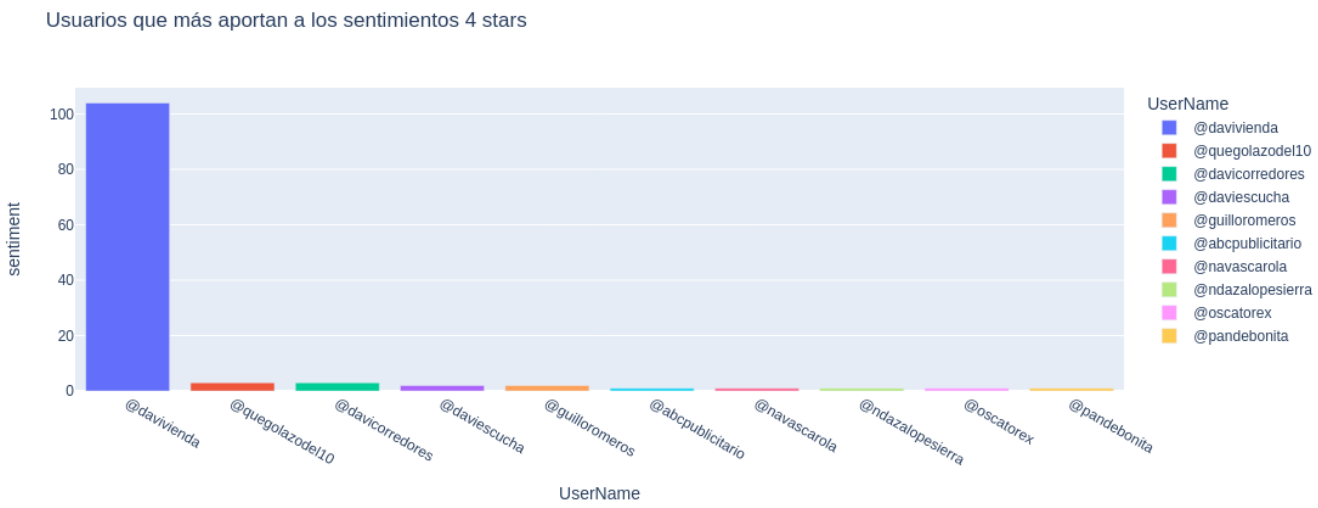
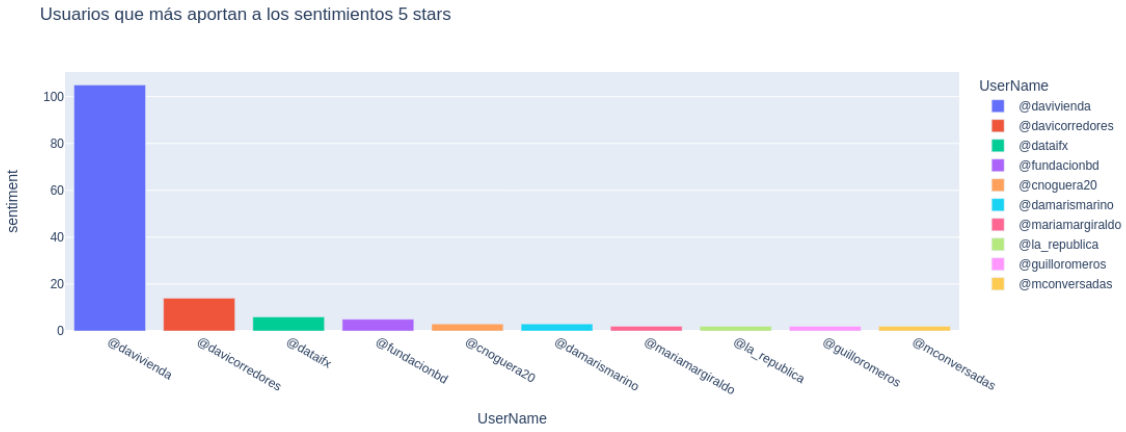


