Dirección de Economía Naranja y Desarrollo Digital

Unidad de Científicos de Datos

El futuro es de todos

DNP
Departamento
Nacional de Planeación

ALERTAS TEMPRANAS II

INFORME FINAL

Dependencias y entidades involucradas	Departamento Nacional de Planeación • Dirección de Economía Naranja y Desarrollo Digital - Unidad de Científicos de Datos • Dirección de Seguridad, Justicia y Gobierno • Dirección de Gobierno DD.HH y Paz
Sector	Planeación
Tecnologías utilizadas	Python
Fuentes de datos	Twitter

Contenido

1. Presentación

Este informe presenta la herramienta de visualización que contiene los resultados de la extracción y análisis de tweets de interés, los cuales se obtuvieron a partir de términos claves pertenecientes a distintos temas que se extrajeron de cuentas de Twitter relevantes para el proyecto. Esta es la segunda versión de la herramienta (la primera se desarrolló en 2021) y contiene gráficos y tablas para analizar las tendencias de Twitter en Colombia con respecto a los siguientes temas: (1) Acceso a justicia, (2) Cultivos ilícitos, (3) Protección de áreas protegidas, (4) Relación estado-ciudadano y (5) Seguridad y paz. La herramienta se accede desde la intranet de DNP y se actualiza automáticamente cada día.

This document presents the visualization app that shows the results of the extraction and analysis of tweets, which were obtained by finding in them keywords belonging to different subjects. The tweets belong to accounts chosen by the dependencies that will use the app. This is the second version of the application (the first one was developed in 2021) and contains graphs and tables that analyze Twitter trends in Colombia related to the following subjects: (1) access to justice, (2) illicit crops, (3) protection of protected areas, (4) State-citizen relationship and (5) peace and security. The visualization is accessed in DNP's intranet and updates automatically every day.

2. Objetivos del proyecto

2.1. General

Realizar un diagnóstico de tendencias identificados en Twitter a partir de 5 temas principales definidos por los solicitantes de este proyecto mediante análisis descriptivo y de texto. En esta segunda etapa se continuó con el desarrollo del tablero de visualización del proyecto Alertas Tempranas de 2021, el cual extrae información de Twitter a partir de actores y términos clave y genera resultados. Se trabajó en la mejora y automatización del proceso de extracción de información y visualización de resultados.

2.2. Específicos

1. Generar la extracción de información con la librería Tweepy de Python
2. Almacenar la información extraída de Twitter usando un motor de bases de datos relacionales (SQL)
3. Ampliar los actores y/o términos clave de insumo para la extracción de información relevante en Twitter
4. Automatizar el proceso de extracción, almacenamiento y visualización de resultados.
5. Incluir gráficos para alimentar el análisis de alertas de la información extraída de Twitter
6. Implementar un mecanismo de generación de alertas

3. Metodología

De manera general, la metodología de desarrollo de la herramienta incluye, por un lado, la extracción de la información de interés a partir de temas y términos clave y de actores específicos, lo cual es el insumo para la elaboración de resultados. Por el otro lado, se utiliza este insumo para desarrollar la herramienta de visualización, con gráficos y tablas que se actualizan automáticamente. A continuación, se presenta la metodología en 5 pasos, desde la definición de términos claves y usuarios de Twitter para saber qué información extraer hasta la creación de las tablas y gráficos de la herramienta.

3.1. Definición de términos clave y usuarios de Twitter

La información de interés extraída de Twitter fue definida por las direcciones solicitantes del proyecto y está compuesta por cinco temas principales. Estos son: (1) Acceso a justicia, (2) Cultivos ilícitos, (3) Protección de áreas protegidas, (4) Relación estado-ciudadano y (5) Seguridad y paz. Cada uno de estos temas tiene sus propios términos clave, los cuales se utilizan para extraer los tweets de una serie de cuentas también definidas por las direcciones.

Los términos pueden ser una sola palabra o varias. En caso de contener varias, se buscan si todas hacen parte de un tweet y en caso positivo se extrae el texto e información asociada a la publicación. En varios casos se buscan las raíces de las palabras para facilitar la extracción de información relevante. La Tabla 1 de los anexos contiene los

términos clave para cada tema y la Tabla 2, también de los anexos, presenta los usuarios de quienes se extrae la información.

