

Resultados Proceso Text Mining

A planes de gobierno candidatos

Presidencia Colombia 2022

Análisis estadístico descriptivo a documentos planes de gobierno de:
Gustavo Petro, Sergio Fajardo, Rodolfo Hernandez y Federico Gutierrez.

Frederick Salazar Sanchez

@FrederickSalazar

Data Engineer - Data Scientist

Msc BigData Analytics (En curso)

Universitat Oberta de Catalunya

Versión 1.0

Fecha 29 Abril de 2022

Declaración Previa

El desarrollo de este proyecto es con fines totalmente educativos, no tiene como objetivo principal definir cuál es el mejor o peor plan de gobierno, ni emitir juicios de valor sobre el contenido de los mismos. El desarrollo de este proyecto tiene como objetivo principal: realizar un proceso de Text Mining sobre los planes de gobierno de diferentes candidatos a las elecciones presidenciales de Colombia en el año 2022, a fin de presentar información estadística desconocida sobre los documentos y que sirva para comprender mejor a título personal e interesados el campo de estudio del Text Mining y Natural Language Processing. Cualquier uso adicional, indiscriminado o fuera del marco de este documento a la información expuesta, queda bajo la responsabilidad del lector.

Muchas Gracias.

Recursos Usados

Recursos usados para el proyecto:

Plan gobierno Gustavo Petro: <https://gustavopetro.co/descarga-programa-de-gobierno/>

Plan gobierno Sergio Fajardo: <https://sergiofajardo.co/propuestas/>

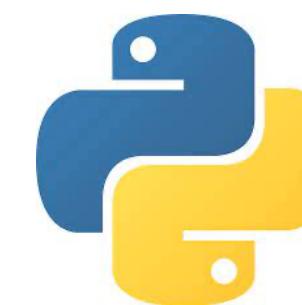
Plan gobierno Rodolfo Hernandez: <https://www.ingrodolfohernandez.com>

Plan Federico Gutiérrez: <https://federicogutierrez.com/programa-de-gobierno/>

Desarrollo y publicación del Proyecto

Este proyecto ha sido desarrollado usando herramientas Open Source, su publicación es bajo la licencia GPL3.

- Python 3.9.12
- Pdfminer
- Nltk
- Matplot lib
- Spacy
-



matplotlib

spaCy

El proyecto, recursos y el código fuente se encuentran disponibles en el siguiente repositorio Github, siéntase libre de aportar al proyecto bajo los términos de la licencia GPL y reglas open source



<https://github.com/fredericksalazar/textAnalysis-ColombianElections>

Licencias del proyecto

Este documento es desarrollado y publicado haciendo uso de la licencia Creative Commons CC BY-SA para mas información acerca del alcance de la licencia visite la siguiente url

<https://creativecommons.org/licenses/by-sa/4.0/deed.es>



Reconocimiento-CompartirIgual
CC BY-SA

EL código fuente desarrollado en el proyecto ha sido licenciado bajo la licencia copy left GPL V3:

CopyRight @ Frederick Salazar 2022

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

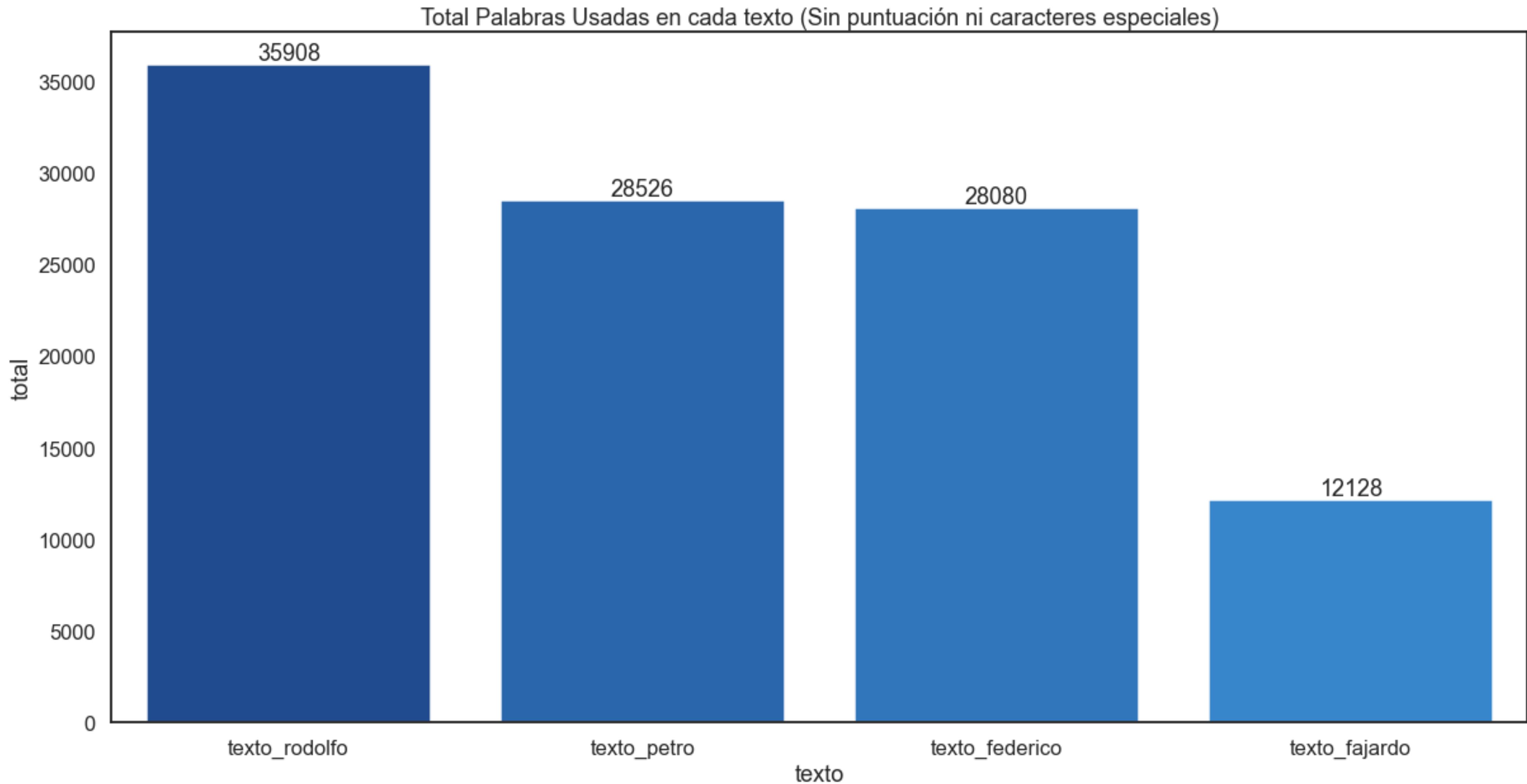
This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <<https://www.gnu.org/licenses/>>.

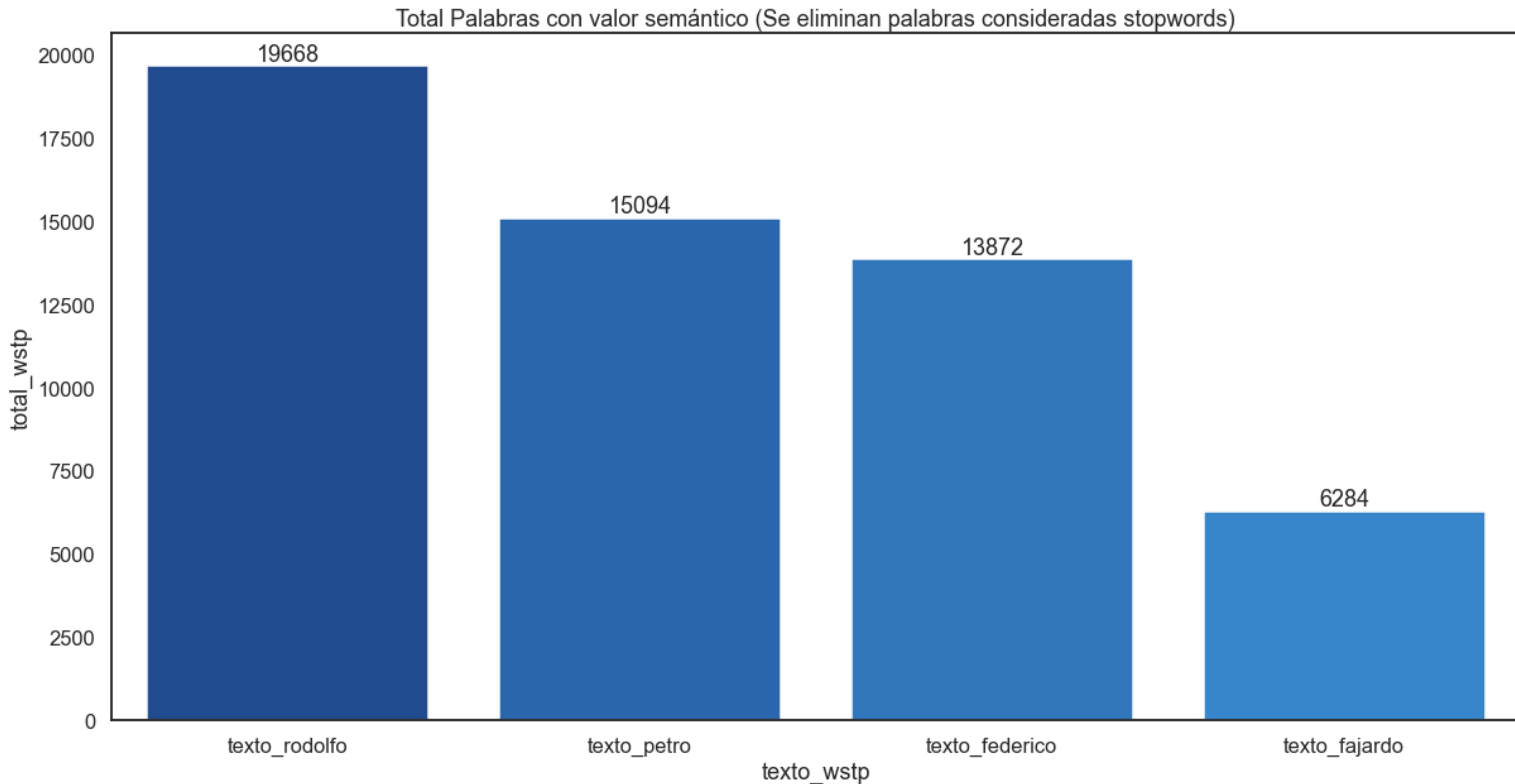
<https://www.gnu.org/licenses/licenses.es.html>



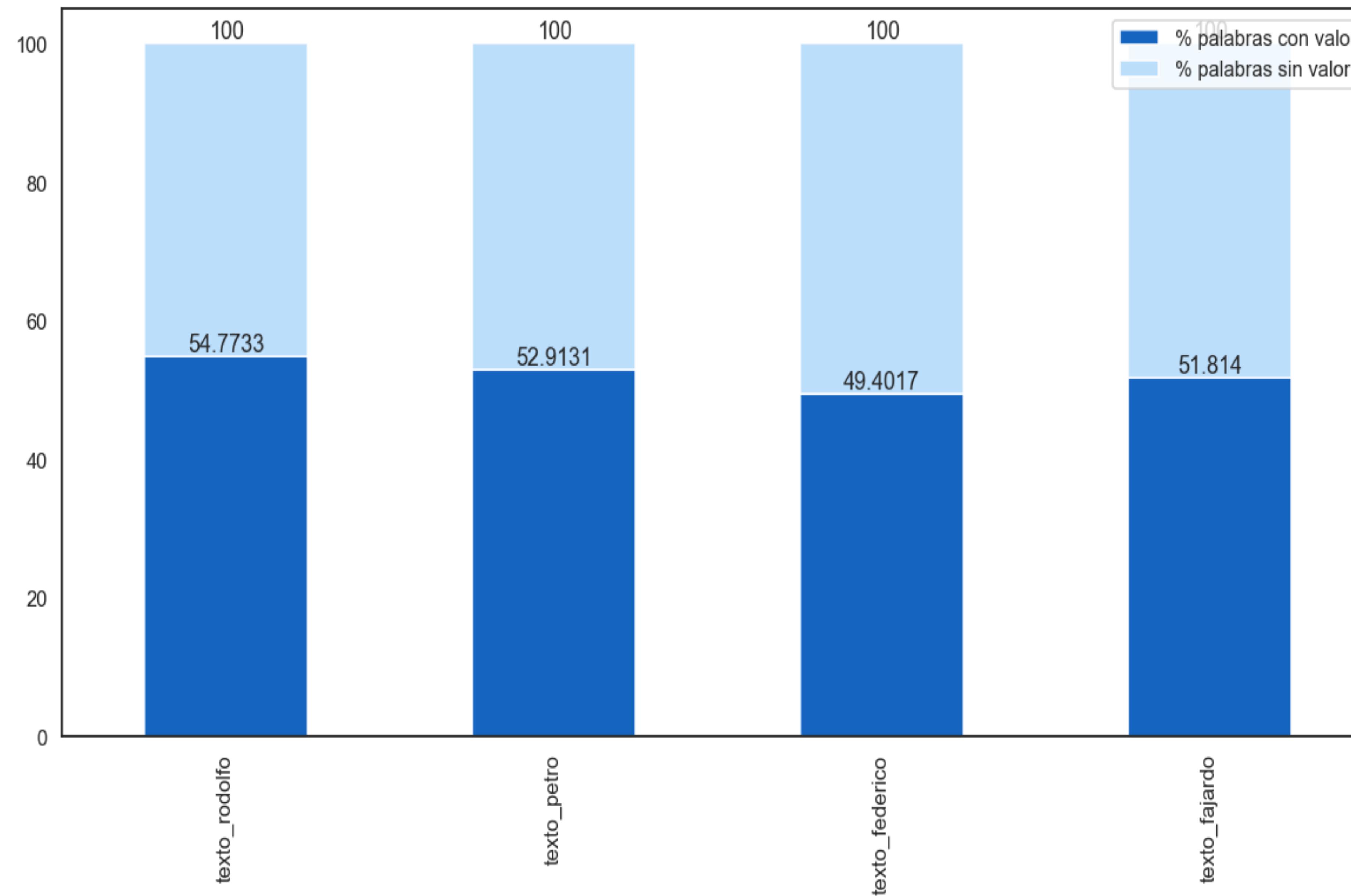
Total Palabras por Documento



Total Palabras con valor semántico Por Documento



% Composición texto con y sin valor semántico



Del total de palabras usadas en cada documento se eliminan palabras que no aportan valor semántico denominadas Stop Words, a partir de la diferencia con el total general se calcula el porcentaje de uso de palabras con y sin valor semántico

WordCloud Palabras Documento Gustavo Petro



Se toman las 50 palabras con mayor frecuencia de uso dentro del documento a fin de identificar su orden descendente, en el caso de la nube de palabras entre más grande sea la palabra más frecuencia de uso dentro del documento corresponde.



