PRUEBA DE CONOCIMIENTOS DATOS NO ESTRUCTURADOS II

Hebert Gomez

Procesamiento y Análisis de Datos de Tweets

Carga y Preprocesamiento de Datos

Inicialmente se analiza el contenido del archivo CSV, en donde se encuentra como principal fuente de información para el análisis los mensajes contenidos en la columna `Embedded_text`.

	UserScreenI	UserName	Timestamp	Text	Embedded_text	Emojis	Comments	Likes	Retweets	Image link	Tweet URL
0	Andrés Langeba	@ALangebaek	2021-12-01T20:	Andrés Langeba @ALangebaek · 1 dic.	La confianza se afectó. El indicador de confianza Davivienda tuvo una leve 1 7 19		1	7	19	['https://pbs.twim	https://twitter.com/A
1	Plaza Futura	@plaza_futura	2021-12-01T21:	Plaza Futura @plaza_futura · 1 dic.	Buscamos la accesibilidad y mejor ate Banco Cuscatlán Banco Fedecrédito Davivienda Bancoagrícola Banco Promerica	✅ ✅ ✅ ✅ ✅				['https://pbs.twim	https://twitter.com/p
2	Julián Martinez	@JulianM998	2021-12-01T22:	Julián Martinez @JulianM998 · 1 dic.	Señores @Davivienda no he podido ingresar a mi app davivienda, ingreso la cédula y me dice verif 1 1		1		1	[]	https://twitter.com/J

Se inicia cargando un archivo CSV que contiene datos de tweets utilizando Pandas. Este proceso asegura que los datos estén listos para su análisis.

Se define una función `f_preprocess_df` para realizar la limpieza del texto en los tweets. Esto incluye la conversión a minúsculas, eliminación de caracteres especiales, URLs, puntuaciones, dígitos, y caracteres individuales, además de manejar espacios múltiples.

Preparación del DataFrame

La función de preprocesamiento se aplica a la columna `Embedded_text`, que contiene el texto de los tweets procesados.

Se eliminan columnas que no son relevantes para el análisis posterior, como nombres de usuario, texto completo del tweet, marcas de tiempo y otros metadatos.

Se eliminan las filas donde `Embedded_text` comienza con 'respuesta', ya que estos mensajes no hacen parte de la opinión o expresión de un usuario sino de la respuesta del banco.

Este flujo de trabajo garantiza que los datos de los tweets se limpien y se preparen adecuadamente para el análisis subsiguiente, asegurando que solo la información relevante y útil se utilice para la extracción de insights y la toma de decisiones estratégicas.

Procesamiento de Texto para Análisis

Para mejorar la calidad y la eficiencia del análisis, se implementa el filtrado de palabras con **stopwords,** ya que se encuentran palabras (como artículos, preposiciones, pronombres) que no aportan significado semántico relevante para el análisis de los mensajes.

Lista actualizada de stopwords: ['de', 'la', 'que', 'el', 'en', 'y', 'a', 'los', 'del', 'se', 'las', 'por', 'un', 'para', 'con', 'una', 'su', 'al', 'lo', 'como', 'más', 'pero', 'sus', 'le', 'ya', 'o', 'este', 'sí', 'porque', 'esta', 'entre', 'cuando', 'muy', 'sin', 'sobre', 'también', 'me', 'hasta', 'hay', 'donde', 'quien', 'desde', 'todo', 'nos', 'durante', 'todos', 'uno', 'les', 'ni', 'contra', 'otros', 'ese', 'eso', 'ante', 'ellos', 'e', 'esto', 'mí', 'antes', 'algunos', 'qué', 'unos', 'yo', 'otro', …… 'tuvierais', 'tuvieran', 'tuviese', 'tuvieses', 'tuviésemos', 'tuvieseis', 'tuviesen', 'teniendo', 'tenido', 'tenida', 'tenidos', 'tenidas', 'tened', 'si']

Obteniendo así, unos mensaje sin ruido y facilitando la interpretación para los modelos de procesamiento de texto.