3.2. Extracción de información de Twitter

La extracción de tweets relevantes se hace con la librería Tweepy de Python. Con un usuario y código (token) gestionado por las direcciones solicitantes de este proyecto, se activan los permisos para extraer tweets desde el código de Python. Para ello se itera sobre cada uno de los usuarios de Twitter y se extraen todos sus tweets publicados 3 días antes de la fecha actual. Luego se filtran los tweets de acuerdo con cada tema y término clave, junto con información de insumo para la posterior creación de tablas y gráficos de la herramienta de visualización. Las tablas de cada tema se guardan en un archivo ".db", el cual es una tabla de datos de SQL. Las columnas de las tablas almacenadas ahí (una tabla por tema), son las siguientes:

1. Texto del tweet publicado
2. Fecha de publicación
3. URL de la publicación en Twitter
4. Fecha de extracción
5. Número de *retweets*
6. Número de *replies*
7. Número de *likes*
8. Nombre de usuario

El código de la herramienta de visualización y el proceso de extracción se encuentra en los servidores del DNP. Desde ahí se realiza el proceso de extracción de información y de generación de resultados cada día.

3.3. Preprocesamiento y filtro de tweets relevantes

De las 8 columnas extraídas de Twitter, la que contiene el texto de los tweets es la única que necesita preprocesamiento y de la cual se extrae más información. El preprocesamiento consiste simplemente en pasar el texto a minúsculas y quitar acentos de las palabras. Luego se buscan los términos clave (se pueden ver en la Tabla 1 de los Anexos), se eliminan las filas que no contengan ningún término y se crean columnas para cada término, las cuales indican qué término se encontró en cada tweet. Estos resultados se guardan en bases de datos ".db", de SQL.

3.4. Creación de tablas, gráficos y la herramienta de visualización

La base de datos ".db" con la información completa de los tweets es el insumo principal para la creación de las tablas y gráficos que se presentan en la herramienta de visualización. La herramienta se construyó con la librería de Python llamada Dash y contiene una ventana por cada tópico de análisis. Los gráficos y tablas que se pueden consultar en la herramienta son los siguientes:

- Tabla histórica de alertas
- Gráfico de barras de frecuencia de términos clave
- Tabla de datos
- Nube y tabla de *trending topics*
- Nube y tabla con las palabras más frecuentes de tweets
- Red y matriz de coocurrencias
- Gráfico de líneas de frecuencias históricas

Todas las tablas y gráficos se desarrollaron con librerías de Python. Cada uno de estos se presentará en la sección de resultados. A continuación, se explicará la metodología de creación de alertas, ya que el algoritmo fue definido por los miembros solicitantes del proyecto.

3.5. Elaboración de metodología para generar alertas

La generación de alertas (cuyos resultados se presentan en la Sección 4.1) se desarrolló con el fin de resaltar los días en los que hubo tendencias más grandes de lo normal en alguno de los 5 temas de interés. Para generar las alertas se crea un indicador sencillo de tendencias para cada tema (suma de *likes, retweets* y respuestas) y se calculan su media y desviación estándar históricas. Posteriormente se define una variable límite como la suma de la media con tres desviaciones estándar. Al final se revisan los días en los cuales los indicadores superan este límite y, en caso afirmativo, se genera una alerta.

4. Resultados

A través del desarrollo metodológico descrito en la Sección 3, se obtuvieron los resultados que se presentan a continuación. Toda retroalimentación desde un punto de vista experto o de usuario por parte de la Dirección de Seguridad, Justicia y Defensa y la Dirección de Gobierno, DDHH y Paz es bienvenida. Este insumo será de gran ayuda para mejorar la calidad y utilidad de los resultados obtenidos, de manera que agreguen mayor valor.

A continuación, se presentan los resultados de la herramienta de visualización, una sección por cada ventana de la herramienta.

4.1. Tabla de alertas históricas

La primera vista al abrir la herramienta es la tabla de alertas históricas. Contiene el día donde se generó la última alerta, el tema al que pertenece y los números de *likes*, *retweets* y repuestas totales relacionadas con el tema. También resalta la última alerta generada. La Imagen 1 presenta cómo se ve la ventana de alertas en la herramienta de visualización.