WordCloud Palabras Documento Sergio Fajardo

Se toman las 50 palabras con mayor frecuencia de uso dentro del documento a fin de identificar su orden descendente, en el caso de la nube de palabras entre más grande sea la palabra más frecuencia de uso dentro del documento corresponde.

WordCloud Palabras Documento Rodolfo Hernandez



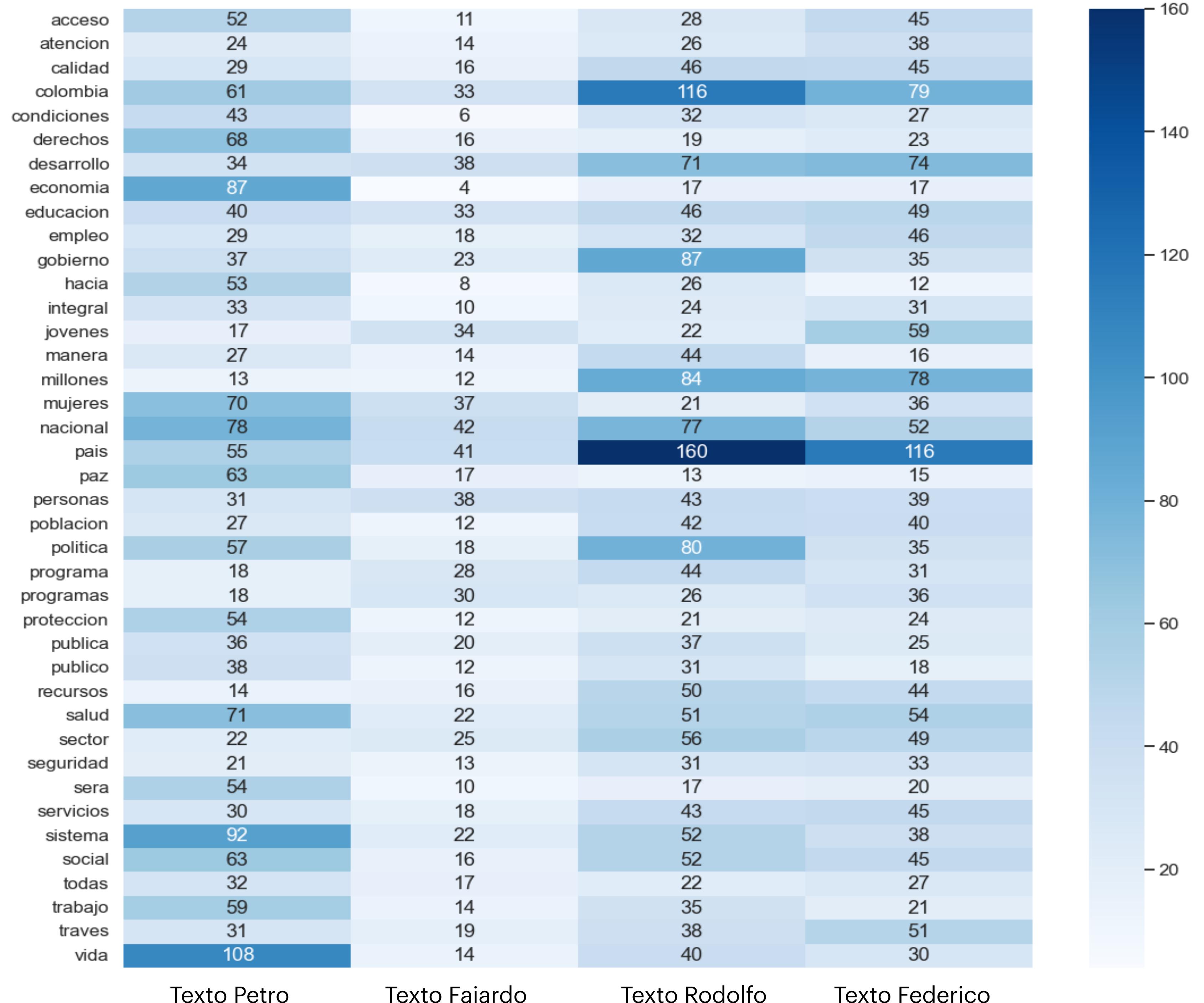
Se toman las 50 palabras con mayor frecuencia de uso dentro del documento a fin de identificar su orden descendente, en el caso de la nube de palabras entre más grande sea la palabra más frecuencia de uso dentro del documento corresponde.



WordCloud Palabras
Documento Federico

Se toman las 50 palabras con mayor frecuencia de uso dentro del documento a fin de identificar su orden descendente, en el caso de la nube de palabras entre más grande sea la palabra más frecuencia de uso dentro del documento corresponde.

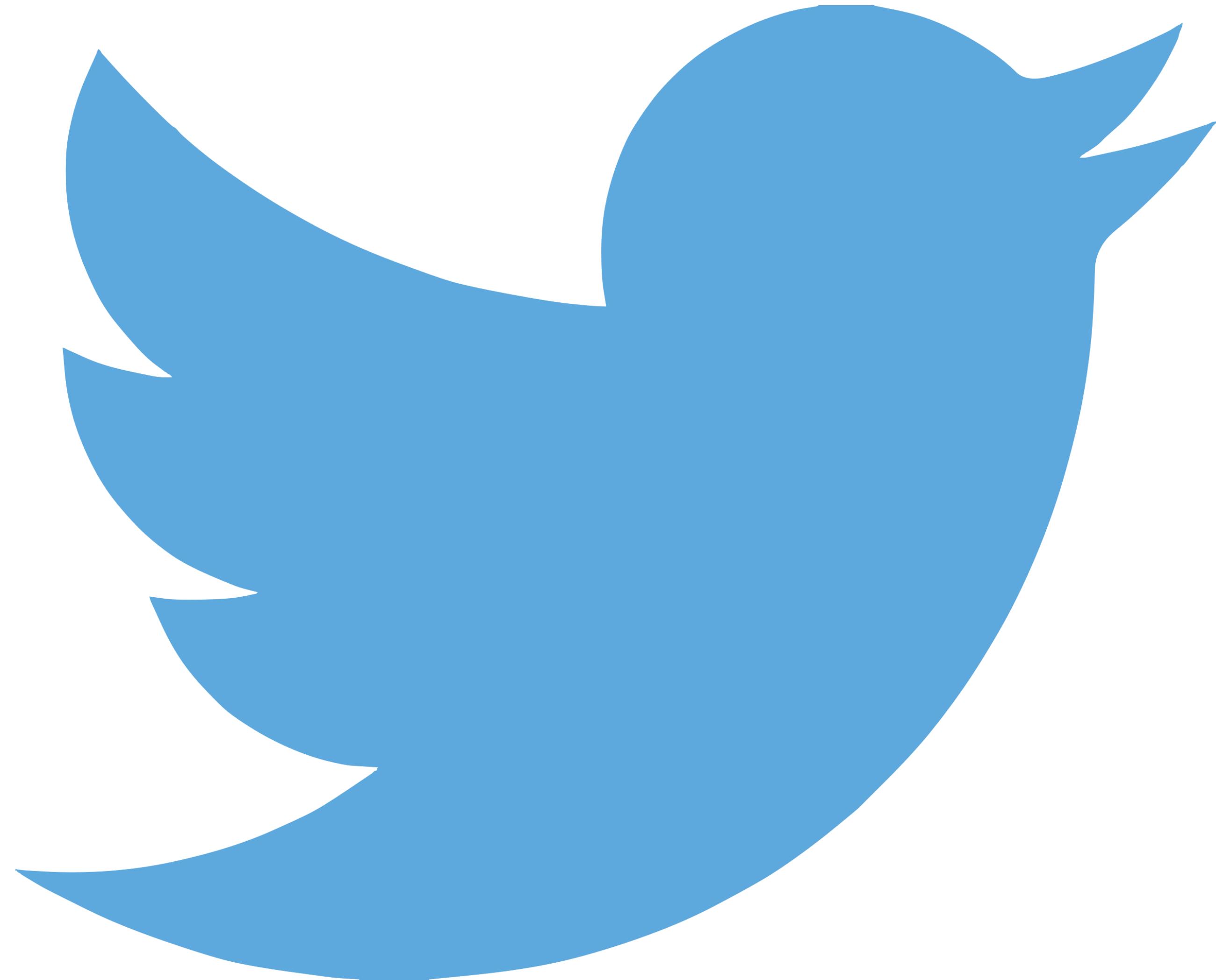
HeatMap palabras comunes más usadas en todos los documentos



A partir del corpus general de todas las palabras comunes más usadas, se construye este mapa de calor que identifica y compara el uso de palabras comunes en cada texto, permite visualizar de acuerdo a cada palabra cuantas veces es usada por cada documento.

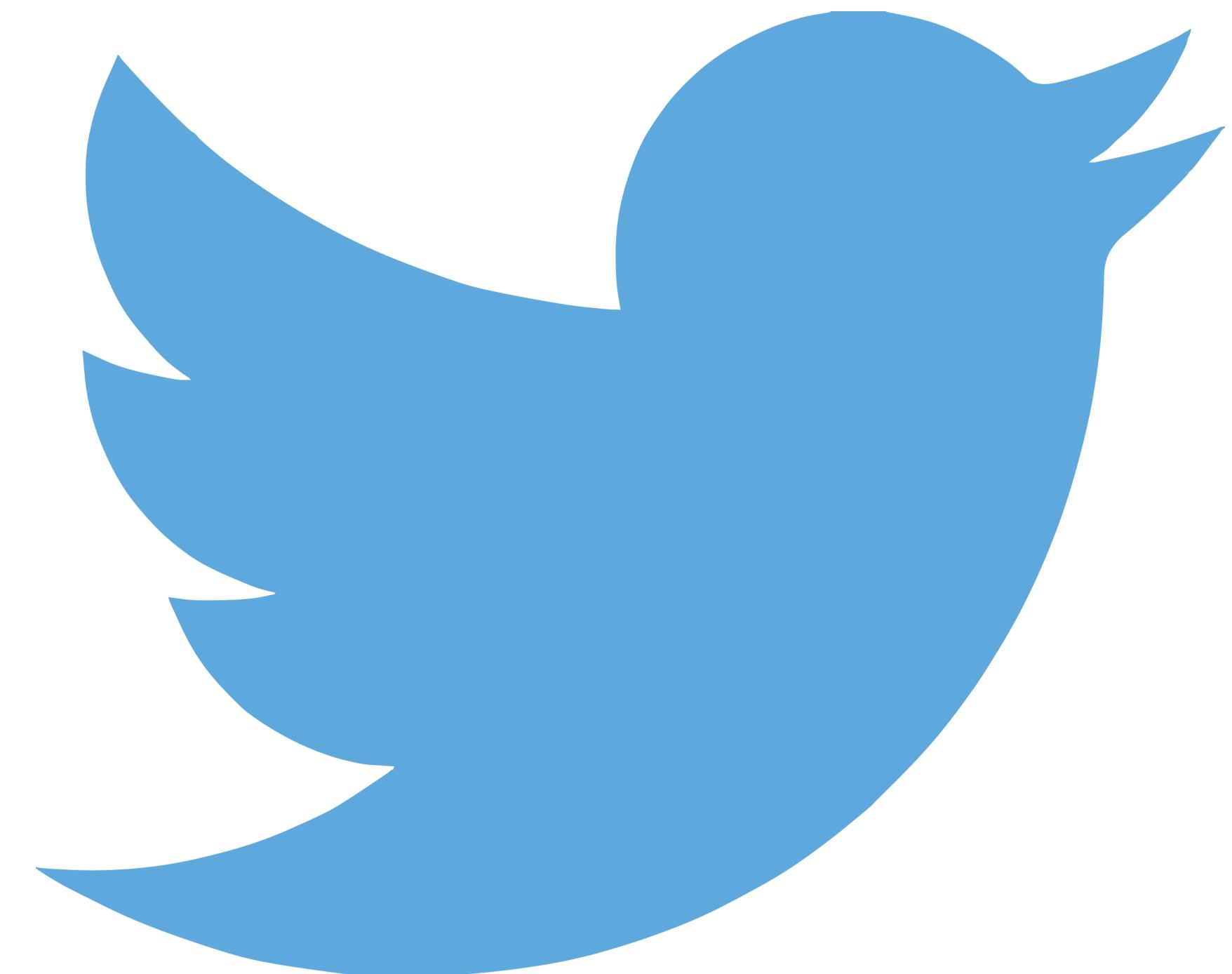
Fase 2 - Análisis perfiles Twitter candidatos

Se generarán estadísticas de Uso de la plataforma twitter, se extraerán las palabras mas usadas y se predice el sentimiento de los tweets



Análisis sobre perfiles twitter candidatos Presidenciales Colombia 2022

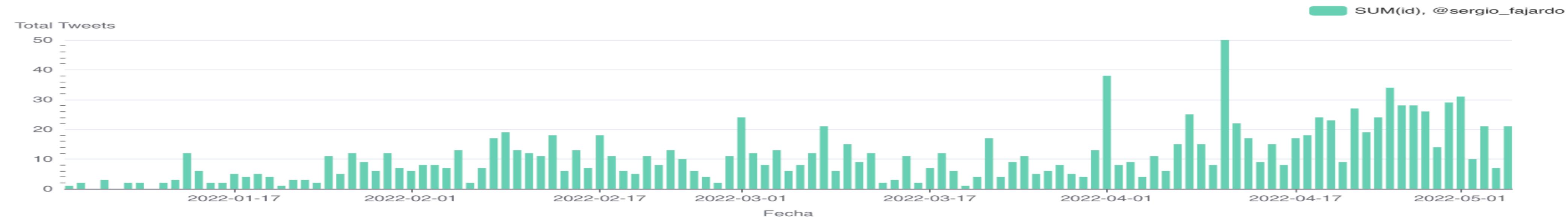
Se desarrolla un proceso de minería de textos sobre los tweets escritos por los candidatos presidenciales desde el 01 de enero de 2022, se analizarán aspectos como frecuencia de publicación de tweets, palabras mas usadas y un análisis de sentimiento sobre los textos de los mismos.