Análisis de Frecuencia de Palabras

Para facilitar la identificación rápida de patrones y temas predominantes, se utilizan herramientas como **CountVectorizer** convirtiendo datos de texto en una matriz de términos y calculando la frecuencia de las palabras. Luego, se crea un DataFrame ordenado por estas frecuencias con el fin de generar una nube de palabras para visualizar las más comunes.

De esta forma, se genera una representación gráfica intuitiva de los temas principales en los tweets.

Con las palabras clave identificadas, se procede a crear una agrupación de términos por área de servicio:

Atención al cliente: Incluye palabras clave como "respuesta", "línea", "problema", "caso", entre otras.

Soporte: Contiene palabras clave como "app", "web", "acceso", "ingresar".

Marketing: Engloba palabras clave como "comprar", "crédito", "cuenta", "beneficios".

Productos: Agrupa palabras clave específicas como "daviplata", "corredores", "segurosbolivar", "cajero".

Analizador de sentimientos

Se aplica el analizador de sentimientos (sentiment) a cada entrada de texto en la columna 'Embedded_text' del DataFrame. Esto se realiza utilizando apply() junto con una función lambda: lambda x: sentiment.sentiment(x).

En este contexto, x representa cada cadena de texto en la columna 'Embedded_text'. El método sentiment.sentiment(x) calcula y asigna puntuaciones de sentimiento o etiquetas basadas en el tono emocional y el contenido del texto.

Este proceso ayuda a comprender el contexto emocional o la polaridad del sentimiento asociado con los datos textuales, facilitando análisis o insights adicionales sobre el conjunto de datos.

Sumarización de Texto

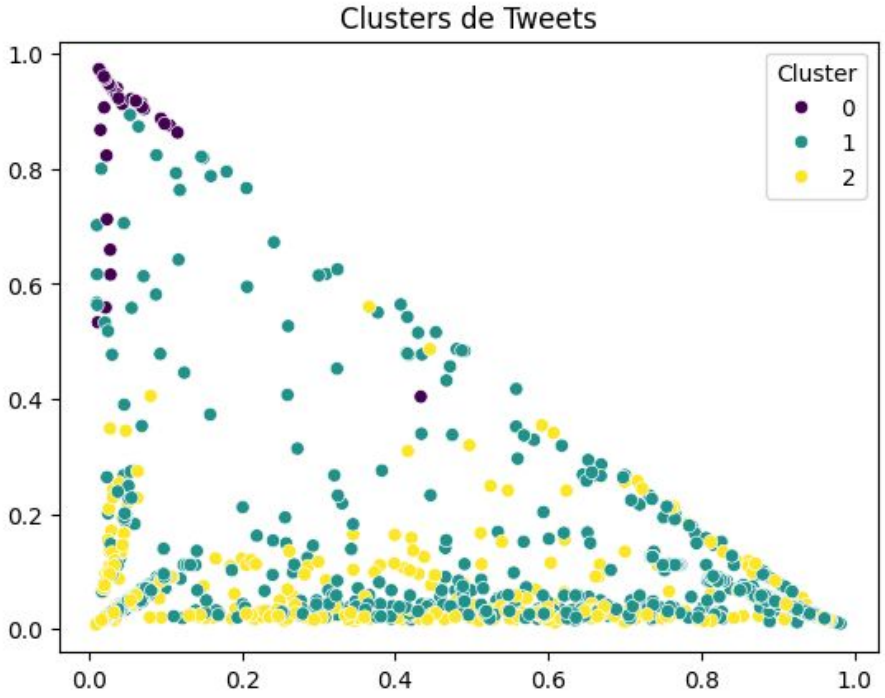
La sumarización de texto mediante el algoritmo Luhn, es una técnica eficaz para condensar información clave de documentos extensos. Con esto en mente, se procede a extraer las ideas principales y frases relevantes de los tweets.

