	Fecha: 30/03/2022
Solicitud de Registro de Tema para Titulación	

Quien suscribe, estudiante regular de la Maestría en Ciencia de Datos e Información

Nombre:	Dulce Maria Reyes Lucas
----------------	--------------------------------

Solicitamos ante el Coordinador Académico de la Maestría en Ciencia de Datos e Información, la autorización y registro para titulación en la Maestría, en la modalidad:

	Reporte Analítico de Experiencia Laboral
	Propuesta de Intervención
x	Implementación de un Proyecto
	Solución Estratégica

1.- INFORMACIÓN REFERIDA AL TEMA

TITULO TENTATIVO DEL PROYECTO	Análisis de sentimiento y extracción de tópicos en comentarios de aplicaciones móviles del sector bancario.
--------------------------------------	---

2.- RESUMEN

2.1. **Introducción y antecedentes del tema** (completar por el alumno en digital, máximo 20 líneas)

<p>En los últimos años se ha presenciado el increíble crecimiento y uso de las redes sociales, blogs y demás medios que guardan principalmente textos, estos datos no estructurados se han convertido en el mayor interés de empresas principalmente, debido a que los datos estructurados no son capaces de mostrar tal juicio o sentimiento como se describe en los comentarios de las personas que expresan su sentir respecto a una infinidad de temas.</p> <p>La pandemia obligó a todos los sectores a evolucionar y ofrecer mejores servicios y a la distancia de un clic, poder realizar pagos, recargar saldo, consultar los movimientos de su cuenta, obtener un estado de cuenta, hacer una cita, renovar un servicio, pero principalmente a los bancos, de poder ofrecer la mayoría de los servicios sin necesidad de acudir a una sucursal, por ello, se requiere de una aplicación móvil cada vez más robusta, intuitiva, que cubra estas necesidades y que deje satisfecho al usuario.</p>

Solicitud de Registro de Tema para Titulación

Derivado de esta necesidad se presenta un proyecto que analizará los comentarios para identificar las debilidades de la aplicación móvil de Banco Santander.

2.2. Planteamiento del problema. (Completar por el alumno en digital, máximo 20 líneas)

Aquí se debe explicar muy claramente el problema que se pretende investigar.

Analizar las reseñas de aplicaciones móviles bancarias (Santander) que son escritas por los usuarios en las tiendas de aplicaciones (Google y Apple store) realizando la extracción de tópicos que muestren las debilidades y fortalezas de la aplicación móvil, además, realizar un análisis de sentimiento para determinar la polaridad de los comentarios presentados por los usuarios.

2.3. Objetivo General (máximo 5 líneas) y **Objetivos Específicos** (máximo 10 líneas), (completar por el alumno en digital)

Aquí se debe indicar muy claramente el objetivo general y los objetivos específicos

OBJETIVO GENERAL

Implementar un análisis de sentimiento y modelos de texto de las reseñas en Google Play y Apple Store orientado a la aplicación móvil de Santander para detectar las debilidades y fortalezas de las funcionalidades de dicha aplicación.

OBJETIVOS ESPECÍFICOS

- 1) Realización de web scrapping hacia las tiendas de aplicaciones con Python y posteriormente el procesamiento para obtener un conjunto de documentos listos para poder analizarlos de manera correcta.
- 2) Aplicación del algoritmo de la distancia de Levenshtein para corregir las palabras mal escritas por el usuario.
- 3) Creación de modelos de texto para detectar los puntos de dolor en las funcionalidades de las aplicaciones móviles bancarias.
- 4) Generación de un modelo de análisis de sentimiento para detectar la polaridad de los comentarios de las tiendas de aplicaciones dado que la calificación por estrellas no aporta mucho sentido ni significado.
- 5) Categorización de los tópicos generados por funcionalidad.

Solicitud de Registro de Tema para Titulación

6) Presentar los resultados y conclusiones a través de un dashboard en power bi.

3. Resultados Esperados

3.1. Resultados a los que se pretende llegar con este trabajo (completar por el alumno, digital, máximo 10 líneas)

Con el proyecto propuesto se pretende obtener una generación de tópicos que describan las funcionalidades de las aplicaciones móviles, sus fortalezas y debilidades para potenciales mejoras, además lograr una calificación del sentimiento expresado en los comentarios para poder enfocar los esfuerzos en aquellos tópicos que tienen una polaridad mayormente negativa. Los resultados de esta investigación ayudarán a enfocar las estrategias de mejora de la aplicación del sector bancario.

4. Programa de Trabajo

4.1 Cronograma de actividades estimado para el desarrollo del tema.

Aquí deben especificarse las fechas estimativas para el desarrollo del proyecto

--

5. Metodología

5.1. Metodología por utilizar para el desarrollo del tema (Completar por el Alumno, en digital, máximo 10 líneas).

En mi experiencia estoy familiarizada con la metodología SEMMA, sin embargo, me gustaría utilizar la metodología KDD (Knowledge Data Discovery).

Breve descripción de uso de la metodología:

Selección de los datos: Los datos que usaré en este proyecto se encuentran en las tiendas de aplicaciones de Google y Apple Store, la forma de extracción será a través de una técnica llamada Web Scrapping y podremos extraer datos históricos. El proceso ya lo tengo listo.

Preprocesamiento de los datos: En esta parte de la metodología se realizará la limpieza de los datos, en la que se realizarán tareas como conversión a minúsculas, eliminación de acentos y caracteres especiales, stemming, stopwords, aplicación del algoritmo de Levenshtein para la corrección de las malas escrituras, dejar palabras con

Solicitud de Registro de Tema para Titulación

una longitud mínima de 5 caracteres e identificar entidades. Generar diccionarios propios de stopwords y si es posible de sinónimos.

Transformación de los datos: Generar y entrenar modelos de extracción de tópicos y modelos de clasificación de sentimiento para alcanzar los objetivos.

Interpretación y evaluación: Para este paso se compararán los resultados obtenidos en los modelos de sentimiento con matriz de confusión o alguna otra métrica de desempeño. Mostrar los tópicos finales que tengan una interpretación más clara. Los resultados serán mostrados en un dashboard en la herramienta de Power BI.