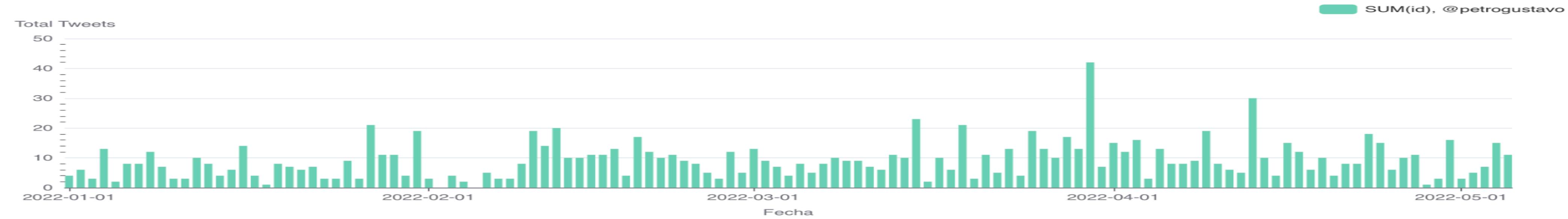
Tweets Diarios Por Candidato



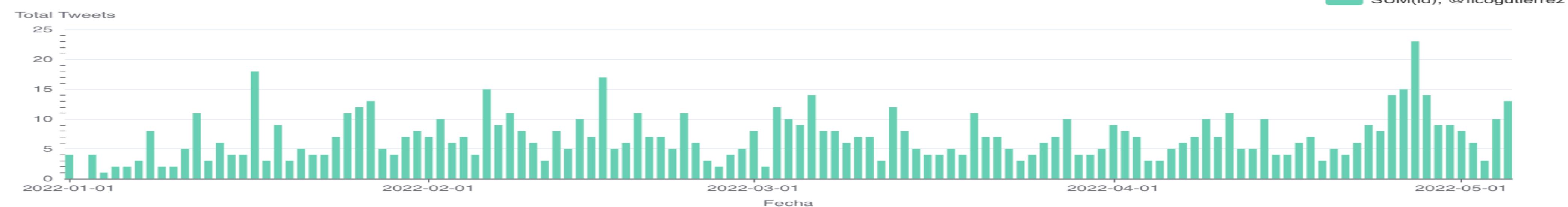
1344 tweets



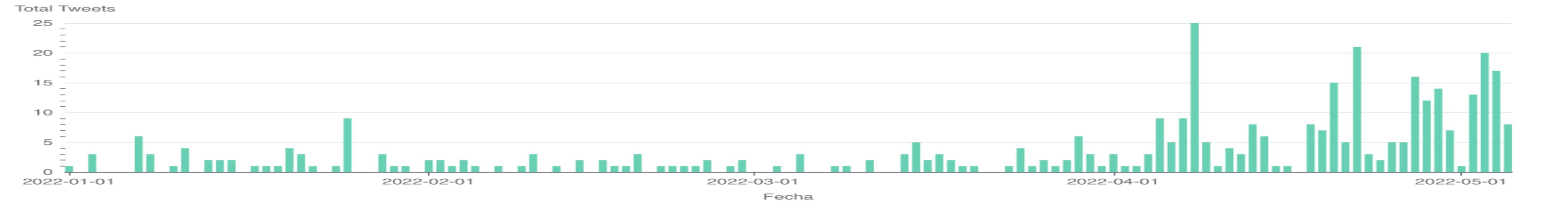
1145 tweets



864 tweets



392 tweets

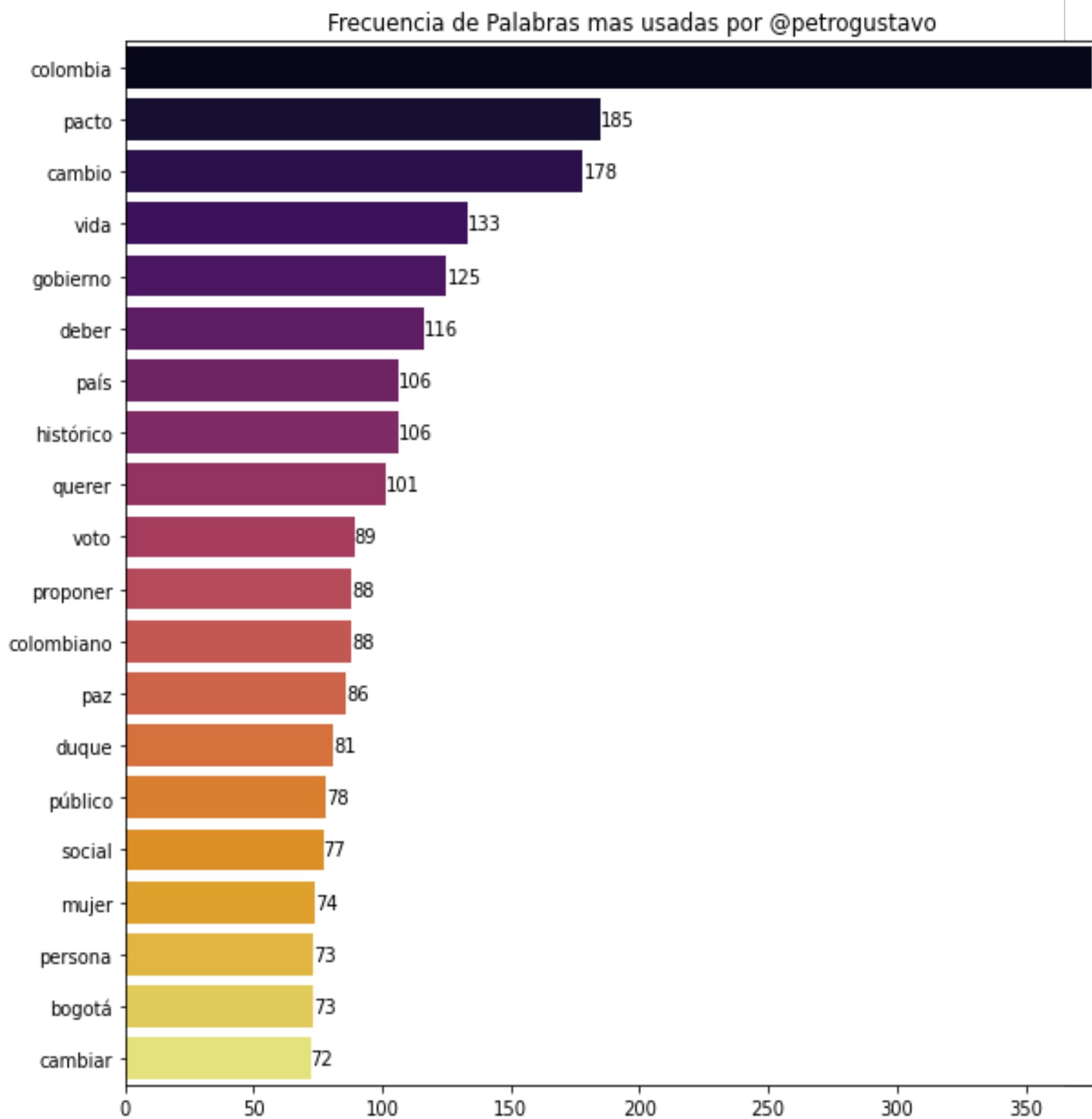




@petrogustavo

4.8 Millones de seguidores

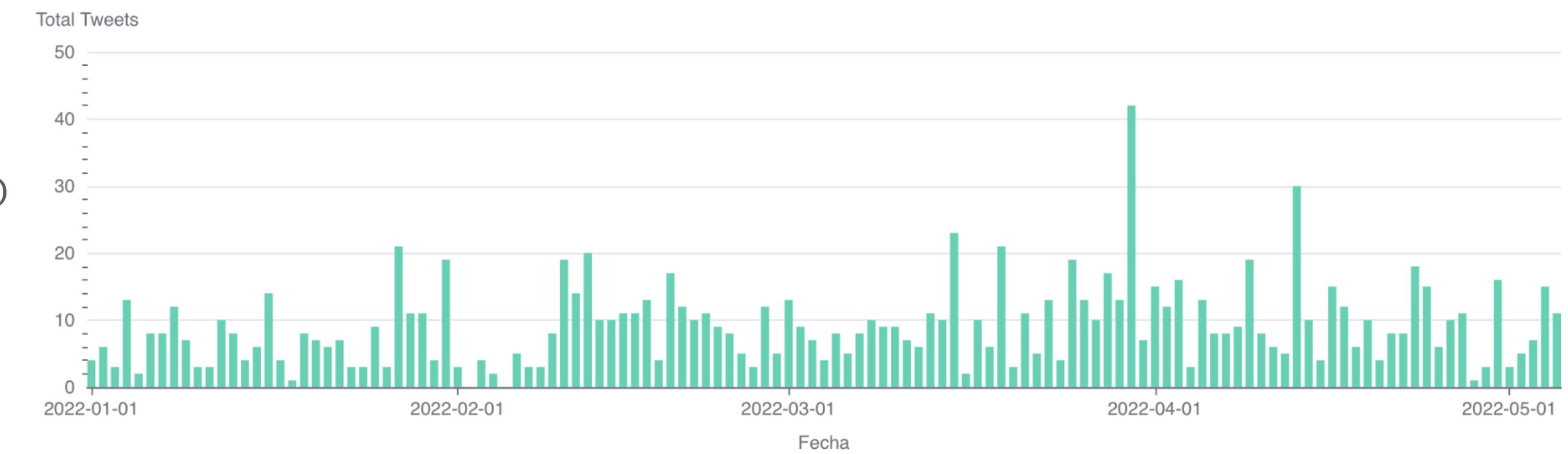
1.145 tweets escritos desde el 01 de enero de 2022 con un promedio de **9.31** tweets por día aprox.





4.8 Millones de seguidores

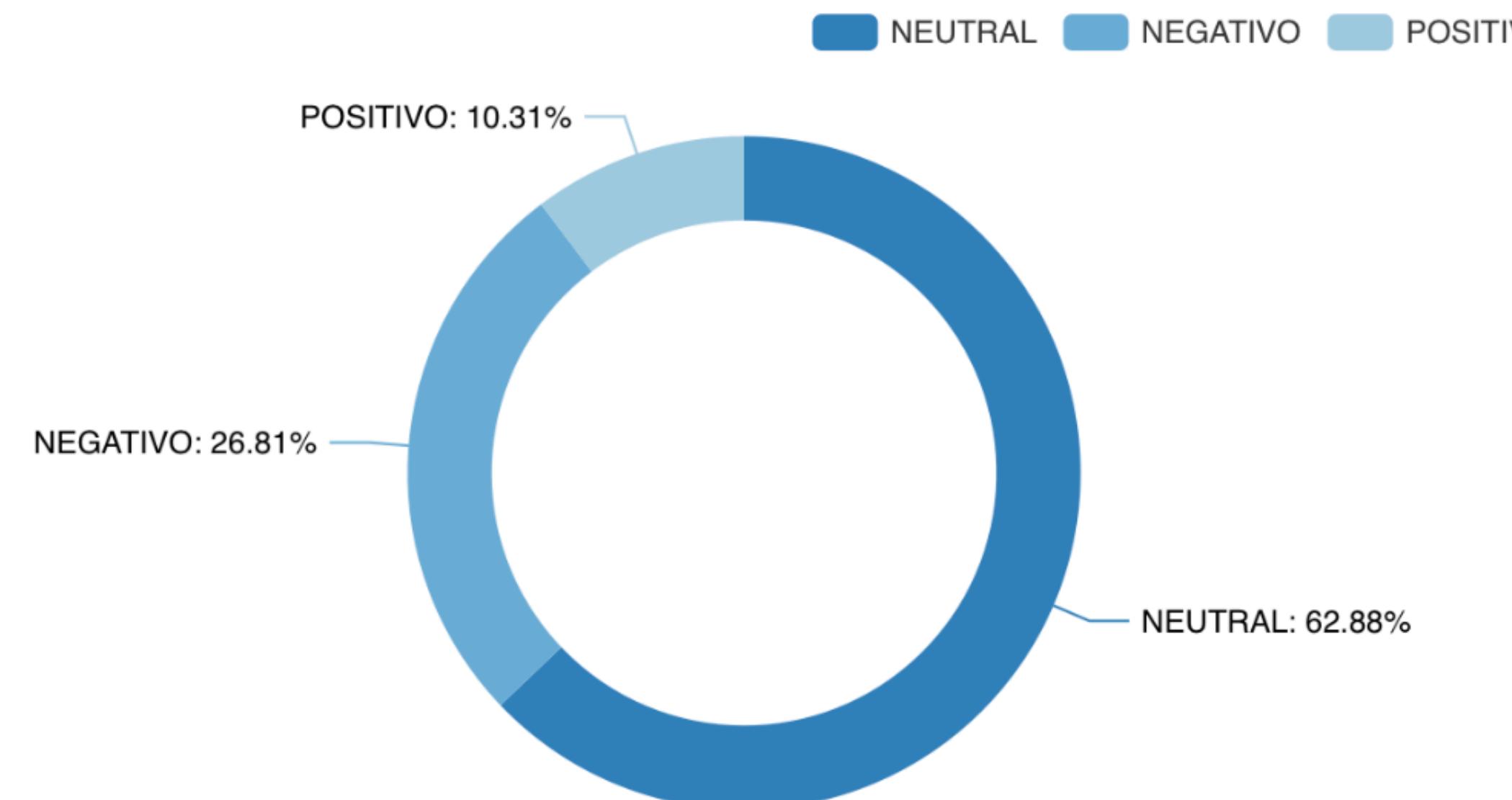
1.145 tweets escritos desde el 01 de enero de 2022 con un promedio de **9.31** tweets por día aprox.