Embedded_text	Embedded_text_Sumy
a gente está cansada del terrorista petrogustavo engaña bobos por eso en san gil le dieron en la jeta ojalá se siga replicando nivel nacional demostrarle que nadie lo quiere que el solo quiere destruir al pais palmaedwin eres buen dirigente pero estas como davivienda	la gente está cansada del terrorista petrogustavo engaña bobos por eso en san gil le dieron en la jeta ojalá se siga replicando nivel nacional demostrarle que nadie lo quiere que el solo quiere destruir al pais palmaedwin eres buen dirigente pero estas como davivienda
os cajeros de davivienda en ibagué inservibles qué hacen los gerentes de esas entidades que no se inmutan ante la urgencia de dinero entretanto la gente recurre los de las otras entidades para que le asalten el bolsillo	los cajeros de davivienda en ibagué inservibles qué hacen los gerentes de esas entidades que no se inmutan ante la urgencia de dinero entretanto la gente recurre los de las otras entidades para que le asalten el bolsillo
en respuesta pulzo ése hp habla mierda con esos bancos davivienda que son de éste señor cuánto no roba cuánto no tendrá comprado amigos del gobierno	en respuesta pulzo ése hp habla mierda con esos bancos davivienda que son de éste señor cuánto no roba cuánto no tendrá comprado amigos del gobierno
kfccolombia son unos hijos de putahice una compra por valor de me la cargaron ellos por error dos veces mi cuenta de davivienda no me devuelven el dinero ahira se hacen los locos para responder en único de cali estoy de paseo sicsuper ayuda por favor	kfccolombia son unos hijos de putahice una compra por valor de me la cargaron ellos por error dos veces mi cuenta de davivienda no me devuelven el dinero ahira se hacen los locos para responder en único de cali estoy de paseo sicsuper ayuda por favor
en respuesta subirath lineauribista como cada jefe toma sus propias decisiones en su entorno estasra esta como la publicidad de davivienda en el lugar equivocado	en respuesta subirath lineauribista como cada jefe toma sus propias decisiones en su entorno estasra esta como la publicidad de davivienda en el lugar equivocado
davivienda que pésimo servicio el de los cajeros de davivienda los cajeros del minuto dios quirigua titán dañados imposible sacar plata	davivienda que pésimo servicio el de los cajeros de davivienda los cajeros del minuto dios quirigua titán dañados imposible sacar plata
henryfloresst por favor alcalde venga poner orden por medio del camdesantatecla la santateclasv frente al banco davivienda estos vendedores están reventando cohetes habiendo viviendas adultos mayores que no pueden estar tranquilos además en la bodega de tienda godo	henryfloresst por favor alcalde venga poner orden por medio del camdesantatecla la santateclasv frente al banco davivienda estos vendedores están reventando cohetes habiendo viviendas adultos mayores que no pueden estar tranquilos además en la bodega de tienda godo

Se encuentra que este método no corresponde al enfoque correcto, debido a la naturaleza de los mensajes y su longitud. Esto se observa con los resultados de la columna Embedded_text_Sumy, en dónde no existe diferencia apreciable con el texto original.

Análisis de Temas (Topic Modeling) con LDA

Con esta técnica, se procede a explorar y comprender los temas predominantes en un conjunto de datos de texto, específicamente tweets en este caso, y agruparlos en clusters para una comprensión intuitiva y visual de los datos.



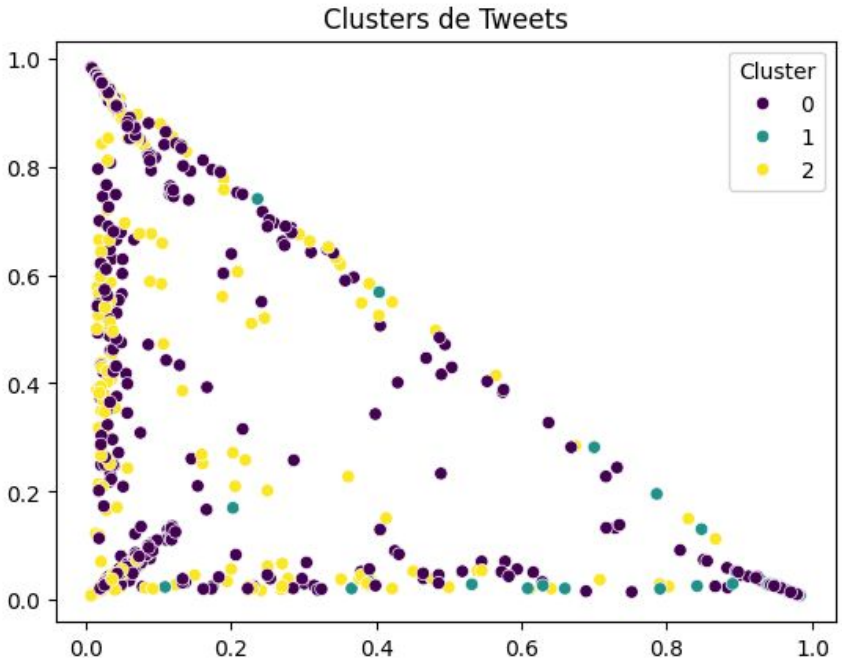
Análisis de Temas (Topic Modeling) con LDA

