Minería de textos

Tweets sobre Davivienda
(Diciembre 2021)

Presentado por: Juan Manuel Romo

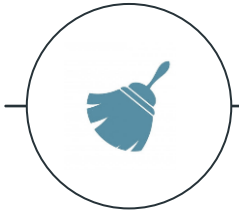
Dataset (Conjunto de datos)

Se revisó un conjunto de **1811 tweets** que de alguna manera están vinculados con la cuenta de Twitter del Banco Davivienda.

- Los tweets corresponden a las primeras tres semanas de **diciembre del año 2021**.
- Hubo **1145 usuarios diferentes** que tuitearon sobre el Banco.
- Por medio de librerías de uso libre, se emplearon técnicas de **análisis de textos y lenguaje (NLP) para extraer la mayor cantidad de información** de los tweets.

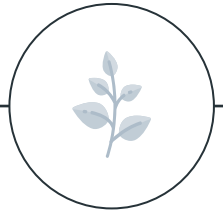
Preprocesamiento de los textos

Limpieza



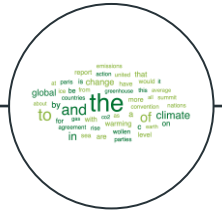
Se remueven caracteres como **números y puntuaciones.**

Estemarizar



Se empleó el stemmer Snowball para combinar palabras con la **misma raíz**.

StopWords



Se eliminan **palabras sin significado propio**, en particular, se incluyó la palabra “davivienda”

Tokenizar

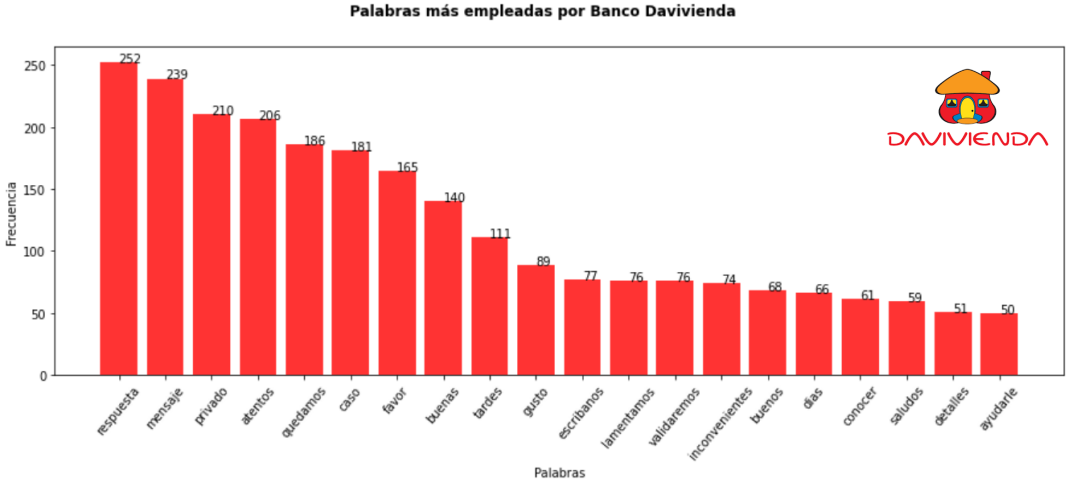
Separar cada palabra del texto.