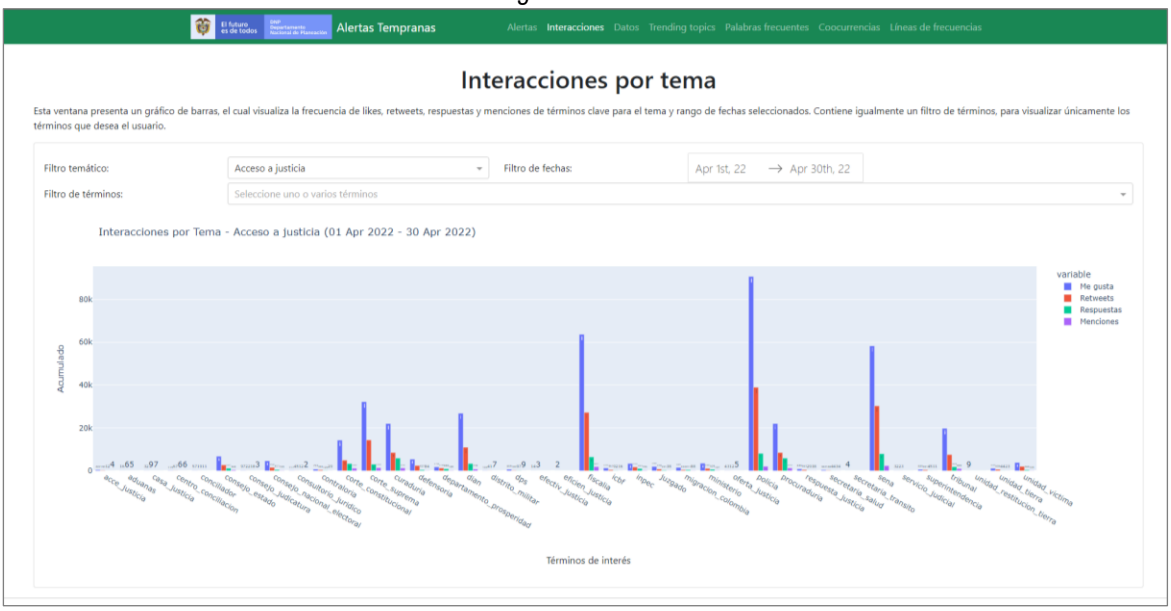
Imagen 1. Tabla de alertas históricas

Fuente: herramienta de visualización

4.2. Gráfico de barras

El gráfico de barras presenta el número de interacciones (*likes*, *retweets*, *respuestas* o menciones de Twitter) agrupadas por tema y términos de interés en un rango de fechas seleccionados por el usuario. El eje X del gráfico contiene los términos clave del tema escogido y el eje Y el número de *likes*, *retweets*, *respuestas* o menciones de Twitter, los cuales se representan en distintas barras de colores, según la leyenda que se encuentra al lado derecho del gráfico. Es posible oprimir sobre el gráfico y ampliar la vista de una porción específica, recortando la porción deseada. La Imagen 2 presenta un ejemplo de cómo se ve el gráfico de barras.

