Minado de tópicos y clasificación de posturas relacionadas al Asalto al Capitolio de los Estados Unidos de 2021 utilizando tweets

Aguilar Luna Gabriel Daniel, García Racilla Sandra, Rosales Almazán Laura Angélica, Suárez Espinoza Mario Alberto

1. Introducción

En el presente proyecto se aplica minería de tópicos para extraer diferentes argumentos involucrados en la noticia sobre el Asalto del Capitolio Estadounidense 2021, posteriormente, se aplican métodos de clasificación de textos para clasificar la postura de los argumentos. Los tópicos y posturas son extraídos de tweets de personas que se expresan acerca del Asalto al Capitolio EUA 2021.

1.1. Justificación

En la actualidad se genera una alta cantidad de texto en internet, principalmente de redes sociales. A partir de este texto es que se puede extraer información valiosa como la opinión que tienen las personas con respecto a un producto, o bien, los argumentos y posturas que las personas expresan con respecto a un acontecimiento. Esta información puede ayudar a la toma de decisiones de equipos de marketing o simplemente puede servir para entender de forma general cómo es que piensan ciertos sectores de la sociedad.
El asalto al Capitolio de Estados Unidos de 2021 fue un evento insólito por suceder en medio de una pandemia, con una sociedad americana dividida, y que nunca en la historia de Estados Unidos había sucedido algo similar. El extraer argumentos de las personas y las diferentes posturas que estas tienen permite agregar un factor más al estudio del Asalto al Capitolio de Estados Unidos como un fenómeno social.

1.2. Planteamiento del problema

El asalto al Capitolio de los Estados Unidos fue un evento que generó diversas opiniones, tales como la que expresa la periodista del New York Times, Lindsay Crouse, en su artículo: *Dejemos de pretender que "esto no es quienes somos"*. En su artículo (disponible aquí) ella da argumentos sobre porqué los americanos siempre han tenido una naturaleza violenta. Así como Lindsay existen muchas personas que comparten su idea, sin embargo está también la contraparte de aquellos que argumentan que en realidad los actos violentos en el Capitolio Americano son sólo reflejan a una minoría del país.
Argumentos y posturas como éstas son demasiados, de tal forma que detectarlos manualmente puede llegar a ser complicado.
Esto conduce a la pregunta de investigación:

¿Cuáles son los argumentos involucrados en el Asalto del Capitolio de Estados Unidos de 2021 y cuál es la postura de las personas ante estos?

1.3. Hipótesis

La gran mayoría de la gente está en contra de las acciones vandálicas ocurridas en el Asalto al Capitolio de los Estados Unidos en el año 2021.

1.4. Descripción del resto del artículo

En el apartado *Marco Teórico* se presentan los *Antecedentes* a este proyecto. Nos basamos principalmente en dos trabajos que describen diferentes métodos para la clasificación de posturas.
También como parte del *Marco Teórico* se presenta el *Estado del Arte*, el cual contiene la teoría en la que basamos nuestro trabajo. Se presenta una descripción general de los algoritmos utilizados.
En la *Configuración del Experimento* se describen las herramientas tecnológicas que utilizamos, así como cada uno de los pasos que realizamos para obtener los tópicos (argumentos) y la postura de los tweets, los resultados obtenidos y la evaluación de los mismos.
Por último en el apartado *Conclusiones* se presenta un análisis e interpretación de los resultados obtenidos. Se recapitula los problemas que se presentaron y algunas ideas para trabajos futuros.

2. Marco teórico

2.1. Antecedentes

Hay dos trabajos que han intentado resolver el problema, los cuáles se explican a continuación.

Transfer Learning in NLP for Tweet Stance Classification (Prashanth Rao)

En este trabajo se comparan dos aprendizajes de transferencia moderna: ULMFiT y OpenAI GPT mostrando cómo pueden ser afinados con facilidad para llevar a cabo tareas de clasificación enfocándose a clasificar la postura de Tweets hacia un tema objetivo. En su modelado podemos observar la variación de los resultados de dicha clasificación según la arquitectura del modelo utilizada.
Para el entrenamiento de los modelos se ocupa un conjunto de datos pre etiquetados
Método 1: ULMFiT. Consta de tres etapas:

1. Entrenamiento del modelo de lenguaje en un corpus de dominio general que captura características de lenguaje natural de alto nivel
2. Ajuste del modelo de lenguaje previamente entrenado en datos de tareas de destino
3. Ajuste del clasificador en los datos de la tarea de destino

El mejor puntaje F macro promedio de 0.65 usando ULMFiT en todos los temas se obtuvo usando el enfoque y los parámetros siguientes:

1. Ajuste del modelo de lenguaje en un vocabulario de Twitter aumentado.
2. Capacitar a 5 clasificadores distintos (es decir, una tarea de capacitación separada para cada tema durante la clasificación) y luego combinar los resultados para compararlos con la referencia de oro.
3. Una tasa de aprendizaje óptima de 1e-03 para el paso de ajuste fino del modelo de lenguaje
4. Una tasa de aprendizaje óptima en el rango de 1e-05a 1e-03 para el clasificador, con descongelación gradual

Método 2: OpenAI GPT. Los siguientes pasos se utilizan para entrenar el transformador OpenAI:

1. Entrenamiento previo sin supervisión.
2. Ajuste fino supervisado

La mejor puntuación F macro de 0,69 obtenida de GPT en todos los temas se obtuvo utilizando el enfoque y los parámetros siguientes:

1. Ajustar el modelo de lenguaje solo con los Tweets de capacitación proporcionados (2914 de ellos).
2. Entrenamiento de un solo clasificador para todos los temas a la vez
3. Una función de ponderación de modelado de lenguaje (lambda) de 0.5 según el documento OpenAI.
4. Abandonos de 0.1 en todas las capas (incluida la capa de clasificación)

Una vez entrenados los modelos y comparándolos por su f-score se muestra que OpenAI GPT produce el mejor resultado.

Method	Macro F-Score (All Topics)
MITRE	0.67
ULMFiT	0.65
OpenAI Transformer	**0.69**

Imagen
Mejores resultados en comparación con el índice de referencia de MITRE

La siguiente imagen compara el puntaje F (FAVOR) y el puntaje F (EN CONTRA) de nuestros dos enfoques con los de MITRE, esta vez por tema. Al observar estos resultados, una vez más está claro que OpenAI GPT supera claramente a ULMFiT en la mayoría de los temas, en cualquier clase, lo que explica la puntuación F general más alta.

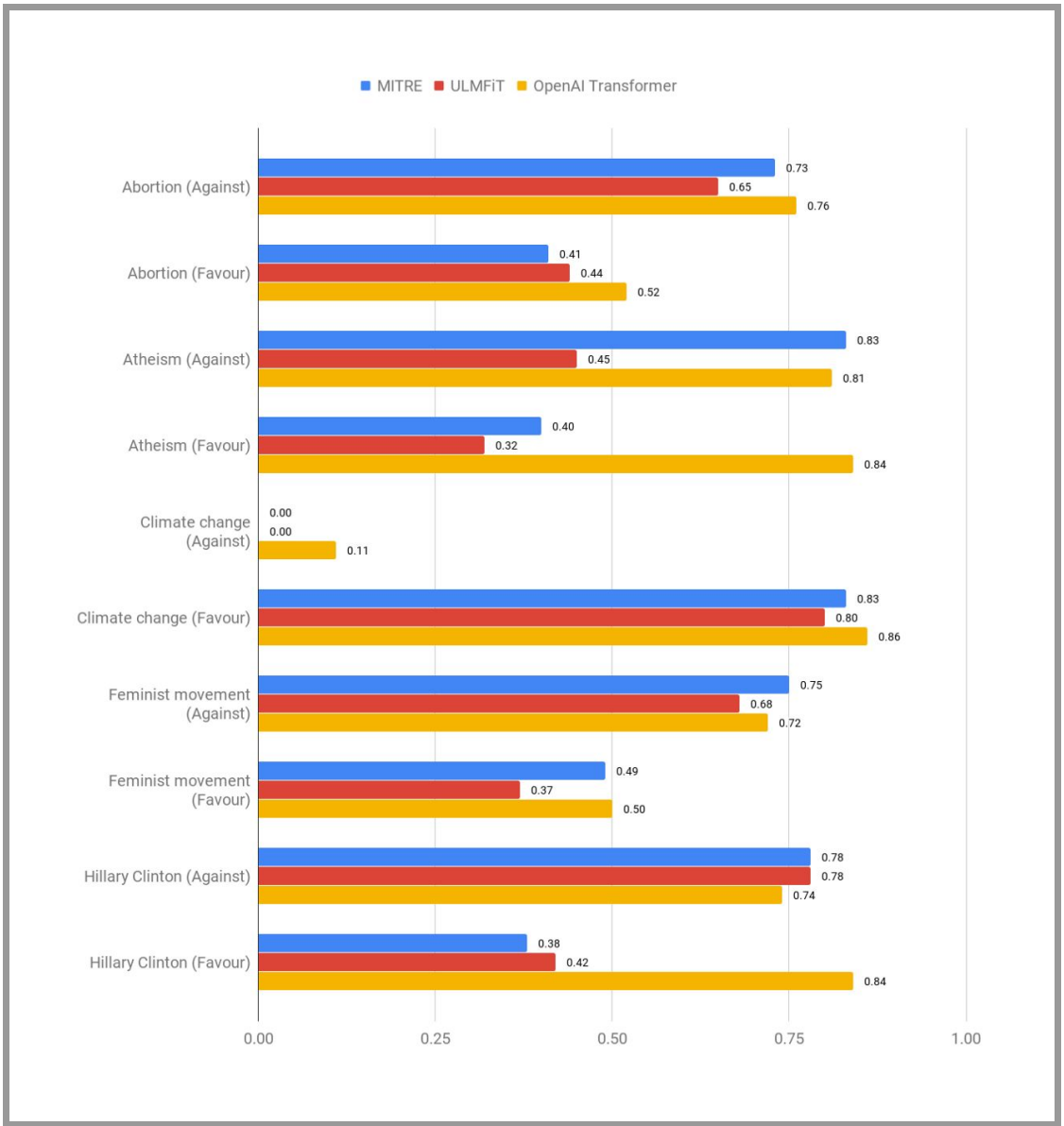


Imagen
Comparación de las puntuaciones F por tema de ULMFiT y el transformador OpenAI frente a las de MITRE