6. Fuentes de Información

6.1 Bibliografía física y electrónica mínima a utilizar (completar por el alumno, en digital, mínimo citar 5 referencias).

Comentarios de tiendas de aplicaciones de Apple Store y Google Play Store.

- SuperMóvil. <https://play.google.com/store/apps/details?id=mx.bancosantander.supermovil&hl=es>
- SuperMóvil. <https://apps.apple.com/mx/app/santander-superm%C3%B3vil/id498944221>
- Chong. Sentiment Analysis and Topic Extraction of the Twitter Network of #Prayforparis
- Yeun Kim. Modeling Topic Extraction-based Sentiment Analysis Based on User Reviews
- Gabriele Pergola_, Lin Gui, Yulan He. A Topic-Dependent Attention Model for Sentiment Analysis

7. Índice Tentativo (Máximo 6 líneas)

1. Antecedentes del análisis de sentimiento y extracción de tópicos
2. Preprocesamiento y exploración de los datos
3. Modelos de texto
4. Modelos de análisis de sentimiento
5. Resultados del caso de análisis
7. Bibliografía

ANTECEDENTES

APRENDIZAJE SUPERVISADO

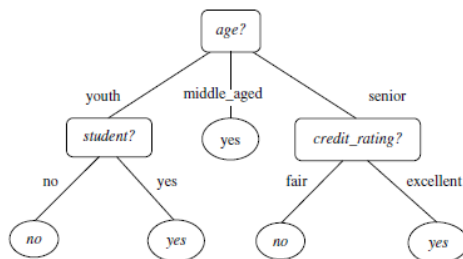
El aprendizaje supervisado proviene de los ejemplos etiquetados en un conjunto de datos de entrenamiento en donde el analista provee al algoritmo un dataset conocido con entradas y salidas deseadas. El caso más simple es un dataset de clases positivas y negativas, después vienen las etiquetas multiclase, por ejemplo, positivo, negativo y neutro; y al final regresiones que vienen con entradas de datos continuos.

Por ejemplo, si queremos aprender la clase C de una familia de coches, nosotros tenemos un conjunto de autos ejemplo entonces encuestamos a un conjunto de personas, estas personas etiquetaran las imágenes, los carros que ellos creen son de la misma familia serán los casos positivos y el resto los negativos. Dadas estas clasificaciones se puede hacer una predicción, dado un ejemplo de carro que no ha sido visto antes por el modelo y basado en lo aprendido podemos identificar la clase a la que pertenece. Para que aprenda el modelo se puede añadir diferentes variables (atributos) para la correcta identificación del modelo.

- **ÁRBOLES DE DECISIÓN**

Un árbol de decisión es una estructura jerárquica que implementa la estrategia de divide y conquista; se trata de un método no paramétrico. En este método la tarea es analizar los datos y clasificarlos, donde el modelo o clasificador es construido para predecir etiquetas de clase (categóricas), tanto como cuan seguro o riesgoso es un caso de fraudes en préstamos bancarios. Estas categorías pueden ser representadas a través de valores discretos donde el orden entre los valores no tiene significado alguno; por ejemplo, los valores 1,2 y 3 podrían representar tratamiento A, tratamiento B y tratamiento C donde no hay un orden implicado entre un grupo de tratamientos. La clasificación de datos es un proceso de dos pasos, consistiendo en un primer paso de “aprendizaje” donde el modelo de clasificación es construido y un segundo paso de “clasificación” donde el modelo es usado para predecir nuevas etiquetas de clase. En el paso de aprendizaje o fase de entrenamiento donde un algoritmo de clasificación construye el clasificador a través de analizar o aprender del conjunto de entrenamiento hecho de un conjunto de tuplas y su asociada etiqueta de clase; una tupla X es representada por un vector de atributos n -dimensional $X=(x_1, x_2, x_3, \dots, x_n)$.

La estructura de un árbol de decisión es similar a un diagrama de flujo en donde cada nodo interno denota una prueba en un atributo, cada rama representa una salida de la prueba y cada nodo hoja mantiene una clase, el nodo más alto es la raíz del árbol. Ejemplo:



El algoritmo comienza con una partición de datos de entrenamiento, una lista de atributos y un método de selección de atributos que representa un procedimiento heurístico para la selección de atributos que mejor discrimine las tuplas dadas acorde a la case, este método puede ser Information Gain o Gini Index; criterios como Gini Index orilla el árbol resultante a ser binario, otros, como Information Gain permiten dos o más ramas crecientes desde un nodo; al aplicar un método de particionamiento este nos dará el mejor atributo para continuar las divisiones. Cuando el árbol de decisión es construido muchas de las ramas reflejarán anomalías en los datos de entrenamiento por ello se debe realizar la poda.

Solicitud de Registro de Tema para Titulación

• CLASIFICADOR DE NAÏVE BAYES

Los clasificadores Bayesianos con clasificadores estadísticos, estos pueden predecir probabilidades, como la probabilidad de que una tupla pertenezca a una clase específica, se asume que el efecto del valor de un atributo en una clase dada es independiente de los valores de otros atributos, este supuesto es llamado independencia condicional de clase.

A continuación, se listan los pasos que hay que realizar para poder utilizar el algoritmo Naive Bayes en problemas de clasificación como el mostrado en el apartado anterior.

- Convertir el conjunto de datos en una tabla de frecuencias.
- Crear una tabla de probabilidad calculando las correspondientes a que ocurran los diversos eventos.
- La ecuación Naive Bayes se usa para calcular la probabilidad posterior de cada clase.
- La clase con la probabilidad posterior más alta es el resultado de la predicción.

Aunque son unos clasificadores bastante buenos, los algoritmos Naive Bayes son conocidos por ser pobres estimadores, la presunción de independencia Naive muy probablemente no reflejará cómo son los datos en el mundo real. Cuando el conjunto de datos de prueba tiene una característica que no ha sido observada en el conjunto de entrenamiento, el modelo le asignará una probabilidad de cero y será inútil realizar predicciones.

