Procesamiento de Lenguaje Natural usando Tweets de Davivienda

Prueba técnica para la selección de Científico(a) de Datos

Nikolás Sebastián Buitrago Vargas
Economista y magister en estadística

Lunes 20 de marzo de 2023

Objetivo: Identificar qué se habla de Davivienda en Twitter

Insumo

Base de datos con:

- **1.811** Tweets
- **1.168** usuarios de Twitter
- Tweets entre el **1 y el 22 de diciembre de 2023**
- **12** variables

Se propone el desarrollo de las siguientes tareas

- Exploración preliminar de la información

  Se encuentran cuentas institucionales de Davivienda (@Davivienda y @davicorredores) que tienen un importante número de Tweets.

- Preprocesamiento de los datos

  - En los textos de los Tweet se encuentra información de número de comentarios, me gusta y retweets. Se limpia esta información.
  - Se identifica el tipo de Tweet, si es propio o "En respuesta a…".
  - Se hace limpieza usual de los textos: todo en minúsculas, eliminar puntuación y caracteres extraños, quitar palabras vacías y tokenizar en términos de palabras.

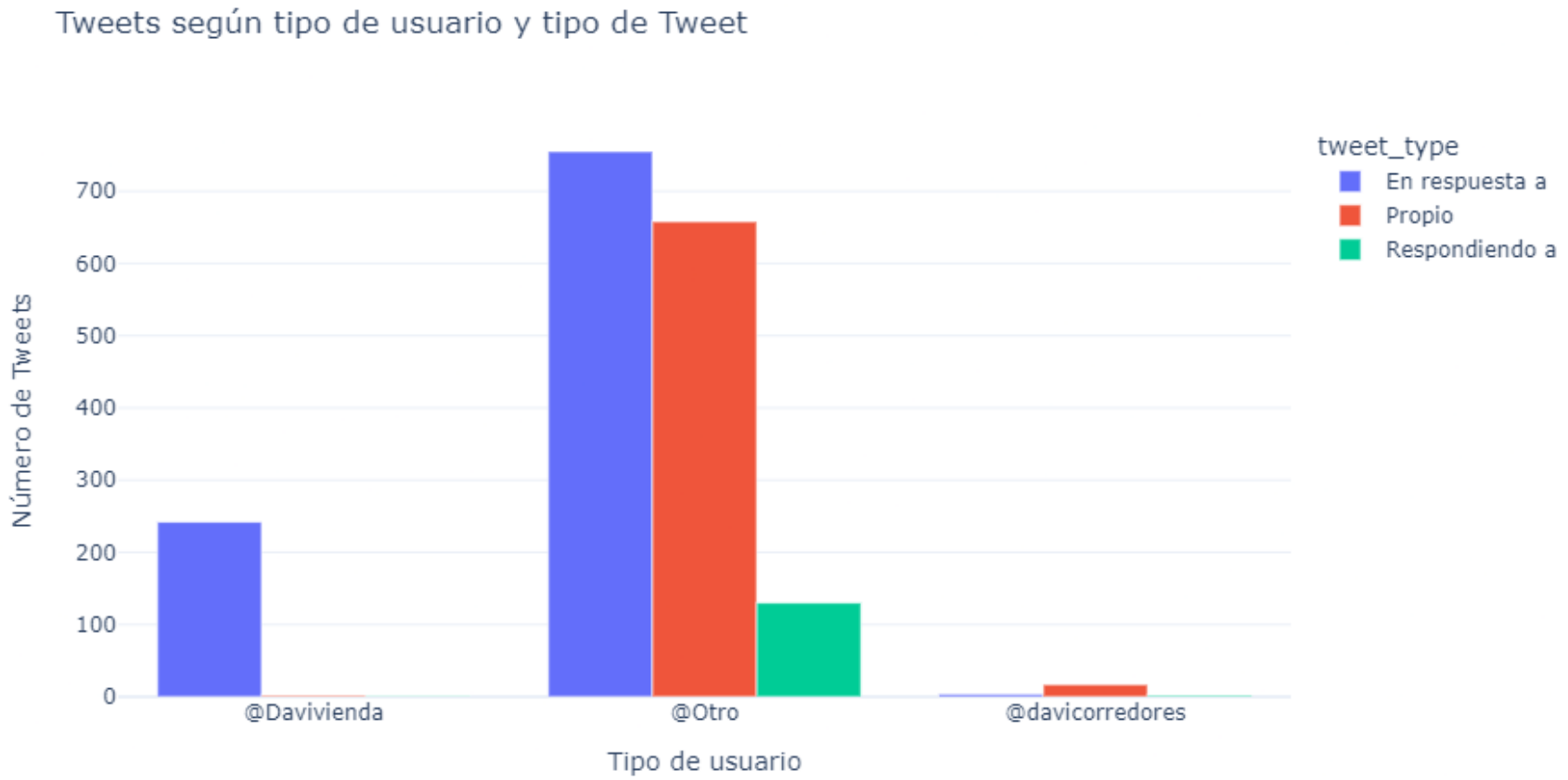
- Propuesta de modelamiento

  Obtener una representación numérica de las palabras en un plano cartesiano usando una arquitectura **CBOW**.

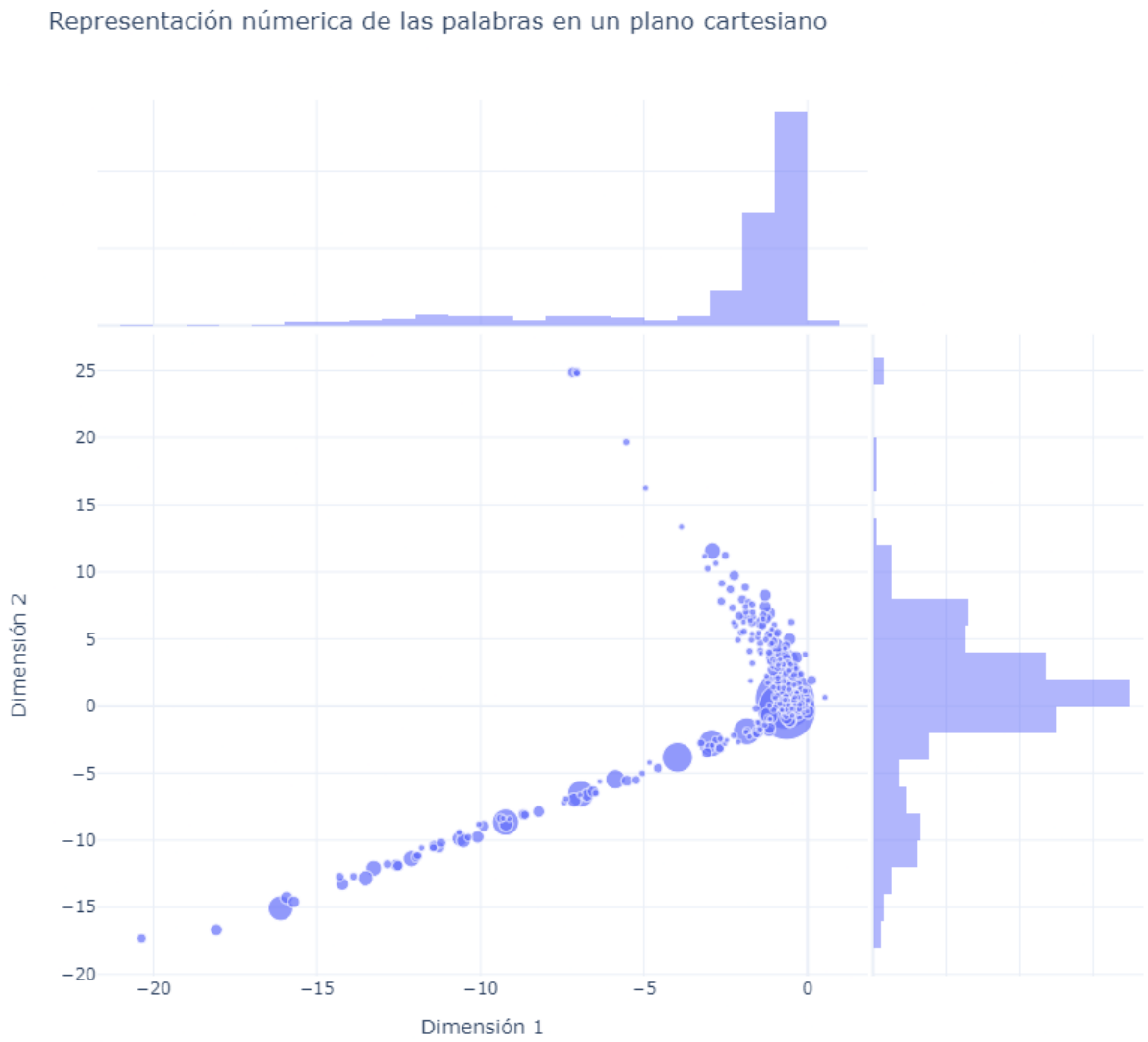
  En este plano identificar patrones de temáticas en los comentarios de Twitter y su relación con sentimientos negativos y positivos.