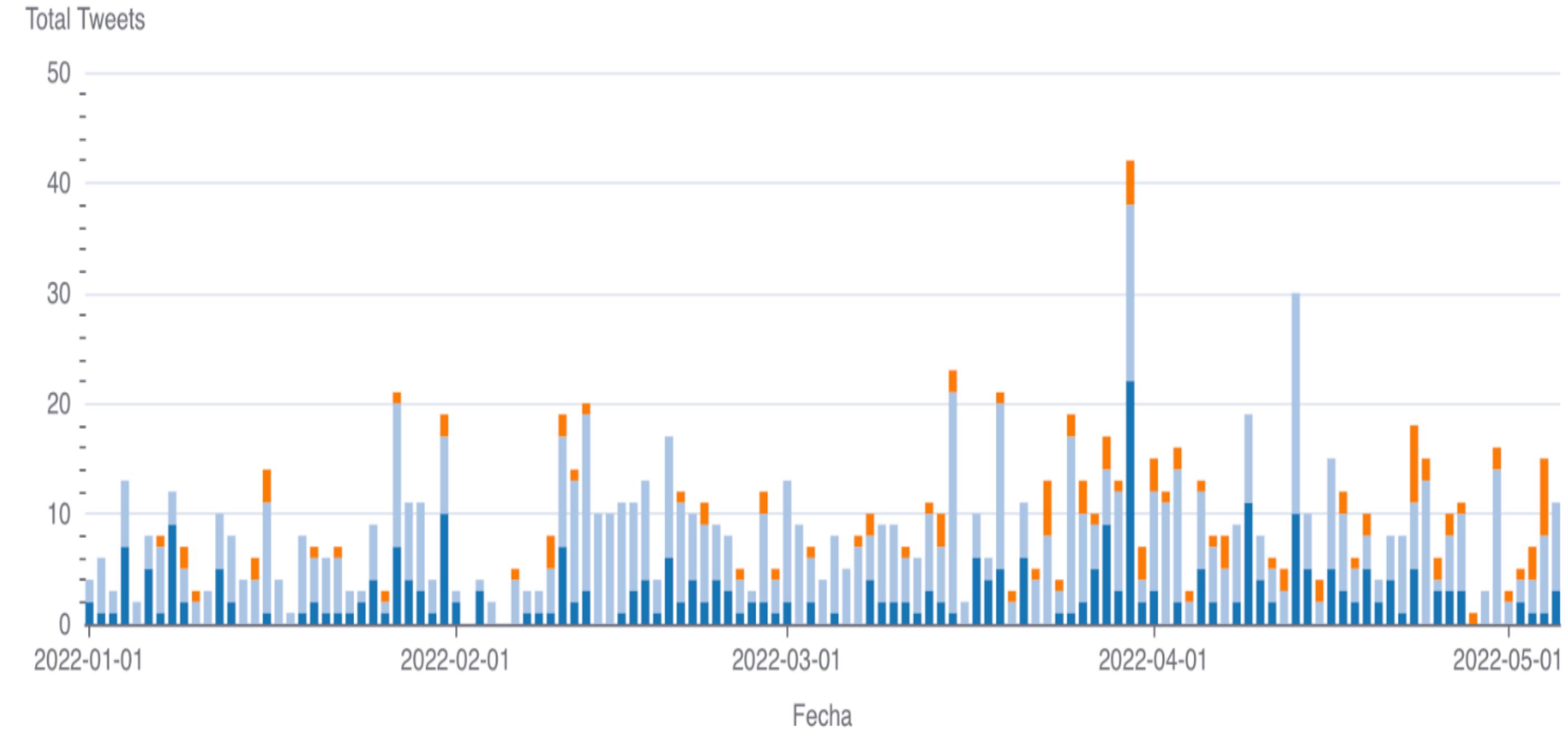
@petrogustavo

El **73.19%** de los tweets escritos por el candidato son neutrales o positivos, el **26.81%** son tweets cuya composición gramatical expresan sentimientos tristes o negativos

count, NEGATIVO count, NEUTRAL count, POSITIVO



Distribución Tweets por tipo de sentimiento



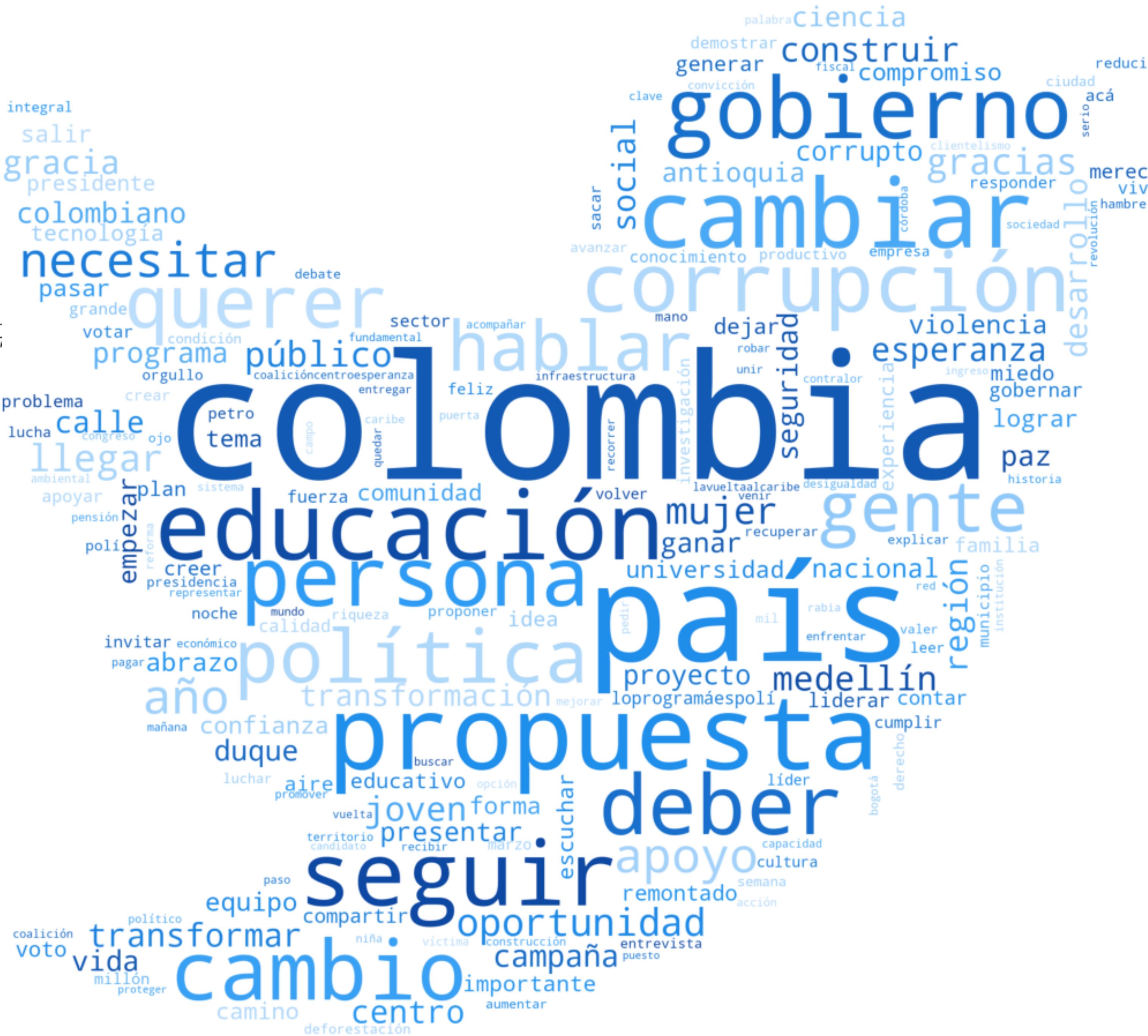
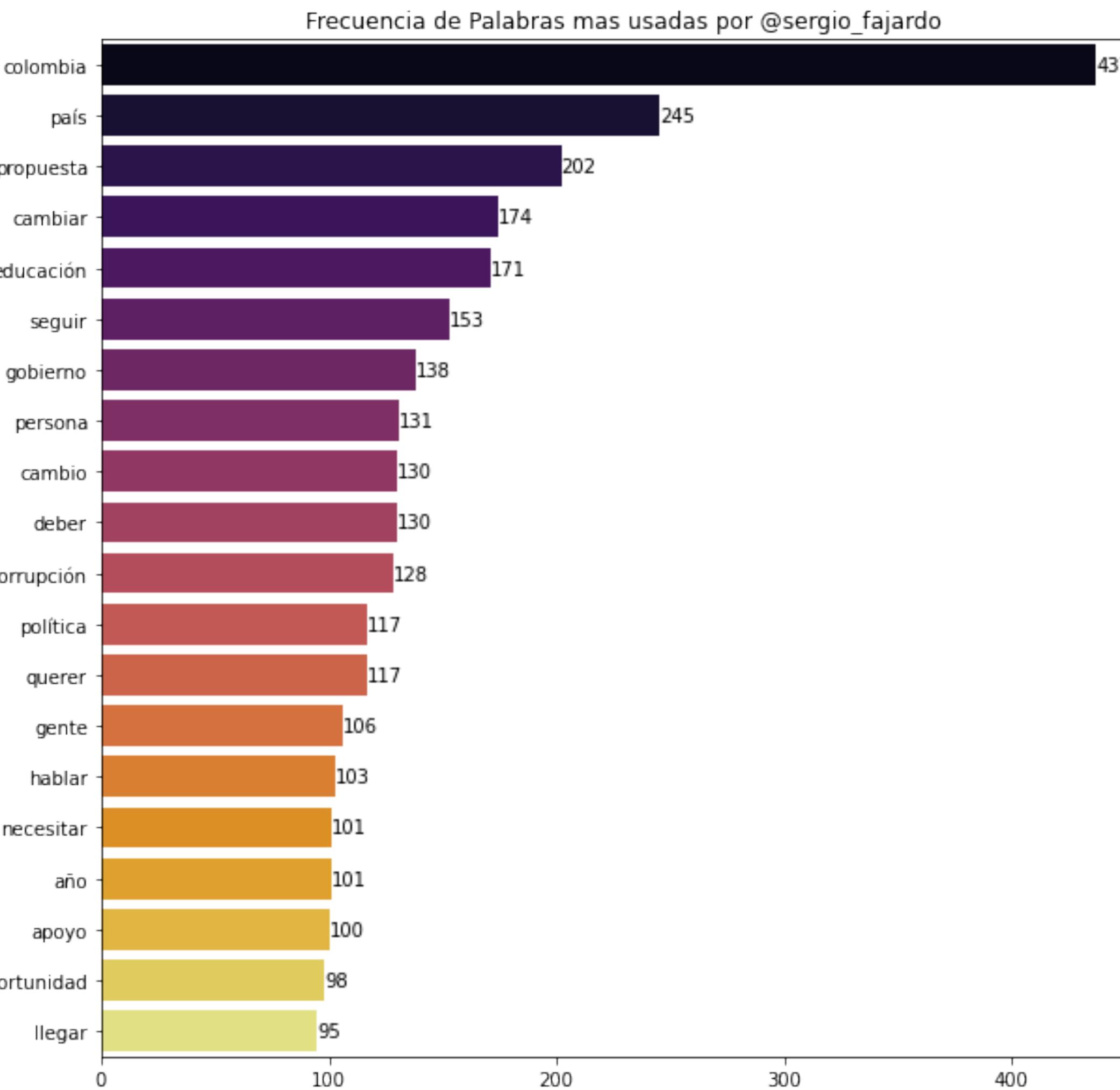
Cantidad de Tweets escritos por día por tipo de sentimiento



@sergio_fajardo

1.6 Millones de seguidores

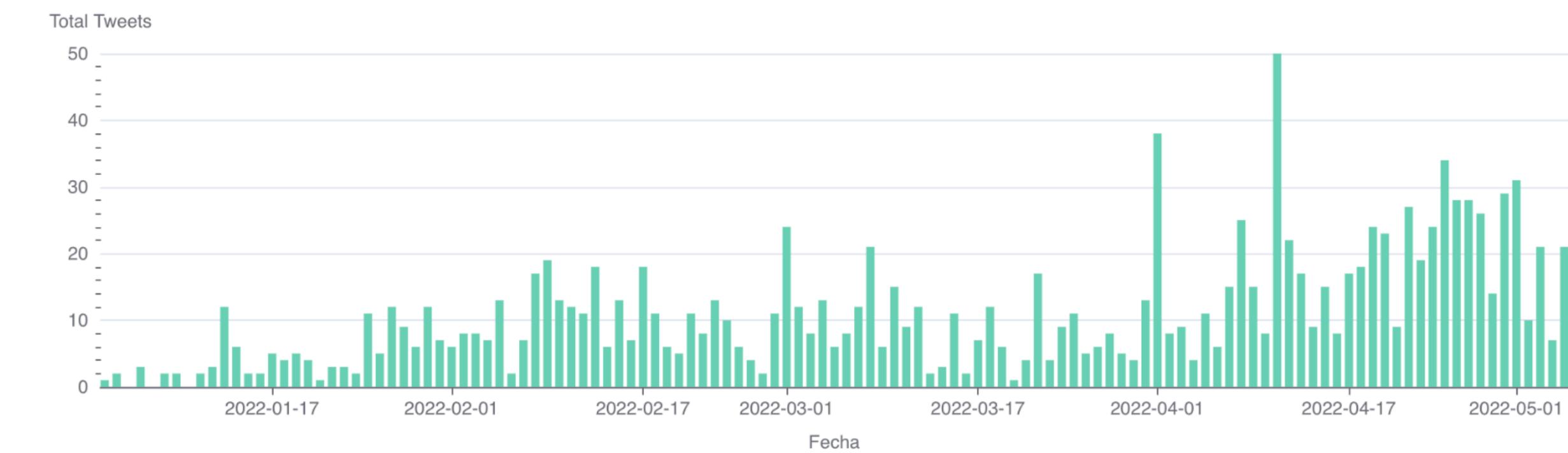
1.344 tweets escritos desde el 01 de enero de 2022 con un promedio de 11.2 tweets por día aprox.