Imagen 2. Gráfico de barras

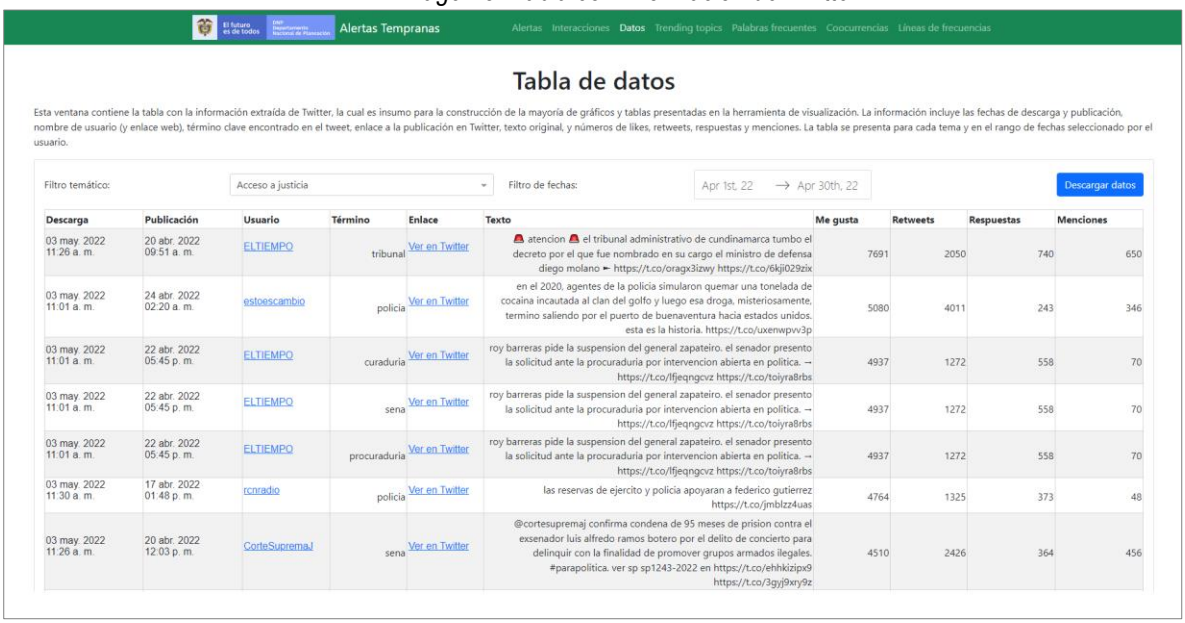


Fuente: herramienta de visualización

4.3. Tabla de datos

La herramienta también permite ver una tabla que contiene la información histórica extraída de Twitter. Esta se puede filtrar con un rango de fechas seleccionado por el usuario. Un ejemplo de esta tabla se encuentra en la Imagen 3. La información incluye las fechas de descarga y publicación, nombre de usuario (y enlace web), término clave encontrado en el tweet, enlace a la publicación en Twitter, texto original y números de *likes*, *retweets*, respuestas y menciones.

Imagen 3. Tabla con información de Twitter

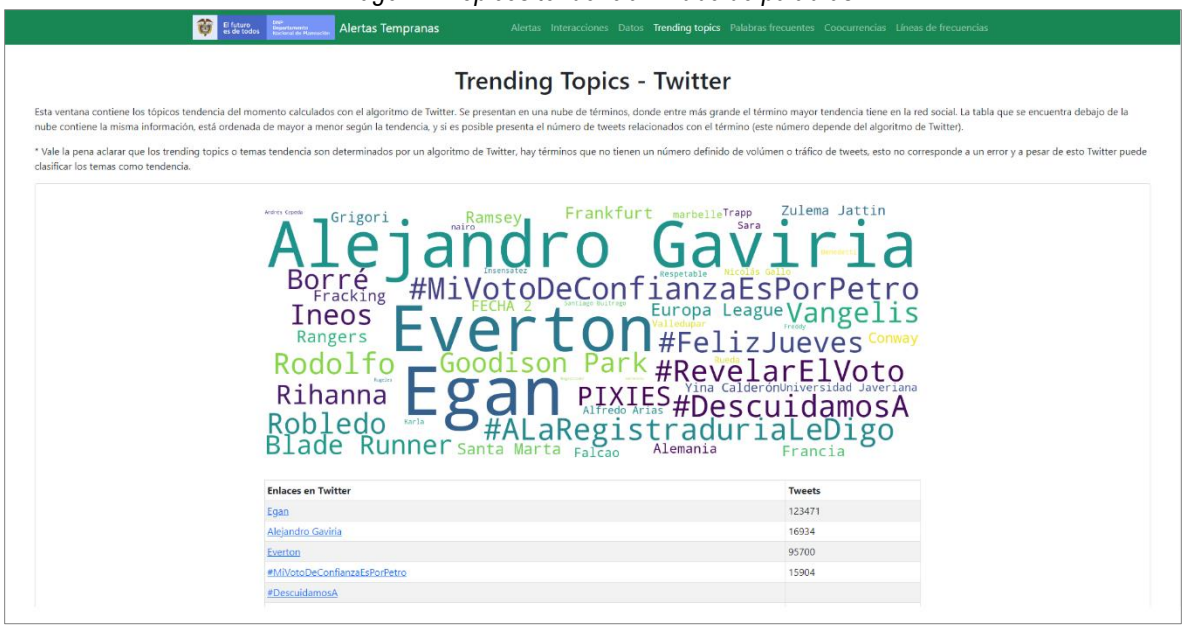


Fuente: herramienta de visualización

4.4. Tópicos tendencia

La herramienta también presenta un análisis de *trending topics* (tópicos tendencia), con el fin de dar una idea general de los temas más publicados y discutidos en el momento de acceder a la herramienta. La Imagen 4 presenta un ejemplo de una nube de palabras que se obtuvo de la herramienta. La nube presenta la importancia de los tópicos tendencia según el tamaño del término en el gráfico, donde entre más grande el término mayor tendencia tiene en la red social.