Este tipo de gráfico permite una visualización rápida y efectiva de cómo los tweets se agrupan en diferentes clusters basados en características específicas.

En este caso se utilizó **n_clusters=3**, basándose en la cantidad de temas:
Tema 0:['davivienda', 'respuesta', 'no', 'respondiendo', 'banco', 'servicio', 'lugar', 'cuenta', 'equivocado', 'app']

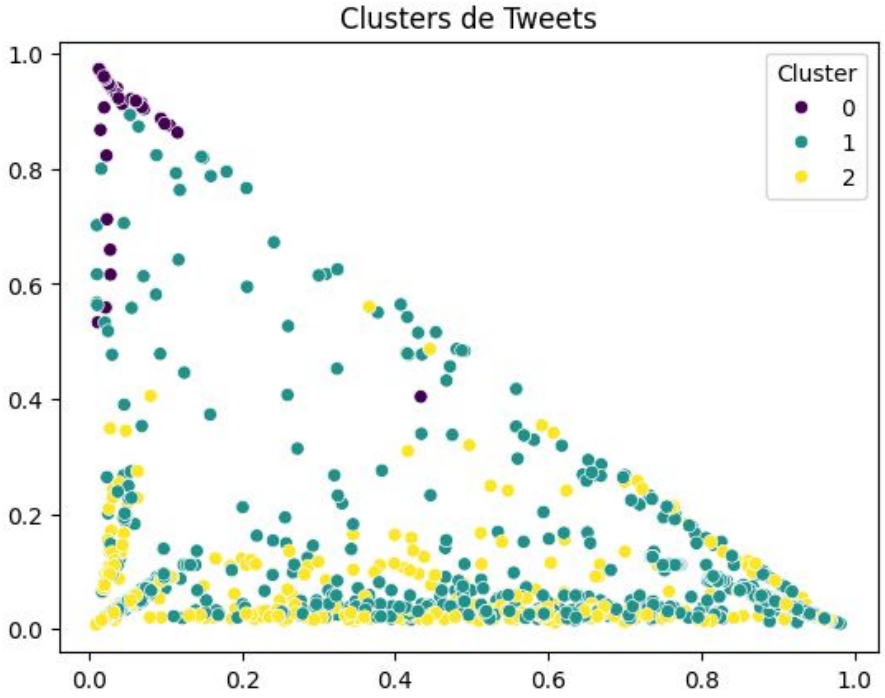
Tema 1:['respuesta', 'mensaje', 'privado', 'caso', 'favor', 'atentos', 'quedamos', 'gusto', 'buenos', 'lamentamos']

Tema 2:['davivienda', 'no', 'respuesta', 'daviplata', 'dinero', 'cuenta', 'banco', 'solución', 'bancolombia', 'sfcsupervisor'].