Figura 9. Distribución de usuarios con comentarios de cinco estrellas

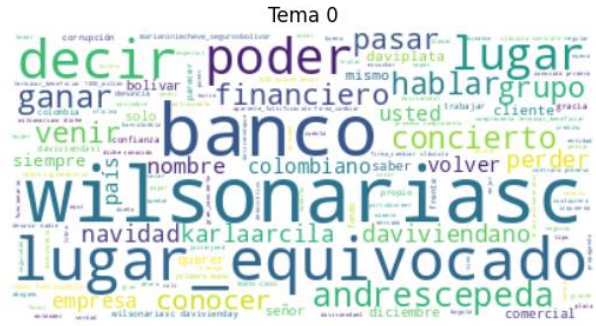


Vemos la relación cordial de @davivienda con sus usuarios empresariales. Tanto para cuatro estrellas y cinco estrellas la mayoría de cuentas son comerciales.

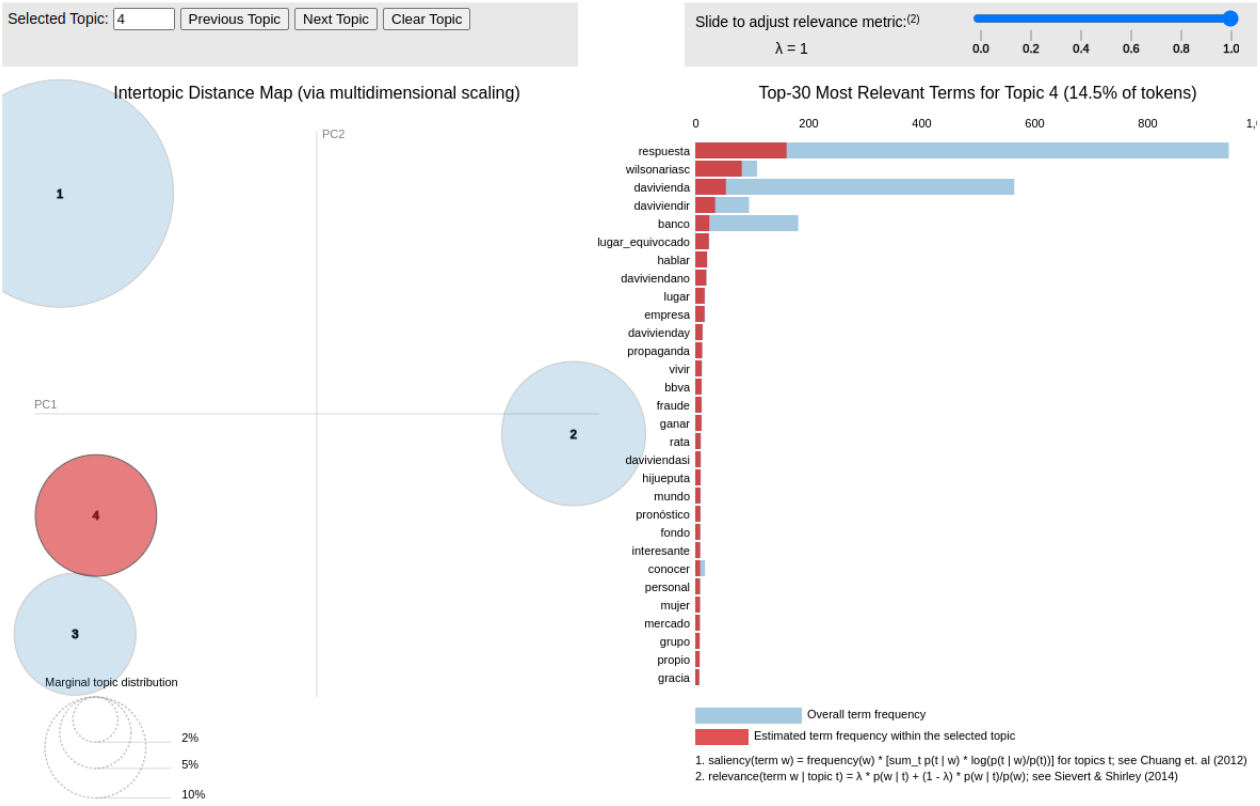
Exploración de temas

En esta sección se presenta una aproximación no supervisada de los posibles temas que se tienen en los tweets. Se uso un modelo LDA para lograrlo y según el coeficiente de coherencia aplicándolo a máximo 10 temas se identificó que la mejor aproximación se encuentra con 4 temas. Según el modelo tenemos:

