

**TESIS DE LICENCIATURA EN CIENCIAS DE LA  
COMPUTACIÓN**

**MINERÍA DE RELACIONES A PARTIR DE OPINIONES Y  
EVENTOS EN TWEETS SOBRE EL CAMBIO CLIMÁTICO**



**UNIVERSIDAD NACIONAL DEL SUR  
DEPARTAMENTO DE CIENCIAS E INGENIERÍA DE LA COMPUTACIÓN**

Autor: Juan Manuel Marten  
Directora: Dra. Ana Gabriela Maguitman

Noviembre de 2023

# Índice

<b>1. Introducción.....</b>	<b>4</b>
1.1. Motivación.....	4
1.2. Objetivos.....	4
<b>2. Conceptos Preliminares.....</b>	<b>5</b>
2.1. Twitter.....	5
2.2. Aprendizaje Automático.....	5
2.3. Aprendizaje No Supervisado.....	5
2.4. Minería de Datos.....	5
<b>3. Herramientas.....</b>	<b>6</b>
3.1. Python.....	6
3.1.1. Pandas.....	6
3.1.2. Lingam.....	6
3.1.3. Tigramite.....	7
3.1.4. Statsmodels.....	7
3.1.5. Efficient-apriori.....	7
3.2. Google Colab.....	7
<b>4. Dataset.....</b>	<b>7</b>
4.1. Tweets.....	8
4.1.1. Geolocalización.....	8
4.1.2. Género.....	9
4.1.3. Postura.....	9
4.1.4. Sentimiento.....	10
4.1.5. Agresividad.....	10
4.1.6. Temperatura.....	10
4.1.7. Tópico.....	11

4.2. Desastres Ambientales.....	11
<b>5. Análisis Exploratorio.....</b>	<b>12</b>
<b>6. Procesamiento de Datos.....</b>	<b>17</b>
6.1. Filtrado.....	17
6.2. Discretización y Agrupamiento.....	18
<b>7. Conjuntos de Ítems Frecuentes y Reglas de Asociación.....</b>	<b>20</b>
7.1. A-Priori.....	21
7.2. Conjuntos de Ítems Frecuentes.....	22
7.3. Reglas de Asociación.....	25
<b>8. Análisis de Causalidad.....</b>	<b>28</b>
8.1. Algoritmos de Causalidad.....	28
8.2. ensemble.....	29
8.3. Curado de Grafos.....	30
<b>9. Conclusión.....</b>	<b>35</b>
<b>10. Trabajo a Futuro.....</b>	<b>36</b>
<b>11. Bibliografía.....</b>	<b>37</b>
<b>Apéndice.....</b>	<b>39</b>
A. Gráficos de tweets por año y por tópico.....	39
Gráficos por año.....	39
Gráficos por tópico.....	46

# 1. Introducción

## 1.1. Motivación

En la era digital, las redes sociales se han convertido en una fuente invaluable de datos que captura las voces y opiniones de millones de personas en todo el mundo. Entre estas plataformas, Twitter ha surgido como un espacio único donde los individuos expresan sus pensamientos, emociones y puntos de vista sobre una amplia gama de temas y, en particular, sobre el cambio climático, un tema que ha generado extensos debates y polémicas durante los últimos años además de ser un tema de importancia crucial para la persistencia de la vida en nuestro planeta.

Teniendo esto en cuenta, el objetivo de este trabajo es extraer conocimiento sobre la forma de opinar de las personas respecto al cambio climático a partir de un gran conjunto de tweets sobre el tema aplicando técnicas de aprendizaje automático no supervisado y minería de datos. De esta manera se busca encontrar patrones y entender el pensamiento de la gente, lo cual proporcionaría información muy valiosa para los responsables de políticas públicas a la hora de tomar decisiones relacionadas al cambio climático.

## 1.2. Objetivos

Tomando como base el dataset “The Climate Change Twitter Dataset” [1], el cual consiste de más de 15 millones de tweets etiquetados con variables tales como sentimiento, postura, agresividad y tópico, entre otras variables, el objetivo de este proyecto es utilizar esta gran cantidad de datos para minar relaciones entre las distintas variables de interés presentes en el dataset.

En una primera instancia se realizará un análisis exploratorio del dataset, realizando estadísticas y gráficos sobre los datos sin procesar, lo que permitirá familiarizarse con la información, empezar a conocer la distribución de los datos y formar algunas primeras hipótesis.

Luego se realizará un análisis de conjuntos de ítems frecuentes y reglas de asociación, en el que se buscará entender cómo se relacionan las distintas variables analizando los tweets a nivel individual.

Finalmente se procederá con un análisis de causalidad, dónde se buscará encontrar relaciones causales entre las distintas variables, se intentará definir cuando un cambio en una variable produce con el pasar del tiempo un cambio sobre otra variable. En este punto se incluirán datos sobre desastres ambientales además de las variables asociadas a los tweets.

## 2. Conceptos Preliminares

### 2.1. Twitter

Twitter es una red social que permite a sus usuarios compartir mensajes breves, conocidos como “tweets”, con una extensión de hasta 280 caracteres. Estos mensajes pueden contener texto, imágenes, enlaces y demás tipos de elementos multimedia. Dicha plataforma se ha convertido en una fuente invaluable para obtener información en tiempo real, donde una enorme cantidad de usuarios expresan sus opiniones, comparten noticias y participan en conversaciones sobre cualquier tipo de temas de manera constante.

Debido a su carácter público, su inmediatez, la gran cantidad de usuarios y su diversidad, Twitter se ha convertido en una herramienta muy importante a la hora de analizar y comprender patrones de comportamiento y opiniones de la población.

### 2.2. Aprendizaje Automático

El aprendizaje automático, o machine learning, es una rama de la inteligencia artificial que se basa en el desarrollo de algoritmos que permiten a las computadoras aprender patrones a partir de datos y realizar tareas para las que no fueron programadas de manera explícita. Estos tipos de sistemas suelen mejorar su rendimiento a medida que crece el volumen de los datos utilizados para su entrenamiento por lo que resultan particularmente eficaces para resolver problemas complejos en los que se cuenta con una gran cantidad de información.

### 2.3. Aprendizaje No Supervisado

El aprendizaje no supervisado es una forma de aprendizaje automático donde los algoritmos se entrena utilizando datos sin etiquetar. A diferencia del aprendizaje supervisado, donde los modelos se entrena con ejemplos etiquetados, el aprendizaje no supervisado busca encontrar patrones y estructuras inherentes en los datos sin la guía explícita de las etiquetas. Este enfoque es muy útil para la exploración y el descubrimiento de información oculta en conjuntos de datos complejos y no estructurados y en situaciones donde resultaría muy costoso realizar un etiquetado manual de los datos.

### 2.4. Minería de Datos

La minería de datos es una disciplina propia de la estadística y las ciencias de la computación que implica la exploración y descubrimiento de patrones, tendencias y

conocimiento en general a partir de grandes conjuntos de datos. La información obtenida resulta particularmente valiosa cuando el gran volumen de datos hace que resulte imposible para un humano extraer este conocimiento sin la ayuda de herramientas computacionales y estadísticas. Hoy en día, debido a la extensión de Internet, la gran cantidad de dispositivos electrónicos que permiten generar o medir datos y las capacidades de las computadoras modernas, capaces de procesar grandes cantidades de información en segundos, este tipo de técnicas resultan muy importantes para solucionar problemas de todo tipo, como recomendaciones en línea, diagnóstico médico, detección de fraudes y análisis de redes sociales, entre otras.

## 3. Herramientas

### 3.1. Python

Para el desarrollo de este trabajo se utilizó Python como lenguaje de programación principalmente debido a que cuenta con una amplia variedad de bibliotecas especializadas para ciencia de datos, las cuales proporcionan herramientas poderosas y eficientes para el análisis de datos, manipulación de datos, construcción de modelos y visualización. Además se destaca su flexibilidad y facilidad de uso que permite desarrollar tareas complejas de manera ágil, centrándose en la lógica del programa en lugar de las complejidades sintácticas o semánticas del lenguaje. Y también cuenta con una gran comunidad activa por lo que resulta sencillo acceder a recursos como documentación, tutoriales, foros de discusión y librerías de la comunidad.

A continuación se describen algunas de las principales librerías utilizadas.

#### 3.1.1. Pandas

Pandas es una poderosa librería para el análisis y la manipulación de datos en Python. Esta proporciona estructuras de datos flexibles y eficientes para almacenar y manipular datos tabulares. Es utilizada en este proyecto para manipular el gran dataset de tweets, calcular estadísticas y realizar filtrados, agregaciones y transformaciones sobre los datos.

#### 3.1.2. Lingam

Lingam es una librería basada en el modelo LiNGAM (Linear Non-Gaussian Acyclic Model), que permite modelar relaciones causales lineales entre variables en un conjunto de datos. En este proyecto se utiliza la componente DirectLiNGAM para extraer relaciones causales dentro del dataset de tweets.

### 3.1.3. Tigramite

Tigramite es otra librería utilizada para modelar relaciones causales entre variables de un conjunto de datos, consta de una colección de métodos para el análisis de causalidad, de los cuales se utilizan en este proyecto PC y PCMCI para tal fin.

### 3.1.4. Statsmodels

Esta librería ofrece una gran cantidad de funcionalidades para la estimación de modelos estadísticos, pruebas estadísticas y exploración de datos. Dentro de las funcionalidades que provee se utilizó el modelo VAR para contribuir al análisis de causalidad.

### 3.1.5. Efficient-apriori

Efficient-apriori es una librería que implementa de manera eficiente el algoritmo Apriori, el cual es una técnica de minería de datos utilizada para obtener conjuntos de ítems frecuentes y reglas de asociación sobre datos representados como listas de transacciones compuestas por ítems.

## 3.2. Google Colab

En este proyecto se utilizó Google Colab como plataforma de desarrollo para el lenguaje Python. Se destacan como ventajas que está basado en la nube, lo que permite trabajar fácilmente desde cualquier equipo sin necesidad de configurar todo un entorno, la utilización de notebooks Jupyter, que permite ejecutar el código de manera interactiva, y la perfecta integración con librerías preinstaladas para ciencia de datos acompañada por una gran facilidad para instalar nuevas librerías.

El código fuente del trabajo realizado puede ser accedido a través del siguiente enlace: <https://colab.research.google.com/drive/1nPOqJ43UiecvZnRQ2IfrS6ESEEdnEh7J?usp=sharing>

## 4. Dataset

Para la realización de este trabajo se utilizó el dataset “The climate change Twitter dataset” [1] el cual consta de una gran cantidad de tweets relacionados con el cambio climático etiquetados de manera manual (estándar de oro) y automática (estándar de plata) con variables tales como sentimiento, postura, agresividad, tópico, geolocalización (en algunos casos), desvíos de la temperatura histórica, tópico asociado e información sobre eventos de desastres ambientales, entre otras variables. A continuación se describe el proceso seguido por los autores del dataset para la obtención y el etiquetado de los datos.

## 4.1. Tweets

Los tweets utilizados para la construcción del dataset se obtuvieron al combinar 3 datasets disponibles públicamente: “Credibility of Climate Change Denial in Social Media Data” [2], “Climate Change Tweets IDs Data” [3] y “Twitter Archive Data” [4].

El dataset “Credibility of Climate Change Denial in Social Media Data” contiene 14.353.859 identificadores de tweets que fueron recolectados entre 2006 y 2018, basándose en un filtro de búsqueda que establece que cada tweet contenga los términos “climate change”, “#climatechange”, “global warming” o “#globalwarming”.

El dataset “Climate Change Tweets IDs Data” se obtuvo del repositorio “Harvard Dataverse”. Este dataset contiene 39.622.026 identificadores de tweets relacionados con el cambio climático que fueron recolectados entre 2017 y 2019 a través de la API de Twitter. Estos tweets fueron filtrados usando los siguientes términos: “#climatechange”, “#climatechangeisreal”, “#climatechangehoax”, “#actonclimate”, “#globalwarming”, “#climatedeniers”, “#climatechangeisfalse”, “#globalwarminghoax”, “#climatechangenotreal”, “climate change”, “global warming”, y “climate hoax”.

Finalmente para llenar un vacío en el último dataset mencionado se recolectaron tweets del Archivo de Internet [4] entre enero del 2019 y octubre del 2019 usando como filtro los términos “#climatechange” y “#globalwarming”.

Basado en los términos utilizados como filtro para obtener los tweets el dataset resultante consta de tweets escritos mayormente en el idioma inglés (un 98% de los tweets), por lo que se puede esperar un sesgo con respecto a la ubicación de las personas que publican los tweets.

Como las reglas de Twitter no permiten publicar el texto de los tweets almacenados en este punto el dataset solo contiene los identificadores de los tweets, por lo que para obtener el texto y los metadatos se lleva a cabo un proceso de “hidratación” a través del cual se recupera la información de cada tweet a partir de su identificador. Debido a que los tweets pueden haber sido eliminados o no estar disponibles por varias razones, no todos los tweets pueden recuperarse. Finalmente, tras completar el proceso de hidratación y aplicar un filtrado por palabras clave, el dataset resultante consta de 15.789.411 tweets entre el 6 de junio de 2006 y el 1 de octubre de 2019.

A partir de los 15.789.411 tweets con su texto y metadatos se procedió con el proceso de etiquetado de los tweets sobre las variables de geolocalización, género, postura, sentimiento, agresividad, temperatura y tópico. A continuación se describe la estrategia seguida para el etiquetado de las distintas variables.

### 4.1.1. Geolocalización

Los autores consideraron crucial tener datos geolocalizados lo cual podría contribuir a obtener hallazgos espaciales. Los tweets pueden incluir información geográfica solo si el usuario habilita esta opción en su cuenta de Twitter. Lamentablemente

solo un pequeño porcentaje de los usuarios tienen esta opción habilitada. En este dataset solo el 0.22% de los tweets contienen información geográfica.

Esta pequeña cantidad de tweets con información geográfica resulta insuficiente para realizar algún tipo de análisis espacial, por lo que se debieron utilizar otros metadatos para extraer información espacial.

Otra fuente de información importante sobre la localización de los tweets es la ubicación del usuario en Twitter, la cual es añadida manualmente a su perfil a fin de enriquecerlo con información personal sobre dónde vive. En este dataset el 75.45% de los tweets están acompañados con este tipo de datos. Esta ubicación pude ser engañosa en algunos casos ya que en general no hace referencia a dónde se encontraba el usuario al momento de publicar un tweet sino a dónde afirma que vive.

A partir de esta información se utilizó la base de datos GeoNames [5], que contiene más de 11 millones de nombres de lugares y sus coordenadas, para establecer la localización geográfica de los usuarios. En el proceso se excluyeron ubicaciones muy generales, ubicaciones escritas completamente en minúsculas y algunas palabras específicas (como Cielo o Tierra). Como resultado se logró establecer la localización de 5.307.538 tweets, quedando indefinida la localización de los 10.481.873 tweets restantes.

#### 4.1.2. Género

Para obtener el género de la persona que publicó un tweet se utilizó un clasificador de género implementado usando LSTM (Long Short-Term Memory) multicapa a nivel de carácter. Este clasificador fue entrenado con datos de la Administración del Seguro Social de Estados Unidos (versión 2017) que incluye 32.634 nombres y un conjunto de datasets publicados por Milos Bejda [6] que incluye unos 130.000 nombres. El clasificador toma el nombre de usuario como entrada y produce como salida la posibilidad de que el usuario sea masculino o femenino. Solo se consideraron los resultados en los que la posibilidad de un determinado género fuera mayor a 95%, caso contrario se etiquetó como indefinido. El modelo entrenado se evaluó con el dataset de la Administración del Seguro Social de Estados Unidos (versión 2020) dando como resultado un 90% de exactitud. De los tweets presentes en el dataset se identificaron 10.307.402 escritos por hombres, 4.895.134 escritos por mujeres y 586.875 indefinidos.

#### 4.1.3. Postura

La postura identifica si la persona que publicó el tweet cree en el cambio climático producido por el hombre (believer), niega el cambio climático producido por el

hombre (denier) o no lo apoya ni lo niega (neutral). Para realizar esta clasificación se utilizó Aprendizaje por Transferencia (Transfer Learning) entrenando al clasificador con un dataset publicado en la plataforma Kaggle [7] que contiene 34.667 tweets etiquetados sobre el cambio climático recolectados entre 2015 y 2018, los cuales fueron etiquetados por 3 anotadores independientes. Se entrenaron y evaluaron varios modelos utilizando este dataset y se utilizó el modelo con mejores resultados (BERT [8]) para clasificar todo el dataset. Como resultado 11.292.424 tweets fueron clasificados como creyentes, 1.191.386 como negacionistas y 3.305.601 como neutrales.

