# **Prueba técnica Departamento de datos no estructurados: Análisis de tweets (NLP)**

Una de las principales aplicaciones del procesamiento natural del lenguaje es extraer automáticamente de grandes volúmenes de texto los temas que discuten las personas. Algunos ejemplos de grandes volúmenes de texto pueden ser noticias de redes sociales, opiniones de clientes, comentarios de usuarios, noticias, etc.

Saber de qué hablan las personas y comprender sus problemas y opiniones es muy valioso para empresas, administradores y campañas políticas. Y es realmente difícil leer manualmente volúmenes tan grandes y recopilar los temas.

Por lo tanto, se necesita un algoritmo automatizado que pueda leer los documentos de texto y extraer automáticamente los temas tratados. Tomando como caso de estudio el análisis de tweets extraídos de la red social de Twitter que contienen la palabra **Davivienda** y con el cual se quiere conocer cuál es la interacción que tienen los diferentes usuarios de la red social con el Banco Davivienda.

## Stack de tecnológico

Para el desarrollo del proyecto se empleó el lenguaje de programación Python en su versión 3.9.12 haciendo uso de librerías como “pandas” y “numpy” para el manejo de matrices. “matplotlib”, “seaborn” y “wordcloud” para ilustraciones y gráficos. Para el preprocesamiento y análisis de texto se emplearon librerías como “re” para expresiones regulares, “spacy” para obtener las “STOP\_WORDS” para el idioma español y “stanza” para la lematización de los tweets en idioma español. “sklearn” para vectorizar texto y realizar análisis de de N-gramas. Para el modelo LDA y la matriz termino documento utilicé la librería “gensim”.

La decisión de trabajar con estas librerías fue por su popularidad y buena documentación que presentan.

## Exploración y Limpieza de Datos

El proyecto comenzó con una exploración detallada de una base de datos denominada davivienda\_tweets.csv, que contiene tweets relacionados con el Banco Davivienda para el mes de diciembre del año 2021, este dataset cuenta con 1811 registros y 12 variables (“Unnamed: 0”, “UserScreenName” , “UserName”, “Timestamp“ , “Text” , “Embedded\_text”, “Comments” ,”Emojis” , “Likes”, “Retweets”, “Tweet URL”, ”Image link”).

Se decide eliminar las columnas “Unamed: 0”, “Tweet URL”, “Image Link” y “Emojis”, principalmente porque no aportan valor semántico para el análisis y la mayoría de los registros son valores NaN.

Posteriormente se realiza la imputación de datos faltantes, comenzando con las variables numéricas (“Likes”, “Retweets”, “Comments”). En este punto se encontró una novedad en la imputación y conversión a tipo de dato numérico pues por naturaleza twitter abrevia el número de comentarios, likes y retweets por miles y le agrega al final el sufijo "mil". En este punto se diseñó una función especial para la imputación y conversión a dato numérico, eliminando el "mil", multiplicar por 1000 y convertir a numérico.

Por último, se eliminan registro duplicados que en este caso fue 0 y la Imputación valores faltantes en “UserScreenName” utilizando valor fijo correspondiente a la columna de UserName. Como resultado de estas operaciones se logró obtener un dataset completo y sin valores faltantes.

### **Análisis de usuarios con mayor número de tweets**

Se realiza un conteo de número de tweets realizados por cada usuario con el fin de obtener el top 10, como lo muestra la siguiente figura.



Figura 1. Top 10 usuarios con mayor número de tweets

Dentro de este top se encuentra en primer lugar el perfil de Davivienda, el cual de acuerdo con la exploración realizada se observa que la mayoría de los tweets son de soporte para los usuarios. Estos tweets no aportan al análisis, debido a que la mayoría repiten las mismas palabras e invitan al usuario a comunicarse con el banco para resolver sus inquietudes.

Esto puede generar ruido para el entrenamiento del modelo LDA y es por esto por lo que se decide eliminar los tweets que provengan de este perfil los cuales en su contenido incluyan la palabra como “En respuesta a”.