• REGRESIÓN POR MÍNIMOS CUADRADOS

La regresión es una herramienta de modelación común para modelar la relación entre algunas variables explicativas y algunas reales o la variable dependiente; el método de los mínimos cuadrados se utiliza para calcular la recta de regresión lineal que minimiza los residuos, esto es, las diferencias entre los valores reales y los estimados por la recta, cuando hay varias variables independientes nos encontramos ante un modelo de regresión lineal múltiple, mientras que cuando hay solo una hablaremos de la regresión lineal simple.

La regresión lineal requiere que la relación entre las variables sea lineal y puede representarse mediante la ecuación de la recta $Y = \beta_0 + \beta_1 X$

son los coeficientes del modelo de regresión. β_0 representa la constante del modelo (también llamada intercepto) y es el punto donde la recta corta el eje de ordenadas (el de las Y, para entendernos bien). Representaría el valor teórico de la variable Y cuando la variable X vale cero.

Por su parte, β_1 representa la pendiente (inclinación) de la recta de regresión. Este coeficiente nos dice el incremento de unidades de la variable Y que se produce por cada incremento de una unidad de la variable X.

El problema es que la distribución de valores no se va a ajustar nunca de manera perfecta a ninguna recta así que, cuando vayamos a calcular un valor de Y determinado (y_i) a partir de un valor de X (x_i) habrá una diferencia entre el valor real de y_i y el que obtengamos con la fórmula de la recta. Ya nos hemos vuelto a encontrar con el azar, nuestro compañero inseparable, así que no tendremos más remedio que incluirlo en la ecuación: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

El componente determinista lo marcan los dos primeros elementos de la ecuación, mientras que el estocástico lo marca el error en la estimación. Los dos componentes se caracterizan por su variable aleatoria, y_i y ε_i , respectivamente, mientras que x_i sería un valor determinado y conocido de la variable X. ε_i representa la diferencia entre el valor real de y_i en nuestra nube de puntos y el que nos proporcionaría la ecuación de la recta (el valor estimado, representado como \hat{y}_i). Podemos representarlo matemáticamente de la siguiente forma y se conoce como residuo: $e_i = y_i - \hat{y}_i$

Solicitud de Registro de Tema para Titulación

• REGRESIÓN LOGÍSTICA

La Regresión Logística Simple, desarrollada por David Cox en 1958, es un método de regresión que permite estimar la probabilidad de una variable cualitativa binaria en función de una variable cuantitativa. Una de las principales aplicaciones de la regresión logística es la de clasificación binaria, en el que las observaciones se clasifican en un grupo u otro dependiendo del valor que tome la variable empleada como predictor.

Si una variable cualitativa con dos niveles se codifica como 1 y 0, matemáticamente es posible ajustar un modelo de regresión lineal por mínimos cuadrados $\beta_0 + \beta_1 x$. El problema de esta aproximación es que, al tratarse de una recta, para valores extremos del predictor, se obtienen valores de Y menores que 0 o mayores que 1, lo que entra en contradicción con el hecho de que las probabilidades siempre están dentro del rango [0,1].

En mi experiencia este tipo de modelo se utiliza mucho para fraudes, o impagos en el sector bancario.

APRENDIZAJE NO SUPERVISADO

Aprendizaje no supervisado es sinónimo de clúster, el proceso de aprendizaje es no supervisado desde que los ejemplos de entrada no están etiquetados, típicamente, nosotros podríamos usar clustering para descubrir clases o grupos en los datos, patrones, por ejemplo, un modelo podría tomar como entrada un conjunto de imágenes de dígitos escritos, suponemos que encuentra 10 clúster tendríamos que describir cada grupo.

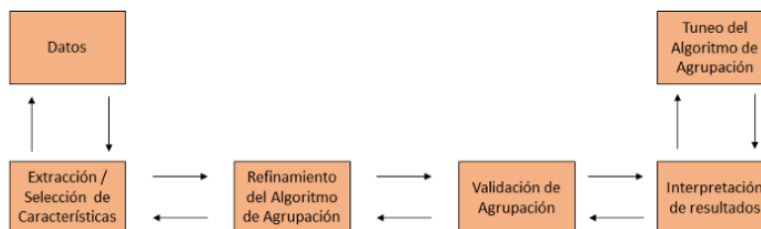
Las principales aplicaciones de aprendizaje no supervisado son:

- Segmentación de conjuntos de datos por atributos compartidos.
- Detección de anomalías que no encajan en ningún grupo.
- Simplificación de datasets agregando variables con atributos similares.

Estas técnicas se pueden condensar en dos tipos principales de problemas que el aprendizaje no supervisado trata de resolver. Estos son los problemas:

- Agrupación
- Reducción de la dimensionalidad

El proceso general que seguiremos al desarrollar un modelo de aprendizaje no supervisado se puede resumir en el siguiente cuadro:



• ALGORITMOS DE CLUSTERING

En términos básicos, el objetivo del clustering es encontrar diferentes grupos dentro de los elementos de los datos. Para ello, los algoritmos de agrupamiento encuentran la estructura en los datos de manera que los elementos del mismo clúster (o grupo) sean más similares entre sí que con los de clústeres diferentes.

Solicitud de Registro de Tema para Titulación

Es el proceso de particionar el conjunto de datos en subconjuntos, cada subconjunto es un cluster, tal que los objetos dentro de un cluster son similares, comparten características y son disimilares entre clusters. Hay varios algoritmos de clustering, por ejemplo:

Métodos de particionamiento: Dado un conjunto de n objetos, un método de particionamiento construye K particiones de los datos, donde cada partición representa un cluster, los métodos de particionamiento adoptan la separación exclusiva, es decir, cada objeto solo puede pertenecer a un cluster, muchos de estos métodos están basados en la distancia.

Métodos jerárquicos: Estos crean una descomposición jerárquica de los datos y puede ser clasificado como aglomerativo o divisivo. El acercamiento aglomerativo comienza con cada objeto formando un grupo separado, este sucesivamente une los objetos o grupos cercanos el uno del otro hasta que todos los grupos son unidos en uno solo o hasta que se cumplan una determinada condición. El acercamiento divisivo comienza con todos los objetos en el mismo cluster y en cada iteración sucesiva un cluster es dividido en clusters más pequeños hasta que eventualmente cada objeto este en un cluster o que las condiciones de terminación se cumplan.