Imagen 4. Tópicos tendencia – nube de palabras



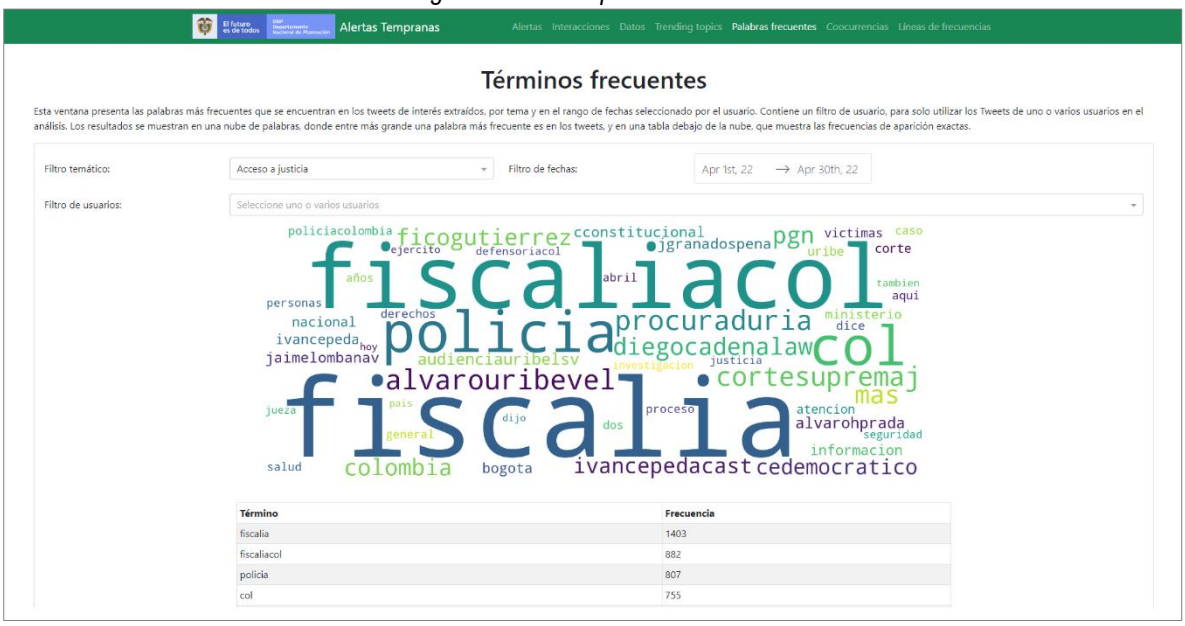
Fuente: herramienta de visualización

Adicionalmente se tiene una tabla con los tópicos tendencia, que se encuentran en la misma ventana de la nube de palabras, ordenados de mayor a menor importancia y si es posible presenta el número de tweets relacionados con el término (este número depende del algoritmo de Twitter).

4.5. Palabras más frecuentes de tweets

Este resultado consiste en presentar las palabras más frecuentes que aparecen en los tweets de interés para cada tema. Los resultados se muestran en una nube de palabras, donde entre más grande una palabra más frecuente es en los tweets, y en una tabla debajo de la nube, que muestra las frecuencias de aparición exactas. La Imagen 5 presenta un ejemplo de esta ventana con la nube de palabras y la tabla debajo.