### **Análisis de interacciones por día**

Gráfico, Gráfico de líneas

Descripción generada automáticamenteOtro tipo de análisis son la cantidad de interacciones (Comentarios, Likes, Retweets) por día, como se muestra en la siguiente ilustración:

Figura 2. Interacciones diarias en tweets

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 3. Promedio de interacciones Diarias en tweets

Los resultados muestran un bajo nivel de interacciones (Comentarios, Likes, Retweets) por día en el periodo de muestra correspondiente al mes de diciembre del 2021. Sin embargo, se grafican dos picos representativos los cuales fueron los días 15 de diciembre y 17 de diciembre de 2021, los cuales fueron excluidos para obtener el promedio pues pueden considerarse como días atípicos, resultando un valor no mayor a 12 interacciones promedio por día, especialmente de Retweets.

Esta es una característica común de los datos de redes sociales, donde algunos contenidos se vuelven virales mientras que la mayoría permanece con bajo nivel de interacción.

En la actualidad se encuentran usuarios que dada la gran cantidad de seguidores que tienen en las redes sociales, al momento de realizar una publicación de cual cualquier temática generan gran impacto en muchos usuarios, como es el caso del senador Wilson Neber Arias Castillo el cual realizó una denuncia publica el 17 de diciembre del 2021.

### **Distribución de la Longitud de los Textos de los Tweets**

Otro aspecto importante es el análisis de la longitud promedio de los tweets, La mayoría de los tweets tienen una longitud promedio de 184 caracteres, lo que es esperado dada la naturaleza concisa de los tweets. Existe una distribución relativamente normal con una cola hacia la derecha, indicando la presencia de algunos tweets más largos, como lo muestra la figura 4.

Gráfico, Histograma

Descripción generada automáticamente

Figura 4. Distribución de la longitud de los tweets

Esto puede ser un indicador de la complejidad o la riqueza del contenido. Al realizar la distribución de la longitud de los textos, puedes identificar grupos de tweets con longitudes similares. Esto puede ayudar en la segmentación de documentos antes de aplicar el modelo LDA, lo que puede mejorar la interpretación y la calidad de los tópicos extraídos.

Los tweets con longitudes extremadamente cortas o largas pueden indicar diferentes cosas. Por ejemplo, tweets muy cortos pueden ser simplemente enlaces o mensajes sin mucho contenido informativo, mientras que tweets muy largos pueden contener discusiones detalladas o análisis. Entender la distribución de la longitud de los textos puede ayudar a discernir la calidad y la complejidad del contenido.

## Preprocesamiento de Texto

Para llevar a cabo el entrenamiento del modelo LDA, es importante procesar el texto de los tweets para que sea eficiente y legible para los modelos. Se ha incluido 4 pasos principales en el preprocesamiento:

* Poner el texto en minúsculas: para que el algoritmo no trate las mismas palabras en distintos casos como diferentes.
* Eliminación de ruido: Todo lo que no está en un número o letra estándar, como puntuación, valores numéricos, texto común sin sentido (/n), texto HTML.
* Tokenización: La tokenización es el término utilizado para describir el proceso de conversión de cadenas de texto normales en una lista de tokens, es decir, palabras que realmente queremos.
* Eliminación de palabras vacías (stop words): Las palabras muy comunes que parecen tener poco valor para ayudar a seleccionar los documentos para le modelo LDA se excluyen completamente del vocabulario. Estas palabras se denominan "stop words".
* Lematización: se decide lematizar las palabras en lugar de hacerles un estematización, ya que la lematización considera la estructura y propiedades gramaticales de las palabras y las convierte a una forma base, mientras que la estematización es una técnica que se utiliza para reducir una palabra hasta la raíz de la palabra. Esto puede resultar en raíces que no son realmente palabras reales. Por lo tanto, la lematización es mejor para la modelización de temas ya que conserva el significado semántico de las palabras. Sin embargo, computacionalmente, es más costoso.