1.6 Millones de seguidores

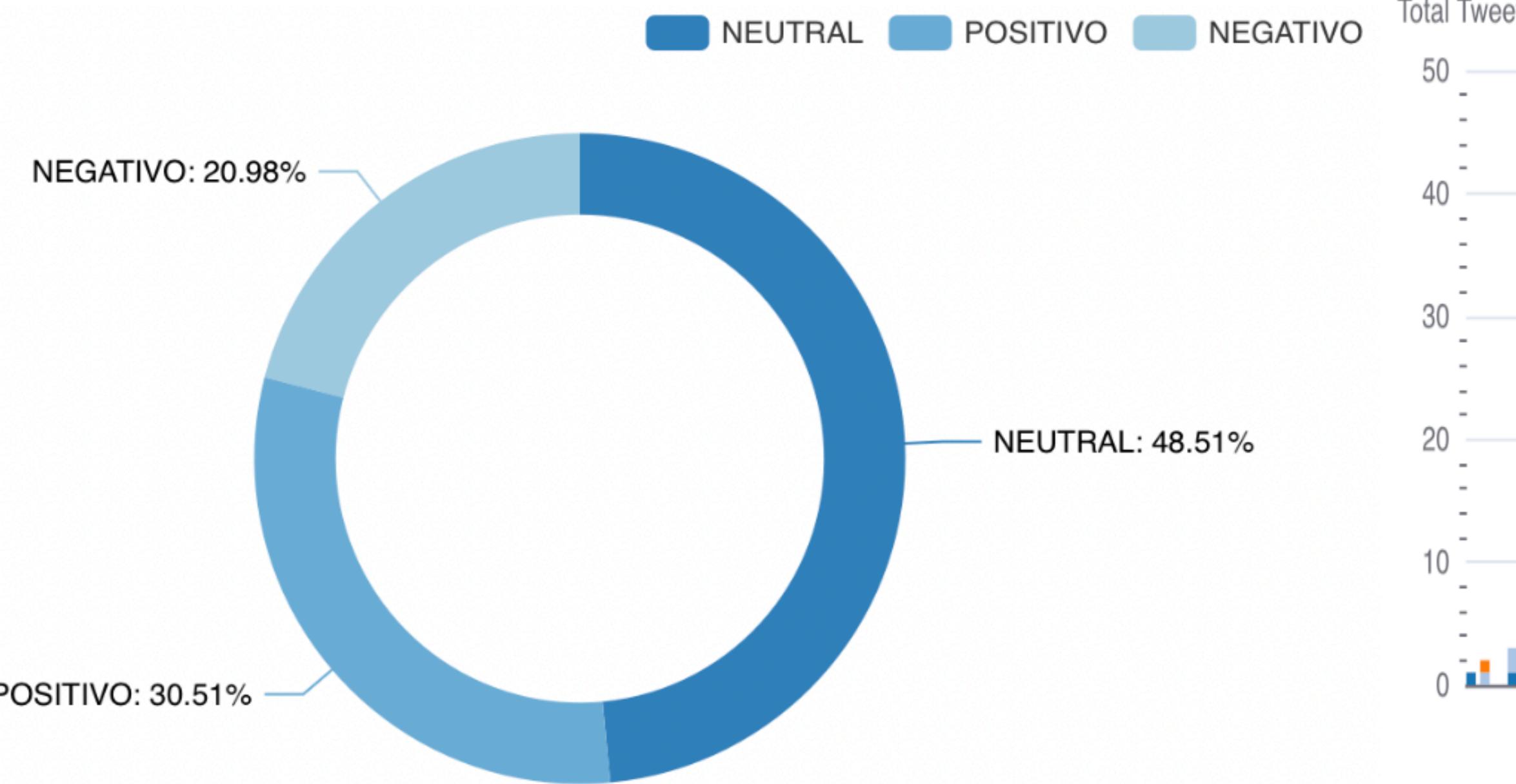
1.344 tweets escritos desde el 01 de enero de 2022 con un promedio de **11.2** tweets por día aprox.



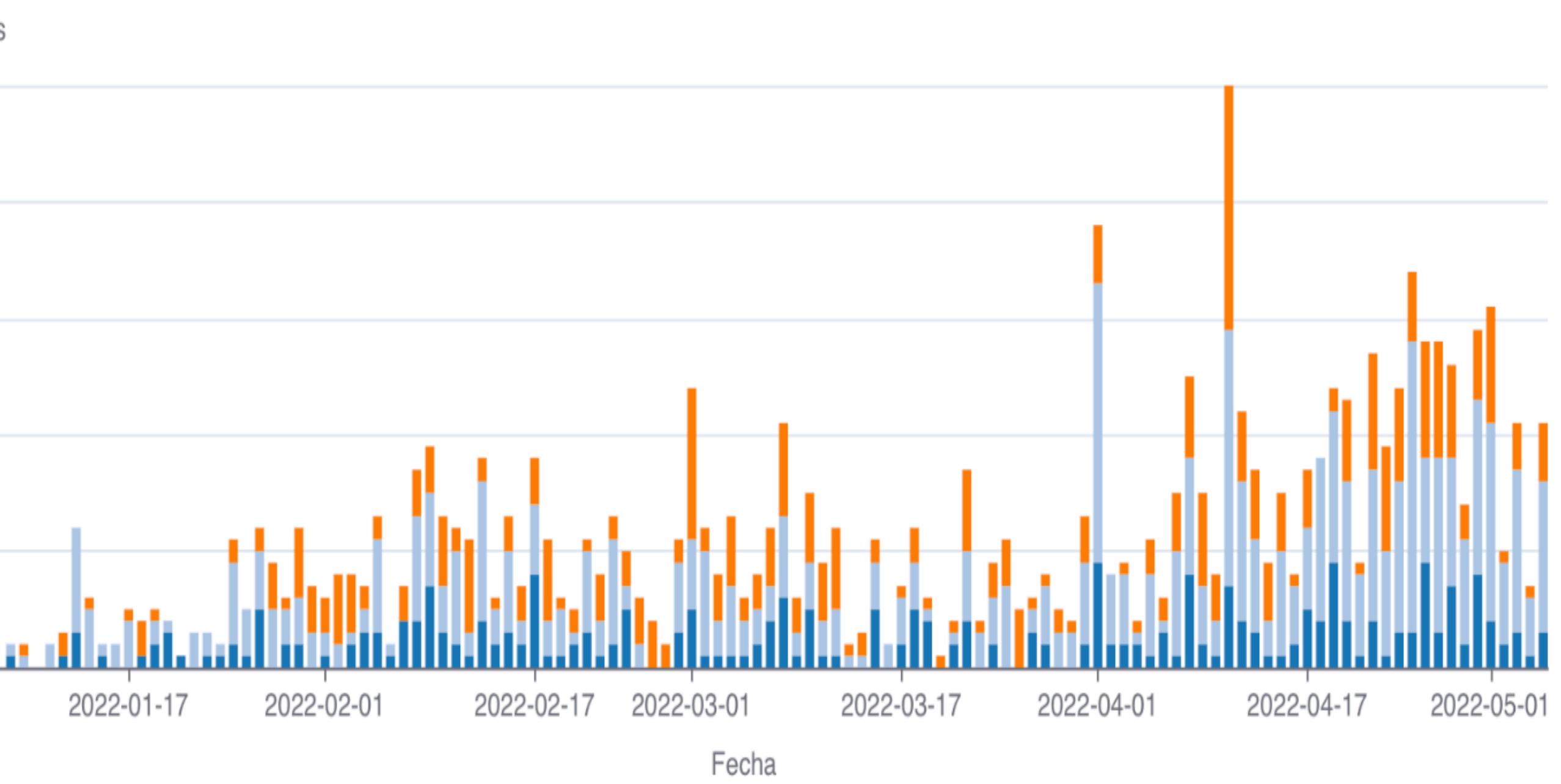
@sergio_fajardo

Cantidad de tweets escritos por día

El **79.02%** de los tweets escritos por el candidato son neutrales o positivos, el **20.98%** son tweets cuya composición gramatical expresan sentimientos tristes o negativos



Distribución Tweets por tipo de sentimiento



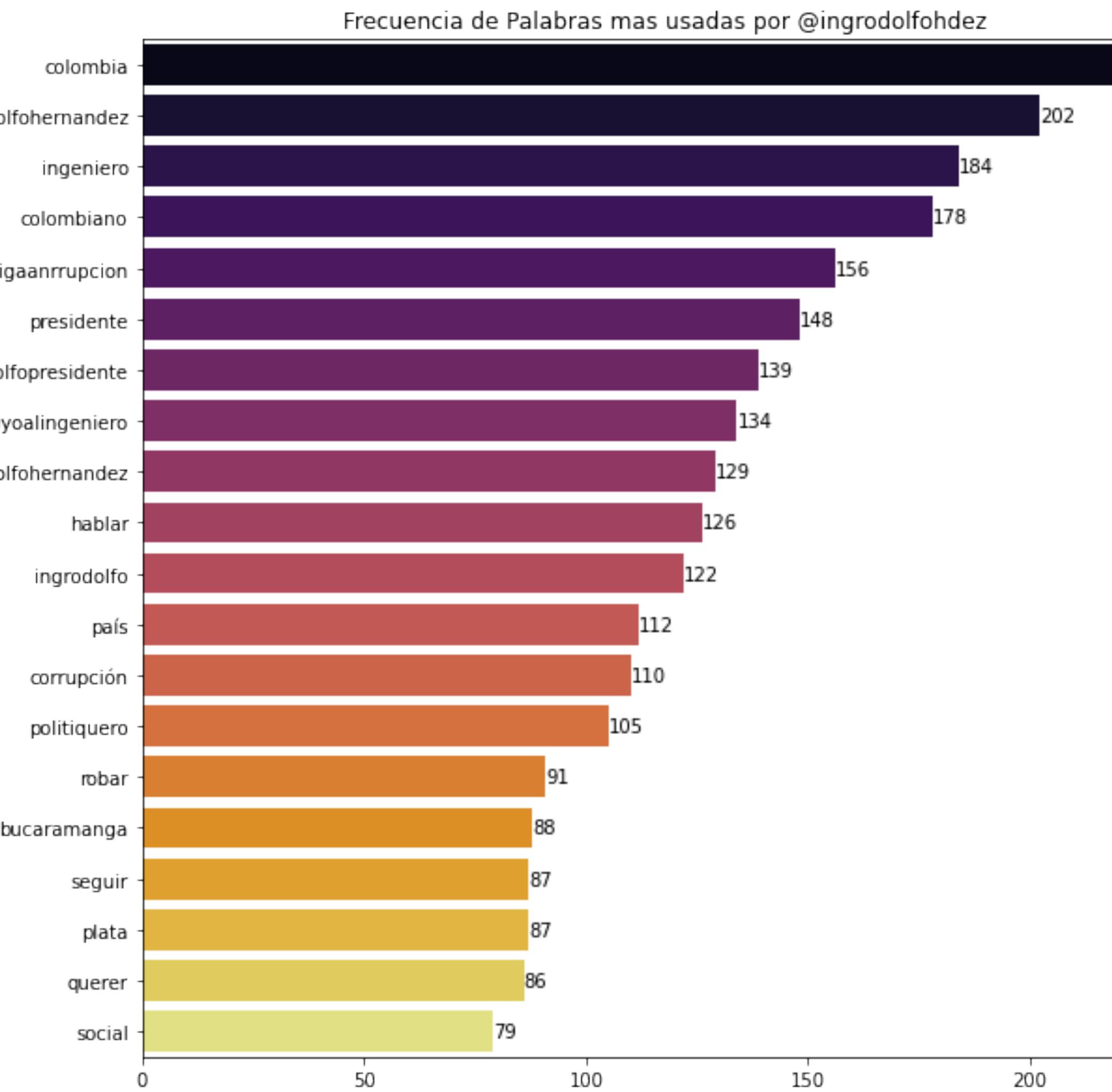
Cantidad de Tweets escritos por día por tipo de sentimiento



@ingrodolfohdez

159 Mil seguidores

392 tweets escritos desde el 01 de enero de 2022 con un promedio de **4.08** tweets por día aprox.