Imagen 5. Ventana palabras más frecuentes

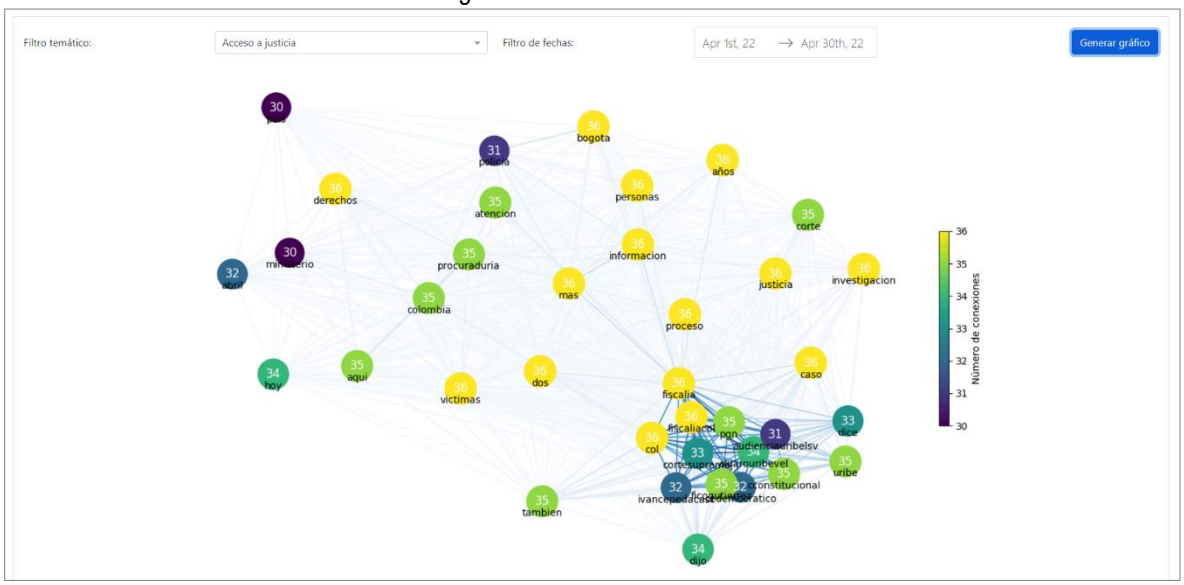


Fuente: herramienta de visualización

4.6. Red y matriz de coocurrencias

La herramienta también presenta una red de coocurrencias, la cual muestra de manera visual la frecuencia de términos dentro de los tweets y su relación con otros términos. Debajo de la red se encuentra una matriz de coocurrencias, para verificar el número exacto de veces que aparecen palabras muy frecuentes en cada tweet y los términos clave. La Imagen 6 presenta un ejemplo de cómo se visualiza la red de coocurrencias.

Imagen 6. Red de coocurrencias



Fuente: herramienta de visualización

La Imagen 7 presenta un ejemplo de la matriz de coocurrencias, que contiene la relación de los términos clave las palabras más frecuentes dentro de los tweets extraídos. Muestra la frecuencia de cada término clave y los términos que se encuentran con más frecuencia junto a ellos.

Imagen 7. Matriz de coocurrencias

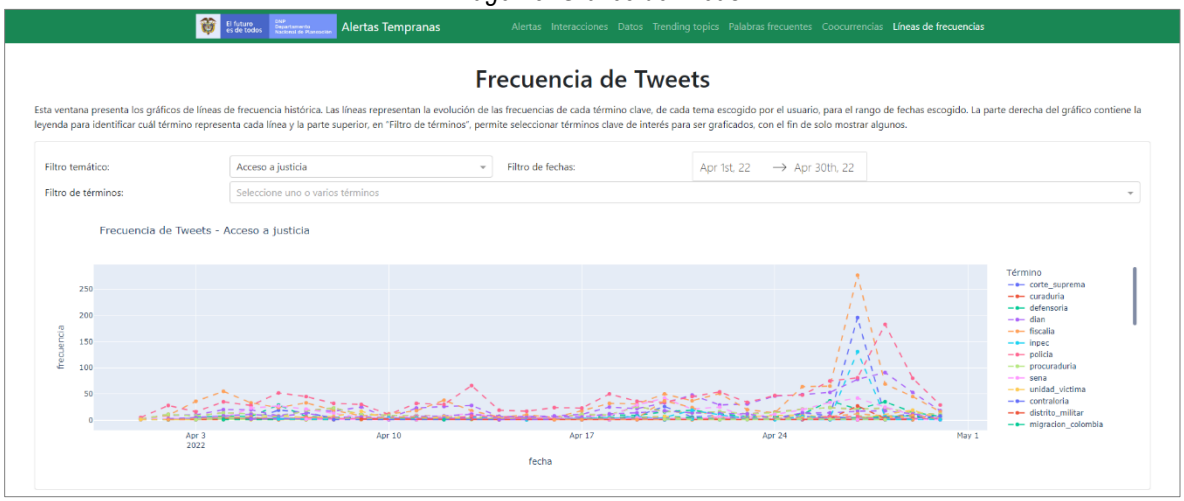
Término	cconstitucional	colombia	corte	cortesupremaj	defensoriacol	fiscalia	fiscaliacol	justicia	ministerio	nacional	policia	policiacolomb	procuraduria	salud	victimas
abril	2	42	10	0	15	28	11	6	28	16	35	7	16	28	6
alvarohprada	411	0	11	572	0	511	573	14	0	0	0	0	4	0	21
alvarouribevel	545	3	25	849	0	699	862	23	2	0	0	0	7	0	29
aqui	3	29	8	10	9	18	15	11	25	10	9	0	143	7	23
atencion	5	17	19	7	13	99	6	11	31	26	20	11	57	41	32
audienciauribelsv	402	0	13	606	0	510	618	14	0	0	0	0	4	0	23
años	3	11	13	10	1	105	17	8	3	9	19	5	17	6	10
bogota	4	14	9	6	3	52	21	4	4	42	156	11	11	32	5
caso	59	13	19	92	2	182	116	18	4	4	23	3	26	0	13
cconstitucional	438	11	16	487	3	443	479	25	1	0	1	0	2	0	12
cedemocratico	411	1	13	616	0	520	623	14	0	0	0	0	4	0	21
col	427	65	13	642	92	548	728	26	30	1	8	105	18	0	27
colombia	11	573	23	14	34	36	23	33	61	49	46	32	66	36	34
corte	16	23	320	14	0	30	18	79	8	1	8	0	12	5	7
cortesupremaj	487	14	14	688	0	610	758	31	0	0	0	0	4	0	25