Un algoritmo muy conocido y común es K means.

- **COMPONENTES PRINCIPALES**

Análisis de componentes principales o PCA por sus siglas en inglés es un método estadístico que permite simplificar la complejidad de espacios muestrales con muchas dimensiones a la vez que conserva su información. Supóngase que existe una muestra con n individuos cada uno con p variables (X_1, X_2, \dots, X_p), es decir, el espacio muestral tiene p dimensiones. PCA permite encontrar un número de factores subyacentes ($z < p$) que explican aproximadamente lo mismo que las p variables originales. Donde antes se necesitaban p valores para caracterizar a cada individuo, ahora bastan z valores. Cada una de estas z nuevas variables recibe el nombre de componente principal.

Cada componente principal se obtiene de una combinación lineal de las variables reales, el primer componente tendrá la mayor varianza, el proceso es, primero se resta a cada valor la media de la variable a la que pertenece para conseguir que tengan una media de cero después se resuelve el problema de optimización con el que se maximiza la varianza, una forma de hacerlo es con el cálculo de los eigenvalores y eigenvectores de la matriz de covarianzas.

Tanto la proporción de varianza explicada como la proporción de varianza explicada acumulada son dos valores de gran utilidad a la hora de decidir el número de componentes principales a utilizar en los análisis posteriores, la suma de la proporción de varianza explicada por los componentes es 1.

Al término del cálculo de los componentes se preguntará cuántos son los ideales y que me reducen la dimensionalidad de los datos, una forma consiste en evaluar la proporción de varianza explicada acumulada y seleccionar el número de componentes mínimo a partir del cual el incremento deja de ser significativo, en general tendremos $n-1$ o p componentes dada una matriz de dimensiones $n \times p$.

- **LDA**

El modelado o extracción de tópicos es método de aprendizaje no supervisado y apoya a la clasificación de documentos capaz de encontrar agrupaciones de palabras, en este modelo no sabemos que vamos a encontrar, aquí el objetivo es descubrir tópicos principales y relevantes de lo que hablan en los comentarios o verbalizaciones.

Solicitud de Registro de Tema para Titulación

El objetivo final de LDA es encontrar la representación óptima de la matriz Documento-Tema y la matriz Tema-Palabra para encontrar la distribución Documento-Tema y Tema-Palabra más optimizada. Como LDA asume que los documentos son una mezcla de temas y los temas son una mezcla de palabras, LDA retrocede desde el nivel del documento para identificar qué temas habrían generado estos documentos y qué palabras habrían generado esos temas.

- **SVD**

Se trata de una factorización de esa matriz (original) en tres matrices $A=UWV^t$. Primero se calculan los valores singulares, haciendo AA^t , después se encuentran los vectores singulares. SVD se puede considerar como un método de proyección en el que los datos con m columnas (características) se proyectan en un subespacio con m o menos columnas, conservando la esencia de los datos originales. El SVD se usa ampliamente tanto en el cálculo de otras operaciones matriciales, como la matriz inversa, como también como un método de reducción de datos en el aprendizaje automático. En relación con el modelado de texto apoya con los problemas de alta dimensionalidad y dispersión; se define A como la matriz de término documento con m documentos y n términos, típicamente habrá más términos que documentos en el corpus, los valores singulares nos dan una medida de la importancia usada para decidir cuantas dimensiones mantener en la matriz.

METODOLOGÍA FUNDAMENTAL PARA LA CIENCIA DE DATOS EN EL PROYECTO

ETAPA 1: COMPRENSIÓN DEL NEGOCIO

Durante esta primera etapa se requiere definir el problema, los objetivos y los requisitos de la solución.

Objetivo general

Implementar un análisis de sentimiento y modelos de texto de las reseñas en Google Play y Apple Store orientado a la aplicación móvil de Santander para detectar las debilidades y fortalezas de las funcionalidades de dicha aplicación.

Objetivos específicos

- 1) Realización de web scrapping hacía las tiendas de aplicaciones con Python y posteriormente el procesamiento para obtener un conjunto de documentos listos para poder analizarlos de manera correcta.
- 2) Aplicación del algoritmo de la distancia de Levenshtein para corregir las palabras mal escritas por el usuario.
- 3) Creación de modelos de texto para detectar los puntos de dolor en las funcionalidades de las aplicaciones móviles bancarias.
- 4) Generación de un modelo de análisis de sentimiento para detectar la polaridad de los comentarios de las tiendas de aplicaciones dado que la calificación por estrellas no aporta mucho sentido ni significado.
- 5) Categorización de los tópicos generados por funcionalidad.

ETAPA 2: ENFOQUE ANALÍTICO

Esta etapa implica expresar el problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático, a continuación, se expresa el problema con estas características.

Analizar las reseñas de aplicaciones móviles bancarias (Santander) que son escritas por los usuarios en las tiendas de aplicaciones (Google y Apple store) realizando la extracción de tópicos mediante un análisis de texto probando tres tipos de modelado, LDA, NMF y el SVD para mostrar las debilidades y fortalezas de la aplicación móvil, además, realizar un análisis de sentimiento a través de un modelo de clasificación para determinar la polaridad de los comentarios presentados por los usuarios.

ETAPA 3: REQUISITOS DE DATOS

Debido a que los modelos a implementar requieren de datos no estructurados, los datos contemplados son las reviews (comentarios) de la aplicación Móvil de Santander México “SuperMóvil”.

ETAPA 4: RECOPIACIÓN DE DATOS

Mediante la técnica de web scraping en Python se apuntará a la tienda de Google y de Apple Store a los ids mx.bancosantander.supermovil y id498944221 respectivamente que pertenecen a la aplicación SuperMóvil de Santander México.

ETAPA 5: COMPRENSIÓN DE LOS DATOS

Después de la recopilación de datos inicial, los científicos de datos suelen utilizar estadísticas descriptivas y técnicas de visualización para comprender el contenido de los datos, evaluar su calidad y descubrir insights iniciales sobre ellos.