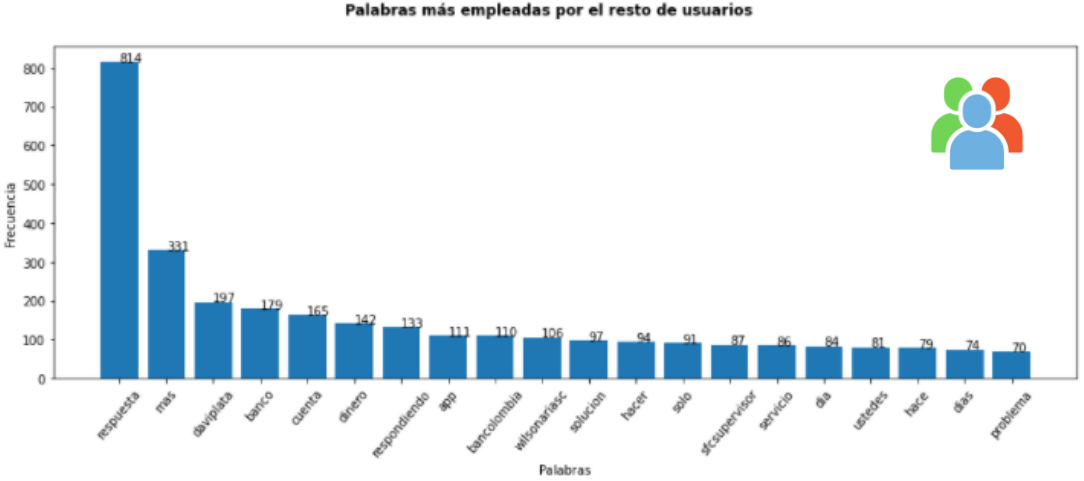
Análisis de los datos

1. Palabras más frecuentes
2. Fechas de actividad
3. Estimación automática de temas con LDA
4. Propuestas para próximos análisis

Palabras más frecuentes

La palabra "respuesta" fue la más repetida. Esta aparece automáticamente al dar una responder algún tweet.