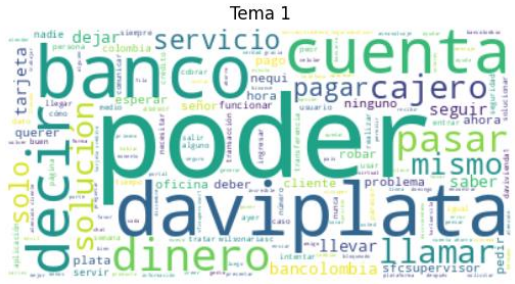
Tema 0



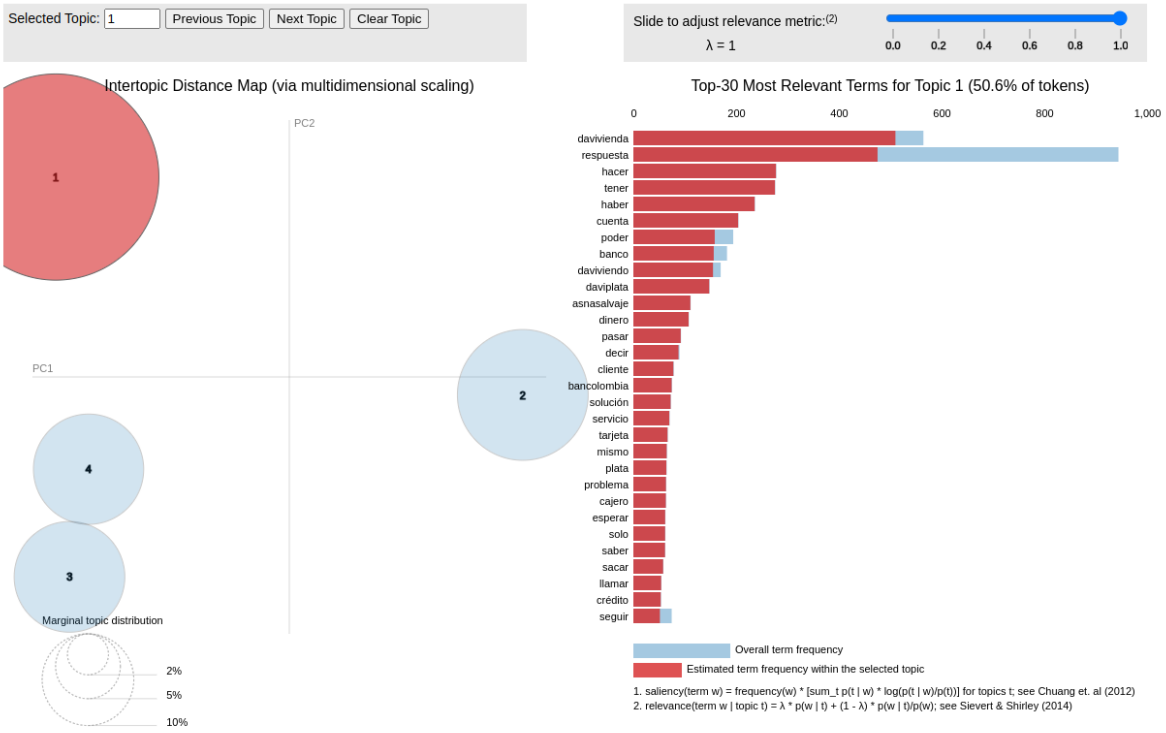
Un tema particularmente interesante, pues entre los tweets más destacados están los que mencionan a @wilsonariasc que hablaba sobre un fraude. Cuando se hace una búsqueda de términos relevantes palabras como propaganda, ganar, hijueputa y rata aparecen.