ETAPA 6: PREPARACIÓN DE DATOS

Esta etapa abarca todas las actividades para construir el conjunto de datos que se utilizará en la siguiente etapa de modelado. Entre las actividades de preparación de datos están la limpieza de datos (tratar con valores no válidos o que faltan, eliminar duplicados y dar un formato adecuado), combinar datos de múltiples fuentes (archivos, tablas y plataformas) y transformar los datos en variables más útiles.

Para esta parte se realizará el preprocesamiento y limpieza, como los datos son no estructurados eliminaré acentos, caracteres especiales, números, convertiré a minúsculas, aplicaré stemming, quitaré stopwords, quitaré las palabras con longitud menor a 3 para eliminar dispersión, aplicaré algún algoritmo con Levenshtein para corregir las palabras mal escritas.

ETAPA 7: MODELADO

La etapa de modelado utiliza la primera versión del conjunto de datos preparado y se enfoca en desarrollar modelos predictivos o descriptivos según el enfoque analítico previamente definido.

Como en esta parte ya hemos comprendido y explorado los datos, se modelarán los datos utilizando los métodos propuestos: LDA, NMF y SVD, ajustarlos y ver cómo se comportan y si los tópicos generados son interpretables o entendibles.

Solicitud de Registro de Tema para Titulación

En la parte del análisis de sentimiento se evaluará un modelo de clasificación, naïve y árboles de decisión, quizá probar otro tipo de modelos y revisar el desempeño. Antes de este paso habrá que calificar manualmente un set de comentarios con el sentimiento positivo y negativo, para que bajo esas etiquetas entrenemos los modelos.

ETAPA 8: EVALUACIÓN

Durante el desarrollo del modelo y antes de su implementación, el científico de datos evalúa el modelo para comprender su calidad y garantizar que aborda el problema empresarial de manera adecuada y completa. La evaluación del modelo implica el cálculo de varias medidas de diagnóstico y de otros resultados, como tablas y gráficos, lo que permite al científico de datos interpretar la calidad y la eficacia del modelo en la resolución del problema.

La evaluación de los tópicos será con respecto a la interpretación de los tópicos generados, es decir, que los términos que compongan a los tópicos sean claros y que evidencien las funcionalidades de la aplicación.

Con respecto a la evaluación del modelo de análisis de sentimiento se evaluarán con una matriz de confusión, claro que un modelo así debe ser entrenado y ajustado frecuentemente.

ETAPA 9: IMPLEMENTACIÓN

La implementación de los modelos generados será a través de un dashboard en Power BI, un informe que contemple los tópicos y el sentimiento de los comentarios clasificados por funcionalidad en la aplicación.

ETAPA 10: RETROALIMENTACIÓN

Al recopilar los resultados del modelo implementado, la organización obtiene retroalimentación sobre el rendimiento del modelo y su impacto en el entorno en el que se implementó.

MATERIALES

- Russ Albright. Taming Text with the SVD.
- Tae-Yeun Kim. Modeling Topic Extraction-based Sentiment Analysis Based on User Reviews
- Miyoung Chong. Sentiment Analysis and Topic Extraction of the Twitter Network of #Prayfor-paris
- Gabriele Pergola, Lin Gui, Yulan He. TDAM: a Topic-Dependent Attention Model for Sentiment Analysis
- Clifford Lewis, Michael Mehmet. Does the NPS® reflect consumer sentiment? A qualitative examination of the NPS using a sentiment analysis approach

HERRAMIENTAS IDENTIFICADAS PARA EL PROYECTO

En este proyecto usaré PowerBi para mostrar los resultados finales, Python para todo el procesamiento y modelos, y R para el discover de los textos, nubes de palabras, etc. Además, las tiendas de aplicaciones de Google Play y Apple Store.

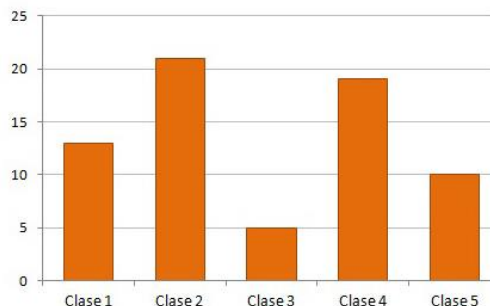
GRÁFICOS PROPUESTOS PARA LA PRESENTACIÓN DE RESULTADOS

- **NUBE DE PALABRAS**



En el proyecto utilizaré el gráfico de nube de palabras, dado que voy a hacer un análisis de datos no estructurados, en específico para la generación de tópicos de los comentarios de las tiendas de aplicaciones de la app “SuperMóvil” de Santander México, para los tópicos descubiertos a través de los métodos no supervisados propuestos se pretende generar un dashboard interactivo en Power Bi, por lo que al darle clic a un tópico se verá la nube de palabras de los comentarios pertenecientes a ese tópico. Por la versatilidad de la herramienta y del tipo de gráfico será muy útil para identificar las palabras más importantes del tópico en un tamaño grande y de ahí en tamaños inferiores de acuerdo con la frecuencia de las palabras.

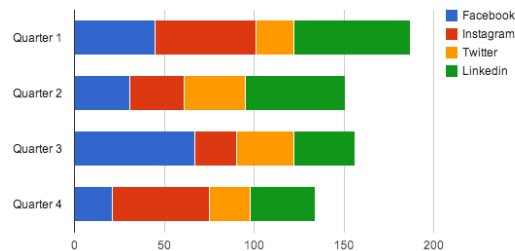
- **GRÁFICO DE BARRAS**



Se hará uso del gráfico de barras simples horizontales para mostrar los totales por tópico, en el eje X se presentará el volumen del tópico generado y en el eje Y se colocarán los diferentes tópicos generados, para que a una vista se pueda observar y de inmediato identificar el tópico con más comentarios clasificados en él, y de ahí en orden descendente de importancia o volumen del tópico.