Fuente: herramienta de visualización

4.7. Gráfico de frecuencias

La Imagen 8 contiene un ejemplo del gráfico de líneas, la cual presenta la evolución de la frecuencia de tweets relacionados a un tema o de un término escogido por un usuario en un rango de fechas específico.

Imagen 8. Gráfico de líneas



Fuente: herramienta de visualización

5. Conclusiones y recomendaciones

A partir de la metodología desarrollada y de los resultados obtenidos para cumplir los objetivos de este proyecto, planteados en el plan de trabajo, se presentan a continuación las principales conclusiones obtenidas por el equipo de la UCD y las principales recomendaciones para un mejor uso y aprovechamiento del proyecto.

1. La herramienta de visualización presenta gráficos y tablas que se actualizan automáticamente, con el fin de presentar un análisis de tendencias de 5 temas que fueron escogidos por los solicitantes del proyecto. Todos los resultados presentados provienen de información extraída de Twitter.
2. Durante el transcurso del proyecto se enviaron 2 entregables a las dependencias que solicitaron el proyecto y un manual de usuario de la herramienta de visualización, el cual contiene las instrucciones de ejecución y

uso de la herramienta. Se recomienda leer el manual antes de hacer uso de ella. El presente documento (el informe final) consolida la información de los entregables.

6. Socialización

Los entregables, manual de usuario fueron compartidos con miembros de la DJSD y la Dirección de Gobierno DD.HH y Paz. La herramienta de visualización fue presentada y explicada de manera virtual.

Contacto

Si tiene alguna duda, comentario o sugerencia sobre este proyecto, o si le gustaría conversar con la Unidad de Científicos de Datos sobre la posibilidad de una nueva fase para el mismo, puede comunicarse con nosotros a través del correo electrónico ucd@dnp.gov.co.