#### 4.1.4. Sentimiento

Se utilizaron técnicas de aprendizaje automático no supervisado para determinar el sentimiento, particularmente dos enfoques *lexicon-based* utilizando VADER [9] y Textblob [10], así como un modelo de Red Neuronal Recurrente preentrenado y un modelo BERT [8] preentrenado utilizando el framework Flair [11]. Todos estos enfoques producen como salida un valor continuo en la escala de -1 a 1, donde valores cercanos a -1 representan un sentimiento negativo, valores cercanos a 0 neutralidad y valores cercanos a 1 un sentimiento positivo.

#### 4.1.5. Agresividad

Para medir la agresividad de los tweets se utilizó Aprendizaje por Transferencia de la misma manera que con la postura. Para entrenar a los clasificadores en este caso se combinaron dos datasets, uno proveniente de una competencia sobre Análisis de Sentimiento en Twitter de Analytics Vidhya [12], el cual contiene tweets etiquetados según sean considerados expresiones de odio o no. El segundo dataset se tomó de una tarea de SemEval, el cual consiste de textos cortos etiquetados según su humor y su ofensividad. El dataset resultante consta de 4.500 tweets no agresivos, 2.242 tweets que expresan odio y 2.258 textos cortos ofensivos. Luego de entrenar y evaluar los distintos modelos con este dataset se utilizó el mejor modelo (BERT) para etiquetar los tweets, dando como resultado 4.527.267 tweets agresivos y 11.262.144 no agresivos.

#### 4.1.6. Temperatura

En esta etapa se utilizó el dataset de temperatura de la superficie terrestre de Berkeley Earth [13], el cual provee información detallada en tiempo y espacio sobre la desviación de temperatura mínima, promedio y máxima diaria para cada celda de  $1^\circ \times 1^\circ$  de latitud-longitud sobre la tierra. Dicha desviación se calcula en relación a la temperatura promedio entre 1951 y 1980. A partir de la latitud y longitud de un

tweet y su fecha de publicación, el valor de la desviación de temperatura se puede obtener directamente de este dataset. Al haber sólo 5,307,538 tweets etiquetados con su latitud y longitud, la desviación de temperatura sólo pudo obtenerse para esa cantidad de tweets.

#### 4.1.7. Tópico

Para el modelado de tópicos se utilizó el modelo LDA (Asignación Latente de Dirichlet), el cual busca asignar tweets a tópicos, donde un tópico consiste de combinaciones probabilísticas de palabras que representan tendencias de co-ocurrencia de palabras. Para cada tweet se asignan múltiples tópicos con una determinada probabilidad, el tópico con mayor probabilidad será el asignado a cada tweet.

Para el correcto modelado de tópicos fue importante realizar un preprocesamiento de los textos, comenzando por la eliminación de URLs, caracteres especiales, apariciones de “www”, menciones de usuarios, direcciones de email y palabras de un único carácter. Luego se convirtieron los tweets a minúsculas y se eliminaron las palabras vacías estándar del inglés más otras palabras definidas por los autores (say, get, know, may, etc.). También se eliminaron palabras con menos de 10 ocurrencias a lo largo del dataset. Finalmente las palabras restantes fueron lematizadas y los unigramas y bigramas fueron alimentados al algoritmo de LDA.

A la hora de ejecutar el algoritmo de LDA el parámetro más importante es el número de tópicos  $k$  que el modelo descubrirá. En principio no se conoce cuál será el mejor valor de  $k$  debido a que un  $k$  muy pequeño puede producir que varios tópicos bien diferenciados se combinen en uno y un  $k$  muy grande puede introducir más tópicos de los que realmente tienen sentido. Por lo tanto se hicieron pruebas con 10, 15 y 20 tópicos, obteniendo mejores resultados con 10. Los 10 tópicos encontrados y asignados a los correspondientes tweets del dataset son: 'Weather Extremes', 'Importance of Human Intervention', 'Seriousness of Gas Emissions', 'Ideological Positions on Global Warming', 'Impact of Resource Overconsumption', 'Global stance', 'Politics', 'Significance of Pollution Awareness Events', 'Donald Trump versus Science' y 'Undefined / One Word Hashtags'.

## 4.2. Desastres Ambientales

Además de los tweets sobre el cambio climático, los autores también publicaron un listado de desastres ambientales con datos provenientes de la Base de Datos Internacional de Desastres (EM-DAT) [14], la cual contiene más de 24.000 desastres históricos organizados por categorías. De esta base de datos se tomaron solo los desastres ambientales ocurridos en el mismo período abarcado por los tweets del dataset, llegando a un total de 4.913 eventos, los cuales incluyen datos como: tipo de desastre, nombre del evento, país, coordenadas, fechas de inicio y fin, número de

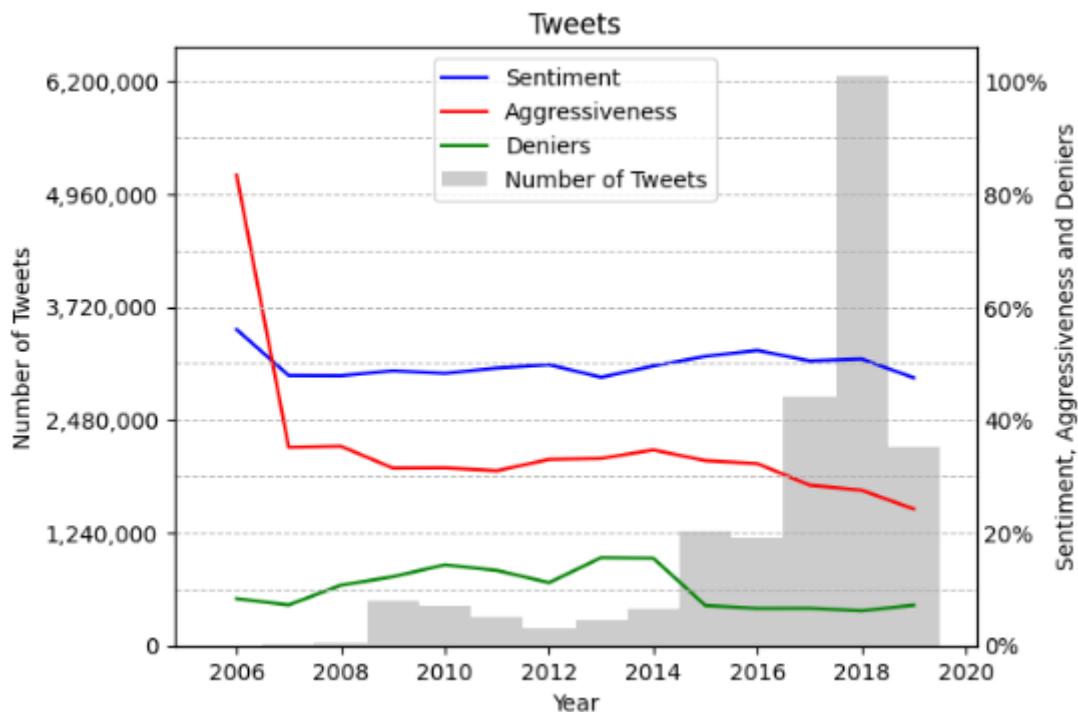
muertes, número de afectados y daño económico en dólares. Los tipos de desastre incluidos son: terremoto, tormenta, actividad volcánica, inundación, deslizamiento de tierra, temperatura extrema, movimiento en masa, incendio forestal y sequía.

## 5. Análisis Exploratorio

La primera etapa de este trabajo consiste en un análisis exploratorio de los datos, a través del cual se busca empezar a entender los datos, su distribución y plantear algunas hipótesis. Para comenzar se analizan los posibles valores de las variables de interés dentro del dataset, los cuales coinciden con los descritos en la sección anterior.

A continuación se desarrolló una serie de gráficos que buscan mostrar los valores del sentimiento promedio, porcentaje de tweets agresivos, porcentaje de negacionistas y cantidad de tweets, en relación a las fechas de los tweets y a los distintos tópicos asociados.

En un primer gráfico se grafican la cantidad de tweets y los valores de las variables de interés a lo largo de los años.



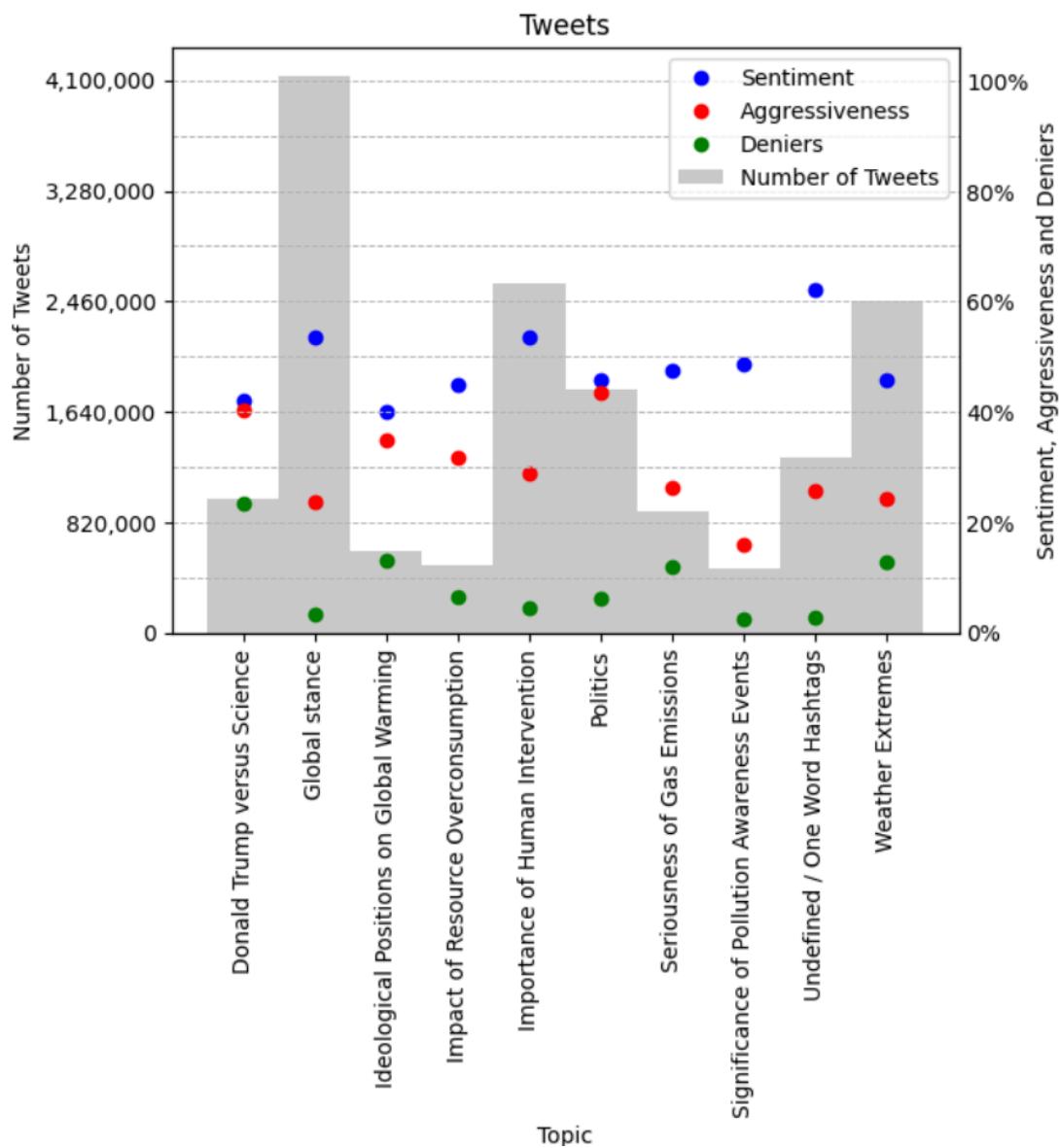
En este gráfico se puede ver la cantidad de tweets, representada por las barras grises, y los años sobre el eje X. Además se muestran los valores promedio de las distintas variables a través de los años, el sentimiento promedio en azul, la cantidad promedio de tweets agresivos en rojo y el promedio de tweets negacionistas en verde. Nótese que los valores de sentimiento fueron mapeados directamente del rango [-1, 1] al rango [0%, 100%].

Este gráfico nos permite apreciar que el sentimiento promedio se ha mantenido generalmente neutral con el paso del tiempo, mientras que la agresividad y el porcentaje de negacionistas ha caído en los últimos años.

Otro dato que llama la atención es el crecimiento en la cantidad de tweets sobre el cambio climático en los últimos años, con un pico particular en 2018 superando los 6 millones de tweets, aunque esto probablemente se deba a la forma en la que se recolectaron los tweets, ya que una gran cantidad de tweets provienen del dataset “Climate Change Tweets IDs Data”, que consta de una extensa lista de tweets pero acotados al rango de tiempo entre 2017 y 2019.

Además se ve un pico destacable en el porcentaje de tweets agresivos en 2006, superando el 80%, pero esto se debe a que sólo se tienen 11 tweets pertenecientes a ese año, por lo que la muestra no resulta significativa.

Luego se graficaron los tweets agrupados por tópico en lugar de año.



Aquí se puede ver que el tópico Postura Global (Global Stance) es el tópico con mayor cantidad de tweets asociados, más de 4 millones, seguido por la Importancia de la Intervención Humana (Importance of Human Intervention) y Extremos Climáticos (Weather Extremes) con cerca de 2,5 millones de tweets cada uno.

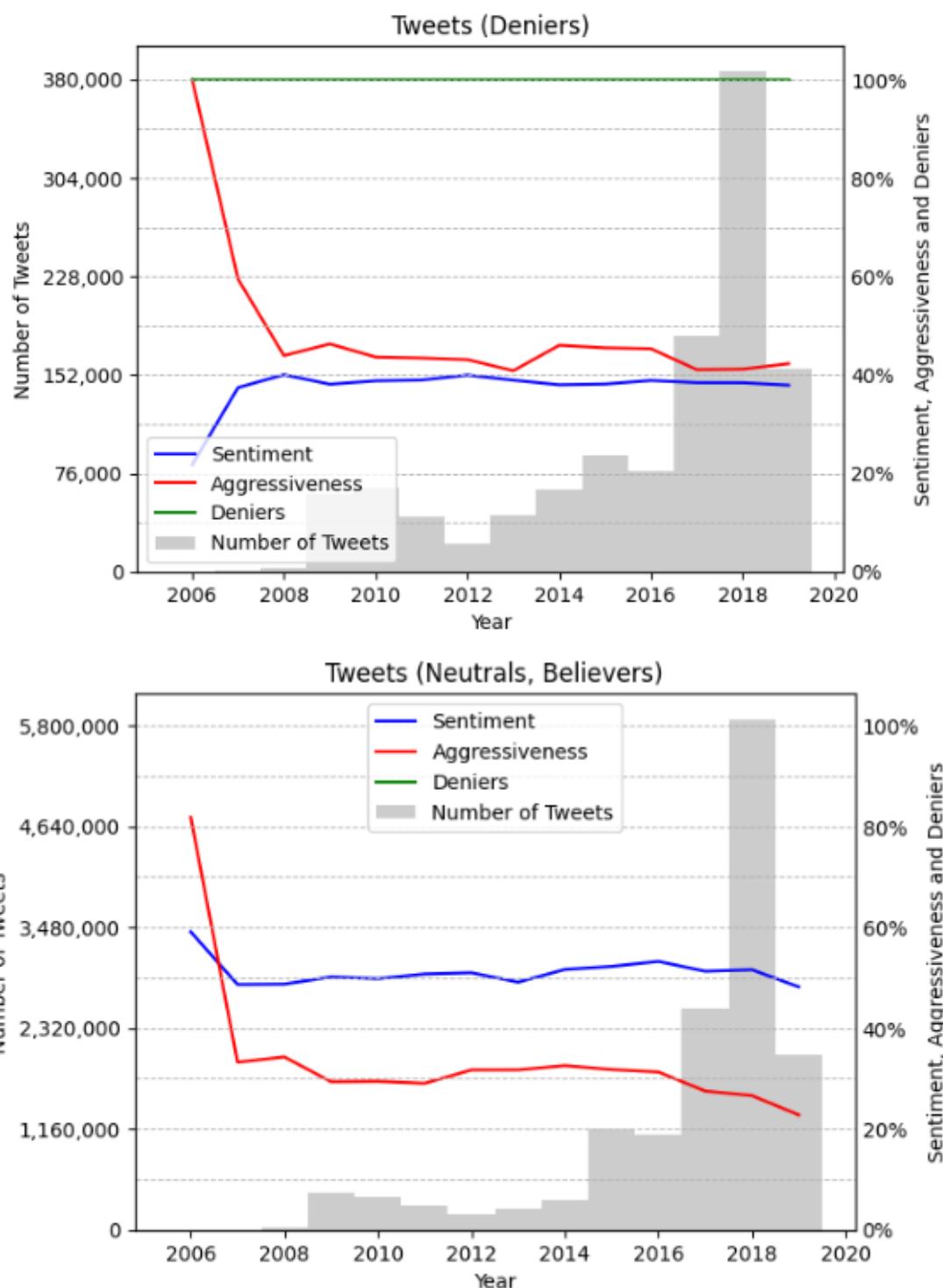
Se puede apreciar que el tópico con mayor porcentaje de agresividad es Política (Politics), por encima del 40%, lo cual parece lógico ya que no es extraño que las personas aumenten su nivel de agresividad al hablar de política, seguido por el tópico Donald Trump versus la Ciencia (Donald Trump versus Science) que es a su vez el tópico con mayor porcentaje de negacionistas, lo cual también es esperable ya que Donald Trump es un reconocido negacionista del cambio climático.