Un primer patrón de tipo de Tweet y el tipo de usuario

La cuenta @Davivienda se caracteriza por responder a otros usuarios. El resto de usuarios tiene un mayor balance entre Tweet propios y en respuesta a otros



Resultados de la modelación



Se obtiene una representación numérica de las palabras en los Tweets.

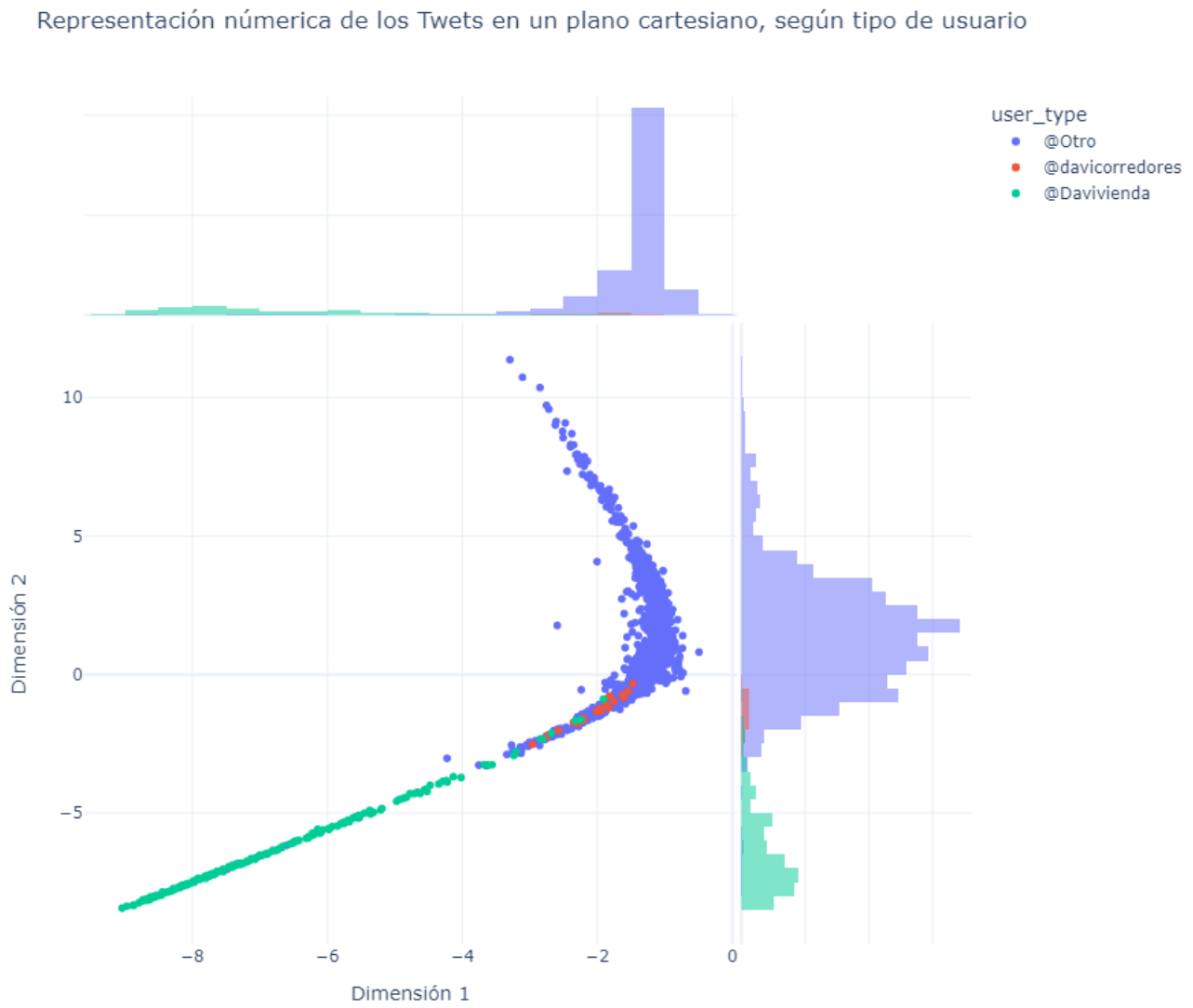
El tamaño de la burbuja indica el número de ocurrencias de la palabra.