### **Análisis de N-gramas**

Para analizar el contenido de cada tweet se opta por extraer las características de N-gramas.

los N-gramas capturan la estructura del lenguaje desde el punto de vista estadístico, como qué letra o palabra es probable que siga a la dada. Cuanto más largo sea el n-grama (cuanto mayor sea n), más contexto tendrá para trabajar. si los n-gramas son demasiado cortos, es posible que no capten diferencias importantes. Por otro lado, si son demasiado largos, puede que no capte el "conocimiento general" y se limite a casos particulares, como se observa a continuación:

Gráfico

Descripción generada automáticamenteInterfaz de usuario gráfica

Descripción generada automáticamente con confianza bajaGráfico, Gráfico de barras

Descripción generada automáticamente

Figura 5. Top 20 Unigramas, Bigramas, Trigramas

Del análisis se puede inferir que las preocupaciones o temas comunes incluyen transacciones monetarias, problemas con servicios digitales o transacciones entre plataformas, atención al cliente, términos de contratos y posiblemente errores o fraudes.

Dada la presencia de palabras como “lugar equivocado” son indicadores del gran éxito que presentó esta campaña publicitaria por parte del banco Davivienda. Donde las personas utilizan estos términos para referirse a problemas con sus productos como tarjetas de crédito, cuentas y cajeros.

### Análisis de Nube de Palabras

La nube de palabras proporcionada refleja términos y frases que se usaron frecuentemente en tweets relacionados con el banco Davivienda en diciembre de 2021. Estas son útiles para visualizar de forma rápida y directa los términos más mencionados dentro de un gran conjunto de datos de texto.

Algunas observaciones y análisis basados en esta nube:

**Posibles problemas de servicio**: Palabras como "problema", "esperar", "necesitar", "funcionar" y "solución" sugieren que los usuarios podrían estar discutiendo problemas con los servicios del banco y buscando asistencia o soluciones a esos problemas.

**Interacciones con el cliente**: La frecuencia de palabras como "llamar", "asesor", "ayudar", "atender" y "gracias" sugiere que hay una cantidad significativa de comunicación entre los clientes y el banco. Esto podría incluir solicitudes de ayuda, agradecimientos por el servicio o necesidad de asistencia de un representante del banco.

**Transacciones y servicios**: Términos específicos de transacciones como "transferencia", "pagar", "cuenta", "saldo", y "crédito" sugieren que los tweets también incluyen discusiones detalladas sobre los servicios financieros proporcionados por el banco.

La naturaleza de las palabras sugiere que podría haber una preocupación general sobre la eficiencia del servicio, la seguridad y la atención al cliente.

Texto

Descripción generada automáticamente con confianza media

Figura 6. Nube de palabras

## Matriz termino-documento

Tras la fase inicial de preprocesamiento, se procede a preparar los datos para el entrenamiento del modelo LDA, esto significa en que se debe transformar el texto en un vector (o matriz) de números con sentido. Esto puede hacerse mediante la técnica de Matriz termino-documento, la cual es una representación del texto que describe la aparición de palabras en un documento.

De esta manera, un término que aparece con frecuencia en el corpus de texto tendrá una puntuación mucho más alta, y un término que aparece raramente en el corpus de texto tendrá una puntuación mucho más baja. Se descarta cualquier información sobre el orden o la estructura de las palabras en el documento y el modelo sólo se preocupa de si las palabras conocidas aparecen en el documento, no de dónde aparecen en el documento.