También se puede ver que el tópico Indefinido / Hashtags de Una Palabra (Undefined / One Word Hashtags) es el tópico con sentimiento promedio más positivo, aunque resulta complicado de analizar ya que es un tópico que no tiene un significado asociado, seguido por Postura Global y la Importancia de la Intervención Humana.

Además pareciera haber algún tipo de relación entre el porcentaje de tweets agresivos y el porcentaje de negacionistas, ya que ambas parecen moverse de manera similar, con la excepción del tópico Política, donde la agresividad se dispara sin un porcentaje importante de negacionistas.

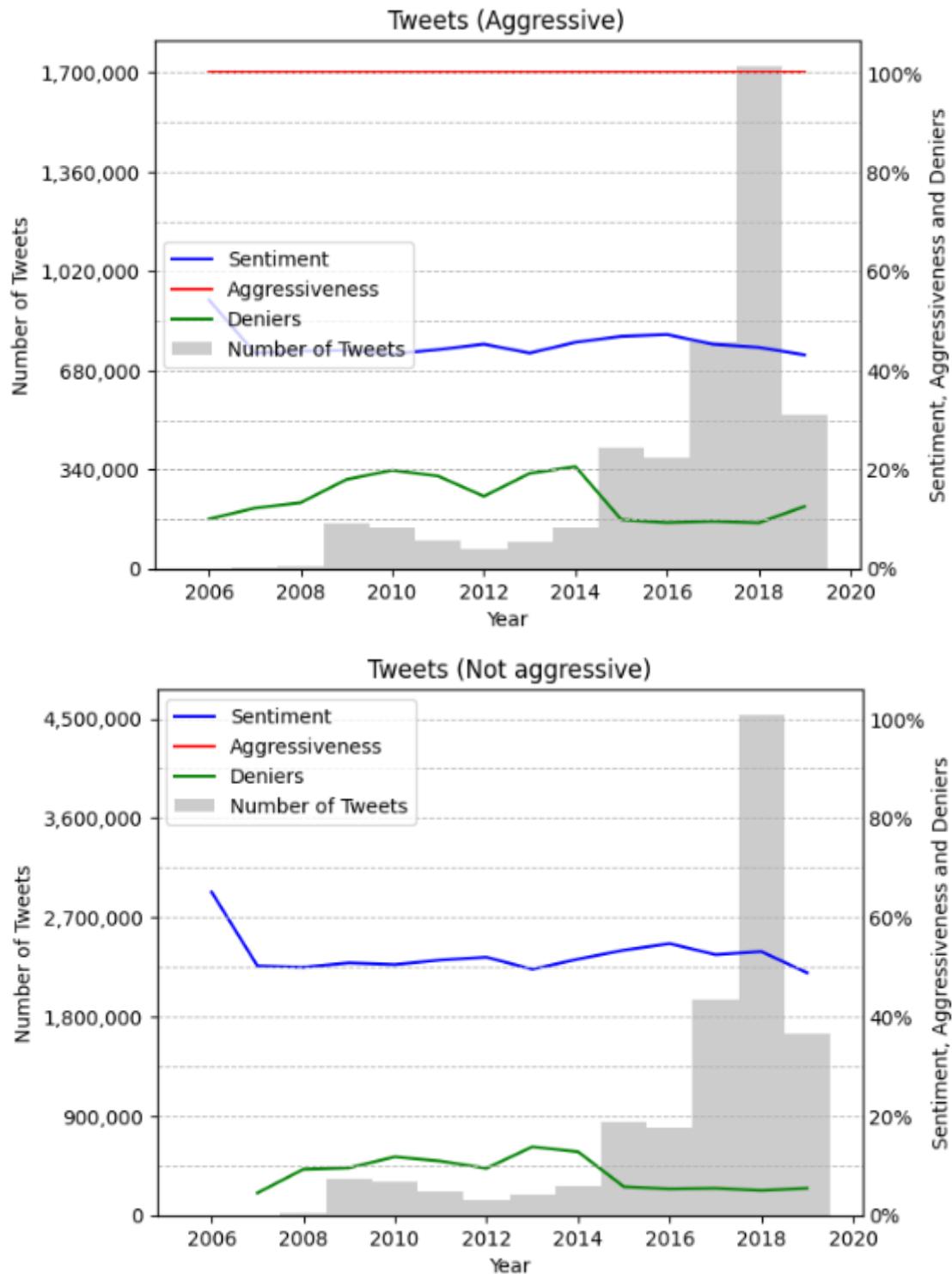
Luego se realizaron gráficos trabajando con subconjuntos de los datos, mostrando los valores de las variables a través del tiempo en detalle para cada año por separado y para cada tópico. Estos gráficos no se incluyen en esta sección debido a su gran cantidad, pero pueden encontrarse en el Apéndice A.

A continuación se realizó una comparación entre los tweets de postura negacionista contra los tweets de postura neutral o creyente.



Comparando estos gráficos se puede ver que los negacionistas (primer gráfico) se han mantenido más agresivos a lo largo de los años, entre un 40% y 50% de tweets agresivos, mientras que los creyentes y neutrales (segundo gráfico) han mantenido alrededor de un 30% de tweets agresivos llegando a bajar casi a un 20% en los últimos años. Además se ve que los negacionistas tienen un sentimiento promedio más negativo que los creyentes/neutrales.

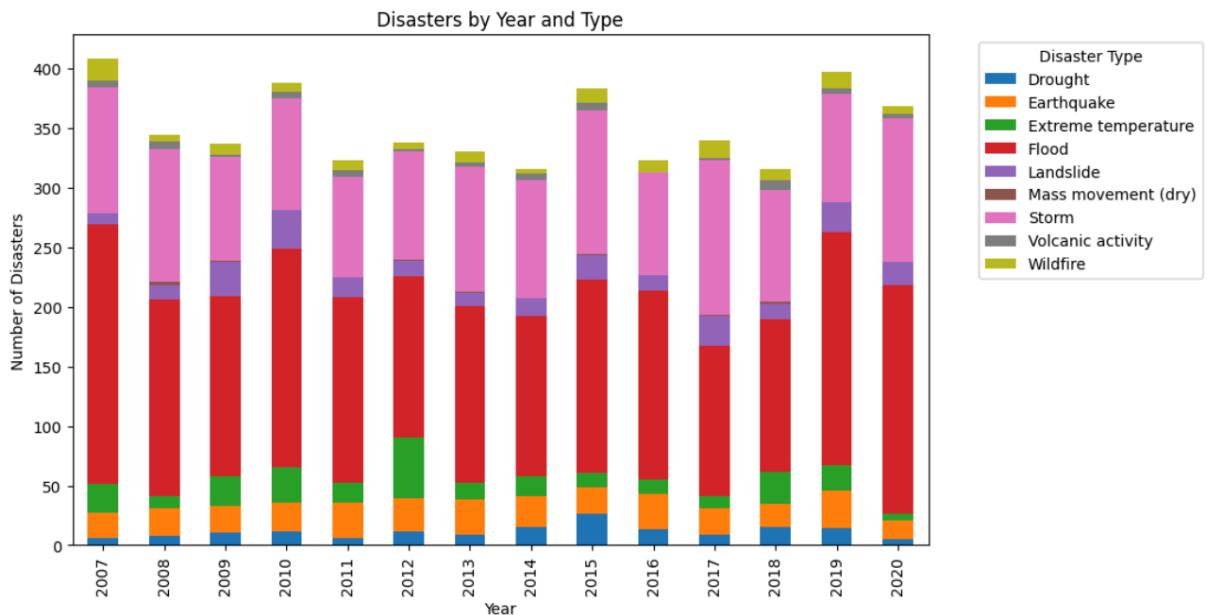
También se realizó el mismo tipo de análisis comparando los tweets agresivos y los tweets no agresivos.



Los resultados obtenidos de este par de gráficos son similares a los obtenidos en el anterior, se puede apreciar que los tweets agresivos tienen un mayor porcentaje de negacionistas que los no agresivos y que su sentimiento es ligeramente más negativo.

Para el dataset de desastres ambientales también se generó un gráfico simple que permite ver la distribución de los distintos tipos de desastres ambientales contemplados

a lo largo de los años. Nótese que para este gráfico, por simplicidad, sólo se consideró la fecha de inicio del desastre, a pesar de que un desastre puede haber abarcado más de un año.



De este gráfico se destaca que se cuenta con una cantidad considerable de desastres por cada año, superando todos los años los 300 desastres ambientales. También se puede ver que en todos los años hay una alta cantidad de tormentas e inundaciones, siendo ambos tipos de desastres que suelen relacionarse con el cambio climático.

## 6. Procesamiento de Datos

Luego de este primer análisis realizado durante la etapa anterior se prosiguió con una etapa de procesamiento de datos, en la cual primero se buscó filtrar los datos que no se consideraron relevantes o que podrían introducir ruido de manera innecesaria y luego se manipularon los datos a fin de llevarlos a un formato que facilite su uso en etapas posteriores, mediante un proceso de discretización y agrupamiento.

### 6.1. Filtrado

En el primer paso de filtrado se eliminaron las variables referentes a la latitud, longitud y desviación de temperatura debido al gran porcentaje de valores nulos y el género debido a que no se consideró relevante para este trabajo.

Luego se descartaron los tweets del año 2006, debido a que solo se cuenta con 11 tweets de este año, estando muy por debajo de la cantidad de tweets recolectada en los otros años, por lo que se consideró una muestra no significativa que podría introducir ruido innecesariamente para las siguientes etapas.

También se descartaron los tweets del tópico Indefinido / Hashtags de Una Palabra debido a que como tópico no tiene un significado relevante y que al contener tweets

que constan de un hashtag de una única palabra las predicciones realizadas por los modelos para las distintas variables podrían no ser muy precisas.

Luego sobre el dataset de desastres ambientales se decidió eliminar los desastres ambientales ocurridos antes del 2007 o después del 2019, para que de esta manera coincida el período de tiempo en el cual se tiene información sobre los tweets y sobre los desastres ambientales.

Además se descartaron los desastres ambientales que se consideraron no relacionados directamente al cambio climático: terremotos, actividad volcánica, deslizamientos de tierra y movimientos en masa, conservando los desastres que podrían relacionarse al cambio climático: tormentas, inundaciones, temperaturas extremas, incendios forestales y sequías.

Finalmente, debido a que el dataset cuenta con desastres ambientales de cualquier escala producidos en cualquier parte del mundo y que los tweets, al ser todos tweets escritos en idioma inglés, provendrían en su mayoría de países donde este idioma tenga mayor influencia, se consideró que desastres ambientales de pequeña escala en cualquier parte del mundo probablemente no tengan impacto sobre la forma de opinar de la mayoría de la gente respecto al cambio climático. Por esto se decidió filtrar los desastres ambientales de pequeña escala usando como referencia el número de muertos, el número de personas afectadas y el costo total de los daños producidos. En un principio se definieron umbrales para cada una de estas variables, descartando los desastres ambientales que no superaran los valores definidos para alguna de las variables, pero esto producía un incremento en la duración promedio de los desastres, debido a que los desastres de mayor duración tienden a producir un mayor número de muertos, afectados y daños. Como se prefirió trabajar con desastres ambientales de una duración no tan extensa, ya que sería difícil analizar en qué momento un desastre que duró varios meses pudo haber tenido impacto sobre la opinión de las personas en Twitter, se modificó la estrategia de filtrado para que en lugar de comparar el valor de una determinada variable contra un umbral definido se compare el valor de la variable dividido por la duración del desastre contra dicho umbral, de esta manera se logró que el filtrado produzca un conjunto de desastres ambientales de gran escala con una duración promedio de 5 días.

## 6.2. Discretización y Agrupamiento

Luego del filtrado se continuó con un proceso de discretización y agrupamiento, necesario para facilitar el trabajo en las etapas posteriores. En primer lugar se tomó la variable sentimiento, que es la única de las variables de interés con valores continuos, y se transformaron sus valores a un rango discreto. Se mapearon los valores del rango [-1, -0.35] a 'Sentimiento Negativo', del rango (-0.35, 0.35) a 'Sentimiento Neutral' y del rango [0.35, 1] a 'Sentimiento Positivo'.

Luego se agruparon los tweets por día, buscando obtener una serie de tiempo que caracterice de forma sencilla la forma de opinar de la gente en cada día. Por simplicidad se decidió que los valores de esta serie de tiempo sean binarios (0 o 1) y se definieron las siguientes variables:

- Sentimiento Positivo: la cual valdrá 1 en un determinado día si el porcentaje de tweets etiquetados como 'Sentimiento Positivo' para ese día supera un umbral (40%) y 0 en caso contrario.
- Sentimiento Negativo: la cual valdrá 1 en un determinado día si el porcentaje de tweets etiquetados como 'Sentimiento Negativo' para ese día supera un umbral (40%) y 0 en caso contrario.
- Negacionistas: la cual valdrá 1 en un determinado día si el porcentaje de tweets etiquetados como 'Negacionista' para ese día supera un umbral (20%) y 0 en caso contrario.
- Agresividad: la cual valdrá 1 en un determinado día si el porcentaje de tweets etiquetados como 'Agresivo' para ese día supera un umbral (40%) y 0 en caso contrario.
- Además para cada tópico se creó una variable (por ejemplo '[T] Extremos Climáticos'), la cual valdrá 1 en un determinado día si el porcentaje de tweets asignados al tópico correspondiente para ese día supera un umbral (25%) y 0 en caso contrario. A las variables correspondientes a los tópicos se les agregó el prefijo “[T]” para diferenciarlas fácilmente del resto.

Los valores que se decidieron usar como umbrales para cada variable fueron establecidos a partir de los valores promedio de cada una de las variables, posicionándolos un poco por encima de la media. De esta manera si, por ejemplo, un día tiene un valor 1 para la variable Sentimiento Positivo esto indica que el porcentaje de tweets con sentimiento positivo ese día supera con cierto margen el porcentaje promedio de tweets con sentimiento positivo a lo largo de todo el dataset.

Luego se generó una serie de tiempo para caracterizar los desastres ambientales ocurridos en cada día. Nuevamente se decidió trabajar con valores binarios por simplicidad. En este caso las variables definidas para esta serie de tiempo fueron basadas en los distintos tipos de desastres ambientales incluidos en el dataset: tormentas, inundaciones, temperaturas extremas, incendios forestales y sequías. Luego, para cada día en la serie de tiempo la variable asociada a cada tipo de desastre ambiental valdrá 1 si durante ese día se estaba desarrollando un desastre de ese tipo o 0 en caso contrario. A estas variables se les agregó el prefijo “[D]” para diferenciarlas. Finalmente se combinaron las dos series de tiempo en una única serie de tiempo que caracteriza para cada día tanto la forma de opinar de las personas en Twitter como los desastres ambientales en desarrollo.

	Sentimiento Positivo	Sentimiento Negativo	Negacionistas	Agresividad	[T] Climas Extremos	[T] Postura Global	:	[T] Política	[T] Donald Trump versus la Ciencia	[D] Tormenta	[D] Inundación	[D] Temperatura Extrema	[D] Incendio Forestal	[D] Seguía
06/01/2007	0	1	0	1	1	0	...	0	1	1	0	0	1	0
07/01/2007	0	1	1	1	1	0	...	1	0	1	0	0	0	0
08/01/2007	1	0	1	0	0	1	...	1	0	1	0	1	0	0
09/01/2007	0	0	0	0	1	1	...	0	0	0	1	0	0	1
10/01/2007	1	1	1	1	0	1	...	0	0	0	0	0	0	0
...							...							