Como era de esperarse, la palabra Davivienda es la que más ocurrencias tiene.

La cercanía de la palabras indica que tienden a usarse en un mismo contexto.

Con lo cual al explorar en detalle (la gráfica en el cuaderno es interactiva) se pueden identificar **temáticas** al interior del plano.

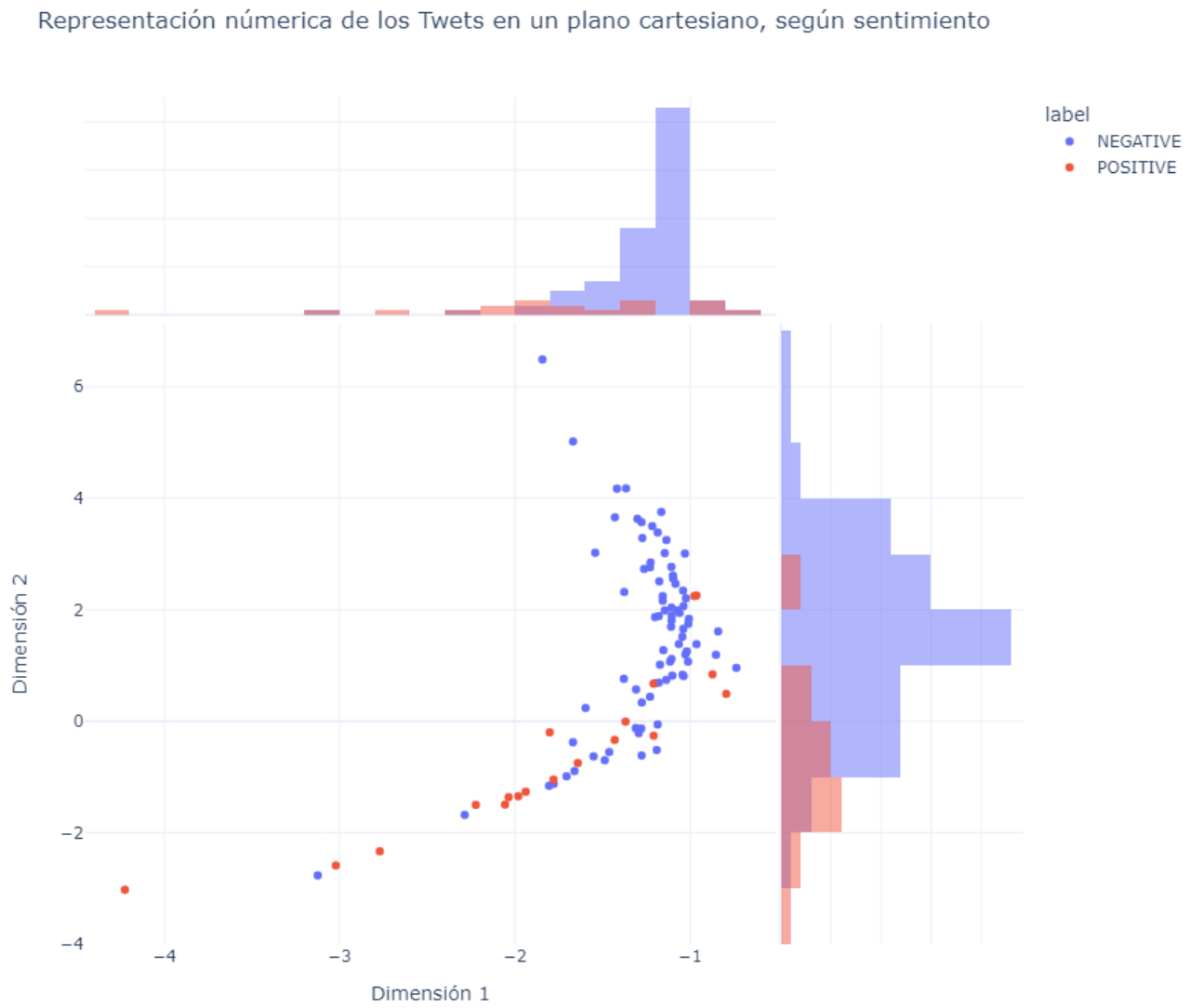
Resultados de la modelación



Los Tweets también se pueden representar en el plano promediando las coordenadas de las palabras al interior del Tweet.

En la imagen es muy claro cómo **el patrón de temáticas de las cuentas institucionales de Davivienda en marcadamente distinto al patrón de las temáticas del resto de los usuarios.**

Resultados de la modelación



Mediante la librería Transformers de Hugging Face se puede identificar el **sentimiento positivo o negativo de los Tweets.**

Sin embargo, es costoso en tiempo. Por lo cual sólo se identificó el sentimiento para los 100 primeros Tweet.

En el plano tenemos una primera idea de en dónde se ubican temáticas positivas y negativas.

Trabajo futuro

- Tener sentimiento de todos los Tweets

  Se requieren cerca de 3-4 horas tener el sentimiento de los poco más de 1.800 Tweets

- Identificar temáticas en el plano

  Dado el patrón de ubicación detectado en las palabras usar métodos de agrupamiento para identificar temáticas. Un método ideal en este contexto puede ser DBSCAN

- Identificar entidades

  Con la librería Transformers de Hugging Face se pueden identificar entidades, con ello se establecer la relación entre una entidad y los sentimientos

- Identificar patrones de acuerdo a otras variables

  Identificar si las temáticas cambian a través del tiempo o interacciones en Twitter, el número de likes, por ejemplo

- Tablero de consulta

  Diseñar una tablero que permitan identificar temáticas bajo ciertas condiciones, tipo de Tweet hora del Tweet, entre otros

Procesamiento de Lenguaje Natural usando Tweets de Davivienda

Prueba técnica para la selección de Científico(a) de Datos

Nikolás Sebastián Buitrago Vargas
Economista y magister en estadística

Lunes 20 de marzo de 2023