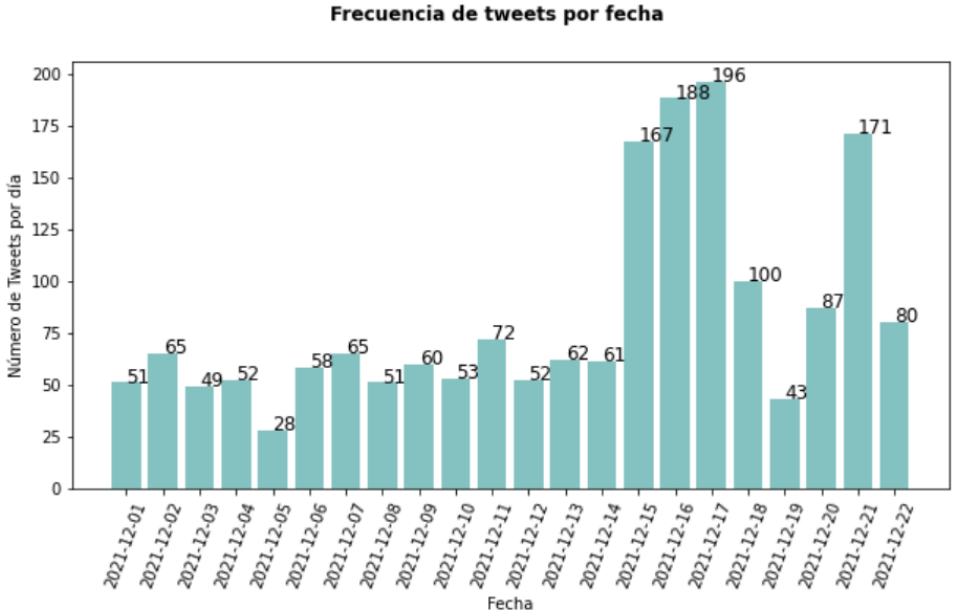




Fechas de actividad

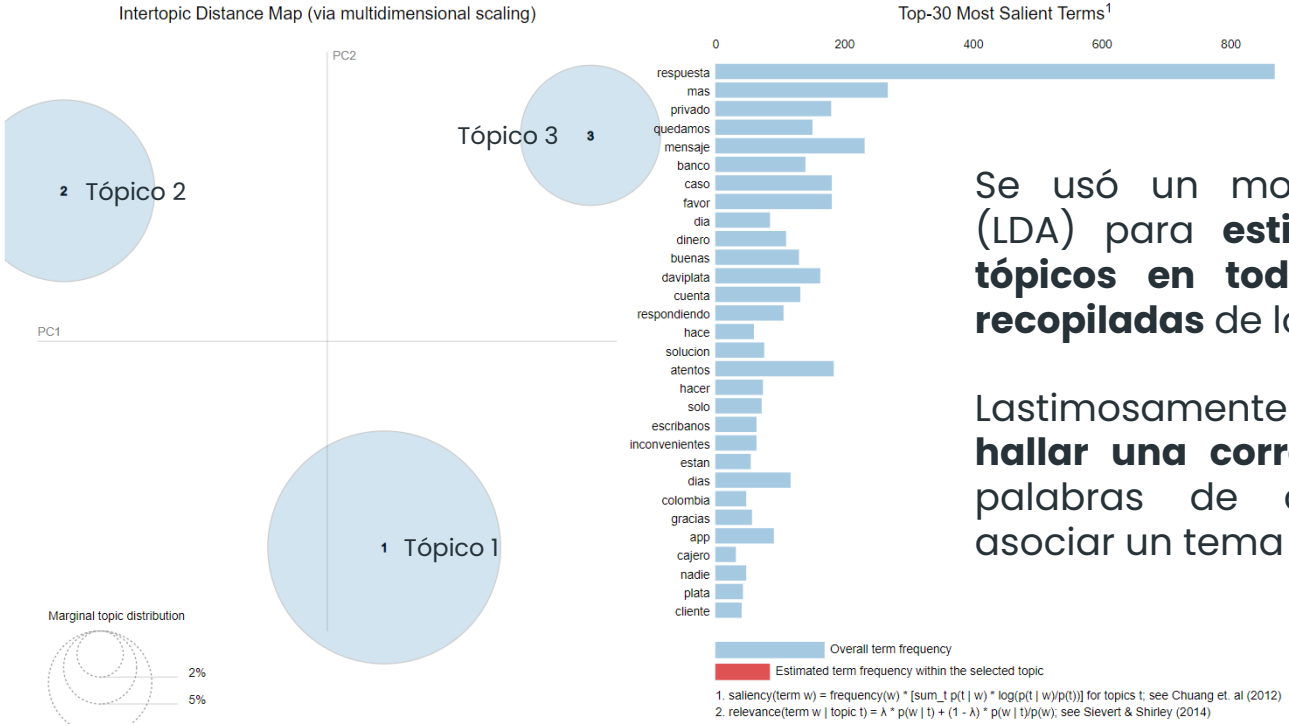
	Semana 1	Semana 2	Semana 3	total_count
respuesta	161	207	694	1062
mas	81	80	178	339
mensaje	19	20	249	288
privado	5	15	213	233
favor	21	20	186	227
caso	14	15	196	225
quedamos	6	9	182	197
daviplata	42	39	114	195
banco	47	46	86	179
cuenta	30	44	91	165
buenas	7	18	139	164
atentos	4	5	143	152
dinero	38	26	78	142
dias	28	24	88	140
respondiendo	23	33	78	134
tardes	7	12	102	121
app	50	22	41	113
bancolombia	31	14	63	108
wilsonariasc	0	0	106	106
gusto	2	6	87	95

- **Matriz Término-Documento para repetición de palabras por semana**



- La **mayor actividad** se presentó en la **tercera semana**.
- El nombre de usuario "**wilsonariasc**" tuvo gran participación la tercera semana, al igual que "**daviplata.**"

Estimación automática de temas



Se usó un modelo automático (LDA) para **estimar 3 temas o tópicos en todas las palabras recopiladas** de los tweets.

Lastimosamente **no fue posible hallar una correlación** entre las palabras de cada grupo, ni asociar un tema a cada uno.

- **Distribución de los tópicos estimados.**

Observaciones generales

Quejas y reclamos

Corresponden a los tweets con mayor interacción en la red. Véase caso Wilson Arias (2500 likes) o Edison Mejía .

Problemas con Daviplata

Se destacan en la tercera semana con el usuario Carlos Noguera.

Participación del banco

Gran parte de los tweets corresponden a respuestas del Banco a quejas y reclamos. Se busca validar o contactar por interno.

Próximos análisis

Dividir los tweets

Entrenar un modelo de LDA para los **tweets del Banco** y otro para los **de los usuarios.**

Optimizar el filtro de palabras

Examinar si hay palabras frecuentes sin mayor información y descartarlas, por ejemplo “respuesta.”

Filtrar por frases

Realizar la tokenización en frases y no por palabras.

Satisfacción del usuario

Emplear otras técnicas de aprendizaje automático (ej. SVM) para **conocer la satisfacción del usuario en un tweet**.