En este ejemplo se puede ver el formato de la serie de tiempo resultante. A partir de este fragmento se entiende que, por ejemplo, el día 06/01/2007 hubo un alto porcentaje de tweets con sentimiento negativo y también un alto porcentaje de tweets agresivos, que se habló mucho sobre los tópicos 'Extremos Climáticos' y 'Donald Trump versus la Ciencia' y que se estaba ocurriendo al menos una tormenta y un incendio forestal.

También se crearon otras series de tiempo similares siguiendo el mismo proceso pero agrupando los tweets por semana o por mes en lugar de por día.

## 7. Conjuntos de Ítems Frecuentes y Reglas de Asociación

Para esta etapa se realizó un análisis de conjuntos de ítems frecuentes y reglas de asociación sobre el conjunto de tweets, mediante el cual se buscó encontrar relaciones entre las distintas variables que muestren qué conjuntos de etiquetas suelen estar presentes simultáneamente en un tweet de manera frecuente y cuándo la presencia de una etiqueta o conjunto de etiquetas implica un incremento en la posibilidad de que otra etiqueta determinada esté presente.

Para este análisis se analizaron los tweets de manera individual, es decir sin realizar ningún agrupamiento de tweets por período de tiempo.

El análisis de conjuntos de ítems frecuentes suele presentarse en el contexto de carritos de compras de distintos tipos de mercados, donde un carrito (o transacción) representa una compra en particular, compuesta por los distintos ítems o elementos comprados. En este contexto el objetivo del análisis de conjuntos de ítems frecuentes es encontrar qué conjuntos de elementos suelen comprarse juntos en una gran cantidad de ocasiones. Para el caso de estudio de este trabajo los tweets serían equivalentes a los carritos y las distintas etiquetas (sentimiento positivo, agresivo, negacionista, etc.) equivaldrían a los ítems, por lo que el objetivo será encontrar que conjuntos de etiquetas suelen estar presentes en gran cantidad de tweets y cómo se relacionan entre sí.

Dado un conjunto de ítems  $C$ , su apoyo es la cantidad de transacciones que contienen a  $C$ .  $C$  será un conjunto de ítems frecuentes si su apoyo supera un determinado umbral de apoyo  $a$ .

Las reglas de asociación relacionan un conjunto de ítems  $C$  con un ítem  $i$ , esta asociación, notada como  $C \rightarrow i$ , implica que si el conjunto de ítems  $C$  está presente en una transacción, entonces es considerablemente probable que también esté presente  $i$ .

Se define la confianza de una regla  $C \rightarrow i$  como la relación entre el apoyo de  $C \cup \{i\}$  y el apoyo de  $C$ . Por ejemplo, si la regla {'Sentimiento Negativo', 'Negacionista'} → 'Agresivo' tiene una confianza de 0.35, esto implicaría que de los tweets etiquetados como 'Sentimiento Negativo' y 'Negacionista' el 35% también estará etiquetado como 'Agresivo'.

## 7.1. A-Priori

A la hora de definir algoritmos para encontrar conjuntos de ítems frecuentes y reglas de asociación a partir de estos conjuntos, el uso de memoria del algoritmo toma gran relevancia, debido a que puede ser altamente costoso mantener en memoria la cuenta de ocurrencias para todos los posibles conjuntos de ítems. Es por esto que en muchos casos, el uso de algoritmos triviales para contar las apariciones de los distintos conjuntos de ítems no serán viables, debido a que el uso de memoria puede sobrepasar fácilmente la cantidad de memoria disponible en el sistema. El algoritmo A-Priori resuelve el problema de encontrar conjuntos de ítems frecuentes haciendo un uso eficiente de la memoria del sistema, almacenando la cuenta de ocurrencias de cada conjunto de ítems de manera inteligente y evitando utilizar espacio para contar conjuntos de ítems que no podrían ser frecuentes. En este trabajo se utilizó una implementación de este algoritmo publicada en la librería de python efficient-apriori, la cual calcula tanto conjuntos de ítems frecuentes como reglas de asociación. Para la ejecución del algoritmo se estableció un apoyo mínimo del 1% y una confianza mínima del 35%, esto permite disminuir el número de conjuntos y reglas que producirá como

salida el algoritmo, dejando fuera conjuntos y reglas con poco apoyo o confianza que se considerarían menos interesantes.

## 7.2. Conjuntos de Ítems Frecuentes

El resultado de la ejecución del algoritmo A-Priori sobre el dataset de tweets es un listado de conjuntos de ítems frecuentes. Para cada conjunto de ítems se detalla el conjunto en sí, definido en este caso como una tupla de etiquetas, su tamaño y su apoyo, que representa qué porcentaje del total de tweets se encuentran etiquetados simultáneamente por todas las etiquetas del conjunto. A continuación se analizarán los distintos conjuntos de ítems frecuentes encontrados, divididos por tamaño.

En primer lugar se pueden ver los conjuntos de tamaño 1, los cuales si bien no son muy relevantes para analizar relaciones entre las distintas etiquetas, permiten ver qué porcentaje de tweets abarca cada etiqueta por separado.

Conjunto	Apoyo
No Agresivo	0,7106
Creyente	0,7071
Sentimiento Neutral	0,4118
Sentimiento Negativo	0,3119
Agresivo	0,2894
[T] Postura Global	0,2855
Sentimiento Positivo	0,2763
Neutral	0,2131
[T] Importancia de la Intervención Humana	0,1792
[T] Extremos Climáticos	0,1702
[T] Política	0,1249
Negacionista	0,0797
[T] Donald Trump versus la Ciencia	0,0688
[T] Seriedad de las Emisiones de Gas	0,0624
[T] Posiciones Ideológicas sobre el Calentamiento Global	0,0416
[T] Impacto del Sobreconsumo de Recursos	0,0343
[T] Importancia de Eventos de Concientización sobre la Contaminación	0,0331

Con tamaño 1 se han encontrado 17 conjuntos de ítems frecuentes, lo cual coincide con la cantidad de etiquetas originales ya que todas las etiquetas están presentes en al menos un 1% de los tweets, es decir que superan el apoyo mínimo establecido para la ejecución del algoritmo.

Se puede ver que las etiquetas 'No Agresivo' y 'Creyente' son las dos más comunes, estando presentes cada una en más del 70% de los tweets. También se puede apreciar la distribución de tópicos, siendo 'Postura Global' el tópico más común. Por último se

destaca que la cantidad de tweets etiquetados como 'Negacionista' es relativamente baja, apenas por debajo del 8%.

A continuación se muestran algunos de los conjuntos de ítems frecuentes de tamaño 2 con más apoyo. Estos conjuntos sí podrían resultar útiles para comenzar a analizar relaciones entre las distintas etiquetas ya que incluyen más de una.

Conjunto	Apoyo
Creyente, No Agresivo	0,5105
Sentimiento Neutral, No Agresivo	0,293
Creyente, Sentimiento Neutral	0,288
Creyente, [T] Postura Global	0,236
No Agresivo, Sentimiento Positivo	0,2213
No Agresivo, [T] Postura Global	0,2179
Creyente, Sentimiento Negativo	0,2143
Creyente, Sentimiento Positivo	0,2048
Agresivo, Creyente	0,1966
Sentimiento Negativo, No Agresivo	0,1963
Neutral, No Agresivo	0,1539
Creyente, [T] Importancia de la Intervención Humana	0,1427
No Agresivo, [T] Extremos Climáticos	0,1288
No Agresivo, [T] Importancia de la Intervención Humana	0,1277
Agresivo, Sentimiento Neutral	0,1188
Sentimiento Neutral, [T] Postura Global	0,118
...	...

En este caso se han encontrado 79 conjuntos de tamaño 2 que superan el 1% de apoyo. Viendo los conjuntos del comienzo de la tabla se puede apreciar que las etiquetas 'No Agresivo' y 'Creyente' están presentes en los 14 conjuntos con mayor apoyo, lo cual es esperable ya que ambas etiquetas por separado aparecen en más del 70% de los tweets, pero no se considera que brinden información muy relevante ya que, por ejemplo, el conjunto {'Creyente', '[T] Postura Global'} tiene un apoyo del 23% pero es difícil determinar solo con este análisis si esto se debe a que existe una relación entre estas etiquetas o sólo a que la etiqueta 'Creyente' está presente en más del 70% de los tweets y por lo tanto la combinación de esta etiqueta con otras etiquetas frecuentes también es relativamente frecuente.

Debido a esto se filtraron los conjuntos de ítems que contenían alguna de las 3 etiquetas más comunes ('No Agresivo', 'Creyente' y 'Sentimiento Neutral'), quedando finalmente 39 conjuntos de ítems frecuentes de tamaño 2 y 16 de tamaño 3.

Conjunto	Apoyo
Agresivo, Sentimiento Negativo	0,1156
Sentimiento Positivo, [T] Postura Global	0,1014
Agresivo, [T] Postura Global	0,0676
Sentimiento Negativo, [T] Postura Global	0,0661
Neutral, [T] Extremos Climáticos	0,065
Sentimiento Negativo, [T] Extremos Climáticos	0,0633
Neutral, Sentimiento Positivo	0,0632
Sentimiento Positivo, [T] Importancia de la Intervención Humana	0,0617
Agresivo, Neutral	0,0592
Sentimiento Negativo, Neutral	0,058
Agresivo, Sentimiento Positivo	0,055
Agresivo, [T] Política	0,0542
Agresivo, [T] Importancia de la Intervención Humana	0,0515
Sentimiento Negativo, [T] Política	0,0425
Sentimiento Negativo, [T] Importancia de la Intervención Humana	0,042
Agresivo, [T] Extremos Climáticos	0,0414
...	...

Conjunto	Apoyo
Agresivo, Sentimiento Negativo, [T] Postura Global	0,0218
Agresivo, Sentimiento Negativo, [T] Política	0,0217
Neutral, Sentimiento Positivo, [T] Extremos Climáticos	0,021
Agresivo, Sentimiento Negativo, Neutral	0,0204
Agresivo, Negacionista, Sentimiento Negativo	0,0185
Agresivo, Sentimiento Negativo, [T] Extremos Climáticos	0,018
Sentimiento Negativo, Neutral, [T] Extremos Climáticos	0,0179
Agresivo, Sentimiento Positivo, [T] Postura Global	0,0179
Agresivo, Sentimiento Negativo, [T] Importancia de la Intervención Humana	0,0165
Agresivo, Neutral, [T] Extremos Climáticos	0,0159
Agresivo, Sentimiento Negativo, [T] Donald Trump versus la Ciencia	0,0156
Agresivo, Sentimiento Positivo, [T] Importancia de la Intervención Humana	0,0135
Agresivo, Neutral, Sentimiento Positivo	0,0121
Neutral, Sentimiento Positivo, [T] Postura Global	0,0118
Agresivo, Neutral, [T] Postura Global	0,0113
Negacionista, Sentimiento Negativo, [T] Extremos Climáticos	0,0105

Si bien los resultados obtenidos con conjuntos de ítems de tamaño 2 y 3 permiten realizar algún análisis sobre cómo se relacionan las distintas etiquetas, resulta complicado sacar conclusiones categóricas ya que un apoyo elevado de por sí sólo no

determina una relación fuerte entre las etiquetas del conjunto. Por ejemplo, el segundo conjunto de tamaño 2 con mayor apoyo luego de aplicar el filtrado es {'Sentimiento Positivo', '[T] Postura Global'} con un apoyo de 0.1014, pero es difícil determinar sólo con esta información si es el caso de que ambas etiquetas están fuertemente relacionadas o son dos etiquetas independientes que son frecuentes por sí solas y también lo son en conjunto. Este conjunto llama particularmente la atención porque dos lugares más abajo se encuentra el conjunto {'Sentimiento Negativo', '[T] Postura Global'} con un apoyo de 0.0676, lo cual puede resultar extraño ya que, si bien es posible, sería esperable que al estar la etiqueta '[T] Postura Global' fuertemente relacionada con 'Sentimiento Positivo' entonces no lo esté con su contraparte 'Sentimiento Negativo'.

Es por este motivo que tal vez el análisis de conjuntos de ítems frecuentes por sí solo no sea muy adecuado para analizar relaciones entre las distintas etiquetas, pero es un paso necesario para el descubrimiento de reglas de asociación, las cuales brindan información mucho más concreta para este caso de estudio y se presentarán a continuación.

### 7.3. Reglas de Asociación

La implementación del algoritmo A-Priori utilizada produce, además de los conjuntos de ítems frecuentes, las reglas de asociación generadas a partir de los conjuntos encontrados.

Para cada regla de asociación  $C \rightarrow i$  se incluyen varias medidas que permiten describir que tan fuerte es esa regla:

- Apoyo: mide la frecuencia de aparición del conjunto  $C \cup \{i\}$
- $$Apoyo(C \rightarrow i) = \frac{\text{Ocurrencias de } C \cup \{i\}}{\text{Tamaño del dataset}}$$
- Confianza: porcentaje de veces que la regla es verdadera dentro del dataset.
- $$Confianza(C \rightarrow i) = \frac{\text{Ocurrencias de } C \cup \{i\}}{\text{Ocurrencias de } C}$$
- Elevación: relación entre las ocurrencias observadas de  $C \cup \{i\}$  y las ocurrencias que se esperarían en caso de que  $C$  e  $i$  fueran independientes
- $$Elevación(C \rightarrow i) = \frac{Apoyo(C \rightarrow i)}{Apoyo(C) * Apoyo(i)}$$

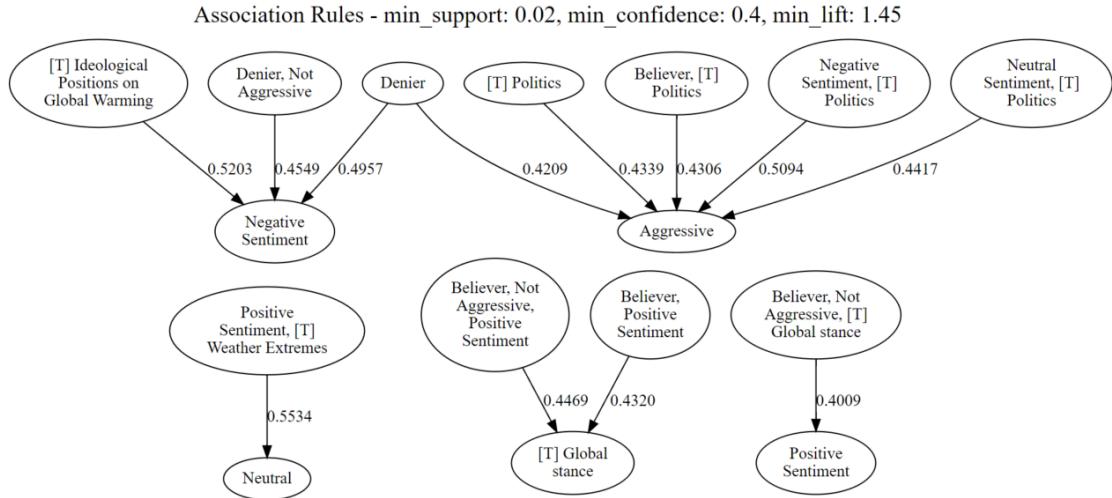
Cabe destacar que una elevación igual a 1 implica que el antecedente y el consecuente son independientes, una elevación mayor a 1 implica una dependencia positiva y una elevación menor a 1 implica que el antecedente tiene un impacto negativo sobre el consecuente. Por ejemplo, una elevación de 1.5 implica que dado  $C$ ,  $i$  tiene un 50% más de probabilidades de darse que en caso de ser independientes.

Las reglas de asociación encontradas se dividieron en 3 grupos de interés en función de los distintos valores para las medidas descritas previamente. Estos grupos de reglas

se muestran a continuación utilizando una representación de grafo, donde los nodos origen representan antecedentes, los nodos destino representan consecuentes y los pesos asociados a los arcos representan la confianza de las reglas.

El primer grupo de reglas incluye a todas las reglas que cumplen las siguientes condiciones:

- Apoyo > 0.02: el antecedente y el consecuente están presentes en simultáneo dentro del dataset en al menos el 2% de los casos.
- Confianza > 0.4: la regla es verdadera en al menos el 40% de los casos.
- Elevación > 1.45: dado el antecedente, el consecuente es al menos un 45% más probable que en caso de ser independientes.