From Argumentation Mining to Stance Classification (Parinaz Sobhani, Diana Inkpen, Stan Matwin)

En este artículo, el objetivo fue investigar la postura sin considerar la estructura conversacional que no siempre está disponible.
Su esquema de anotaciones constaba de dos tareas: clasificación de la postura y etiquetado de argumentos para cada comentario. Para la clasificación de la postura, estaban interesados en la posición general del comentarista hacia la investigación médica objetivo que es el artículo de BMJ sobre la detección del cáncer de mama (Miller et al., 2014). Se consideraron dos posibles posiciones hacia este estudio relacionado con la salud:

1. A favor / De acuerdo / Soporte

2. En contra / En desacuerdo / Oposición

Además de la postura general (a favor o en contra), se interesaban por la fuerza del puesto de comentaristas hacia la investigación de destino. Así, los anotadores tenían cinco opciones para elegir: "Fuertemente a favor", "A favor", "Otro", "En contra" y"Fuertemente en contra". Aquí, "Otro" puede corresponder a comentarios neutrales, ambiguos o irrelevantes. Podemos observar los resultados en la siguiente imagen.

Argument	Strongly For	For	Against	Strongly Against	Total
Argument about the study	0	1	1	1	3
The quality of the study	5	7	35	43	90
Financial benefit of the study	0	0	4	6	10
Study is an effort to cut the expenses for Insurance companies	0	2	22	26	50
Study is an effort to cut the expenses for governments/Obamacare	0	2	26	41	69
Argument about the mammography	2	1	0	0	3
Mammo is not effective in breast cancer treatment	5	9	1	2	17
Mammo may cause cancer	9	1	0	0	10
Mammo can cause cancer by its radiation	42	23	1	1	67
Mammo's accuracy	2	7	0	2	11
Over-diagnosis and over-treatment that may cause because of false positive	51	36	0	0	87
False Negative	13	17	1	0	31
Mammo may detect cancer early and save life or result in better quality of life	0	8	63	175	246
Financial benefit of mammo for medical industry	47	53	1	0	101
Argument about manual exam	20	29	10	9	68
Other/None	118	204	179	168	699
Total	314	400	344	504	1562

Imagen
Distribución de etiquetas de argumentos para diferentes etiquetas de postura en el corpus

En minería de opinión y análisis de sentimientos, es fundamental reconocer de qué se trata la opinión, que se denomina "objetivo de opinión".

Su corpus tiene 1063 comentarios en total.

En este artículo, un marco para el etiquetado de argumentos es introducido. La principal ventaja de este marco es que los datos etiquetados no son necesarios. En este enfoque, NMF se aplica primero a los datos no etiquetados para extraerlos. Posteriormente, los datos se agrupan en función de estos temas. Cada publicación puede pertenecer a ese grupo de temas si su probabilidad de generar a partir de ese tema es más que un cierto umbral.

Cada publicación se representó mediante el término frecuencia de documento inversa (TF-IDF) esquema de ponderación sobre su bolsa de palabras estándar; se eliminaron las palabras vacías en inglés. Además, se eliminaron las palabras vacías específicas del corpus descartando términos que han aparecido en más de veinte por ciento de los documentos.

Para la clasificación de la postura, el argumento predicho, las etiquetas de la sección anterior se aprovecharon para la clasificación de la postura. Su clasificador de posturas propuesto implementa el mismo conjunto de funciones TF-IDF; Adicionalmente,utiliza las etiquetas de argumento predichas como características ycomo método de clasificación, se emplea SVM lineal.Estos métodos se comparan con otras dos clasificaciones: una SVM lineal con TF-IDF como características, y un clasificador de clases de mayoría simple como base. Los resultados se muestran en dos configuraciones que podemos ver en las siguientes imágenes.

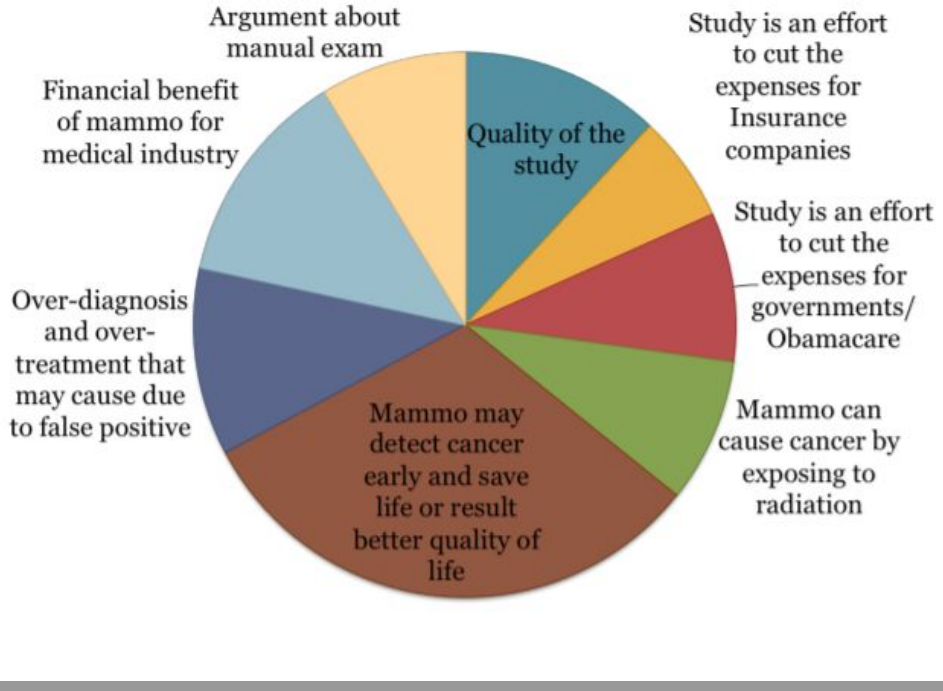


Imagen
El resumen de argumentos basados en datos anotados

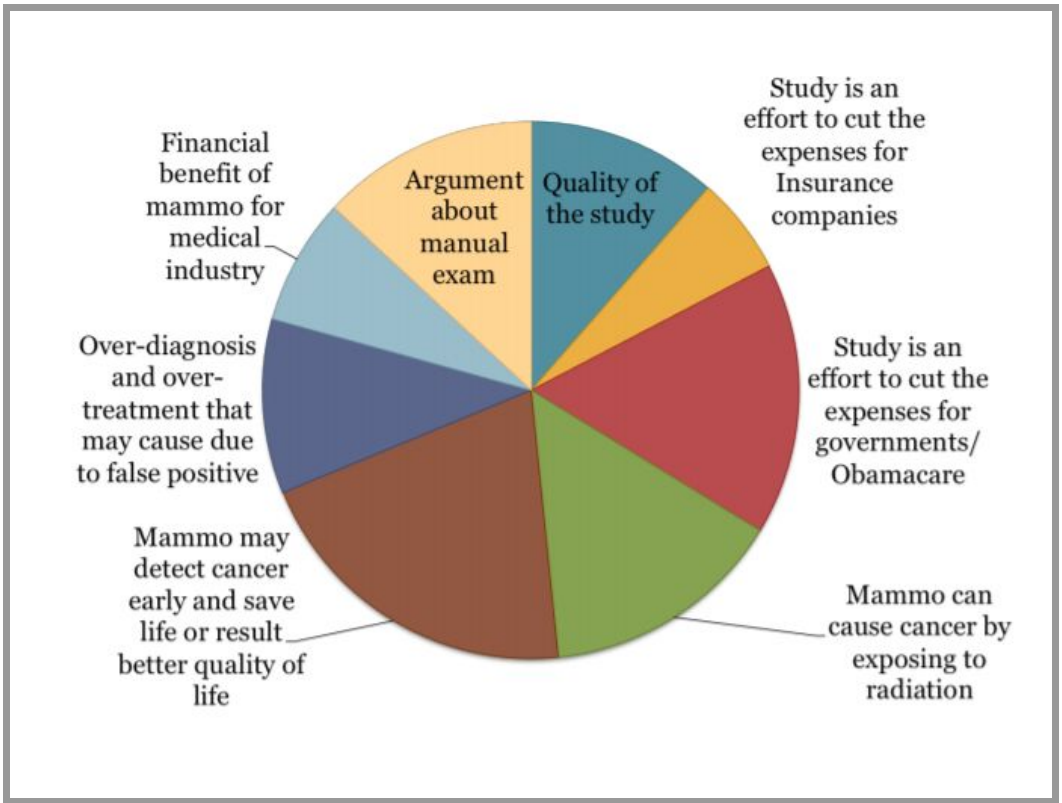


Imagen
El resumen de argumentos basados en datos predichos

Se demostró que el uso de NMF para agrupar compuestos basado en sus argumentos es significativamente mejor que emplear LDA. Esto se puede observar en las palabras clave principales extraídas de los temas.
Se especula que la razón es la brevedad de los comentarios, ya que LDA normalmente funciona mejor para textos más largos. La razón puede ser el hecho de que todos estos datos son sobre el mismo tema general, la detección del cáncer de mama y LDA no puede distinguir entre subtemas.