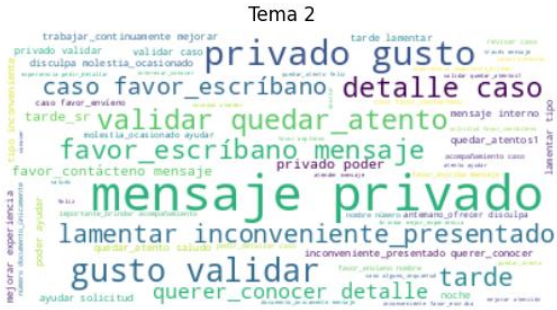
Tema 1



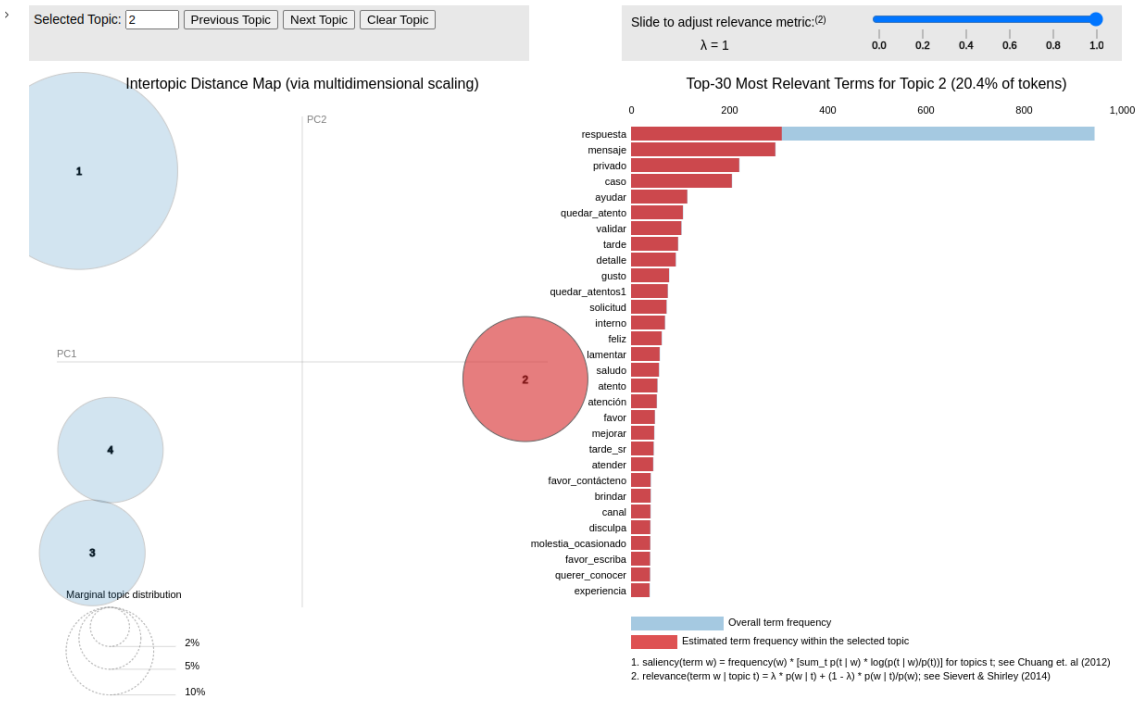
En este tema se agrupan tweets que indican dificultades con el acceso al dinero. Bien sea por transferencia o por dificultades con otras plataformas se muestran palabras como: daviplata, dinero, respuesta, banco, bancolombia, servicio, entre otras.