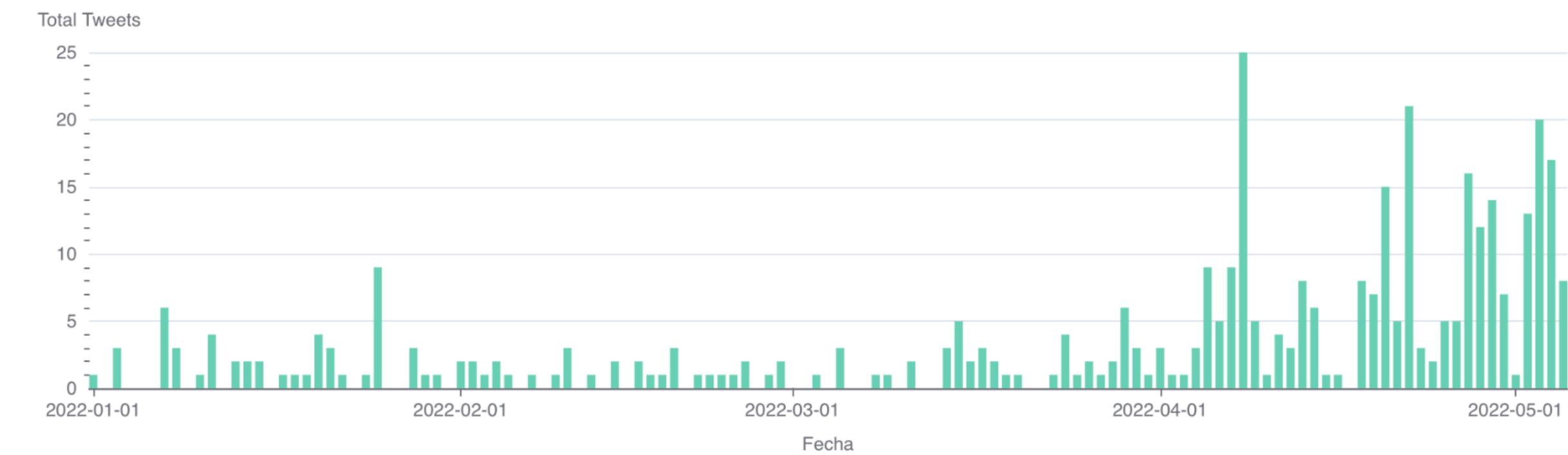




159 Mil seguidores

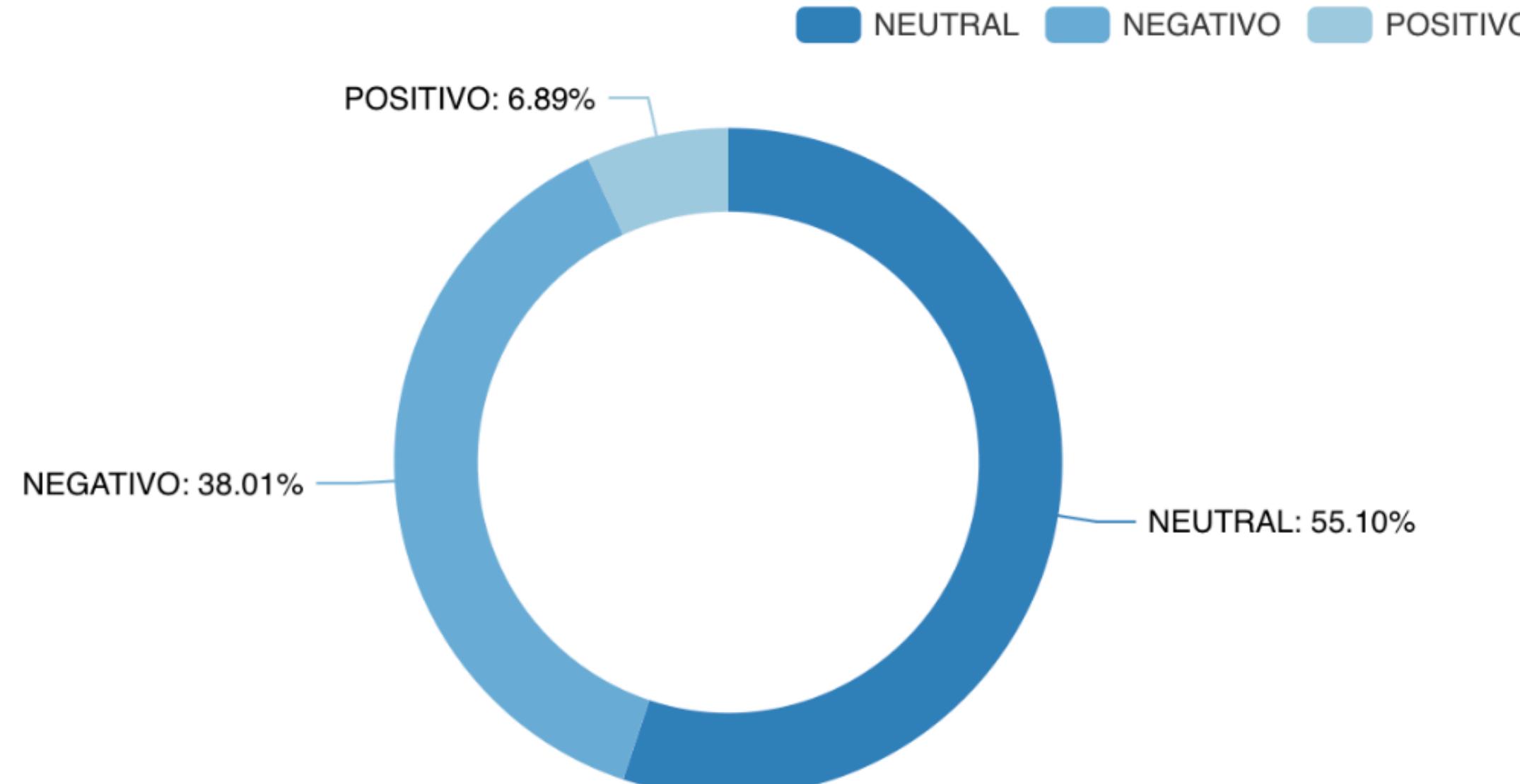
392 tweets escritos desde el 01 de enero de 2022 con un promedio de **4.08** tweets por día aprox.

@ingrodolgohdez

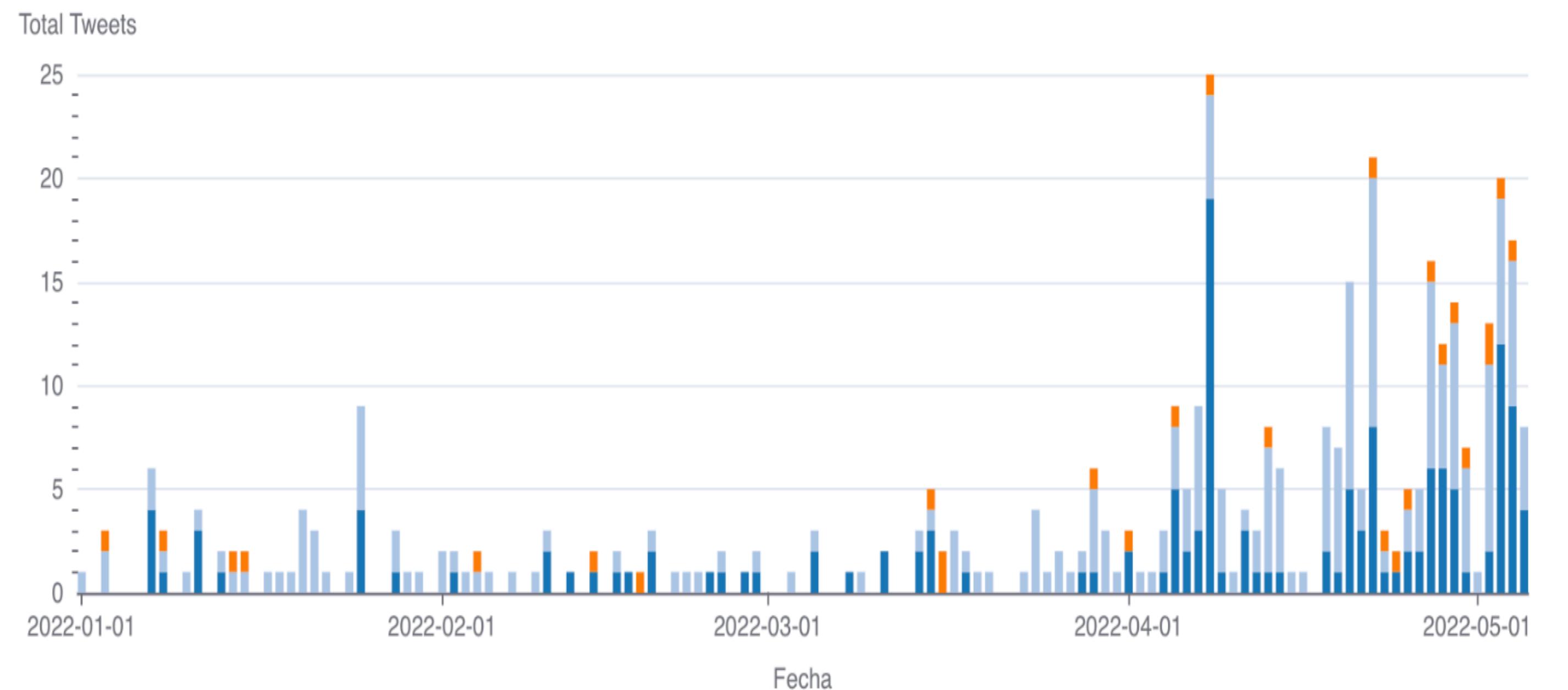


Cantidad de tweets escritos por día

El **61.99%** de los tweets escritos por el candidato son neutrales o positivos, el **38.01%** son tweets cuya composición gramatical expresan sentimientos tristes o negativos



Distribución Tweets por tipo de sentimiento



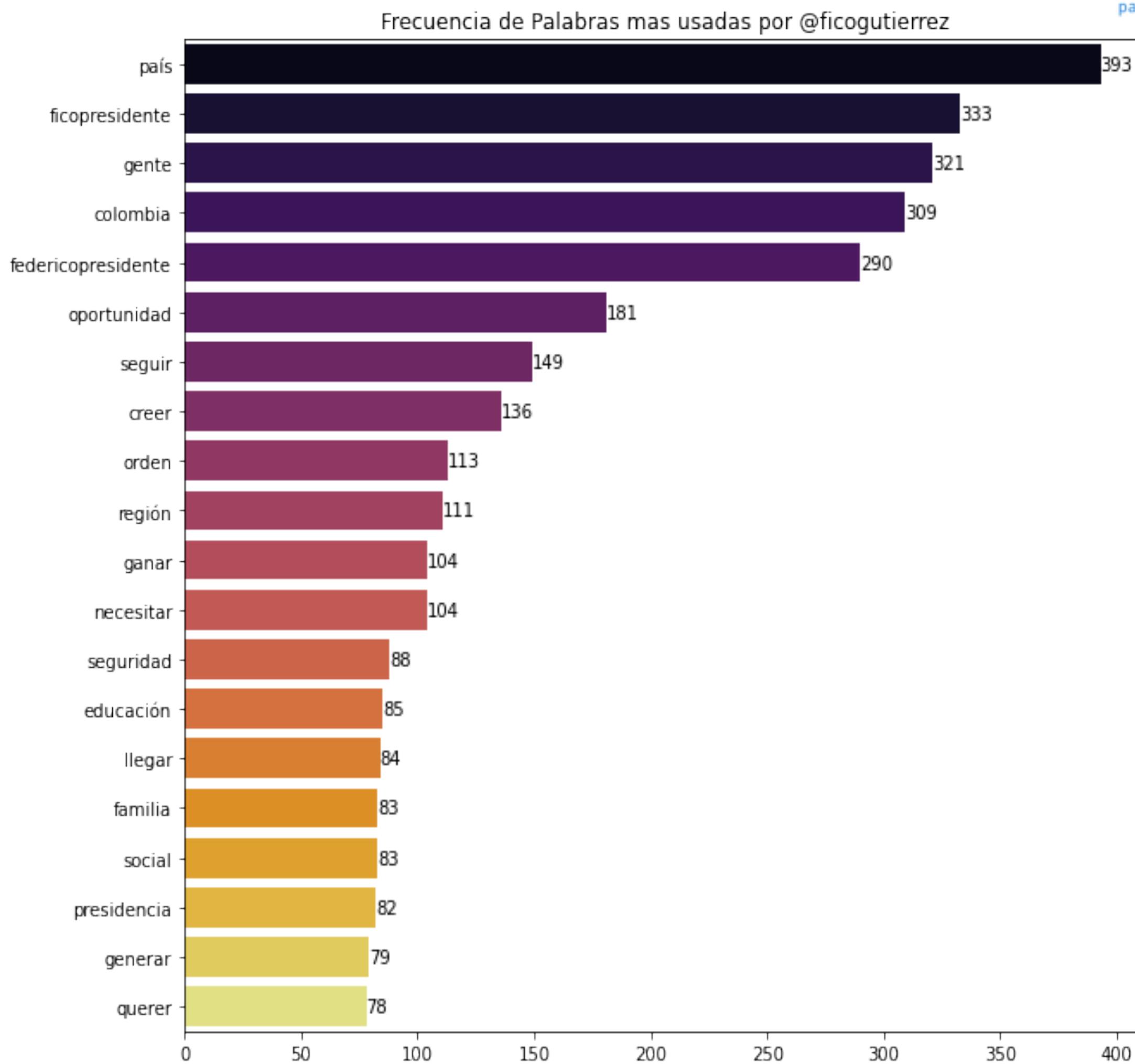
Cantidad de Tweets escritos por día por tipo de sentimiento



@FicoGutierrez

870 Mil seguidores

864 tweets escritos desde el 01 de enero de 2022 con un promedio de **6.97** tweets por día aprox.