2.2. Teoría utilizada (Estado de la técnica)

Minado de tópicos
El minado de tópicos permite identificar el tema principal que se trata en un texto. Este análisis se realiza en diferentes niveles, ya sea dentro de una oración o dentro de un artículo de investigación. El análisis de tópicos se ha utilizado en gran cantidad de aplicaciones, de particular interés es su uso en el análisis de artículos de investigación para identificar las áreas en que se ha realizado investigación en distintas épocas.

Latent Dirichlet Allocation (LDA)
La Asignación Latente de Dirichlet (ALD) o Latent Dirichlet Allocation (LDA) es un modelo generativo que permite que conjuntos de observaciones puedan ser explicados por grupos no observados que explican porqué algunas partes de los datos son similares.
Es un modelo comúnmente utilizado para el minado de tópicos.

Non-negative Matrix Factorization (NMF)
Factorización matricial no negativa (NMF o NNMF), también aproximación matricial no negativa, es un grupo de algoritmos en análisis multivariado y álgebra lineal donde una matriz V se factoriza en (habitualmente) dos matrices W y H, con la propiedad de que las tres matrices no tienen elementos negativos.
En este caso es utilizado para la agrupación de documentos.

Term Frequency - Inverse Document Frequency (TF-IDF)
Medida estadística que evalúa qué tan relevantes es una palabra en el documento o en la colección de documentos. El cálculo se hace a través de la multiplicación de dos métricas: cuántas veces la palabra aparece en el documento, y la frecuencia inversa del documento de la palabra en un conjunto de documentos

Minado de opiniones
El minado y análisis de opiniones y sentimientos permite identificar y extraer estos elementos, presentes en gran cantidad de documentos. Las técnicas desarrolladas permiten identificar al sujeto que detenta una opinión en un texto, así como el ente sobre el que se opina (objetivo de la opinión) y la opinión en sí. También es posible determinar el contexto en el que se emite la opinión. Tomando en cuenta todo lo anterior, se deduce el sentimiento asociado a esa opinión, ¿es una opinión positiva o negativa? Las principales aplicaciones realizadas se encuentran en los sistemas de recomendación, principalmente los orientados a promover productos comerciales, por ejemplo, los distintos modelos de teléfonos celulares.

Clasificación de posturas
La detección de posturas es una subcategoría de la minería de opiniones , donde la tarea es determinar automáticamente si el autor de un texto está a favor o en contra de un objetivo determinado.
La clasificación supervisada es una de las tareas que más frecuentemente son llevadas a cabo por los denominados Sistemas Inteligentes. Por lo tanto, un gran número de paradigmas desarrollados bien por la Estadística (Regresión Logística, Análisis Discriminante) o bien por la Inteligencia Artificial (Redes Neuronales, Inducción de Reglas, Árboles de Decisión, Redes Bayesianas) son capaces de realizar las tareas propias de la clasificación.
Paso previo a aplicar un método de clasificación, es la partición del conjunto de datos en dos conjuntos de datos más pequeños que serán utilizadas con los siguientes fines: entrenamiento y test . El subconjunto de datos de entrenamiento es utilizado para estimar los parámetros del modelo y el subconjunto de datos de test se emplea para comprobar el comportamiento del modelo estimado. Cada registro de la base de datos debe de aparecer en uno de los dos subconjuntos, y para dividir el conjunto de datos en ambos subconjuntos, se utiliza un procedimiento de muestreo: muestreo aleatorio simple o muestreo estratificado. Lo ideal es entrenar el modelo con un conjunto de datos independiente de los datos con los que realizamos el test.

Clasificación por Support Vector Machine
Las Máquinas de Vectores Soporte (creadas por Vladimir Vapnik) constituyen un método basado en aprendizaje para la resolución de problemas de clasificación y regresión. En

ambos casos, esta resolución se basa en una primera fase de entrenamiento (donde se les informa con múltiples ejemplos ya resueltos, en forma de pares {problema, solución}) y una segunda fase de uso para la resolución de problemas. En ella, las SVM se convierten en una “caja negra” que proporciona una respuesta (salida) a un problema dado (entrada).

Imagen. Máquinas de Vectores Soporte

Clasificación por Naive Bayes

Naive Bayes es un modelo de predicción basado en la probabilidad Bayesiana. El modelo es muy simple, pero poderoso, en cuanto que es resultado directo de los datos y su tratamiento con simple estadística bayesiana de probabilidad condicionada. Hay que tener en cuenta que se asume, por simplificación, que las variables son todas sucesos independientes.

El modelo bayesiano de probabilidad condicionada se representa como:

$$P(A|B) = P(A \cap B)/P(B)$$

Clasificación por Logistic Regression

La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras.

Este modelo logístico binario se utiliza para estimar la probabilidad de una respuesta binaria basada en una o más variables predictoras o independientes. Permite decir que la presencia de un factor de riesgo aumenta la probabilidad de un resultado dado un porcentaje específico.

3. Configuración del experimento

3.1. Datos

Conjunto de Datos Objetivo

El conjunto de datos objetivo consiste en Tweets reales que fueron recopilados mediante Twitter API v2 y posteriormente convertidos y agrupados en un sólo CSV.

Cabe resaltar que a pesar de que Twitter API permite aplicar filtros de búsqueda como eliminación de tweets con ciertas palabras, eliminación de tweets con contenido multimedia, o eliminación de retweets, algunos de los tweets obtenidos no necesariamente poseen opiniones o posturas acerca del Asalto al Capitolio Estadounidense 2021, por lo que el

conjunto de datos que logramos recopilar posee algunos registros basura que no pudimos depurar por completo.

	data__id	data__lang	data__text
2419	1347561138553114624	en	@smvn65 @ABC @ThisWeekABC To be honest, I don&#65533;...
1020	1349221017680449537	en	@RepBoebert So you're cool with people bringin...
1688	1347525987366481921	en	@MZHemingway @johnddavidson I reject the premi...
269	1349219788124512256	en	@eoinburgin @Hbomberguy My favorite bizarre fl...
2258	1347558664022986760	en	@realDonaldTrump You lost my support when you ...

Imagen
Datos objetivo (muestreados aleatoriamente)

Conjunto de datos para entrenamiento

Para realizar el entrenamiento de los modelos de clasificación se necesitan datos etiquetados. No podíamos etiquetar manualmente los tweets extraídos porque ese era justo el trabajo que se buscaba automatizar.
Para poder lograr el entrenamiento del modelo utilizamos los datos proporcionados por Sem-Eval 2016 para la detección de posturas en Tweets, los cuales están disponibles aquí y fueron utilizados también en el trabajo Transfer Learning in NLP for Tweet Stance Classification de Prashanth Rao.
Este conjunto de datos pertenece a los siguientes cinco temas:

1. Ateísmo
2. El cambio climático es una preocupación
3. Movimiento feminista
4. Hillary Clinton
5. Legalización del aborto

Los datos etiquetados proporcionados consisten en un tema objetivo, un Tweet que le pertenece y la postura del Tweet hacia el objetivo. Los datos ya están divididos en un conjunto de entrenamiento (que contiene 2914 tweets) y un conjunto de prueba (que contiene 1249 tweets). La postura puede ser una de tres etiquetas: "FAVOR", "EN CONTRA" y "NINGUNO”

Target	Tweet	Stance
Legalization of Abortion	on a side note, just because you think smtg is wrong, doesn't mean everyone else have to live accd to your beliefs.. ^^ #SemST	NONE
Climate Change is a Real Concern	We need to work with confidence, transparence and guided by consensus @manupulgarvidal at @UN_PGA event on #action2015 #SemST	NONE
Hillary Clinton	If you're not watching @HillaryClinton's speech right now you're missing her drop tons of wisdom. #SemST	FAVOR
Feminist Movement	The only reason, I stopped at each entrance in #walkingstreet is to have a #LawOfAttraction #viewpoint #dailydevotional #SemST	NONE
Legalization of Abortion	Murdering an unborn child is the crudest form of contraception! #Catholic #Christian #Conservative #feminist #SemST	AGAINST
Legalization of Abortion	Abortion is legal all nine months in Canada and in some parts of the USA. #ProLifeYouth #SemST	AGAINST
Hillary Clinton	Marriage equality a constitutional right! Woot #HRC2016 #hillaryclinton #readyforhillary #vote #hillaryfor2016 ##mpotus #SemST	FAVOR
Legalization of Abortion	Why is bacteria considered life on Mars, but a heartbeat is not considered life on earth? #heartbeat #SemST	AGAINST

Imagen
Datos de entrenamiento de ejemplo (muestreados aleatoriamente)

3.2. Herramientas

Twitter API

Es una herramienta que permite realizar peticiones para obtener tweets, información de usuarios, publicar contenido, enviar mensajes privados, etc. Para poder utilizar esta herramienta tuvimos que aplicar por una cuenta Developer.
En específico utilizamos el endpoint Recent Search que permite obtener un conjunto de tweets dentro de los últimos 7 días.
La desventaja además del rango de días que se tiene para poder utilizarlo, es que sólo permite obtener un máximo de 100 tweets por petición, por lo que fue necesario realizar varias peticiones para recabar un conjunto de tweets considerable (2428).