Solicitud de Registro de Tema para Titulación

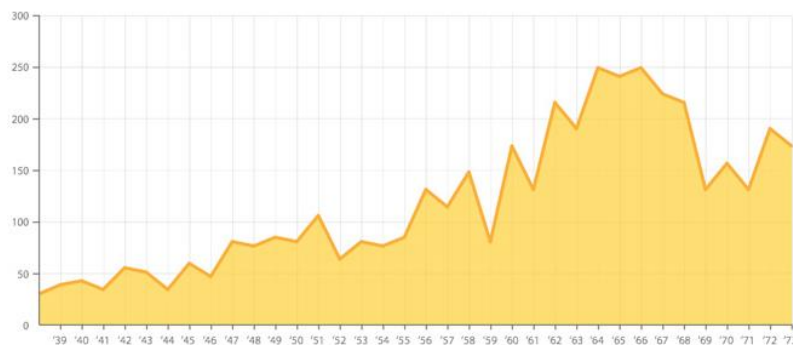
- **GRÁFICO DE BARRAS APILADAS**



También se hará uso del gráfico de barras apiladas para mostrar los resultados del análisis de sentimiento, las categorías serán Positivo, Negativo y Neutro. En este caso primero se entrenará el modelo de clasificación con comentarios etiquetados manualmente, después se harán iteraciones hasta obtener un buen desempeño. En el eje X se pretende colocar el mes de análisis de comentarios y en el eje Y el split por categoría de sentimiento.

Como el dashboard será interactivo al filtrar un tópico podremos ver el detalle en el sentimiento. Adicional, también se probará el gráfico de barras simples para ver el detalle del sentimiento y dependiendo la claridad de los gráficos se decidirá por uno u otro.

- **GRÁFICO DE ÁREA**



Este gráfico se utilizará para mostrar el volumen de comentarios que hay en la aplicación a nivel mensual y diario, para detectar días en los que posiblemente pueda haber alguna falla en la aplicación, o características estacionales.

Solicitud de Registro de Tema para Titulación

AVANCES Y TABLEROS DE CONTROL

ETAPA 3 Y 4: REQUISITOS DE DATOS Y RECOPIACIÓN DE LOS DATOS

Debido a que los modelos a implementar requieren de datos no estructurados, los datos contemplados son las reviews (comentarios) de la aplicación Móvil de Santander México “SuperMóvil”.

WEB SCRAPPING : GOOGLE PLAY STORE

```
#import play_scraper
import pandas as pd

import json

from tqdm import tqdm

import seaborn as sns
import matplotlib.pyplot as plt

from pygments import highlight
from pygments.lexers import JsonLexer
from pygments.formatters import TerminalFormatter

from google_play_scraper import Sort, reviews, app, reviews_all

import datetime
```

```
app_santander="mx.bancosantander.supermovil"
```

```
result, continuation_token= reviews(
    app_santander,
    lang='es',
    country='mx',
    sort=Sort.NEWEST,
    count=30000,
    filter_score_with=None
)
```

```
len(result)
data=pd.DataFrame(result)
data=data[data['content'].notna()]
```

```
data.shape
```

```
(30000, 10)
```

```
data["fecha"]=data["at"].dt.strftime("%Y-%m-%d")
data["hora"]=data["at"].dt.strftime("%H:%M:%S")
data["aniomes"]=data["at"].dt.strftime("%Y-%m")
```

```
data.aniomes.value_counts()

2021-12    8050
2022-01    7251
2022-02    5912
2022-03    5346
2021-11    3441
Name: aniomes, dtype: int64
```

ETAPA 3 Y 4: REQUISITOS DE DATOS Y RECOPIACIÓN DE LOS DATOS

WEB SCRAPPING : GOOGLE PLAY STORE -- RESULTADOS

reviewid	userName	userImage	content	mbstUpCo	viewCreatedVersi	at	plyConte	repliedAt	fecha	hora	aniomes
2	gp:AOqpT Antonio Martinez	https://play	Muy buena app	0	5.62.3	2022-03-27 14:23:23			2022-03-27 14:23:23	2022-03	
3	gp:AOqpT David Orozco	https://play	Lenta y con muchos errores	0	5.62.3	2022-03-27 14:14:56			2022-03-27 14:14:56	2022-03	
4	gp:AOqpT Diana Izbeth Mejia Hernández	https://play	Muy buena	0	5.62.3	2022-03-27 14:11:42			2022-03-27 14:11:42	2022-03	
5	gp:AOqpT Hector Hernandez Cisneros	https://play	Excelente	0	5.62.3	2022-03-27 13:48:36			2022-03-27 13:48:36	2022-03	
6	gp:AOqpT Ana Mora	https://play	Pésima y no me gusta que la app me rastree	0	5.62.3	2022-03-27 13:47:11			2022-03-27 13:47:11	2022-03	
7	gp:AOqpT Jose María Ponce Becerril	https://play	Buena	0	5.62.3	2022-03-27 13:44:32			2022-03-27 13:44:32	2022-03	
8	gp:AOqpT ISRAEL ZAMORANO	https://play	Pesi servicio en sucursal la app esta peor todo los d	0	5.62.3	2022-03-27 13:41:15			2022-03-27 13:41:15	2022-03	
9	gp:AOqpT Cesar Octavio Loza Saucedo	https://play	muy mala, pesima	0	5.62.3	2022-03-27 13:38:55			2022-03-27 13:38:55	2022-03	
10	gp:AOqpT Daniel Ramirez Gama	https://play	Excelente App	0	5.62.3	2022-03-27 13:35:01			2022-03-27 13:35:01	2022-03	
11	gp:AOqpT Brandon Cruz	https://play	Muy buena y rapida	0	5.62.3	2022-03-27 13:31:57			2022-03-27 13:31:57	2022-03	
12	gp:AOqpT Hec Urizar	https://play	Todo bien	0	5.62.3	2022-03-27 13:30:44			2022-03-27 13:30:44	2022-03	
13	gp:AOqpT Naytal Torreblanca Hernández	https://play	Es una app muy buena y muy confiable	0	5.62.3	2022-03-27 13:29:11			2022-03-27 13:29:11	2022-03	
14	gp:AOqpT Eduardo Lira	https://play	Una magnifica herramienta	0	5.62.3	2022-03-27 13:22:39			2022-03-27 13:22:39	2022-03	
15	gp:AOqpT Iveth García	https://play	Buena	0	5.62.3	2022-03-27 13:22:35			2022-03-27 13:22:35	2022-03	
16	gp:AOqpT Nancy Martinez melquiades	https://play	Excelenteapp	0	5.62.3	2022-03-27 13:15:52			2022-03-27 13:15:52	2022-03	
17	gp:AOqpT Un usuario de Google	https://play	Muy buena aplicación y de fácil uso	0	5.62.3	2022-03-27 13:14:11			2022-03-27 13:14:11	2022-03	
18	gp:AOqpT ESTO PASA EN MÉXICO	https://play	Me parece mal que si quiero hacer una transferenci	0	5.62.3	2022-03-27 13:08:26			2022-03-27 13:08:26	2022-03	
19	gp:AOqpT Rubén dario Bojorquez patron	https://play	Feliz con mi app Santander, una de las mejores	0	5.62.3	2022-03-27 12:52:32			2022-03-27 12:52:32	2022-03	
20	gp:AOqpT AA 2	https://play	Iban tan bien, por qué le tienen que andar moviend	0	5.62.3	2022-03-27 12:50:49			2022-03-27 12:50:49	2022-03	
21	gp:AOqpT Takero bailongo Chávez	https://play	Buena a secas , tarde en arreglar mi problema , un p	0	5.62.3	2022-03-27 12:50:32			2022-03-27 12:50:32	2022-03	