Análisis de Temas (Topic Modeling) con LDA

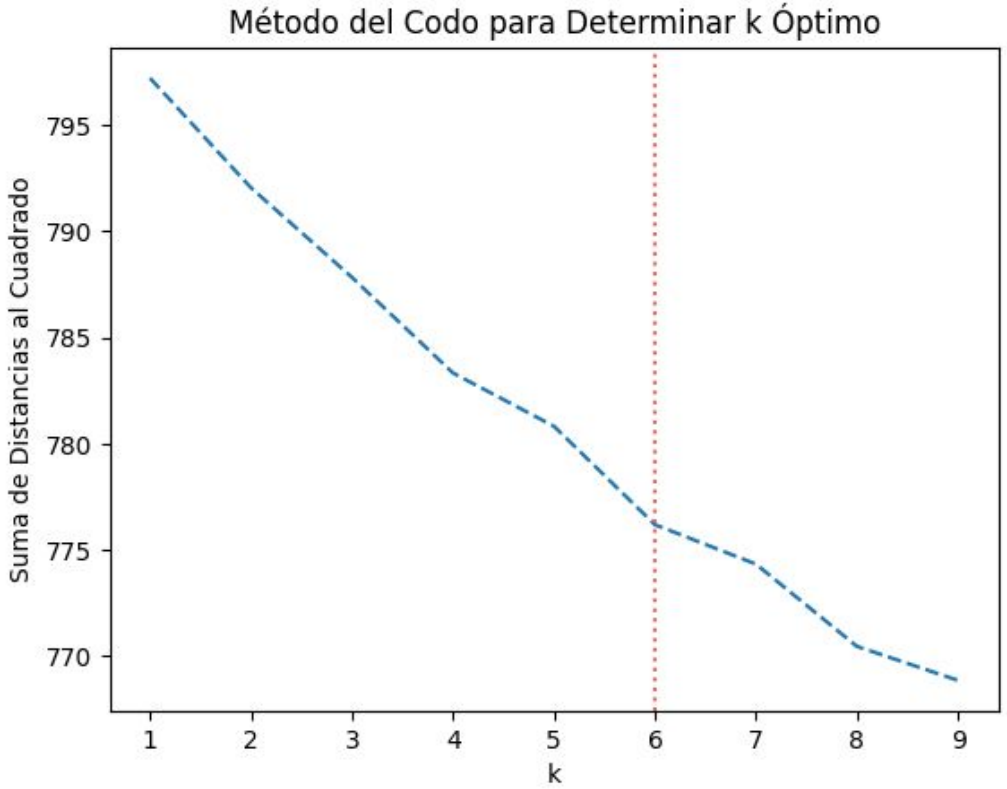
Cabe resaltar que no se observa una agrupación muy clara entre los datos, específicamente para el grupo 1, lo cual puede significar que la data no está lo suficientemente ligada o que se requieren procesos adicionales de limpieza.



Determinación del número óptimo de clusters usando el método del codo

Para lograr una mejor interpretación de los resultados, se procede a determinar el número óptimo de clusters.

Se observa que un punto de inflexión de la gráfica pronunciado se encuentra en 6.



Con **n_clusters=6**, se encuentra la siguiente agrupación

index	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
0	cuenta	bancolombia	respondiendo	daviplata	no	banco
1	davivienda	davivienda	davivienda	solución	davivienda	concierto
2	plata	página	lugar	hermano	dinero	davivienda
3	no	no	equivocado	nadie	app	hoy
4	nequi	caída	noticiasrcn	davivienda	sirve	navidad
5	ahorros	app	día	no	servicio	andrescepeda
6	cuentas	web	no	problema	pasa	httpsbitly
7	hicieron	vez	favor	discapacitado	hacer	saber
8	ustedes	veces	hacer	dan	daviplata	gran
9	cajeros	dos	tarjetas	da	sfcsupervisor	livedataifx
10	sacar	ver	navidad	señores	cajero	loquehoydebesaber
11	respondiendo	ser	app	whatsapp	días	davicorredores
12	dinero	bancos	marianiniecheve	sicsuper	señores	pm
13	daviplata	mes	ustedes	respuesta	puede	corredores
14	responde	hace	crédito	servicio	ingresar	debe