Postman

Es una herramienta que funciona como cliente API para el envío de peticiones REST, SOAP y GraphQL, orientado principalmente al desarrollo y testing de APIs.
Nosotros lo ocupamos como cliente para Twitter API, pues posee una colección que se puede instalar y que contiene la implementación de todas las peticiones de Twitter API disponibles.

JSON to CSV Converter

Permite generar archivos en formato csv a partir de archivos en formato json. Usamos esta herramienta debido a que Twitter API envía respuestas en formato json, y necesitamos archivos en formato csv para facilitar el uso del análisis de datos.

Pandas

Es una biblioteca de código abierto que provee estructuras de datos de alto rendimiento y herramientas para el análisis de datos.
La estructura de datos que utilizamos fue pandas.Dataframe, que representa una tabla bidimensional de tamaño variable y heterogénea.
La ventaja del uso de esta estructura de datos es que permite operaciones a nivel de columnas y filas, lo cual lo hace adecuado para almacenar y operar los datos que provienen de archivos csv.

Scikit-learn

Es una biblioteca para machine learning de software libre en Python. Contiene algoritmos de clasificación, regresión y agrupamiento tales como support vector machine, random forests, gradient boosting, k-means, etc. Está diseñada para operar con bibliotecas numéricas y científicas como NumPy y SciPy.
De esta biblioteca utilizamos los vectorizadores de texto CountVectorizer y TF-IDF para representar a los documentos en vectores de R^n.
Utilizamos también los algoritmos para el modelado de tópicos Non-negative matrix factorization (NMF) y Latent Dirichlet Allocation (LDA).
Por último para la clasificación, realizamos una comparación entre modelos de Naive Bayes, Logistic Regression, Random Forest y Support Vector Machine.

GoogleSearch for Python

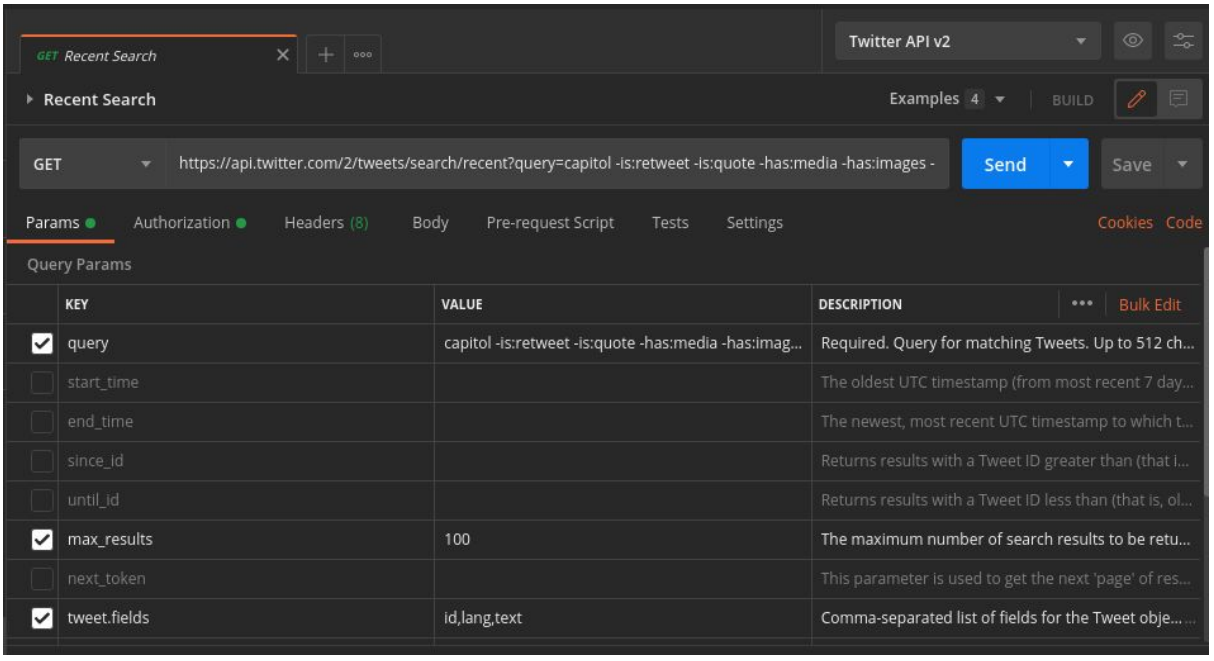
Es una biblioteca que permite realizar búsquedas en Google desde Python. La utilizamos para automatizar el etiquetado de los tópicos obtenidos con NMF y LDA.

3.3. Descripción del experimento

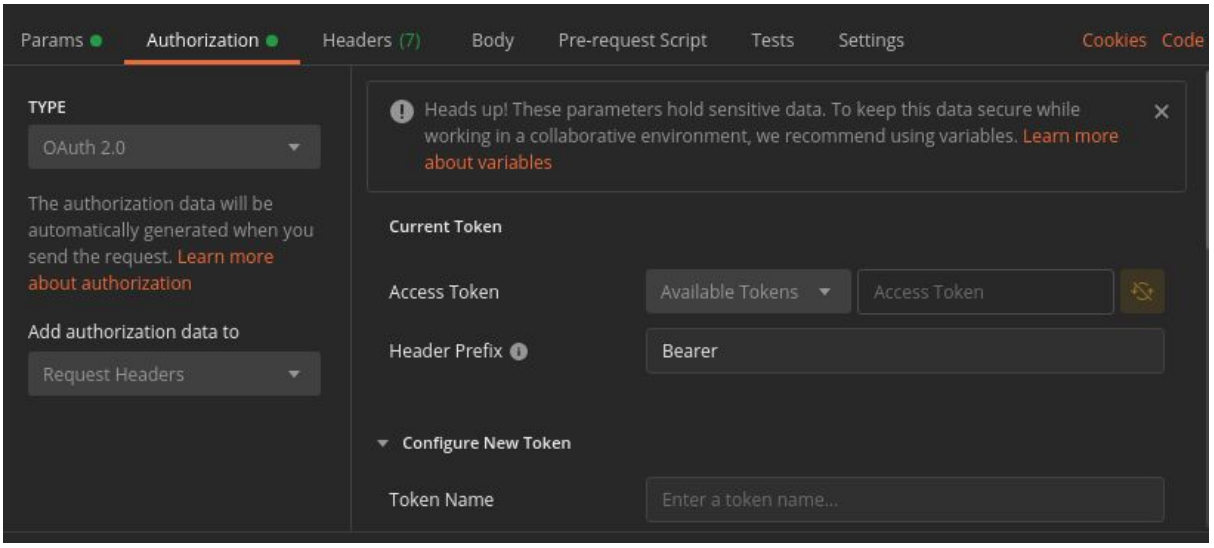
Una postura se refiere a la posición general de una persona acerca de una idea. Para el caso de este estudio se consideran las posturas “A favor” o “En contra”.
Un argumento es la expresión escrita mediante la cual se intenta probar, refutar, o justificar una proposición o tesis. Para el caso de este estudio se trata como tópico principal el Asalto al Capitolio de Estados Unidos 2021, así que los tópicos extraídos se consideran argumentos que intentan justificar lo sucedido en el suceso.