Solicitud de Registro de Tema para Titulación

ETAPA 3 Y 4: REQUISITOS DE DATOS Y RECOPIACIÓN DE LOS DATOS

WEB SCRAPPING : APPLE STORE

```
import json
from app_store_scraper import AppStore
import numpy as np
```

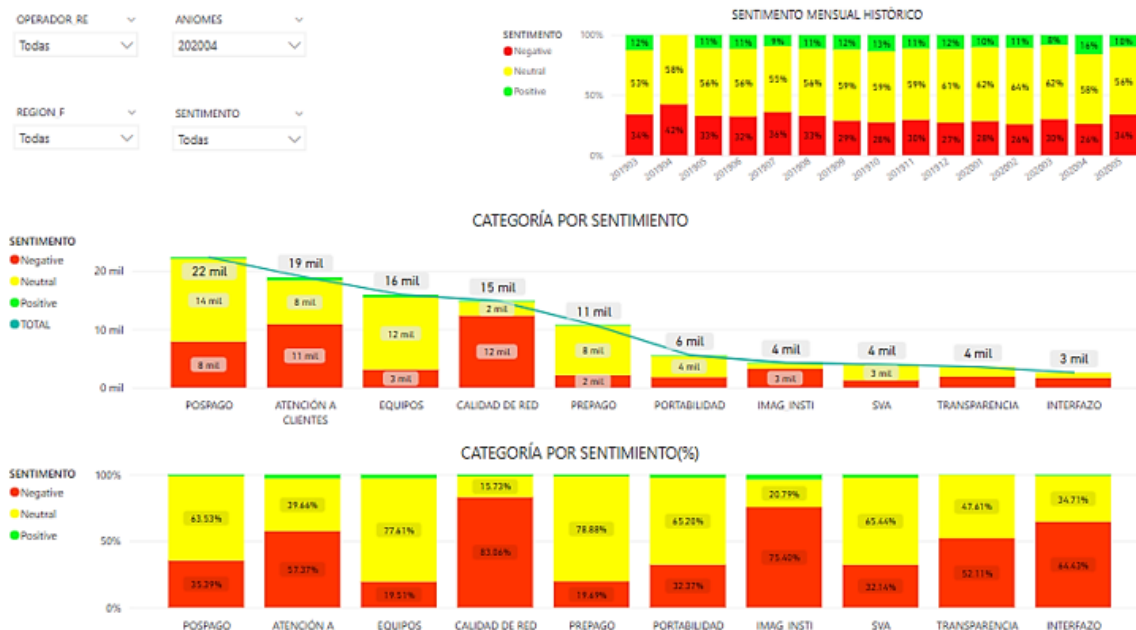
```
sant=AppStore(country='mx', app_name='supermovil-santander-mexico',app_id='498944221')
sant.review(how_many=30000)
sant.reviews
```

```
df=pd.DataFrame(np.array(sant.reviews),columns=['review'])
df2=df.join(pd.DataFrame(df.pop('review').tolist()))
```

```
df2.head(3)
```

	title	date	rating	isEdited	review	userName	developerResponse
0	Notificaciones	2019-01-26 15:12:14	4	False	Porfavor agregen la opción de recibir notifica...	hdhsyehheje	NaN
1	Buena	2017-10-06 15:52:11	4	False	Es buena la aplicación pero falta información ...	Fabby Montes	NaN
2	Solución iPhone 11	2020-01-01 19:59:23	1	False	Para los que tienen el caso de que la App se c...	Arturo26390	NaN

TABLERO PROPUESTO PARA EL RESUMEN DEL ANÁLISIS DE SENTIMIENTO POR FUNCIONALIDAD HERRAMIENTA: POWER BI



Solicitud de Registro de Tema para Titulación

TABLERO PROPUESTO PARA LA PRESENTACIÓN DE TÓPICOS RESULTANTES DE LA APLICACIÓN
HERRAMIENTA: POWER BI

OPERADOR	CATEGORIA	TEMA	SENTIMIENTO	REGION F	SUBTEMA	SENTIMENTOM
Todas	Selección múltiple	Todas	Todas	Todas	Todas	Negative