Figura 7. Matriz termino-documento

## Latent Dirichlet Allocation - Modelo no supervisado de Tópicos.

Ahora que se cuenta con todos los datos necesarios para el entrenamiento del modelo LDA, como el corpus, diccionario y matriz termino-documento, se necesita definir el número de tópicos como se expone en el siguiente bloque de código:

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22 | # establecer parametros para el modelo LDA  num\_topics = **10**  chunksize = **2000**  passes = **20**  iterations = **400**  eval\_every = **1**  # Aplicar modelo LDA  lda\_model\_tuning = gensim.models.ldamodel.LdaModel(  corpus=corpus,  id2word=diccionario,  chunksize=chunksize,  alpha="auto",  eta="auto",  iterations=iterations,  num\_topics=num\_topics,  passes=passes,  eval\_every=eval\_every,  )  # Visualizar los tópicos  lda\_model\_tuning.print\_topics() |

El modelo LDA anterior está construido con 10 temas diferentes donde cada tema es una combinación del top 10 palabras clave y cada palabra clave aporta un peso(importancia) determinado al tema. De tal forma que al observar estas palabras clave se puede inferir cuál es el tema que se está hablando.

En este caso se decide escoger 10 tópicos tomando como base el puntaje de coherencia del modelo, este se obtiene calculando la similitud semántica entre las palabras principales (por tema) es alta.

Se realizaron pruebas construyendo el modelo con un mínimo de 5 temas hasta 25, calculando el puntaje de coherencia, como se muestra en la figura 8.

Esto nos indica que con 10 tópicos se tiene un puntaje de coherencia aproximado a si escogiéramos 16 tópicos e incluso con 20 tópicos. En base a lo anterior si se hubiera escogido 20 o más temas no se podría interpretar de forma sencilla, pues se presentarían solapamientos entre temas y estarían muy cerca uno de otro.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Figura 8. Número óptimo de temas

## Interpretación de Resultados y conclusiones generales.

Ahora que el modelo LDA está listo, el siguiente paso es examinar los temas producidos y las palabras clave asociadas. La herramienta que se escoge es el gráfico interactivo del paquete pyLDAvis. Donde cada burbuja del gráfico de la izquierda representa un tema. Cuanto más grande es la burbuja, mayor es la prevalencia de ese tema. El gráfico de barras horizontales de la derecha representa las palabras más relevantes de cada tema. El gráfico es interactivo, lo que permite seleccionar temas específicos y ver las palabras relacionadas con cada tema, con la esperanza de inferir el significado de cada tema.

Gráfico, Gráfico de burbujas

Descripción generada automáticamente

Del modelo de 10 tópicos se puede interpretar que:

* Tópico 1: "Problemas con la app de Daviviplata"
* Tópico 2: "Problemas con las transferencias de dinero cuentas Davivienda a otros bancos"
* Tópico 3 y 4: "Publicaciones a modo de queja de usuarios con un número considerable de seguidores"
* Tópico 5 y 10: "Problemas con el servicio al cliente, tarjetas y cajeros"
* Tópico 6: "Piloto con la plataforma de Binance y Davivienda"
* Tópico 7: "Referencias a Davivienda utilizando la famosa frase de “En estos momentos su dinero puede estar en el lugar equivocado, tráigalo a Davivienda”
* Tópico 8: "Problemas con el dinero de los clientes y daviplata"
* Tópico 9: "Publicaciones para reportar fraude o estafa por parte de usuario con un número considerable de seguidores"

En general, se evidencia que la mayoría de los usuarios utilizan la plataforma de Twitter para reportar problemas con la app de Daviplata, problemas con las transferencias de dinero a otros bancos, problemas con el servicio al cliente, tarjetas y cajeros, y publicaciones a modo de queja de usuarios con un número considerable de seguidores. Adicionalmente, se evidencia que algunos usuarios utilizan la plataforma para reportar fraude o estafa.

Se probaron distintos métodos de agrupación y búsqueda de temas subyacentes a partir de los tweets del conjunto de datos, como el modelado de temas LDA. Además, se exploraron formas de visualizar los resultados y de compararlos con los temas originales de los tweets.

Sin embargo, probablemente debido a la propia naturaleza del texto (por ejemplo, los tweets son demasiado similares en general, o tienen muy pocas palabras para permitirnos segmentar el tema subyacente), acabamos con algunas distinciones borrosas entre cada tema, como se ve en las palabras principales por cada tópico tras el modelado LDA.