Extracción de tweets

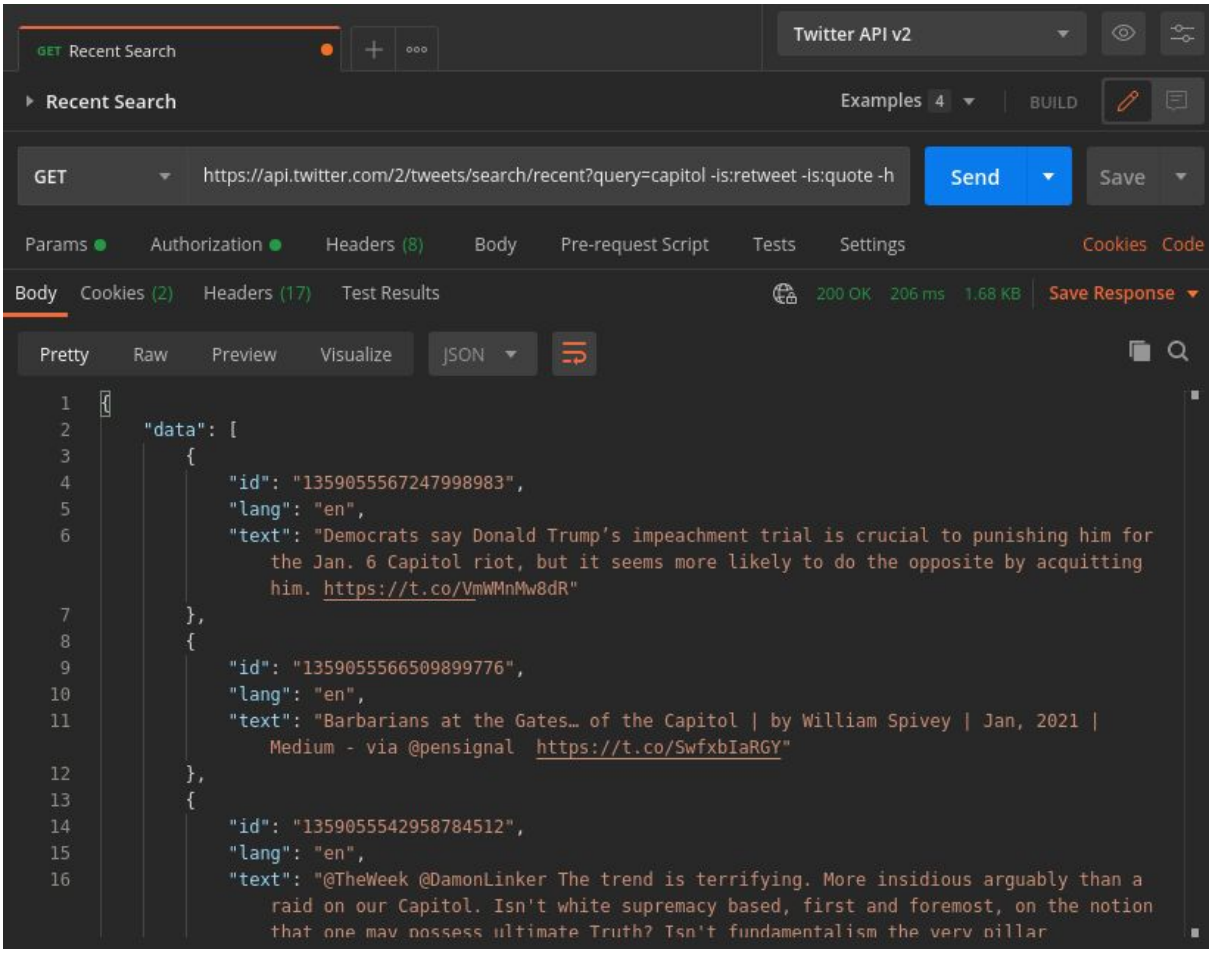
El primer paso fue la extracción de tweets acerca del Asalto al Capitolio de Estados Unidos en 2021. Como ya se mencionó, se utilizó Twitter API v2. Dentro de los parámetros que utilizamos en la petición se encuentran `query`, `max_results` y `tweet.fields`.
En query se especifica el patrón que siguen los tweets, max_results especifica el número máximo de tweets que se retornan por cada petición; este número está limitado hasta 100. tweet.fields contiene la lista de campos del tweet que serán retornados.
Esta petición fue configurada en Postman.



Cabe destacar que para que la respuesta sea exitosa se debió configurar primero las opciones de autorización utilizando las claves que proporciona la cuenta developer de Twitter API.



Un ejemplo de cómo Twitter API envía respuestas es el siguiente:



Una vez extraídos los tweets en formato json, los convertimos a csv utilizando JSON to CSV Converter.

Para las secciones siguientes de la descripción del experimento se describe sólo lo más importante. Para más detalle puede revisarse el cuaderno de python Minería_de_Tópicos.ipynb en el repositorio de GitHub.

Minado de tópicos

El minado de tópicos es un tipo de aprendizaje no supervisado, usado para extraer los temas principales de una colección de documentos. Los temas son representados como un conjunto de palabras.
Se utilizaron dos modelos para poder generar de forma más precisa los temas Tweets que son nuestros documentos.

1. Latent Dirichlet Allocation (LDA)
   Realiza dos suposiciones:
   - Los documentos que tienen palabras similares tienen el mismo tema. En otras palabras, los documentos son distribuciones de probabilidad sobre temas latentes.
   - Los documentos que tienen grupos de palabras que ocurren frecuentemente usualmente tienen el mismo tema; es decir, los temas son distribuciones de probabilidad sobre las palabras.

a. **Importar el conjunto de documentos:** Cada uno de nuestros Tweets representa un documento. Dentro de nuestro archivo csv la columna que contiene los datos referentes al contenido del tweet es “data__text”.

	data__id	data__lang	data__text
0	1347495838839238657	en	@PapaGlider @Jessica26307123 @MontyBoa99 @real...
1	1347495838688440320	en	US Capitol: Police confirms death of officer i...
2	1347495838063284230	en	@HookRocky @NBCNews @NBCNewsTHINK At least we ...
3	1347495834833747969	en	Mike Pompeo Says Capitol Riot Proves U.S. Isn'...

b. **Obtención del vocabulario:** Para poder aplicar LDA es necesario crear una lista con las palabras que conforman el vocabulario de nuestros documentos. Por tal motivo se crea una matriz de los términos del documento, en ella se guardan aquellas palabras que aparezcan al menos en 80% del documento y en al menos 2 documentos. Además de eliminar las palabras que no contengan información relevante para la investigación; tal es el caso de las palabras “https”, “nhttps”, “amps”.
c. **Uso del modelo LDA:** se hace uso del modelo para crear temas sobre la distribución de probabilidad en el vocabulario para cada uno de los temas. Se decidió para esta investigación obtener cinco temas.
d. **Almacenamiento de los datos:** Cada uno de los temas junto con las palabras que lo conforman son guardados en un archivo de texto plano, además se crea un archivo csv el cual agrega una columna adicional “Topic” al csv original, en el que se especifica a cuál tema pertenece cada tweet.

2. Non-Negative Matrix Factorization (NMF)
Esta técnica de aprendizaje no supervisado realiza *clustering*, así como una reducción de dimensiones. Representa la información de un vector de datos como la multiplicación de dos vectores de dimensiones más pequeñas.
    a. **Importar el conjunto de documentos:** Cada uno de nuestros Tweets representa un documento. Dentro de nuestro archivo csv la columna que contiene los datos referentes al contenido del tweet es “data__text”.

	data__id	data__lang	data__text
0	1347495838839238657	en	@PapaGlider @Jessica26307123 @MontyBoa99 @real...
1	1347495838688440320	en	US Capitol: Police confirms death of officer i...
2	1347495838063284230	en	@HookRocky @NBCNews @NBCNewsTHINK At least we ...
3	1347495834833747969	en	Mike Pompeo Says Capitol Riot Proves U.S. Isn'...

b. **Obtención del vocabulario:** NMF utiliza el esquema de TF-IDF para medir la relevancia de cada una de las palabras en el documento, es decir, aquellas palabras que brindan más información sobre el tema del que se está tratando. Además de eliminar las palabras que no contengan información relevante para la investigación; tal es el caso de las palabras “https”, “nhttps”, “amps”.
c. **Uso del modelo NMF:** se hace uso del modelo para crear una matriz de probabilidad la cual contiene la probabilidad de todas las palabras en el vocabulario para todos los temas. Se decidió para esta investigación obtener cinco temas.
d. **Almacenamiento de los datos:** Cada uno de los temas junto con las palabras que lo conforman son guardados en un archivo de texto plano, además se crea un archivo csv el cual agrega una columna adicional “Topic” al csv original, en el que se especifica a cuál tema pertenece cada tweet.

Interpretación de los tópicos obtenidos

Para interpretar los tópicos obtenidos construimos una función para llevar a cabo una búsqueda de las palabras de los tópicos, con la biblioteca Google Search, obtenemos los resultados de ‘googlear’ dichas palabras. Posteriormente obtenemos el título del primer resultado de la búsqueda, el cual asignamos al tópico en cuestión. Lo anterior con el objetivo de facilitar la interpretación de los mismos.

Clasificación de Posturas

El propósito de esta sección es encontrar el mejor modelo para el problema de clasificación de posturas. Se subdivide en las siguientes tareas:

1. **Análisis de datos exploratorio**: Se analiza cómo están conformados los datos de entrenamiento y se aplican las operaciones necesarias para poder generar el modelo. El resultado de esta sección fue un dataframe que contiene las columnas “Stance”, “Tweet” y “stance_id”.