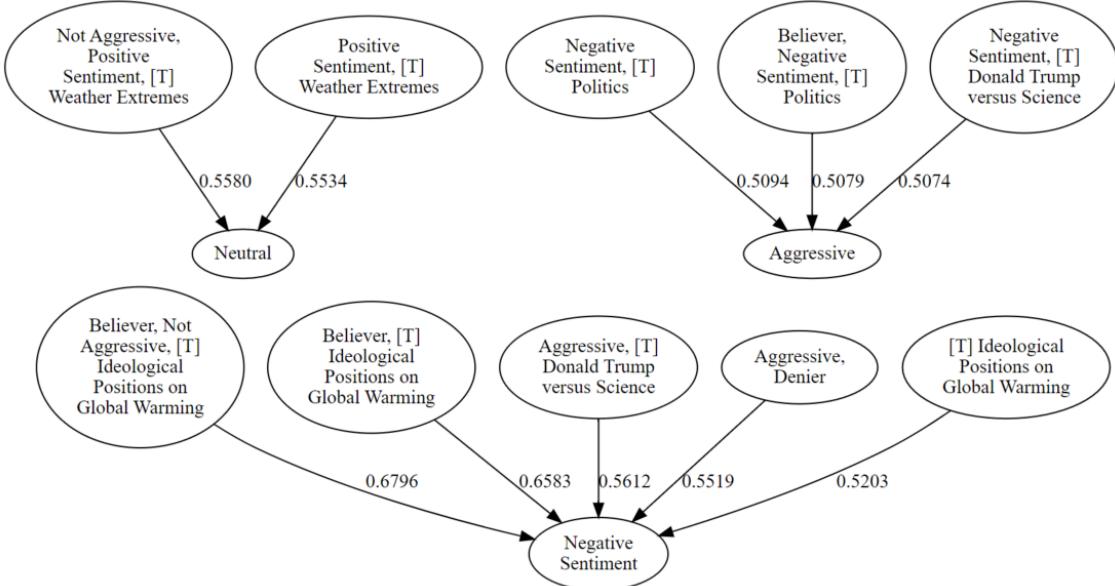
Este grupo incluye reglas relativamente frecuentes y con una elevación que las hace bastante significativas. Se destacan algunas como:

- {'Sentimiento Positivo', '[T] Extremos Climáticos'} → 'Neutral', esta regla tiene una confianza del 55%, es decir que de los tweets etiquetados como 'Sentimiento Positivo' y '[T] Extremos Climáticos' el 55% también están etiquetados como 'Neutral', lo cual es interesante ya que sólo el 20% del total de tweets están etiquetados como 'Neutral'.
- {'Negacionista'} → 'Agresivo' y {'Negacionista'} → 'Sentimiento Negativo', lo que coincide con lo descubierto durante el análisis exploratorio respecto a que los negacionistas suelen ser más agresivos y tener un sentimiento más negativo.
- {'Creyente', 'No Agresivo', '[T] Postura Global'} → 'Sentimiento Positivo', que es la única regla de este grupo con 'Sentimiento Positivo' como consecuente, cuando la mayoría de las reglas encontradas tienen como consecuente 'Sentimiento Negativo' y 'Agresividad'.

Para el segundo grupo de reglas se disminuyó el apoyo mínimo a 0.01 y se aumentó la confianza mínima a 0.5, esto produce como resultado un conjunto de reglas que

pueden no ser tan frecuentes pero que describen reglas en las que cuando se cumple el antecedente hay más de un 50% de probabilidad de que también se cumpla el consecuente.

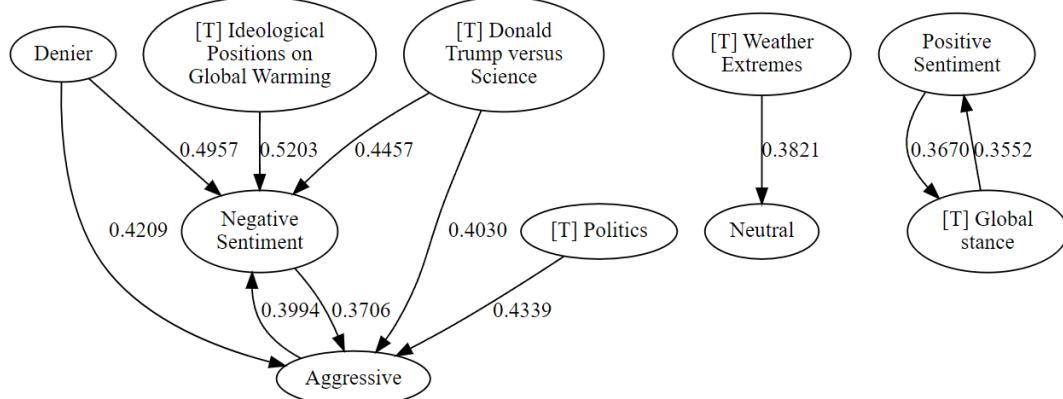
Association Rules - min\_support: 0.01, min\_confidence: 0.5, min\_lift: 1.45



Se destaca particularmente la regla {'Creyente', 'No Agresivo', '[T] Posiciones Ideológicas sobre el Calentamiento Global'}→'Sentimiento Negativo', con una confianza de 0.68 es la regla con mayor confianza dentro de los tres grupos, cumpliéndose en un 68% de los casos.

Por último, para el tercer grupo de reglas, se estableció un apoyo mínimo de 0.02, una confianza mínima de 0.35, una elevación mínima de 1.25 y se agregó la restricción de que el antecedente se conforme por solo una etiqueta. Esto produce un conjunto de reglas que en general pueden no ser tan fuertes pero resultan más fáciles de analizar ya que muestran relaciones entre pares de etiquetas.

Association Rules - min\_support: 0.02, min\_confidence: 0.35, min\_lift: 1.25, max\_size: 2



De este grupo se destaca la relación bidireccional entre 'Sentimiento Positivo' y '[T] Postura Global' y las reglas {'[T] Donald Trump versus la Ciencia'}→'Sentimiento Negativo' y {'[T] Donald Trump versus la Ciencia'}→'Agresividad', que muestran una

comportamiento similar entre el tópico Donald Trump versus la Ciencia y la postura negacionista con respecto al sentimiento negativo y la agresividad. Cabe recordar que el tópico Donald Trump versus la Ciencia es el tópico con mayor cantidad de tweets negacionistas.

## 8. Análisis de Causalidad

En esta última etapa se utilizaron varias técnicas para encontrar relaciones causales entre las distintas etiquetas asociadas a los tweets y los distintos desastres ambientales. A nivel general lo que se buscó con este análisis fue encontrar patrones que indiquen que cuando una determinada variable cambia, esto produce con el paso del tiempo un cambio sobre otra variable.

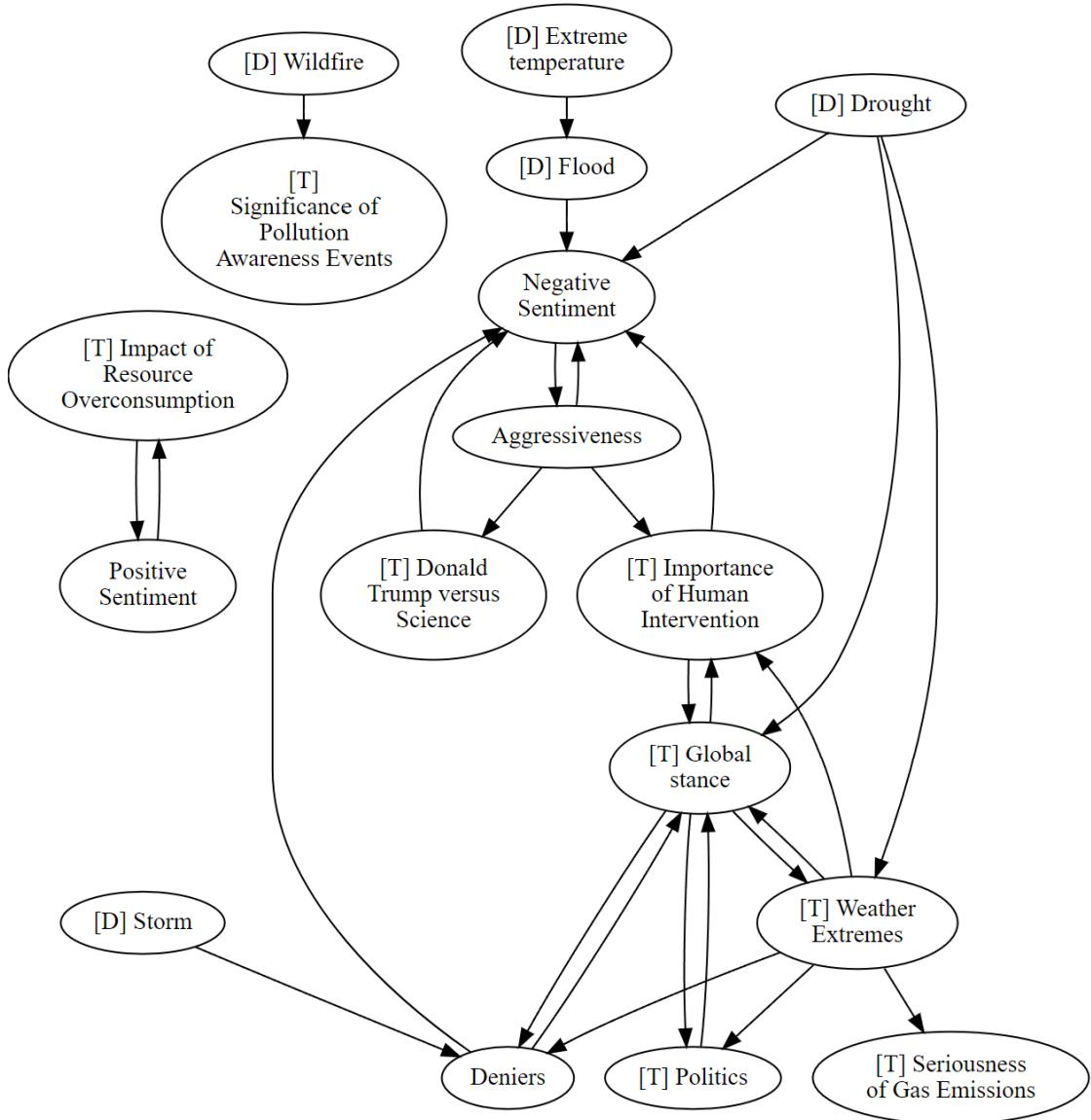
A continuación se describen los pasos seguidos para la obtención de dichas relaciones, partiendo de las series de tiempo obtenidas en la etapa de procesamiento de datos, aplicando sobre ellas varios algoritmos para el descubrimiento de estructuras causales y recopilando sus resultados en una única estructura que modela las relaciones causales en forma de grafo.

### 8.1. Algoritmos de Causalidad

El primer paso es ejecutar varios algoritmos para el descubrimiento de estructuras causales. Para la elección de los algoritmos a utilizar se tomó como base el trabajo publicado por investigadores de la Universidad Nacional del Sur [15], donde se aplican 9 técnicas de aprendizaje de estructuras causales sobre conjuntos de datos sintéticos y se compara el desempeño de cada una. Sobre los datos sintéticos los algoritmos con mejor desempeño general fueron DirectLiNGAM [16], PC [17], PCMCI [18] y VAR [19].

Luego de ejecutar las cuatro técnicas mencionadas sobre la serie de tiempo de tweets y desastres ambientales, se toman los resultados de cada una para construir grafos de causalidad, donde cada nodo representa una etiqueta y los arcos representan las relaciones causales encontradas. Por ejemplo, a continuación se muestra el resultado obtenido por la técnica DirectLiNGAM:

## DirectLiNGAM Causal Graph, granularity=day



En este grafo se pueden ver las relaciones causales encontradas al aplicar la técnica de DirectLiNGAM sobre la serie de tiempo de tweets y desastres ambientales agrupados por día. Cada arco  $A \rightarrow B$  representa una relación causal entre dos variables, donde se entiende que un cambio en A produce con el paso del tiempo un cambio sobre B.

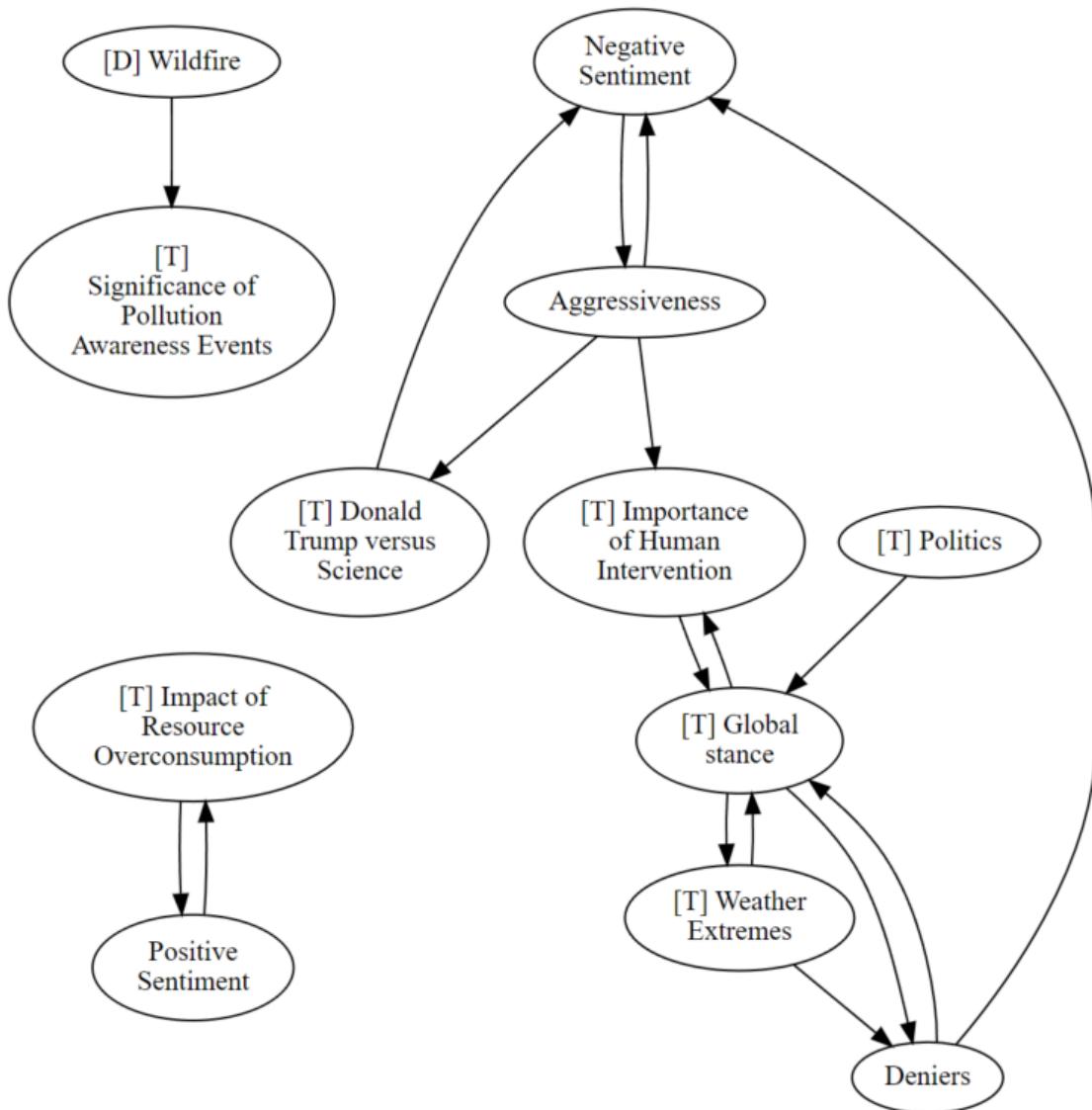
## 8.2. ensemble

El siguiente paso es combinar los cuatro grafos producidos por el paso anterior en uno. Para esto se utiliza una estrategia de votación por consenso basada en el trabajo previamente mencionado, denominada ensemble, que establece que un arco estará presente en el grafo final si ese mismo arco está presente en los cuatro grafos fuente. Con el fin de flexibilizar este método se modificó la definición, estableciendo que un arco estará presente en el grafo resultante si ese mismo arco está presente en al menos  $k$  de los grafos fuentes, para algún  $k$  entre 1 y 4 (cantidad de grafos fuente).

Cabe destacar que utilizar  $k = 4$  es equivalente a la descripción inicial y que  $k = 1$  incluye todos los arcos presentes en cualquiera de los grafos.

Tras aplicar el algoritmo de ensemble con  $k = 4$  sobre los cuatro grafos producidos en el paso anterior se obtuvo el siguiente resultado.

Causal Graph -  $k=4$ , granularity=day



En este grafo se pueden apreciar las relaciones causales en las que las 4 técnicas aplicadas coincidieron, por lo que se consideran las relaciones más fuertes encontradas. Si bien este resultado sería suficiente para analizar las relaciones causales entre las distintas variables, este trabajo se limita a presentar el resultado sin realizar ningún análisis, debido a la falta de conocimiento sobre el tema.