ANEXOS

Tabla 1. Temas y términos clave

Relación Estado Ciudadano	Seguridad y Paz	Acceso a la Justicia	Cultivos Ilícitos	Protección de Áreas Protegidas
licitacion	policia	acce_justicia	antinarcotico	lider_ambient
apropia_indebida	esmad	oferta_justicia	aspersion_aerea	seguridad_climatica
corrupcion_public	desaparicion_forz	servicio_judicial	cultiv_coca	cambio_climatico
transparencia	apoyo_nacional	juzgado	cultiv_ilicit	incendio
soborno	protesta	tribunal	direccion_de_sustitucion	deforestación
impunidad	paro_nacional	defensoria	envio_coca	crimen_ambient
perdida_recurso	bloqueo_via	fiscalia	erradicacion_forz	mebog
elefante_blanco	manifestacion	procuraduria	estrategia_integral	policia_ambient
contratacion_indebida	marcha	unidad_restitucion_tierra	formulario_vinculacion	area_protegida
escandalo	planton	unidad_tierra	fumigacion_coca	sinap
indebida_diligencia	militarizacion	unidad_victima	fumigacion_glifosato	chiribiquete
campaña_politica	victima	contraloria	guerra_droga	parque_natural
presunto_caso	conflict	inpec	microtrafico_cocaina	reserva_natural
clientelismo	reparacion_integral	consultorio_juridico	pasta_coca	santuario_fauna
denuncia	repara_victima	centro_conciliacion	plan_integral	santuario_flora
condena	asesin_lider	conciliador	pnis	pesca_ilegal
renuncia_cargo	homicidio_lider	superintendencia	politica_droga	voladura_oleoducto
anticorrupcion	lider_social	consejo_nacional_electoral	produccion_cocaina	explotacion_ilegal
ente_investigador	reincorpora	casa_justicia	reforma_rural	contamina_ambient
participa_ciudada	reintegra	secretaria_transito	resiembra_coca	trafico_fauna
apropia_recurso	conflict_social	dian	resiembra_cultiv	caza_ilegal
desvi_recurso	acuerdo_paz	secretaria_salud	sustitucion_coca	ecocidio
malversa_recurso	farc	policia	sustitucion_cultiv	runap
organizacion_social	lider_afro	icbf	sustitucion_voluntaria	dicar
sociedad_civil	lider_ambiental	bienestar_familiar	trafico_cocaina	-
estado_transparente	lider_campesin	secretaria_educación	unodc	-
consulta_previa	lider_comunal	aduanas	-	-
trata_persona	lider_comunitari	departamento_prosperidad	-	-
innova_entidad	lider_cultural	sayco	-	-
venezolan	lider_indigena	dps	-	-
migra_venez	lider_juvenil	consejo_judicatura	-	-
estado_eficiente	lider_lgtbi	sena	-	-
innova_estado	lider_sindical	ministerio	-	-
xenofob	lider_victima	migracion_colombia	-	-
innova_public	paz	curaduria	-	-
eficiencia_institucion	restitu_tierra	corte_constitucional	-	-

Relación Estado Ciudadano	Seguridad y Paz	Acceso a la Justicia	Cultivos Ilícitos	Protección de Áreas Protegidas
regula_norma	eln	corte_suprema	-	-
racismo	clan_golfo	oficina_transito	-	-
ocde	disidencia	distrito_militar	-	-
coopera_internacional	bacrim	consejo_estado	-	-
servicio_ciudadano	paramilitar	efectiv_justicia	-	-
consej_planeacion	aguilas_negras	ineficien_justicia	-	-
impacto_norma	postconflicto	eficien_justicia	-	-
ong	pmi	respuesta_justicia	-	-
compra_voto	pdet	-	-	-
eleccion_irregular	zomac	-	-	-
venta_voto	-	-	-	-
vende_voto	-	-	-	-
candidato_corrup	-	-	-	-

Fuente: elaboración propia a partir de términos clave enviados por direcciones solicitantes del proyecto

Tabla 2. Usuarios de Twitter

Usuarios de Twitter			
NoticiasCaracol	ONUHumanRights	MisionONUCol	ideaspaz
NoticiasRCN	MinInterior	TembloresONG	CINEP_PPP
ELTIEMPO	Registraduria	FLIP_org	transparenciaco
elespectador	larepublica_co	MigracionCol	UNPColombia
RevistaSemana	DefensoriaCol	DirectorPolicia	CRIC_Cauca
WRadioColombia	FuerzasMilCol	estoescambio	Indepaz
CaracolRadio	PGN_COL	PnudColombia	RenovacionCo
Citytv	MinjusticiaCo	PnudColombia	colombiacompra
rcnradio	hrw_espanol	ComisionVerdadC	ANZORC_OFICIAL
BluRadioCo	CancilleriaCol	Dejusticia	DefendamosPaz
lafm	CGR_Colombia	PacifistaCol	Stransparencia
NoticiasUno	CorteSupremaJ	ONIC_Colombia	SomosDef
CMILANOTICIA	anticorruption	HRI_ONG	ColombiaCompite
PoliciaColombia	JEP_Colombia	DAFP_COLOMBIA	InstAnticorrup
infopresidencia	RedMasNoticias	OIMColombia	FondoONUCol
lasillavacia	ViceColombia	ConsejeriaDDHH	mesanalvictimas
FiscaliaCol	ComunesCoL	USAID_Colombia	manosporlapaz18
COL_EJERCITO	consejodeestado	ARNColombia	GciaFronterasCo
mindefensa	cuestion_p	parescolombia	SeguridadNalCo
CIDH	ComisionadoPaz	PosconflictoCO	visomutop
BarometroX	CsiviComunes	fcolombiaenpaz	-

Fuente: elaboración propia a partir de usuarios enviados por entidades solicitantes del proyecto