	Stance	Tweet	stance_id
361	AGAINST	Come out of every circle of limitation and aff...	0
366	AGAINST	Didn't do what I came out to do today, but God...	0
976	AGAINST	In so many ways to Christians right here in Am...	0
2123	AGAINST	Insurgent. What will happen if Hillary becomes...	0

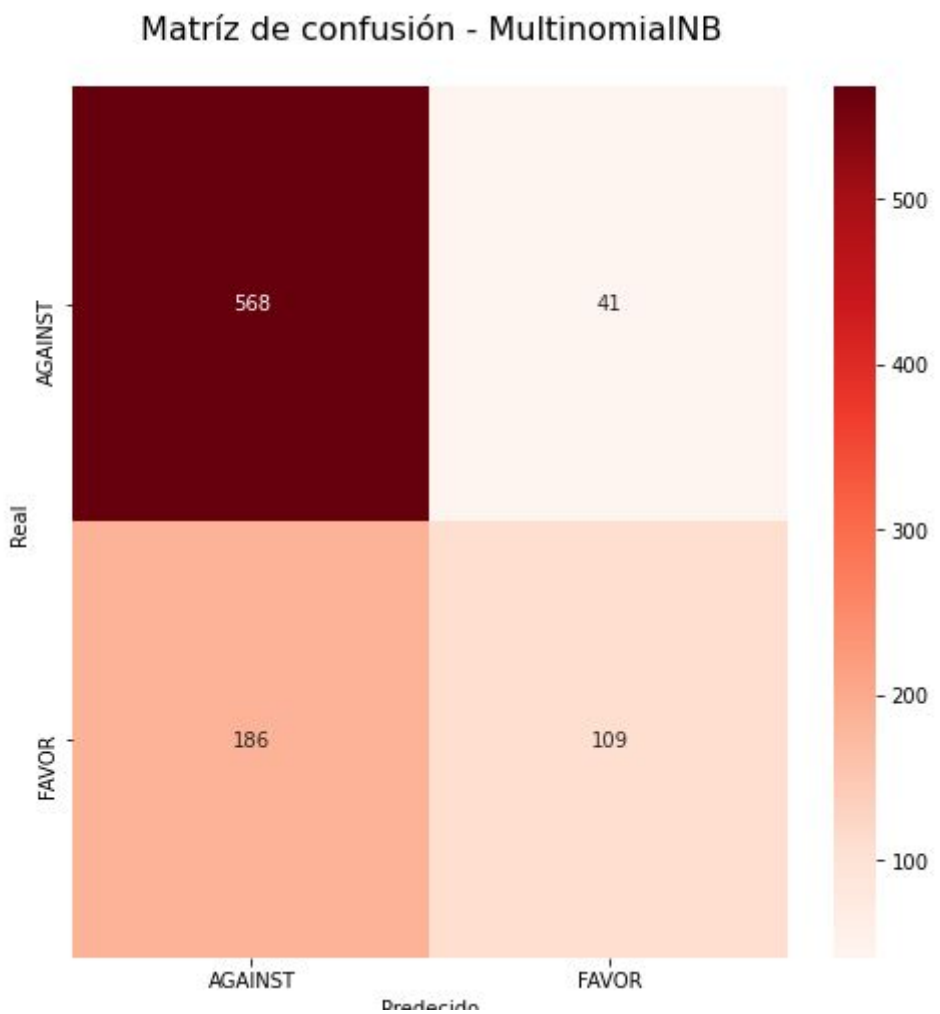
2. **Preprocesamiento del texto**: Para este paso se transforman los tweets a vectores de R^n para que el modelo pueda realizar predicciones. Para ello se utiliza TF-IDF.
3. **Modelos de clasificación**: En esta sección se entrenan seis modelos, y mediante validación cruzada de 5 iteraciones se obtiene la exactitud promedio de cada modelo. Aquel que tiene la mejor exactitud es el elegido como modelo para nuestro caso de estudio. Los modelos que se propone usar son:
    a. Categorical Naive Bayes
    b. Multinomial Naive Bayes
    c. Logistic Regression
    d. Random Forest
    e. Linear Support Vector Machine
    f. Non Linear Support Vector Machine

    El resultado de esta sección fue que Multinomial Naive Bayes era el modelo que tenía la mejor exactitud promedio, así que este es el modelo que se elige para realizar las predicciones.
4. **Evaluación del modelo elegido**: Se evalúa la exactitud(accuracy), precisión (precision), exhaustividad (recall) y valor F (f-score). Los valores obtenidos fueron los siguientes:

MÉTRICAS DE CLASIFICACIÓN

	precision	recall	f1-score	support
AGAINST	0.75	0.93	0.83	609
FAVOR	0.73	0.37	0.49	295
accuracy			0.75	904
macro avg	0.74	0.65	0.66	904
weighted avg	0.74	0.75	0.72	904

Además, se grafica la matriz de confusión, en dónde notamos que el modelo tendía hacia AGAINST debido a la alta cantidad de tweets etiquetados con esta clasificación.



5. **Corrección del modelo**: Se balancea la cantidad de tweets etiquetados con AGAINST y FAVOR para tener un mejor modelo. Las nuevas métricas y matriz de confusión se presentan en la sección 3.5 Evaluación.
6. **Predicción**: Se manipula al modelo para aceptar la entrada de nuevos tweets y obtener como resultado su postura: AGAINST o FAVOR.

Clasificación de Posturas en Tweets sobre el Asalto al Capitolio de Estados Unidos 2021

En esta sección se aplica el modelo obtenido en la sección anterior a los documentos recabados en la extracción de tweets.

Una vez obtenidas las posturas de cada tweet, se analiza el resultado tomando en cuenta también los tópicos extraídos.

3.5. Presentación de Resultados

Minado de tópicos

Tópicos obtenidos con LDA con su interpretación correspondiente

```
LDA Top 10 words for topic #0:
['did', 'make', 'heard', 'peacefully', 'donald', 'people', 'know', 'president', 'building', 'trump']
AP FACT CHECK: Trump's call to action distorted in debate


LDA Top 10 words for topic #1:
['terrorists', 'did', 'weapons', 'don', 'know', 'election', 'police', 'stormed', 'just', 'people']
Who were they? Records reveal Trump fans who stormed Capitol


LDA Top 10 words for topic #2:
['violence', 'took', 'dc', 'pro', 'insurrection', 'supporters', 'police', 'attack', 'riot', 'trump']
Capitol riots: Five takeaways from the arrests - BBC News


LDA Top 10 words for topic #3:
['building', 'storming', 'officer', 'death', 'mob', 'stormed', 'inside', 'police', 'attack', 'trump']
Police officer dies after pro-Trump mob attack on US Capitol | US Capitol breach | The Guardian


LDA Top 10 words for topic #4:
['stop', 'killed', 'congress', 'didn', 'like', 'just', 'building', 'trump', 'police', 'people']
The Capitol Police failure to stop a Trump mob, explained - Vox
```

El resultado de la ejecución de los algoritmos es el datagrama original más la columna Topic, que indica el número de tópico al que se relaciona el tweet.

	data__id	data__lang	data__text	Topic
0	1347495838839238657	en	@PapaGlider @Jessica26307123 @MontyBoa99 @real...	1
1	1347495838688440320	en	US Capitol: Police confirms death of officer i...	3
2	1347495838063284230	en	@HookRocky @NBCNews @NBCNewsTHINK At least we ...	1
3	1347495834833747969	en	Mike Pompeo Says Capitol Riot Proves U.S. Isn'...	1
4	1347495834439606272	en	US Capitol Attack: President Trump Can't Handl...	1
...	...	...	...	...
2423	1347561137278119936	en	Uh oh. Maybe they should've worn masks? \n\nI ...	3
2424	1347561136690843649	und	??\n\nhttps://t.co/FprykuE9bN	0
2425	1347561136242126849	en	Oregon representative allowed protesters into ...	2
2426	1347561136187727873	en	@realDonaldTrump The 5 deaths ARE ALL ON YOUR ...	1
2427	1347561136087052288	en	@RobbieBarstool Or the capitol	0

Tópicos obtenidos con NMF con su interpretación correspondiente

```
NMF Top 10 words for topic #0:
['donald', '2021', 'january', 'make', 'heard', 'peacefully', 'soon', 'voices', 'marching', 'patriotically']
Fact Check: Did Trump Say to 'Peacefully and Patriotically' March to the Capitol?


NMF Top 10 words for topic #1:
['like', 'mob', 'did', 'violence', 'just', 'stormed', 'supporters', 'people', 'amp', 'trump']
What Trump Told Supporters Before Mob Stormed Capitol - The New York Times


NMF Top 10 words for topic #2:
['mo', 'lead', 'paul', 'organizer', 'gosar', 'andy', 'biggs', 'job', 'inside', 'attack']
Paul Gosar, Andy Biggs credited in video with organizing Trump crowd in DC


NMF Top 10 words for topic #3:
['investigation', 'realdonaldtrump', 'murder', 'pro', 'confirms', 'injured', 'riot', 'death', 'officer', 'police']
Trump riots: FBI to investigate death of policeman Brian Sicknick - BBC News


NMF Top 10 words for topic #4:
['know', 'coup', 'taking', 'act', 'coming', 'committing', 'security', 'skirting', 'people', 'weapons']
Invoking Martial Law to Reverse the 2020 Election Could be Criminal Sedition
```

El resultado de la ejecución de los algoritmos es el datagrama original más la columna Topic, que indica el número de tópico al que se relaciona el tweet.