Tema 2

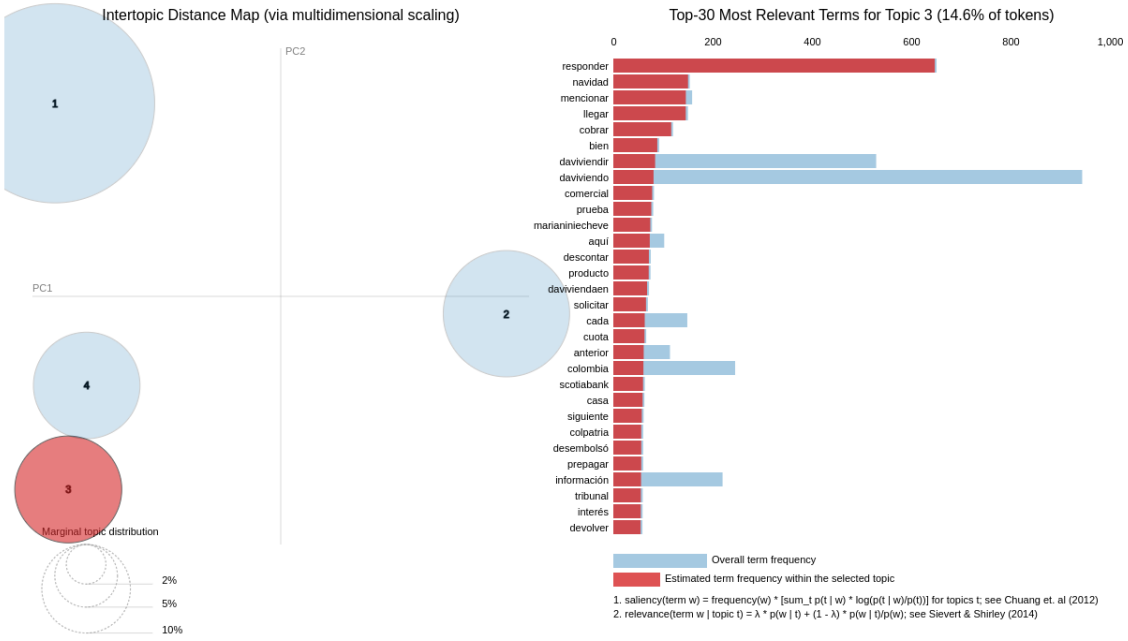


En este tema se presentan palabras asociadas a la validación de información y servicio al cliente. Muestran palabras comúnmente utilizadas en este ejercicio y se identifican palabras que hacen referencia a la comunicación banco cliente: caso, ayudar, mensaje, quedar atento, solicitud, canal.