Analizador de sentimientos

Retomando la categorización de las palabras:
categories = {

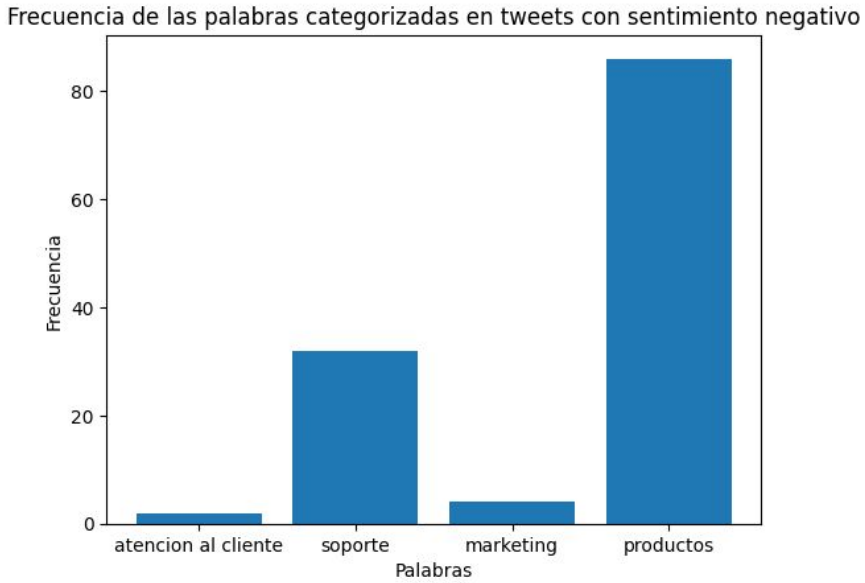
"atencion al cliente": ["respuesta", "linea", "problema", "caso", "servicio", "solicitud", "ayuda", "atencion","hurto"],

"soporte": ["app", "web", "acceso", "ingresar"],

"marketing": ["comprar", "crédito", "cuenta", "beneficios"],

"productos": ["daviplata", "corredores", "segurosbolivar", "cajero"]

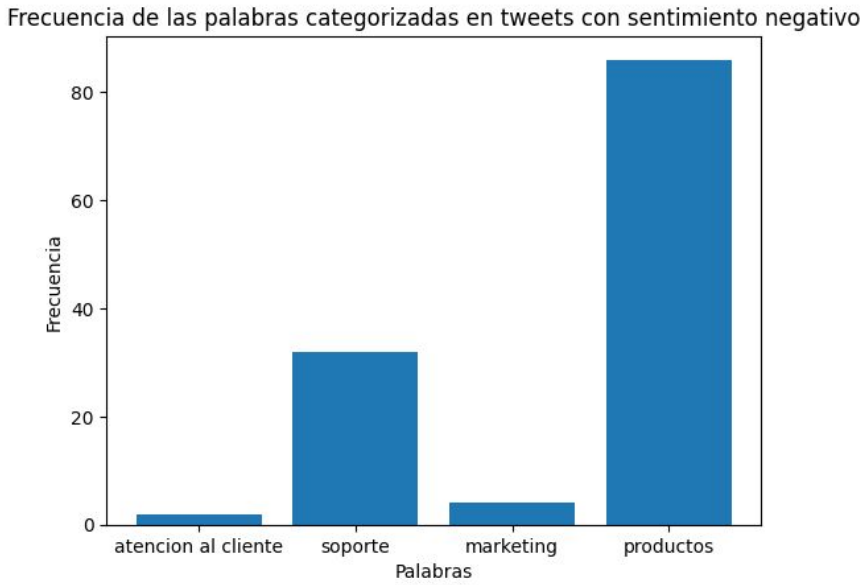
}



Analizador de sentimientos

Se grafica la frecuencia de las palabras categorizadas en tweets con sentimiento negativo.

En donde se puede relacionar cada área de servicio y su frecuencia de menciones negativas en los mensajes, resultando en este caso que los mensajes con más connotación negativa están ligados a temas de productos y soporte.



Resumir una Frase a su Palabra Más Común

Se implementa la función summarize_to_one_word extrae la palabra más relevante o frecuente de cada frase después de eliminar palabras vacías y puntuación, lo cual es útil para resumir el contenido de texto en una sola palabra clave.

Luego se preprocesa el texto dividiéndolo en oraciones, tokenizando palabras, y eliminando puntuación y stopwords.

Siguiente se retorna el modelo Word2Vec entrenado

Se crea un texto que contenga todos los tweets de la columna 'Embedded_text' con el fin de entrenar el modelo Word2Vec y

Con esto, se calcula la similitud entre dos palabras, por ejemplo: la similaridad 'confianza' y 'davivienda' es: 0.018932780250906944, por lo que según el modelo **Word2Vec** utilizado para calcular esta métrica,estas dos palabras tienen una relación bastante baja en términos de contexto y significado