	data__id	data__lang	data__text	Topic
0	1347495838839238657	en	@PapaGlider @Jessica26307123 @MontyBoa99 @real...	1
1	1347495838688440320	en	US Capitol: Police confirms death of officer i...	3
2	1347495838063284230	en	@HookRocky @NBCNews @NBCNewsTHINK At least we ...	1
3	1347495834833747969	en	Mike Pompeo Says Capitol Riot Proves U.S. Isn'...	1
4	1347495834439606272	en	US Capitol Attack: President Trump Can't Handl...	1
...	...	...	...	...
2423	1347561137278119936	en	Uh oh. Maybe they should've worn masks? \n\nI ...	3
2424	1347561136690843649	und	??\n\nhttps://t.co/FprykuE9bN	0
2425	1347561136242126849	en	Oregon representative allowed protesters into ...	2
2426	1347561136187727873	en	@realDonaldTrump The 5 deaths ARE ALL ON YOUR ...	1
2427	1347561136087052288	en	@RobbieBarstool Or the capitol	0

Clasificación de Posturas en Tweets sobre el Asalto al Capitolio de Estados Unidos 2021

Tweets etiquetados por tópico y postura

	data__id	data__lang	data__text	Topic	Stance
0	1347495838839238657	en	@PapaGlider @Jessica26307123 @MontyBoa99 @real...	1	FAVOR
1	1347495838688440320	en	US Capitol: Police confirms death of officer i...	3	AGAINST
2	1347495838063284230	en	@HookRocky @NBCNews @NBCNewsTHINK At least we ...	1	AGAINST
3	1347495834833747969	en	Mike Pompeo Says Capitol Riot Proves U.S. Isn'...	1	FAVOR
4	1347495834439606272	en	US Capitol Attack: President Trump Can't Handl...	1	AGAINST
...	...	...	...	...	...
2423	1347561137278119936	en	Uh oh. Maybe they should've worn masks? \n\nI ...	3	FAVOR
2424	1347561136690843649	und	??\n\nhttps://t.co/FprykuE9bN	0	AGAINST
2425	1347561136242126849	en	Oregon representative allowed protesters into ...	2	FAVOR
2426	1347561136187727873	en	@realDonaldTrump The 5 deaths ARE ALL ON YOUR ...	1	AGAINST
2427	1347561136087052288	en	@RobbieBarstool Or the capitol	0	AGAINST

El número de tweets agrupados en “En contra” y “A favor”

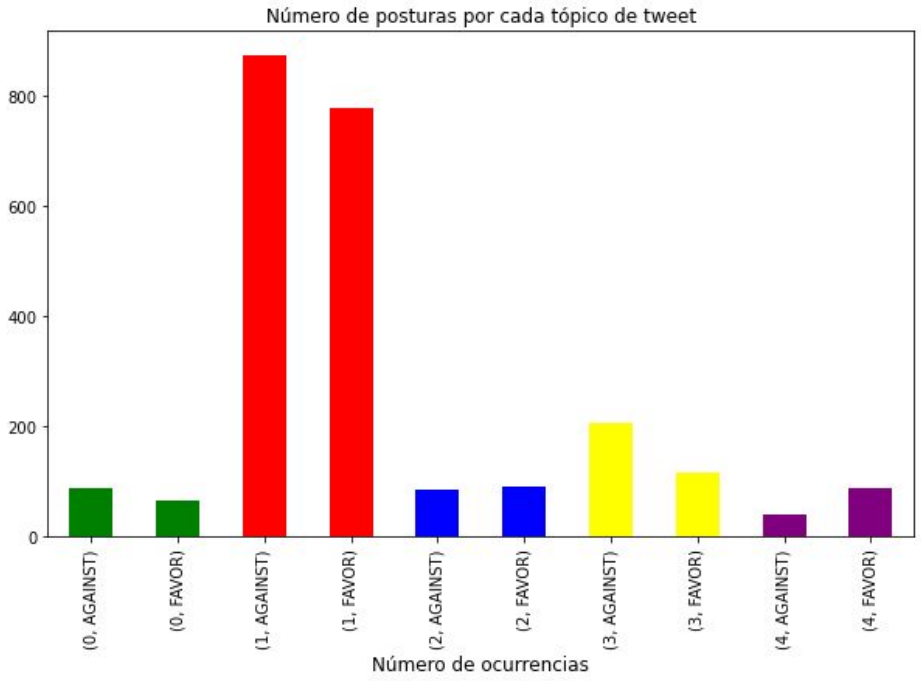
```
Stance
AGAINST    1293
FAVOR      1135
Name: data__id, dtype: int64
```

Identificador del tópico y cantidad de miembros en cada uno

```
Topic
0     152
1    1649
2     175
3     324
4     128
Name: data__id, dtype: int64
```

Cantidad de miembros clasificados en a favor y en contra por tópico

```
Conteo de posturas por tópico LDA:
Topic  Stance
0      AGAINST     87
       FAVOR       65
1      AGAINST    873
       FAVOR      776
2      AGAINST     85
       FAVOR       90
3      AGAINST    207
       FAVOR      117
4      AGAINST     41
       FAVOR       87
Name: data__id, dtype: int64
```



En el gráfico se nota que el tópico más sobresaliente es el número 1. Además

- Para el tópico 0 la mayoría de personas está **en contra.**
- Para el tópico 1 la mayoría de personas está **en contra.**
- Para el tópico 2 la mayoría de personas está **a favor.**
- Para el tópico 3 la mayoría de personas está **en contra.**
- Para el tópico 4 la mayoría de personas está **a favor.**

Algunos resultados de la clasificación de postura y minado de tópicos

Tomando un tweet de muestra, por ejemplo, aquel que tiene el id 1349223551040106497.
Su texto dice “@ProjectLincoln If he were a good American and had dignity, he should resign because he delayed the elections without proof, I questioned the Supreme Court, he excited thousands of fans and fed him to violence by smashing the house of Democracy. Our Capitol”.
Fue clasificado en el tópico 1 con la postura AGAINST.
El tópico 1, utilizando LDA se refiere a “Who were they? Records reveal Trump fans who stormed Capitol”, mientras que con NMF “What Trump Told Supporters Before Mob Stormed Capitol - The New York Times”.
Podemos intuir que este tweet fue agrupado en un tópico que tiene que ver con los fans de Donald Trump y cómo estos fueron quienes “atormentaron” al capitolio.
La postura AGAINST refleja que está en contra de las acciones de los fans de Donald Trump.

Otro tweet de muestra es el que tiene el id 1347551189920002048.

Este tiene el texto “Blm riots caused over 19 deaths and over 14,000 arrests. The capitol building riots caused 4 deaths and 60 arrests. Do not compare. Bad by both sides but are not close statistically.”
Igualmente está agrupado en el tópico 1, pero parece no tener relación con este, sin embargo la postura asignada fue “FAVOR”.

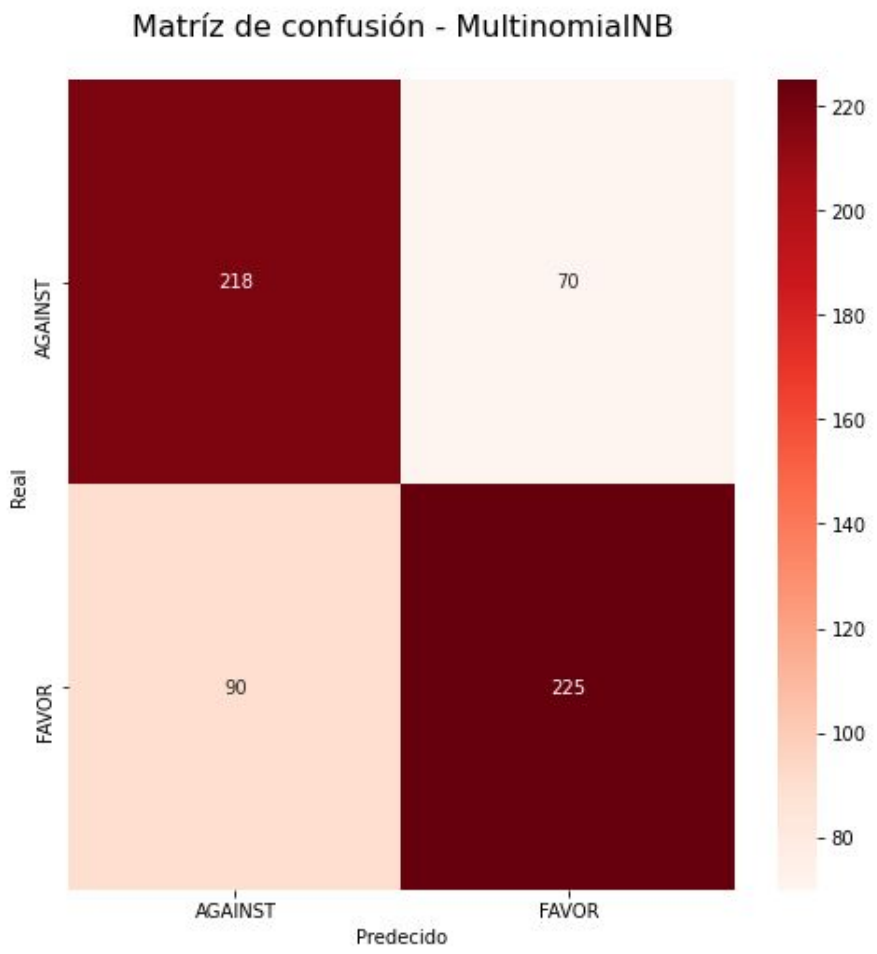
Otro tweet de muestra es el que tiene el id 1347551167195340800.
Este tiene el texto “@PS__Patriot @Spellviper Why didn't the Capitol Police SRT team coming up the stairs not react? Did they take the body away? I'm sure at least one of those guys are TCCC trained. The environment screams fake.....the casualty screams real.”
Está agrupado en el tópico 4, con postura FAVOR.
El tópico 1, utilizando LDA se refiere a “IThe Capitol Police failure to stop a Trump mob, explained - Vox”, mientras que con NMF “Invoking Martial Law to Reverse the 2020 Election Could be Criminal Sedition”.