### 8.3. Curado de Grafos

Durante varias de las pruebas realizadas se observó que algunas de las relaciones causales encontradas parecían no tener demasiado sentido, como alguna relación que

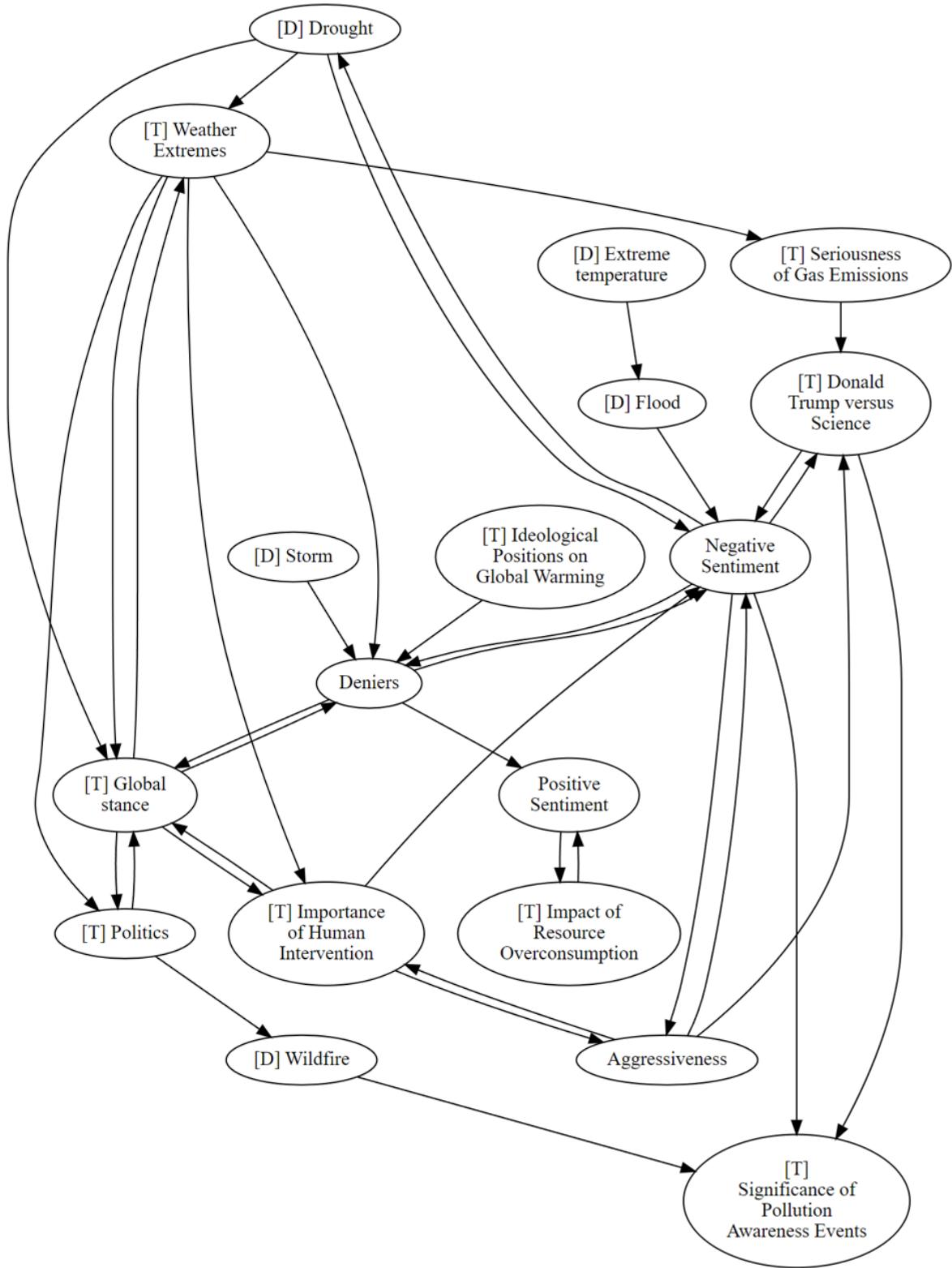
indicaba que algún tipo de sentimiento era causa de algún tipo de desastre ambiental. Por este motivo se decidió aplicar una etapa de curado a los grafos, en la cual se eliminarían arcos que parecieran no tener sentido conociendo el dominio de los datos. Luego de discutir el tema con expertos en el área de causalidad se decidió revertir la eliminación de los arcos y solo notarlos como arcos dudosos, ya que si varios de los algoritmos concluyeron que existe una relación causal entre las distintas variables es muy probable que esta exista por más que parezca extraña.

En el caso del grafo mostrado previamente como resultado de aplicar el algoritmo de ensemble con  $k = 4$  no se presentó ninguna relación de las que se considerarían dudosas por lo que el proceso de curado no produciría ninguna modificación sobre el resultado obtenido.

Luego de este proceso se ensayaron otras configuraciones, modificando el valor de  $k$ , para analizar qué relaciones causales se encontraban al relajar la exigencia del algoritmo, y también modificando la granularidad de la agrupación de los tweets de día a semana, a fin de encontrar relaciones que requieran de una visión más general a lo largo del tiempo.

Tras repetir el proceso de generación del grafo de causalidad utilizando  $k = 3$  se obtuvo el siguiente resultado:

Causal Graph - k=3, granularity=day

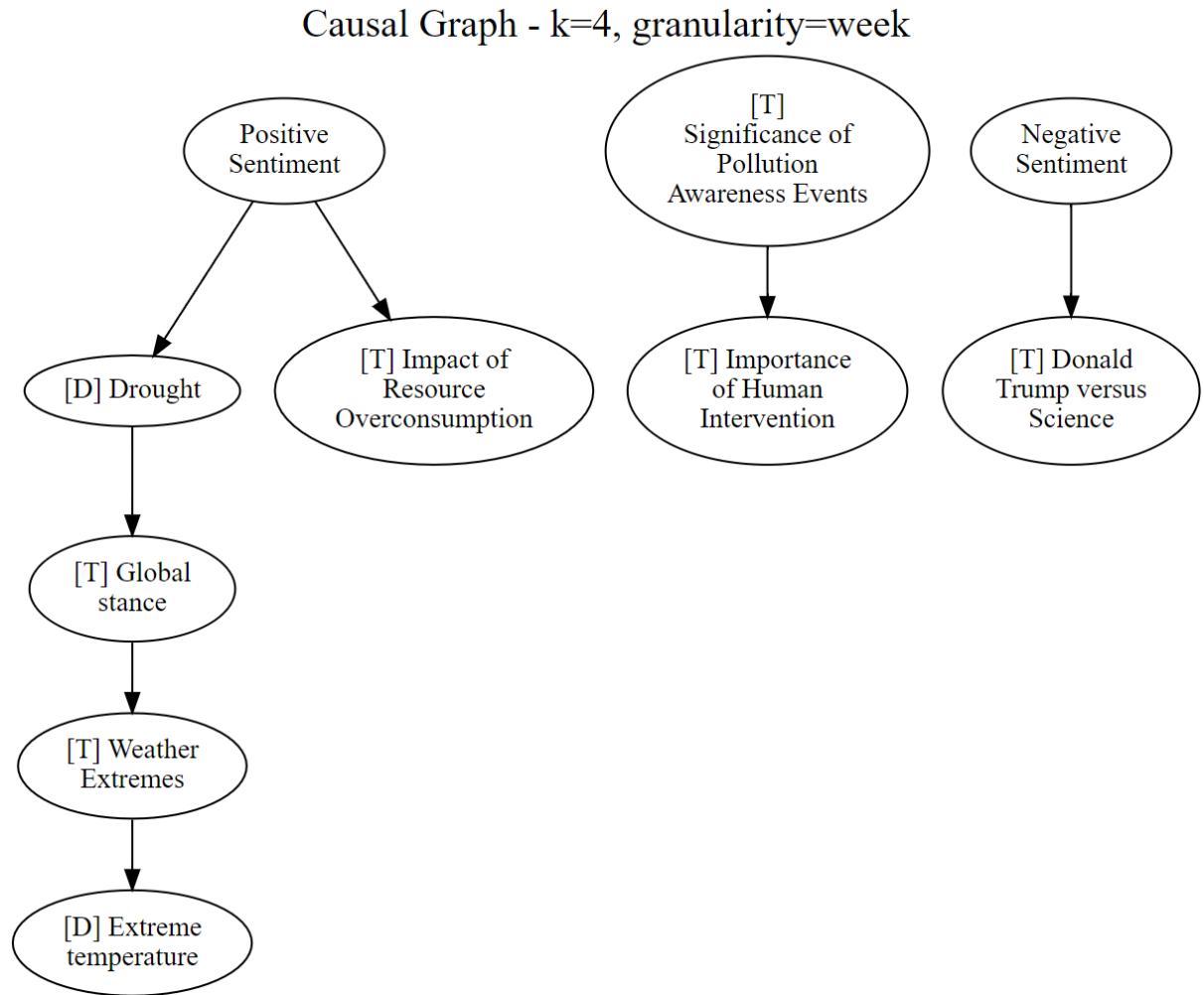


Resulta notable que al disminuir la exigencia del algoritmo de ensemble, permitiendo arcos que estén presentes en sólo 3 de los grafos fuente, la cantidad de arcos crece considerablemente. También se destaca una mayor presencia de nodos relacionados con desastres ambientales, estando presentes todos los desastres ambientales mientras que en la ejecución con  $k = 4$  solo se encontraba presente '[D] Incendio Forestal'. Además

aparecen arcos del tipo mencionado como “dudoso” en la etapa de curado, como '[T] Política' → '[D] Incendio Forestal'.

Al reducir el valor de  $k$  a 2 o 1 los grafos producidos contienen una cantidad demasiado grande de arcos, haciendo que sea imposible analizar el grafo resultante. Es por esto que solo se muestran los resultados producidos con valores de  $k$  4 y 3.

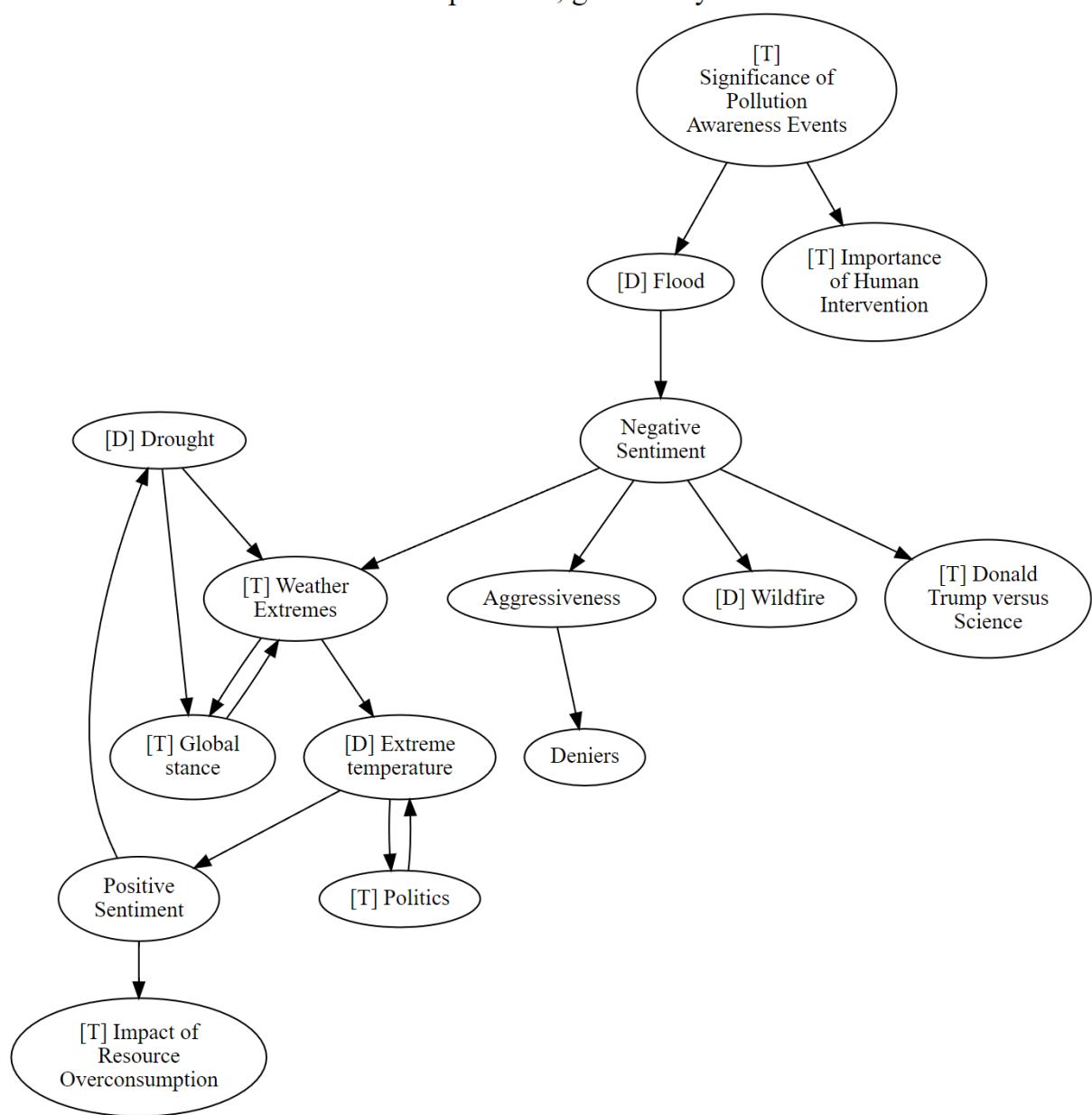
Luego de ejecutar los algoritmos utilizando la serie de tiempo de tweets y desastres ambientales agrupados por día se hicieron pruebas utilizando la serie de tiempo agrupada por semana. Para  $k = 4$ :



Se puede ver que al cambiar la granularidad de la serie de tiempo de día a semana la cantidad de arcos se reduce a menos de la mitad y además aparecen un par de arcos ‘dudosos’: ‘Sentimiento Positivo’ → ‘[D] Sequía’ y ‘[T] Extremos Climáticos’ → ‘[D] Temperatura Extrema’.

Finalmente al cambiar el valor de  $k$  de 4 a 3 se ven efectos similares a los vistos con la granularidad diaria, se incrementan los arcos, aparecen más desastres ambientales y también más arcos “dudosos”.

Causal Graph - k=3, granularity=week



## 9. Conclusión

Este trabajo me permitió aplicar conocimientos obtenidos durante la carrera de Licenciatura en Ciencias de la Computación, así como adquirir nuevos conocimientos y herramientas. Se presentaron varios desafíos propios de trabajar con herramientas y temas desconocidos, los cuales pudieron ser superados gracias a las habilidades desarrolladas en la universidad.

La herramienta desarrollada permite encontrar relaciones entre ciertas variables que describen la forma de opinar de las personas respecto al cambio climático en la plataforma Twitter, tomando también en consideración registros de desastres ambientales. Si bien la herramienta produce varios resultados que permiten analizar el comportamiento de las personas en Twitter con respecto al cambio climático, prefiero no sacar conclusiones respecto a estos resultados debido a que no poseo conocimiento del tema y el objetivo de esta herramienta es brindar información enfocada a expertos para su posterior análisis.

Se destaca como limitación principal de la herramienta desarrollada el hecho de que esta no está preparada para analizar los tweets a partir de su texto original, sino que depende de un dataset donde para cada tweet ya se hayan calculado las distintas variables a analizar.

## 10. Trabajo a Futuro

Si bien existe una gran cantidad de mejoras posibles al trabajo presentado, a continuación se mencionan algunas de las ideas de mejoras o expansiones que surgieron durante el desarrollo, tanto por parte del autor como de expertos que aportaron a este trabajo.