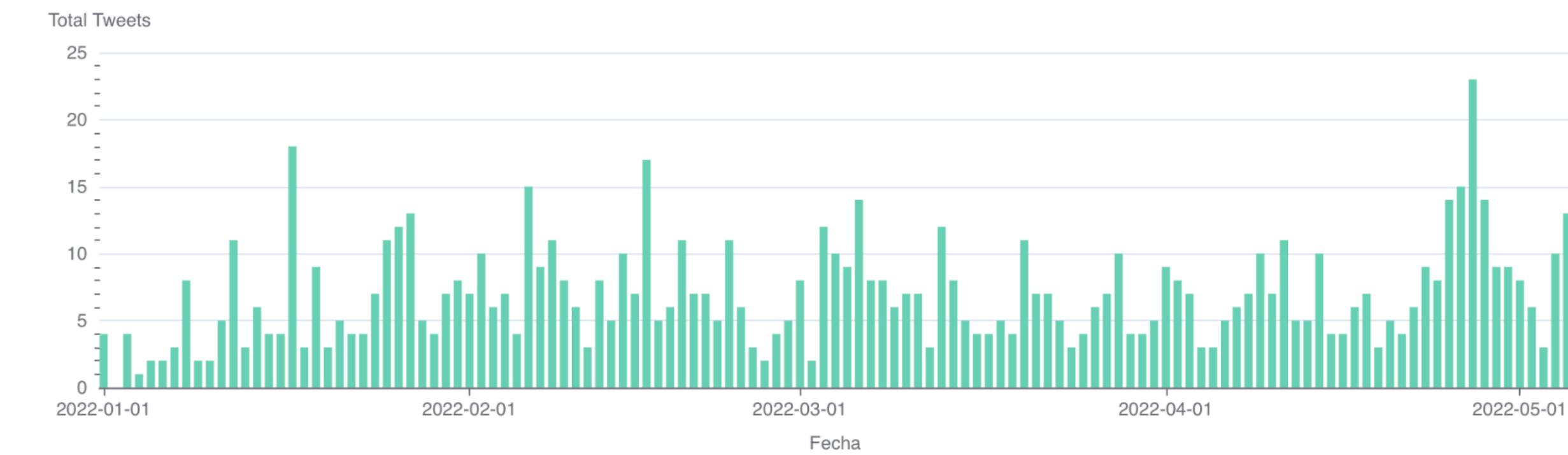




870 Mil seguidores

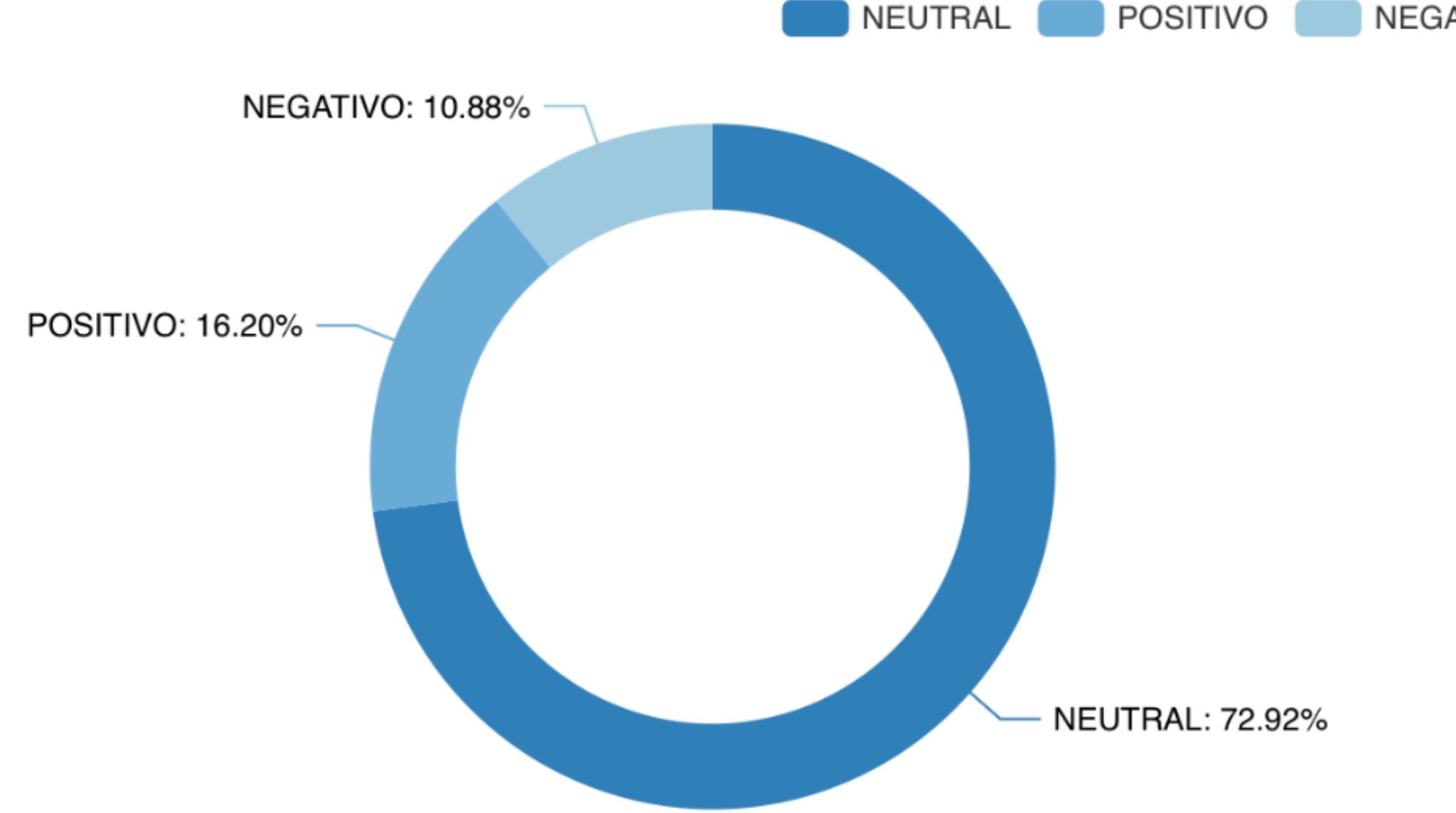
864 tweets escritos desde el 01 de enero de 2022 con un promedio de **6.97** tweets por día aprox.

@FicoGutierrez

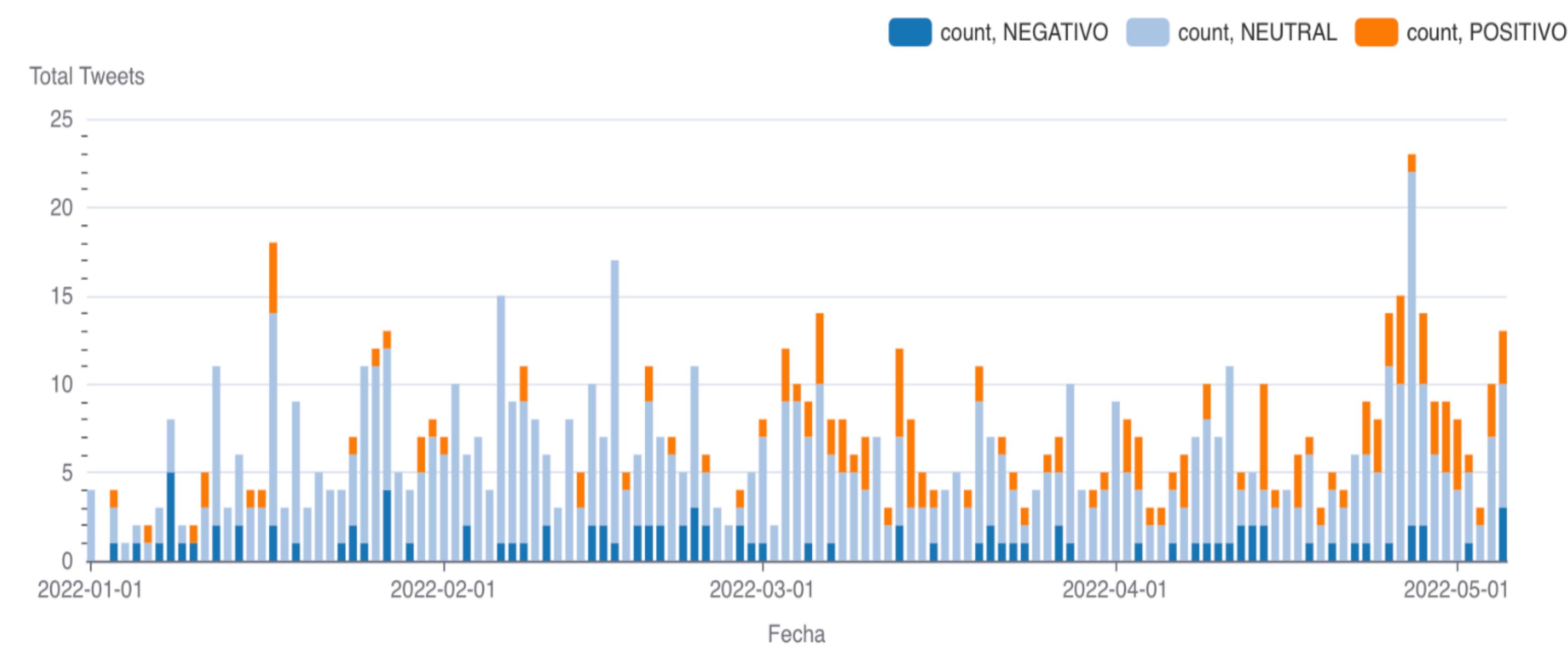


Cantidad de tweets escritos por día

El **89.12%** de los tweets escritos por el candidato son neutrales o positivos, el **10.88%** son tweets cuya composición gramatical expresan sentimientos tristes o negativos