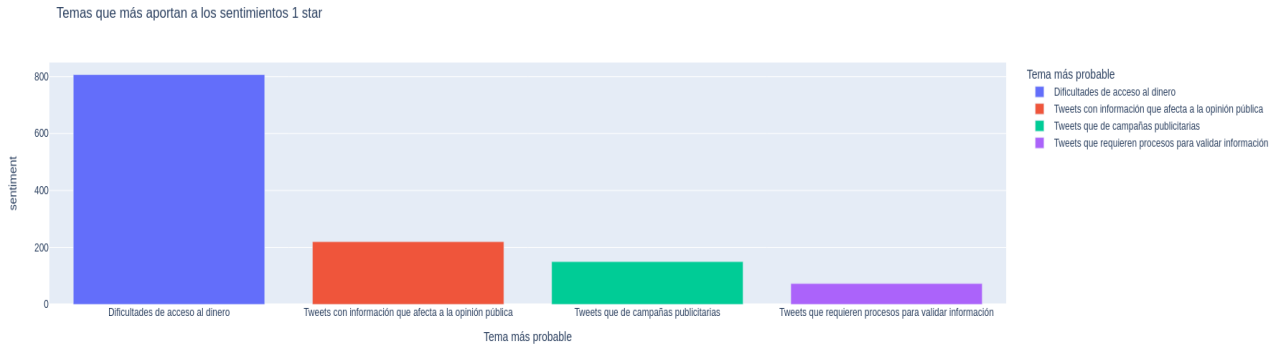
Tema 3

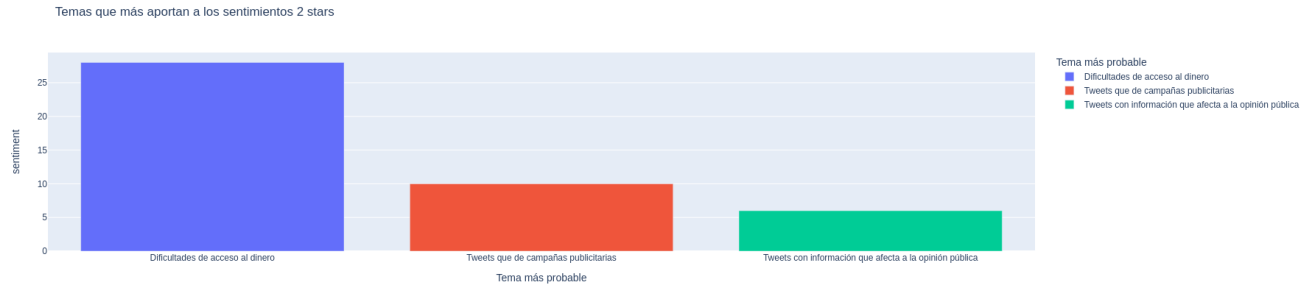
Aquí hay una clara alusión a campañas publicitariasy a relaciones comerciales. Las palabras relevantes que se encuentran en este cluster son: navidad, llegar, comercial, daviviendaen, desembolsó, solicitar.



Relación sentimientos y temas

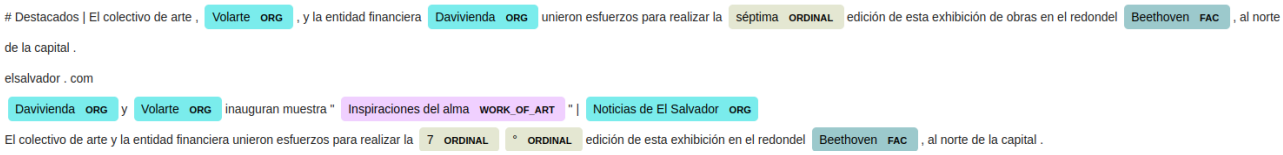
En este aspecto podemos asociar el primer tema como el más negativo, los tweets que tienen que ver con el acceso a servicios y la validación de información resultan positivos, así como los que aluden a campañas publicitarias y relaciones comerciales. No obstante, lo que tiene que ver con la opinión pública está polarizado.





Análisis de entidades

Finalmente se identificaron entidades en cada tweet. Se reconocieron personas, organizaciones, locaciones, obras de arte, entre otros. La información para todos los tweets se puede encontrar en la base de datos Tweets_entidades.csv. Ejemplo de lo que se aplicó y obtuvo se puede ver en el siguiente tweet:



Frente a las organizaciones más mencionadas en los tweets se encontró:

Texto	Menciones
davivienda	207
bancolombia	75
daviplata	40
whatsapp	17
banco davivienda	9
banco de bogotá	8
segurosbolivar	7
pse	7
banco	7
nequi	5
marianiniecheve	5
scotiabank colpatria	5

Respecto a las locaciones:

Texto	Menciones
colombia	47
bogotá	10
cali	7
barranquilla	4
bancolombia	4
canadá	4
el salvador	4
bogota	3
méxico	3
ciudad bolivar	3
misterpan barranquilla	2

Finalmente, respecto a las personas:

Texto	Menciones
marianiniecheve	15
diego	7
laura	7
petro	6
karlaarcila	6
juan	6
efraín forero	5
jorge_montreal	5

Conclusiones

En este apartado se implementaron varias soluciones que ofrece el procesamiento de lenguaje natural para determinar cierta información relevante de los textos, sin embargo, no se hizo un entrenamiento adecuado a las necesidades del negocio. Aun así, la información extraída puede tener valor para diseñar nuevas formas de procesar esta información.

Respecto a lo obtenido vale la pena resaltar la dinámica de los sentimientos que se observó a partir del modelo implementado, aunque no hubo una correlación tan fuerte si se observaron temas que inciden en comentarios negativos en los tweets.

Es fácil identificar tweets tipo denuncia revisando usuarios, tema y menciones. Este trabajo es apenas un abrebocas de muchas formas de establecer relaciones entre usuarios y todos los canales de @davivienda.