- Mejorar la modularización del pipeline y generalizar las operaciones realizadas en cada etapa, de manera que permita adaptar fácilmente la herramienta para tomar como entrada otros tipos de datos distintos y encontrar relaciones entre sus variables.
- Expandir el trabajo realizado en la etapa de Conjuntos de Ítems Frecuentes y Reglas de Asociación, utilizando las series de tiempo agrupadas por día y por semana, para hacer un análisis que incluya los desastres ambientales y considere el estado global de las discusiones sobre el cambio climático en determinado período de tiempo en lugar de sólo considerar los tweets de manera individual independientemente del contexto.
- Agregar a los 4 métodos de descubrimiento de estructuras causales utilizados en la etapa de Análisis de Causalidad un nuevo método que permita, en los casos donde se encuentran arcos bidireccionales, determinar cuál es la dirección donde la relación causal es más fuerte y de esa manera evitar los arcos de ida y vuelta.
- Utilizar otros modelos ensembles que combinen predicciones de manera ponderada.
- Incorporar algún método que permita determinar, dentro de los grafos causales, de qué manera afecta un cambio en una variable causa sobre una variable efecto. Por ejemplo, dada la relación causal  $A \rightarrow B$ , determinar si un incremento en A produciría un incremento o un decremento en B.
- Tomar distintas submuestras para identificar períodos en los que se observan ciertas relaciones causales y otros en los que no.
- Realizar pruebas adicionales con otras granularidades temporales.
- Utilizar otras fuentes de información para complementar la disponible (por ejemplo, si se usa la información de geolocalización cuando está disponible se podría realizar un análisis de tendencias climáticas en una región o áreas de interés).
- Incluir nuevas técnicas de minería de datos para lograr obtener otro tipo de relaciones entre las variables del dataset.
- Consultar con algún experto en cambio climático para que su experiencia pueda ayudar a interpretar los resultados.

## 11. Bibliografía

- [1] Effrosynidis, D., Karasakalidis, A. I., Sylaios, G., & Arampatzis, A. (2022). The climate change Twitter dataset. *Expert Systems with Applications*, 204, 117541.
- [2] Samantray, A., & Pin, P. (2019). Credibility of climate change denial in social media. *palgrave communications*, 5(1).
- [3] Littman, J., & Wrubel, L. (2019). Climate change tweets ids. Harvard Dataverse, 580.
- [4] Internet Archive n.d., Archive Team: The Twitter Stream, <<https://archive.org/details/twitterstream>>.
- [5] GeoNames n.d., GeoNames, <<https://www.geonames.org/>>.
- [6] Milos Bejda n.d., <<https://mbejda.github.io/>>.
- [7] Edward Qian 2019, Twitter Climate Change Sentiment Dataset, <<https://www.kaggle.com/datasets/edqian/twitter-climate-change-sentiment-dataset>>.
- [8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [9] Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225).
- [10] Loria, S. (2018). textblob Documentation. Release 0.15, 2(8), 269.
- [11] Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019, June). FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)* (pp. 54-59).
- [12] Analytics Vidhya 2018, Twitter Sentiment Analysis, <<https://datahack.analyticsvidhya.com/contest/practice-problem-twitter-sentiment-analysis/>>.
- [13] Rohde, R. A., & Hausfather, Z. (2020). The Berkeley Earth land/ocean temperature record. *Earth System Science Data*, 12(4), 3469-3479.
- [14] Centre for Research on the Epidemiology of Disasters n.d., Public EM-DAT platform, <<https://public.emdat.be/data>>.

- [15] Maisonnave, M., Delbianco, F., Tohme, F., Milios, E., & Maguitman, A. G. (2022). Causal graph extraction from news: a comparative study of time-series causality learning techniques. PeerJ Computer Science, 8, e1066.
- [16] Shimizu, S., Inazumi, T., Sogawa, Y., Hyvarinen, A., Kawahara, Y., Washio, T., ... & Hoyer, P. (2011). DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. Journal of Machine Learning Research-JMLR, 12(Apr), 1225-1248.
- [17] Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. Social science computer review, 9(1), 62-72.
- [18] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. Science advances, 5(11), eaau4996.
- [19] Sims, C. A. (1980). Macroeconomics and reality. Econometrica: journal of the Econometric Society, 1-48.

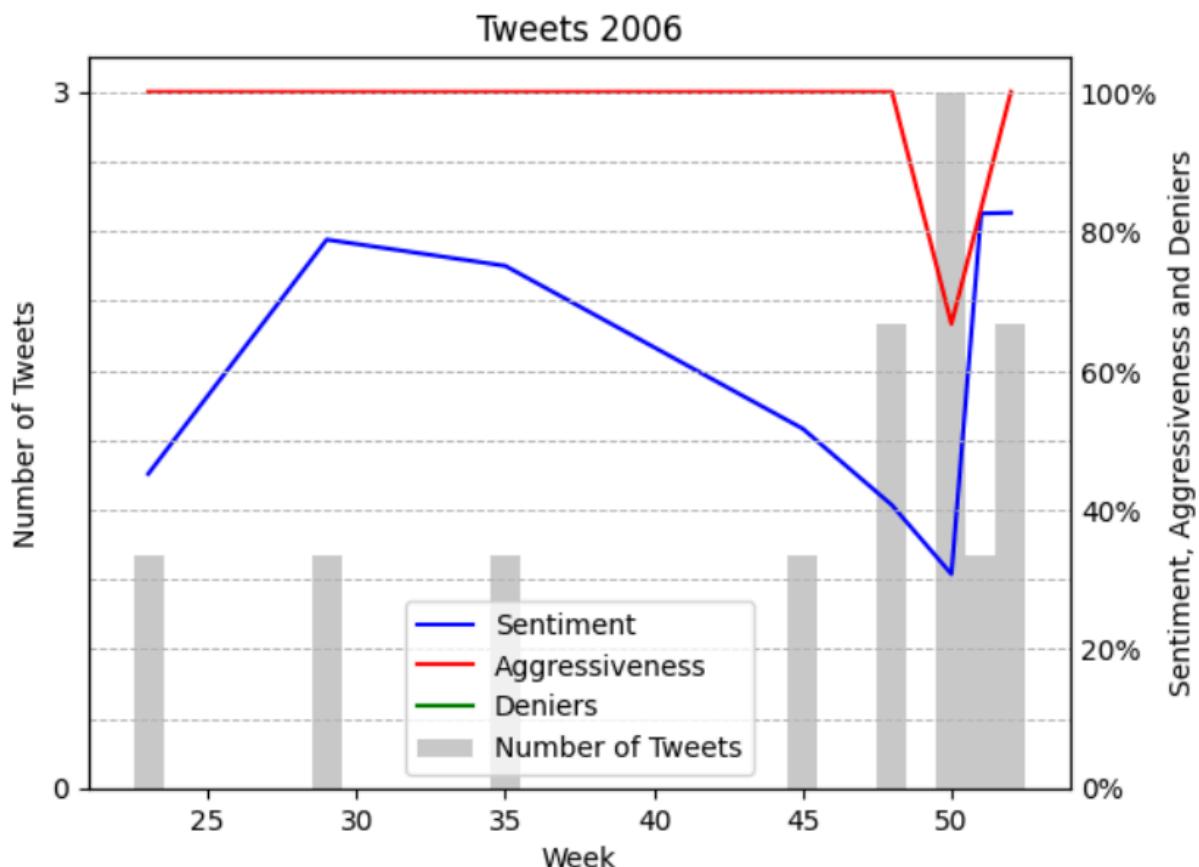
# Apéndice

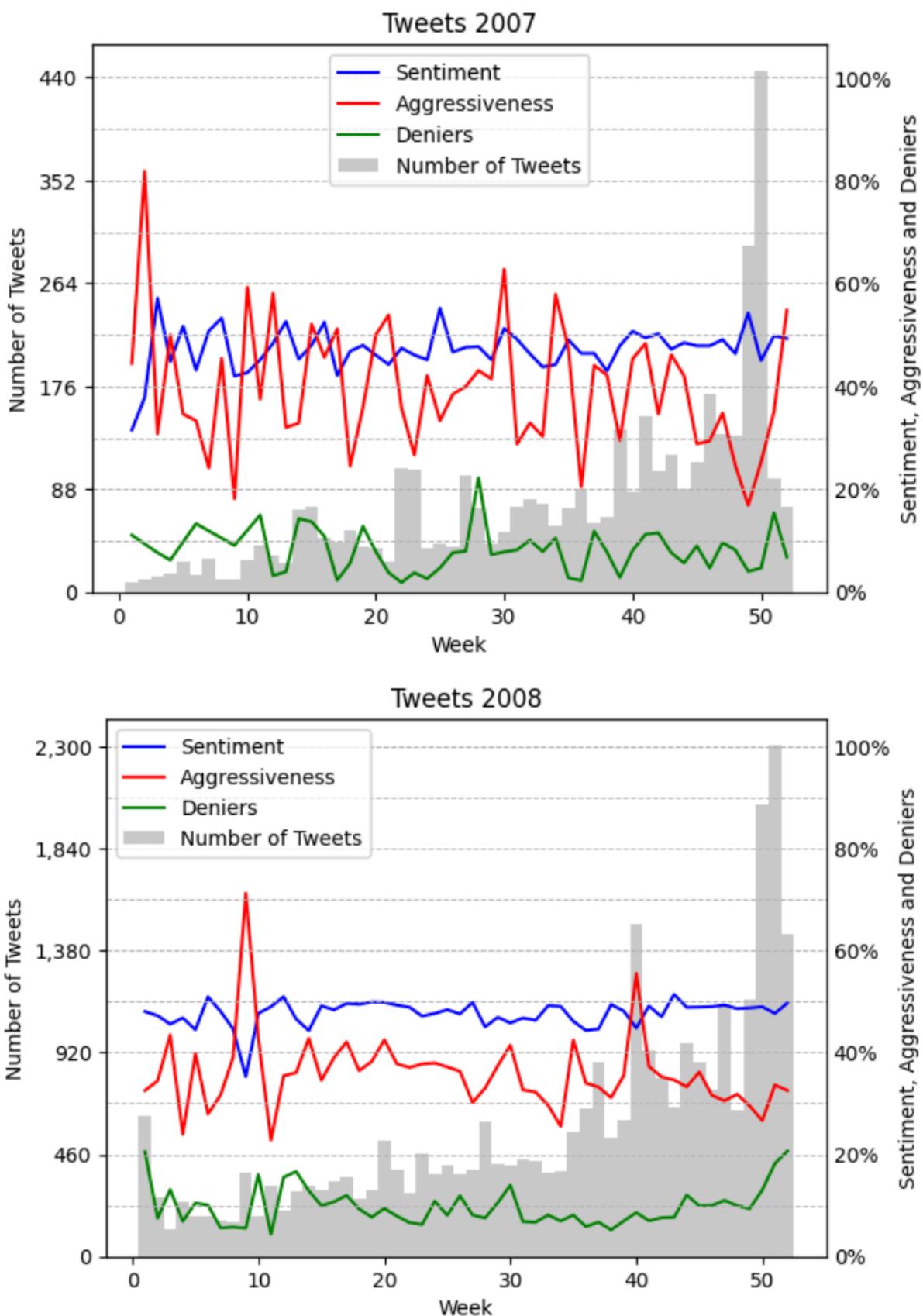
## A. Gráficos de tweets por año y por tópico

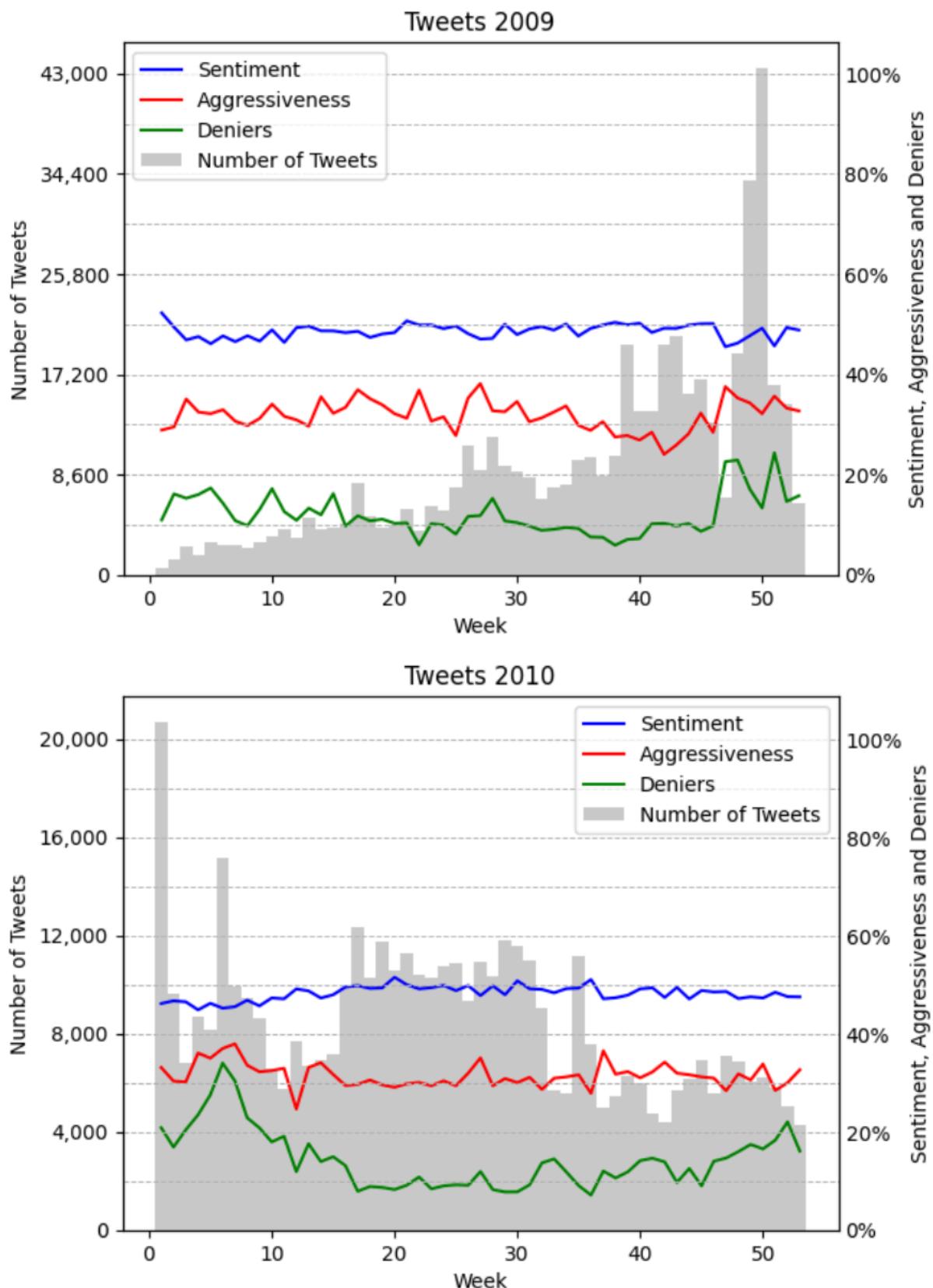
En esta sección se incluyen los distintos gráficos generados en la etapa del análisis exploratorio correspondientes a los tweets agrupados por cada año y por cada tópico. Al igual que en los gráficos vistos previamente se puede ver la cantidad de tweets, representada por las barras grises además de los valores promedio de las distintas variables: el sentimiento promedio en azul, la cantidad promedio de tweets agresivos en rojo y el promedio de tweets negacionistas en verde.