3.6. Evaluación

Se presentan las métricas de clasificación y matriz de confusión del modelo entrenado usando los datos proporcionados en el Sem-Eval 2016.

MÉTRICAS DE CLASIFICACIÓN

	precision	recall	f1-score	support
AGAINST	0.71	0.76	0.73	288
FAVOR	0.76	0.71	0.74	315
accuracy			0.73	603
macro avg	0.74	0.74	0.73	603
weighted avg	0.74	0.73	0.73	603



Cabe resaltar que aunque son buenas métricas, no reflejan el desempeño que se busca tener para la clasificación de tweets acerca de la toma del Capitolio, pues estos fueron entrenados en temas distintos.
Para evaluar el desempeño real necesitaríamos etiquetar manualmente los tweets recopilados.

4. Conclusiones

Uno de los problemas principales que surgió desde el inicio del proyecto fue que la obtención de una base de datos objetivo, pues a pesar de que Twitter API tiene varias opciones para el filtrado de tweets, muchos de los tweets obtenidos eran acerca de noticias o información objetiva que no representa un argumento o postura. Otro inconveniente fue la obtención de una base de datos para la generación de un modelo que clasifique posturas, puesto que esta clasificación depende del tema objetivo.

Para minar los temas de los tweets utilizamos dos modelos diferentes: LDA y NMF; a través de los cuales obtuvimos 5 temas por cada uno. Después de comparar los resultados obtenidos se puede apreciar que los temas coinciden entre sí, es decir, el tema 0 obtenido por LDA es el mismo al tema 0 de NMF.

Topic 0:

```
LDA Top 10 words for topic #0:
['did', 'make', 'heard', 'peacefully', 'donald', 'people', 'know', 'president', 'building', 'trump']
AP FACT CHECK: Trump's call to action distorted in debate

NMF Top 10 words for topic #0:
['donald', '2021', 'january', 'make', 'heard', 'peacefully', 'soon', 'voices', 'marching', 'patriotically']
Fact Check: Did Trump Say to 'Peacefully and Patriotically' March to the Capitol?
```

Si bien observamos que las palabras en cada uno de los vectores son diferentes, al observar la interpretación de Google Search se puede entender de la misma manera “Revisión de hechos: Donald Trump hace un llamado para tomar acciones contra el Capitolio”.
La misma situación se presenta en todos los tópicos.

Topic 1:

```
LDA Top 10 words for topic #1:
['terrorists', 'did', 'weapons', 'don', 'know', 'election', 'police', 'stormed', 'just', 'people']
Who were they? Records reveal Trump fans who stormed Capitol

NMF Top 10 words for topic #1:
['like', 'mob', 'did', 'violence', 'just', 'stormed', 'supporters', 'people', 'amp', 'trump']
What Trump Told Supporters Before Mob Stormed Capitol - The New York Times
```

Ambos tópicos hablan sobre los fans o seguidores de Trump.

Topic 2:

```
LDA Top 10 words for topic #2:
['violence', 'took', 'dc', 'pro', 'insurrection', 'supporters', 'police', 'attack', 'riot', 'trump']
Capitol riots: Five takeaways from the arrests - BBC News

NMF Top 10 words for topic #2:
['mo', 'lead', 'paul', 'organizer', 'gosar', 'andy', 'biggs', 'job', 'inside', 'attack']
Paul Gosar, Andy Biggs credited in video with organizing Trump crowd in DC
```

Los tópicos se refieren a lo que se pudo concluir después de los hechos, como las personas involucradas.

Topic 3:

```
LDA Top 10 words for topic #3:
['building', 'storming', 'officer', 'death', 'mob', 'stormed', 'inside', 'police', 'attack', 'trump']
Police officer dies after pro-Trump mob attack on US Capitol | US Capitol breach | The Guardian

NMF Top 10 words for topic #3:
['investigation', 'realdonaldtrump', 'murder', 'pro', 'confirms', 'injured', 'riot', 'death', 'officer', 'police']
Trump riots: FBI to investigate death of policeman Brian Sicknick - BBC News
```

Se habla sobre la muerte del policía Brian Sicknick después del ataque de los seguidores de Trump.

Topic 4:

```
LDA Top 10 words for topic #4:
['stop', 'killed', 'congress', 'didn', 'like', 'just', 'building', 'trump', 'police', 'people']
The Capitol Police failure to stop a Trump mob, explained - Vox

NMF Top 10 words for topic #4:
['know', 'coup', 'taking', 'act', 'coming', 'committing', 'security', 'skirting', 'people', 'weapons']
Invoking Martial Law to Reverse the 2020 Election Could be Criminal Sedition
```

Ambos cuestionan la forma en la que las autoridades actuaron ante los hechos ocurridos en el Capitolio.

El modelo obtenido para la clasificación de posturas usando la base de datos de Sem-Eval 2016 obtuvo resultados aceptables, con una f-score de 0.73, sin embargo, este resultado es válido sólo para tópicos sobre los que fue entrenado (Hilary Clinton, Legalización del aborto, etc.).
Al momento de clasificar los tweets acerca del Asalto al Capitolio notamos que falla demasiado, pues hay tweets que expresan una posición en contra de los actos vandálicos ocurridos en el Capitolio, pero el modelo termina clasificándolos como “a favor”. Algo más que pudimos notar es que no existen tweets que expresen una clara postura a favor de los actos vandálicos ocurridos en el Capitolio y creemos que fue porque Twitter limitó el alcance de este tipo de tweets y los censuró por promover la violencia, o realizar declaraciones sin argumentos. El caso más sonado fue el cierre de la cuenta de Donald Trump.

De acuerdo a los resultados obtenidos, que establecen que 1293 tweets fueron de posturas en contra y 1135 a favor de los actos vandálicos, se podría pensar que la hipótesis inicial no es cierta (*La gran mayoría de la gente está en contra de las acciones vandálicas ocurridas en el Asalto al Capitolio de los Estados Unidos en el año 2021),* pues a pesar que la mayoría fueron posturas en contra, la diferencia entre cantidad de posturas es baja (de apenas 158, lo que representa un 6.5% del total). Sin embargo, por todos los problemas que mencionamos acerca de los datos extraídos y el modelo de clasificación creemos que los resultados de clasificación no son suficientes para probar o refutar la hipótesis.

Para complementar el estudio iniciado en este proyecto, se podría depurar los datos objetivo manualmente con el propósito de filtrar aquellos tweets que no representan argumentos o posturas. También, pueden implementarse métodos de clasificación más complejos, como el presentado por Prashanth Rao, en el que utiliza redes neuronales pre entrenadas.

Creemos también que el uso de google para facilitar la interpretación de los tópicos es una técnica que puede ayudar en la realización de proyectos similares de otras personas. Algo que se podría mejorar sería no utilizar Google, sino un buscador que permita la automatización de peticiones web, pues en caso de correr demasiadas veces el script, Google bloquea la cuenta porque detecta el uso de bots.

Bibliografía

1. Sobhani, P., Inkpen, D., & Matwin, S. (2015, June). *From argumentation mining to stance classification*. In Proceedings of the 2nd Workshop on Argumentation Mining (pp. 67-77). https://www.academia.edu/download/48947916/thats_a_fact_ACL16.pdf#page=79
2. Rao, P. (2019). *Transfer Learning in NLP for Tweet Stance Classification*. Towards data science. https://towardsdatascience.com/transfer-learning-in-nlp-for-tweet-stance-classification-8ab014da8dde

3. Malik, U. (s.f.). *Python for NLP: Topic Modeling*. Stack Abuse. https://stackabuse.com/python-for-nlp-topic-modeling/
4. Reyes. S. (2019). *Multi-class text classification (TFIDF).* Kaggle. https://www.kaggle.com/selener/multi-class-text-classification-tfidf
5. Velasquez-Ailva, J. D. Extracción de información y conocimiento de las opiniones emitidas por los usuarios de los sistemas web 2.0.
6. Estévez-Velarde, S., & Cruz, Y. A. (2015). Evaluación de algoritmos de clasificación supervisada para el minado de opinión en twitter. Investigación Operacional, 36(3), 194-205.
7. Estupiñan, J. J., Ramírez, D. A. G., & Santa, F. M. (2016). Implementación de algoritmos basados en máquinas de soporte vectorial (SVM) para sistemas eléctricos: revisión de tema. Tecnura: Tecnología y Cultura Afirmando el Conocimiento, 20(48), 149-170.
8. Suárez, E. J. C. (2014). Tutorial sobre máquinas de vectores soporte (SVM). Tutorial sobre Máquinas de Vectores Soporte (SVM), 1-12.