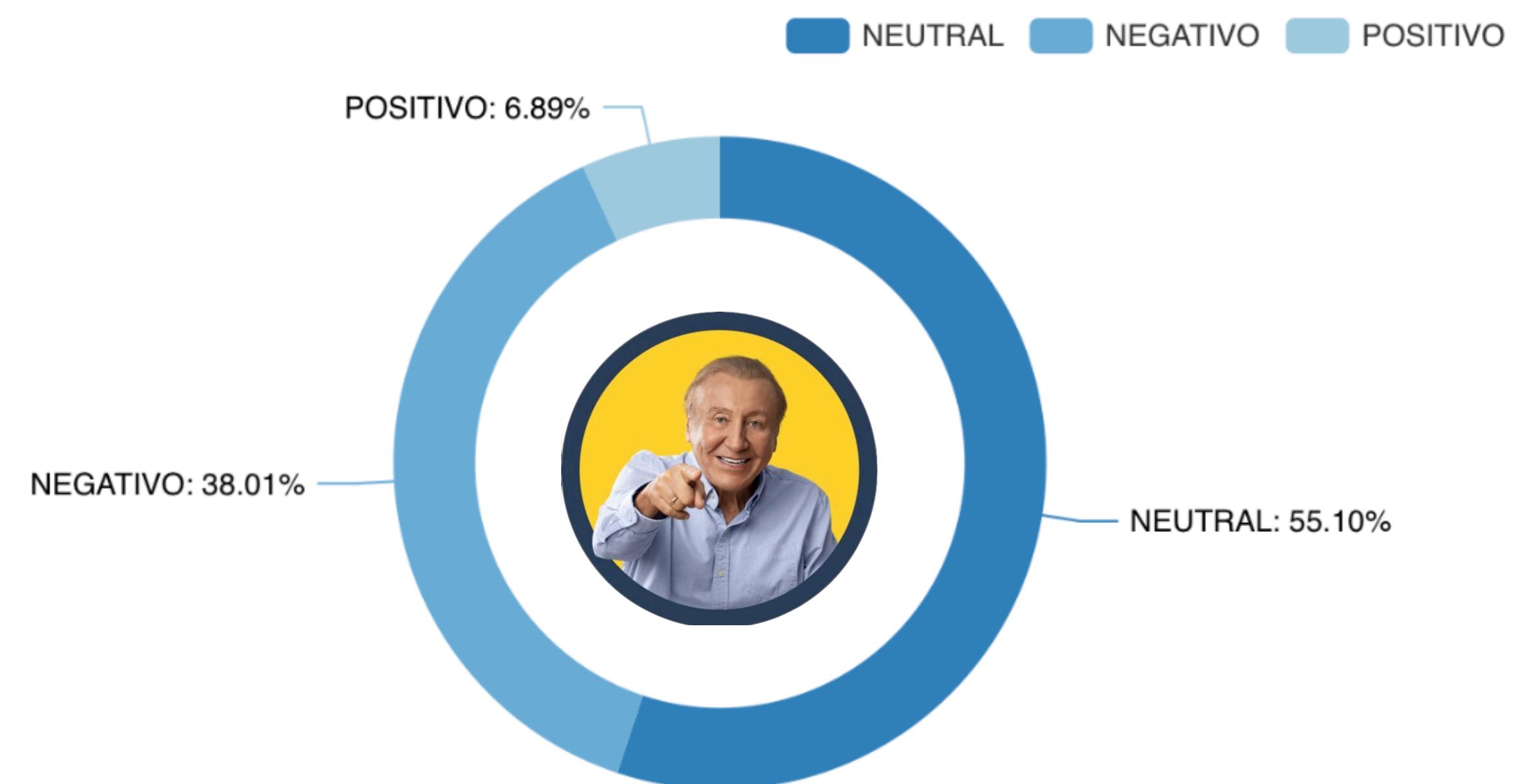
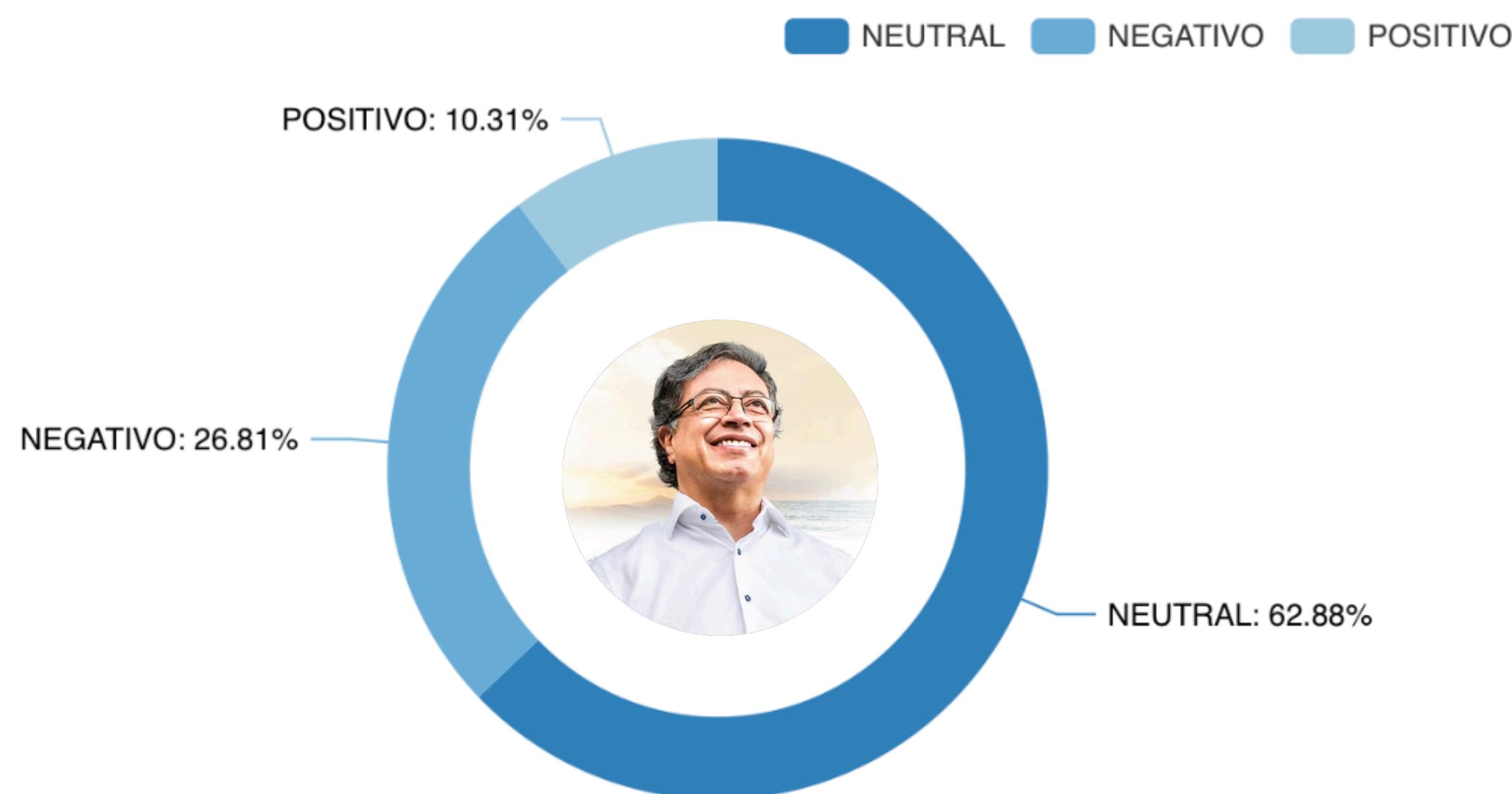
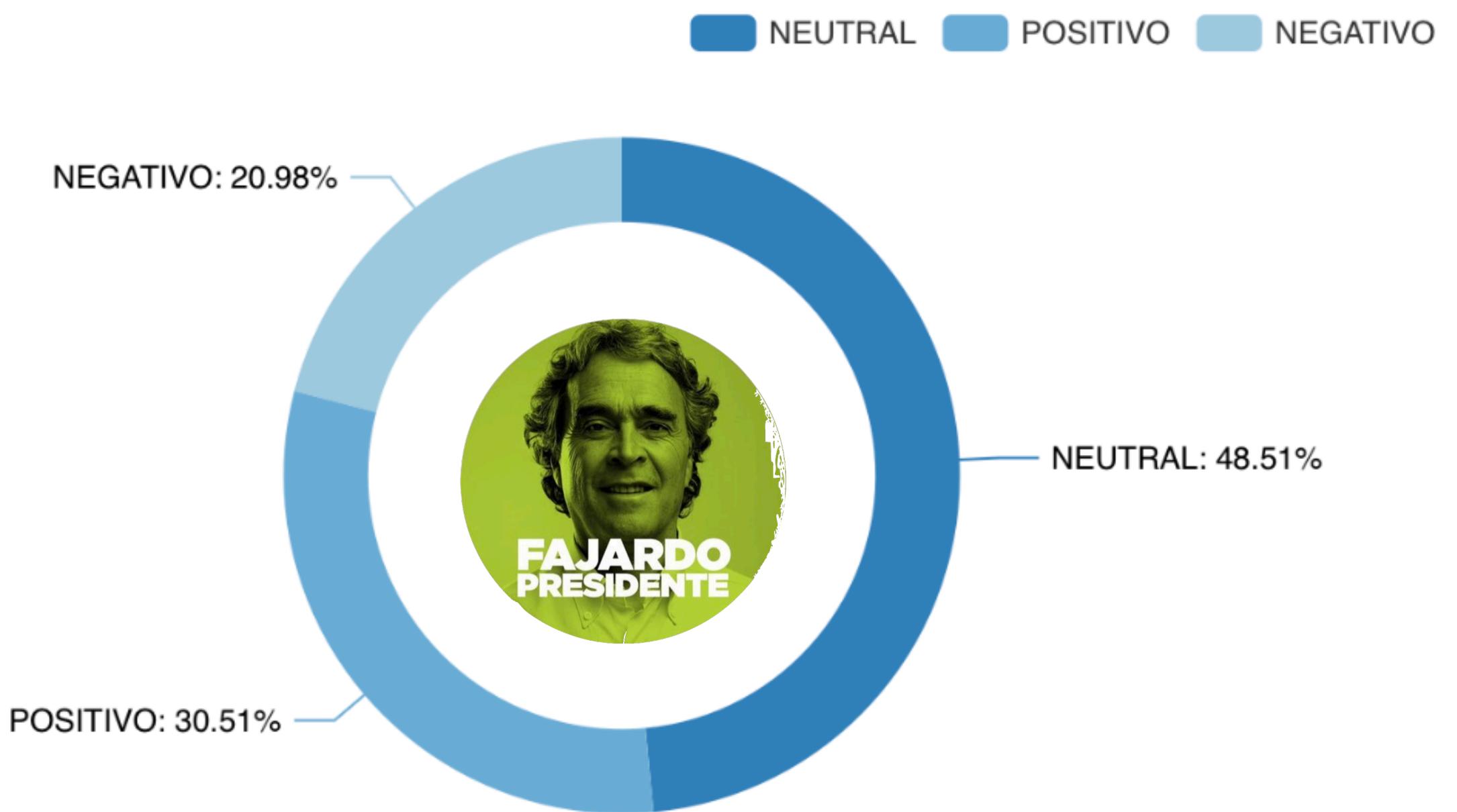
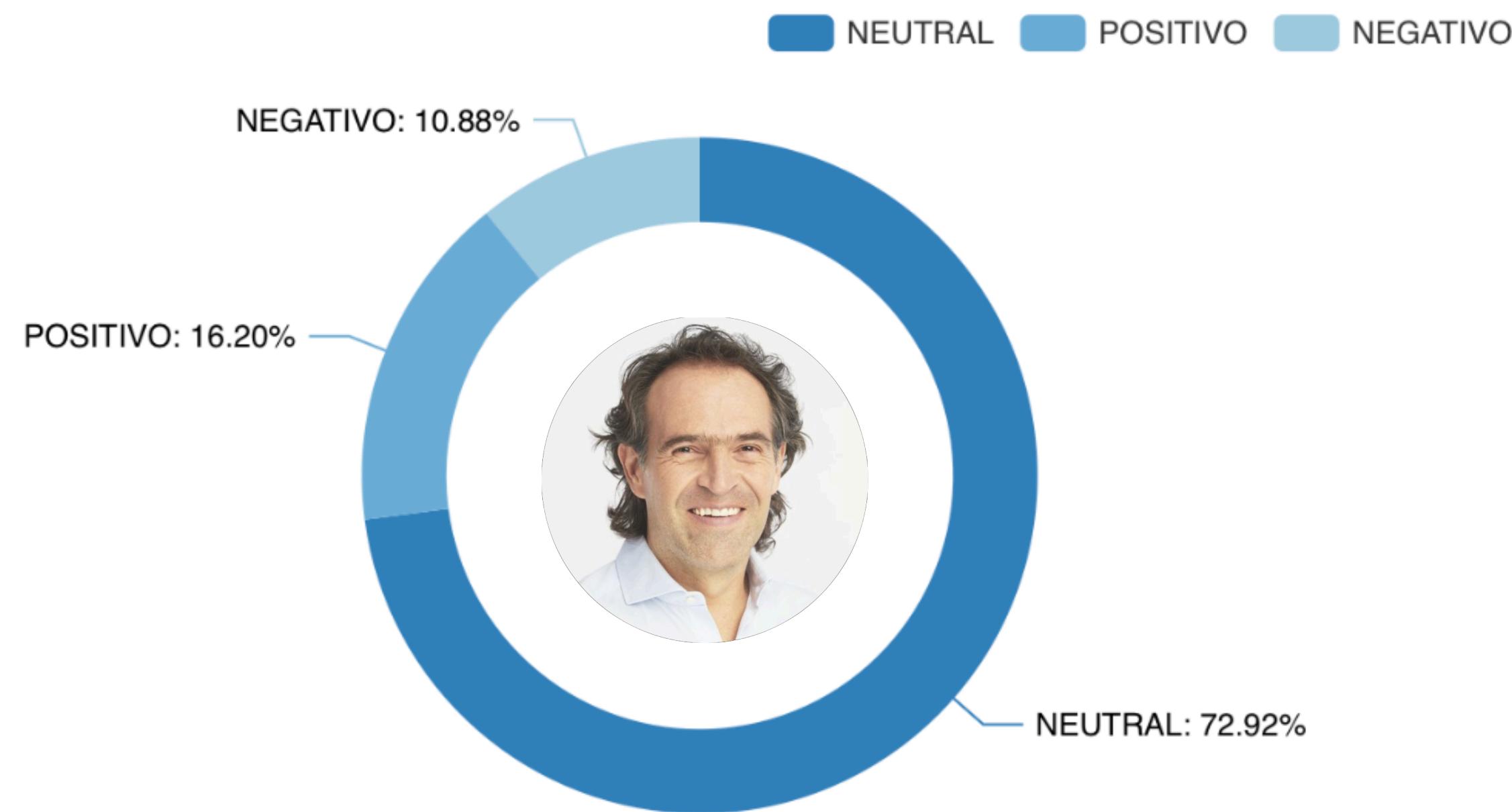


Distribución Tweets por tipo de sentimiento



Cantidad de Tweets escritos por día por tipo de sentimiento

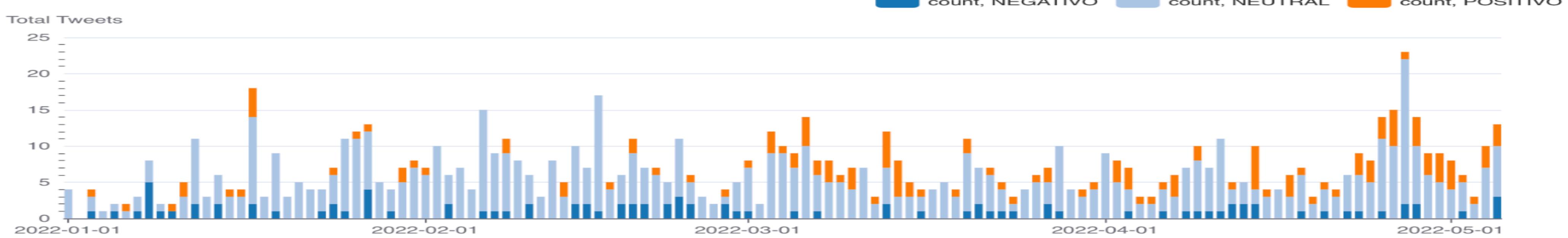
Distribución Tweets por sentimiento (positivo- neutral - negativo)



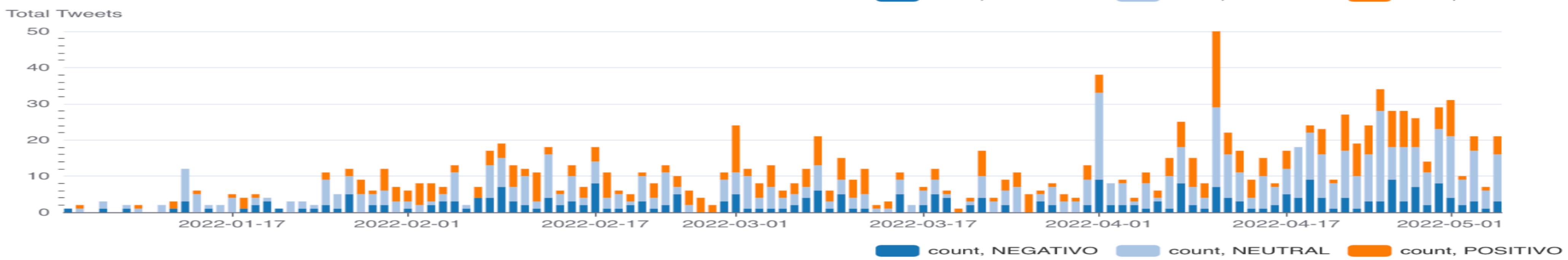
Tweets Diarios por candidato y sentimiento



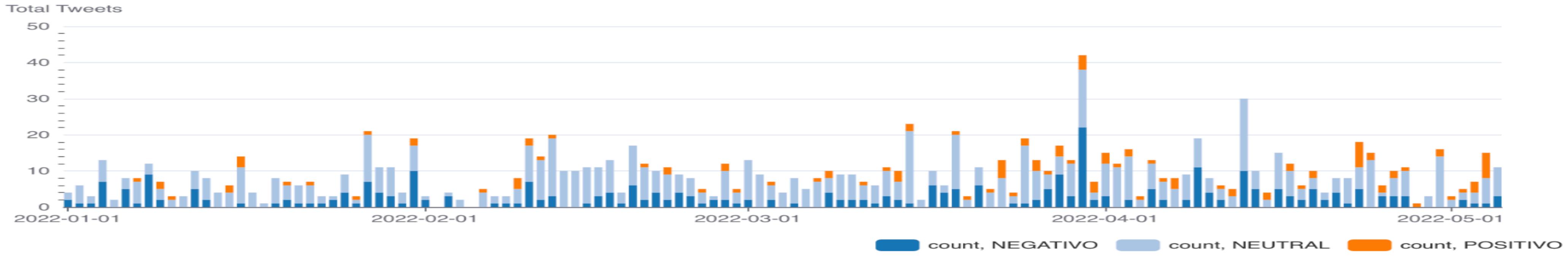
89% Tweets Neutros o Positivos



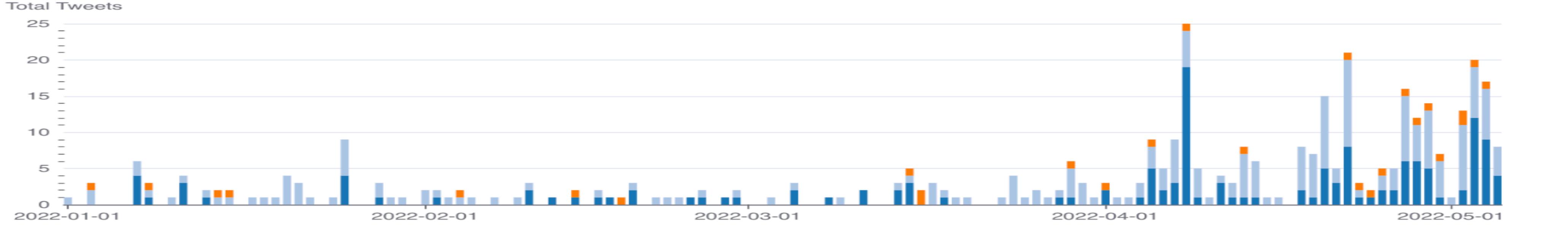
79% tweets neutros o positivos



73% tweets neutros o positivos



62% tweets neutros o positivos



Muchas Gracias

Frederick Salazar Sanchez

@FrederickSalazar

Data Engineer - Data Scientist

Msc BigData Analytics (En curso)

Universitat Oberta de Catalunya

Versión 1.0

Fecha 29 Abril de 2022