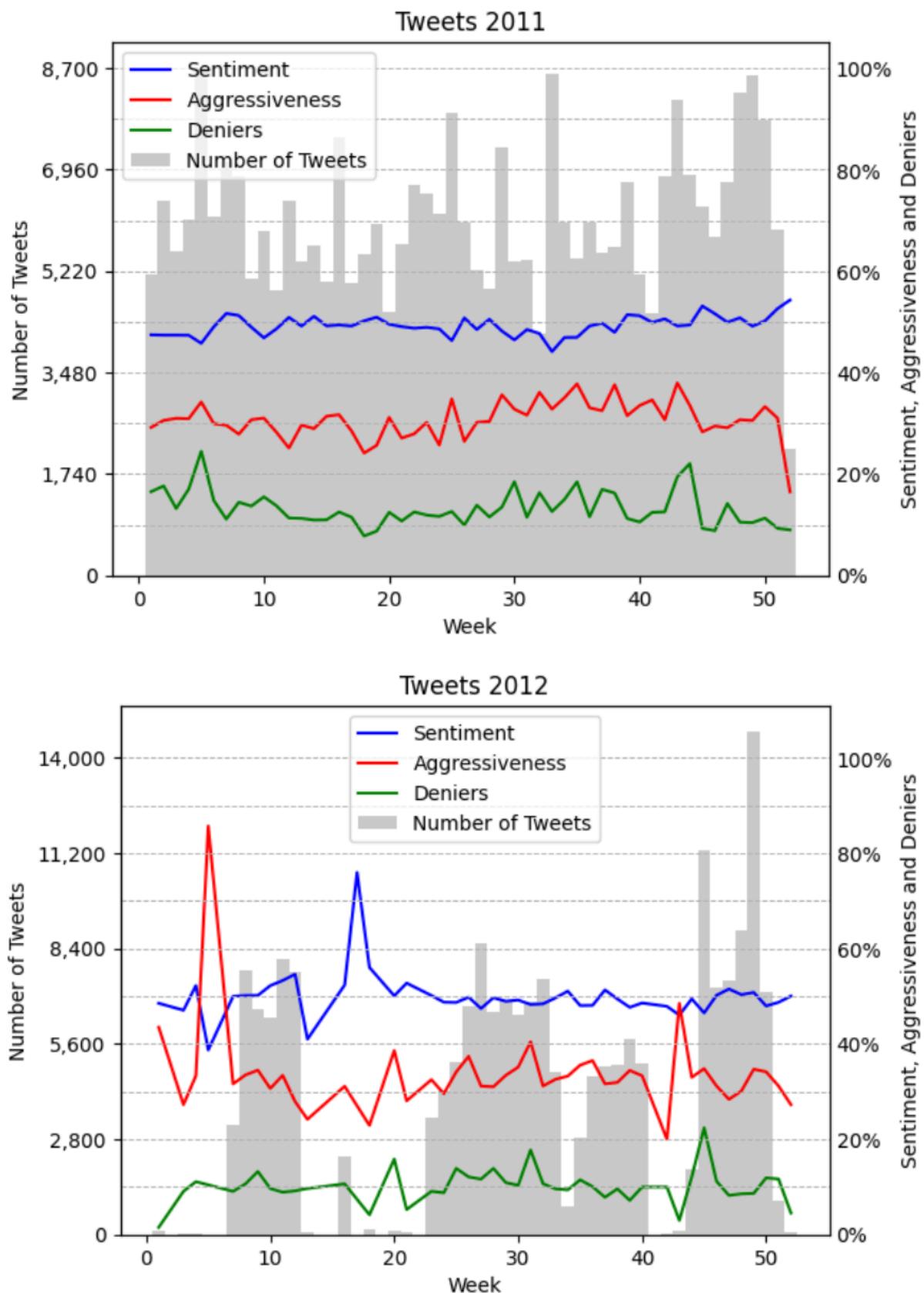
### Gráficos por año

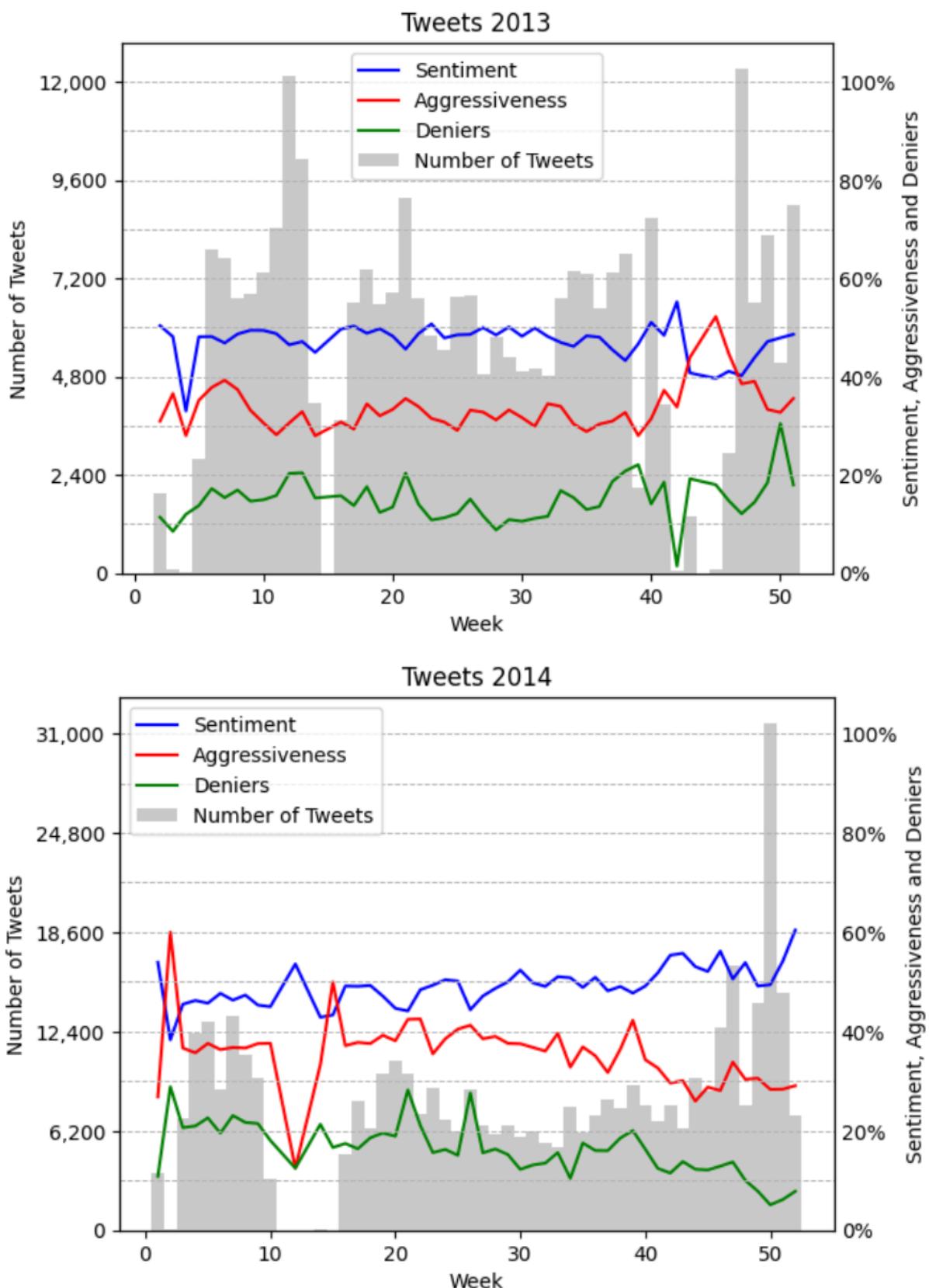
Para el primer grupo de gráficos cada gráfico muestra los datos de un año particular entre 2006 y 2019, en el eje X se pueden ver las semanas del año y en el eje Y las variables previamente mencionadas.

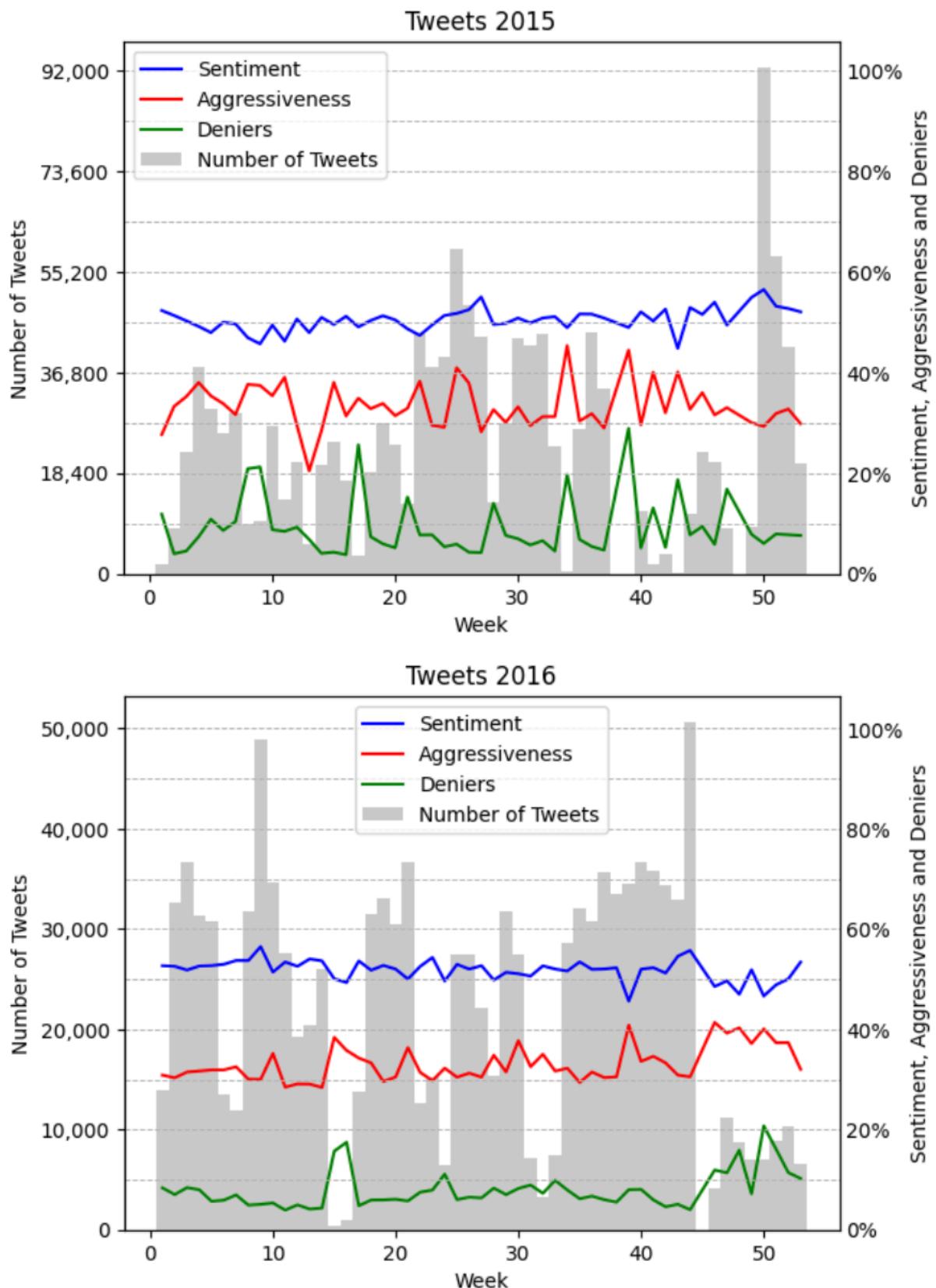


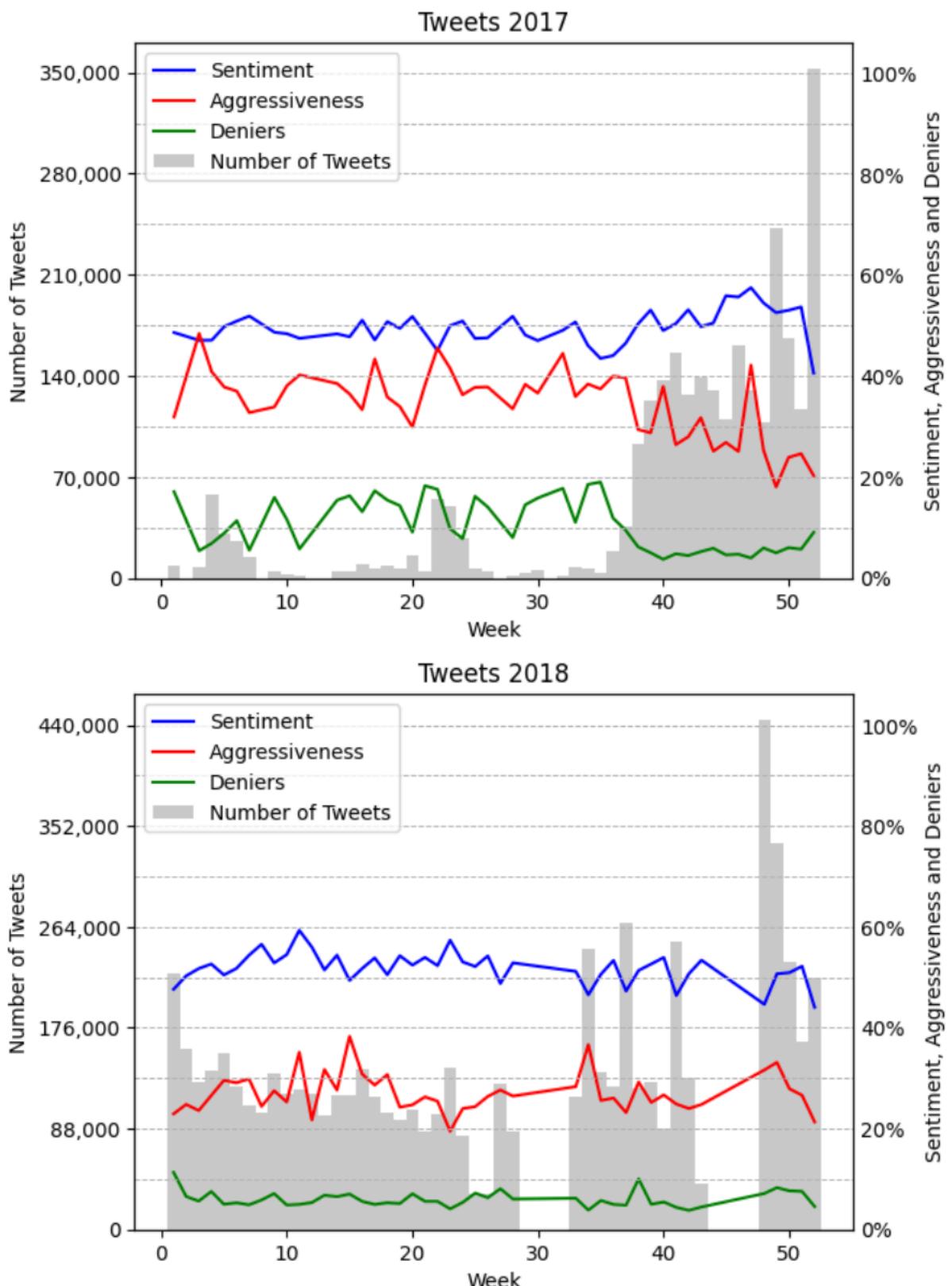


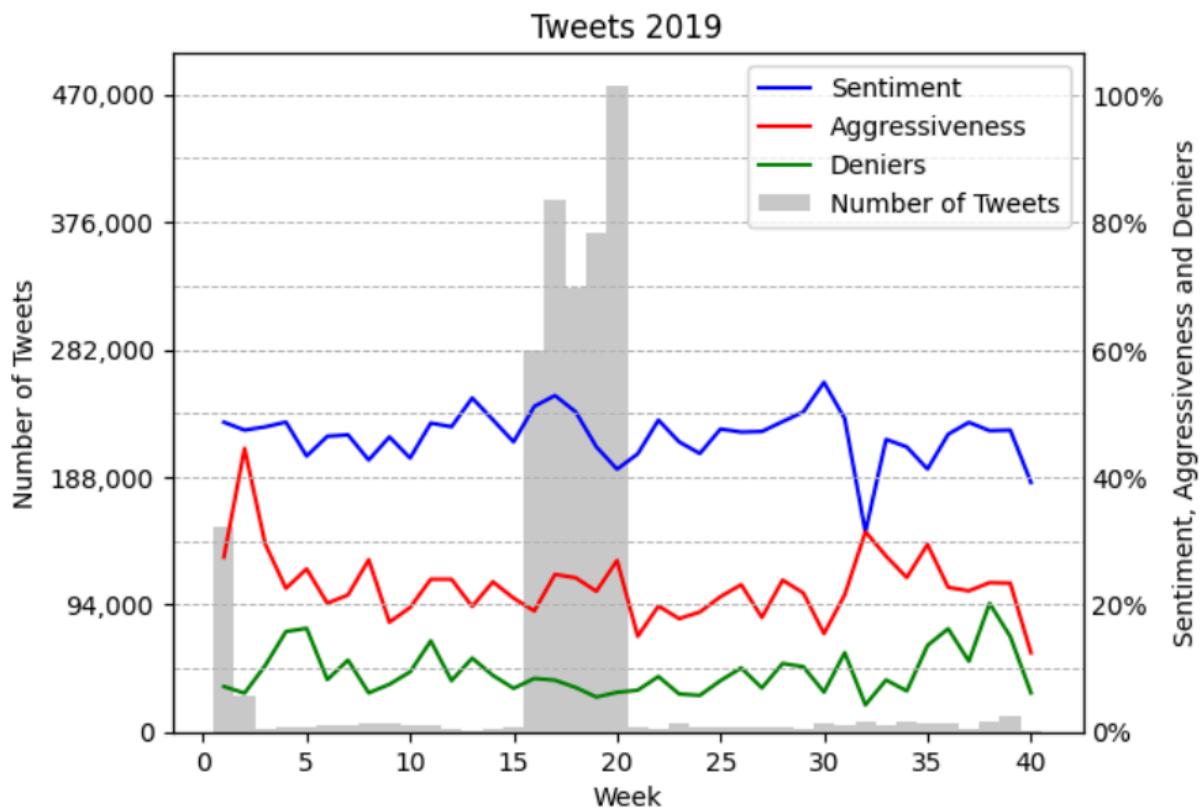












### Gráficos por tópico

En el segundo grupo cada gráfico representa la información de las distintas variables para cada uno de los tópicos, sobre el eje X se muestran los años y sobre el eje Y las distintas variables.