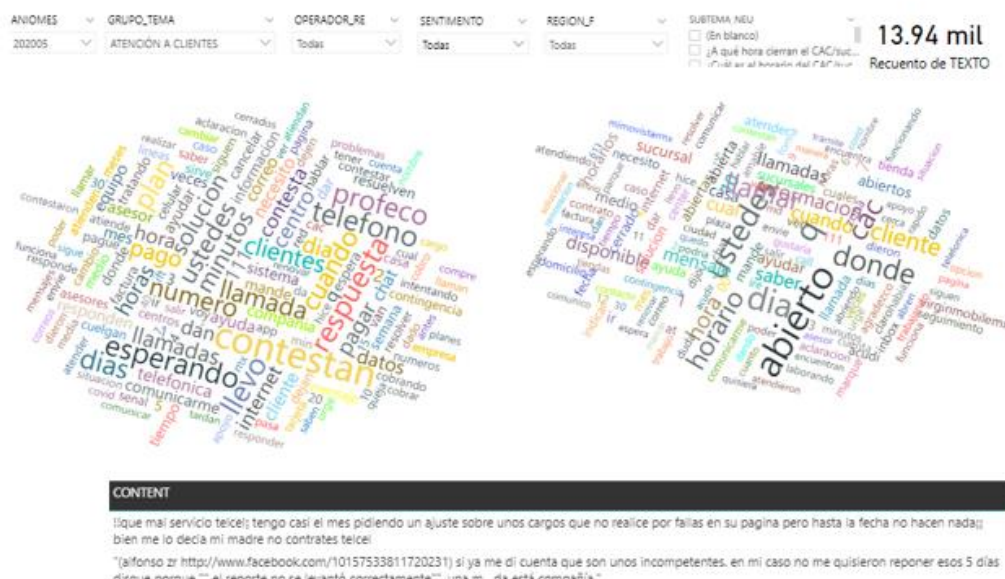
TEMA	201903	201904	201905	201906	201907	201908	201909
Intención de cambio de compañía.	3331	3636	3876	3624	4429	4265	362
No atienden/no contestan.	2572	3247	3640	3334	3173	2768	296
Llamadas no solicitadas.	2828	2792	2662	2821	2650	2347	225
No resuelven/dan solución.	881	879	864	978	615	587	138
Cambio de compañía sin autorización.	1376	1542	1509	1598	711	703	130
Atención sobre la línea	720	1087	935	655	951	810	81
Proceso de portabilidad(compra de sim)	410	502	479	472	497	448	43
Atención CAC/horarios.	253	278	299	269	277	211	22
Sin sistema.	204	330	234	244	246	205	18
Total	13354	15797	15926	15725	15942	14736	1337

TEMA	201903	201904	201905	201906	201907
Intención de cambio de compañía.	25%	23%	24%	24%	28%
No atienden/no contestan.	19%	21%	23%	21%	20%
Llamadas no solicitadas.	17%	18%	17%	18%	17%
No resuelven/dan solución.	4%	12%	12%	13%	11%
Cambio de compañía sin autorización.	10%	10%	9%	10%	10%
Atención sobre la línea	5%	7%	6%	5%	6%
Proceso de portabilidad(compra de sim)	3%	3%	3%	3%	3%
Atención CAC/horarios.	2%	2%	2%	2%	2%
Sin sistema.	2%	2%	1%	2%	2%
Problemas promoción ofrecida de portabilidad	1%	2%	2%	1%	2%
Total	100%	100%	100%	100%	100%

CONTENT


@telcel me llamó pa preguntarme si me quería cambiar de compañía..... les dije q sí..... me preguntaron en q compañía me encuentro..... les dije q en compañía de 30 millones de pesos los q eligieron al peor presidente de la historia y colgaron..... sino me van a ayudar no anden d ofrecidos
 @bienvenidos al infierno!... señor, bienvenido a telcel..... alejate de mí, demonio eterno!... ¿quiere renovar su plan?... chí t
 - buenas tardes, hablamos de telcel (le gustaría cambiar de compañía?... sí, pero por supuesto que sí, ¿me urge?... ¿con quién se encuentra actualmente?... con mi esposa y mi sue
 - buenas tardes, hablamos de telcel (le gustaría cambiar de compañía?... sí, pero por supuesto que sí, ¿me urge?... ¿con quién se encuentra actualmente?... con mi esposa y mi suegra.....ex gobierno federal.
 - buenos días le hablo da telcel para ofrecerle un cambio de compañía... con quién tango el gusto?... de verdad @telcel ?

TABLERO PROPUESTO PARA DETECCIÓN DE PROBLEMAS EFICAZ POR MEDIO DE NUBE DE PALABRAS CON SPLIT POR SENTIMIENTO
HERRAMIENTA: POWER BI



REFERENCIAS BIBLIOGRÁFICAS

- I.V. (2019, 6 noviembre). Tipos de gráficos y diagramas para la visualización de datos. ingenio-virtual.com. Recuperado 18 de marzo de 2022, de <https://www.ingeniovirtual.com/tipos-de-graficos-y-diagramas-para-la-visualizacion-de-datos/>
- Tableau Software. Maila Hardin, Daniel Hom, Ross Perez y Lori Williams. ¿Qué tabla o gráfico es el adecuado para usted? Recuperado 18 de marzo de 2022.
- IBM Analytics. Metodología Fundamental para la Ciencia de Datos. <https://www.ibm.com/downloads/cas/6RZMKDN8>
- IBM Analytics. Metodología Fundamental para la Ciencia de Datos. <https://www.ibm.com/downloads/cas/6RZMKDN8>
- Kulshrestha Ria. Una guía para principiantes sobre la asignación de Dirichlet latente (LDA). <https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2>
- Aashish Nair. Topic Modeling with Latent Semantic Analysis. <https://towardsdatascience.com/topic-modeling-with-latent-semantic-analysis-58aeab6ab2f2>
- 6 Topic modeling. <https://www.tidytextmining.com/topicmodeling.html>
- Part 2: Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim and Sklearn. <https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/>
- Amat Rodrigo. Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. https://www.cienciadedatos.net/documentos/35_principal_component_analysis
- Han, Camber, Pei, Data Mining, concepts and techniques. Third Edition.
- Ethem Alpaydin. Introduction to Machine Learning. Second Edition
- Roman V. Algoritmos Naive Bayes: Fundamentos e Implementación <https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f>
- Molina Manuek. La distancia más corta. El método de los mínimos cuadrados. <https://anestesiari.org/2020/la-distancia-mas-corta-el-metodo-de-los-minimos-cuadrados/#:~:text=El%20m%C3%A9todo%20de%20los%20m%C3%ADnimos%20cuadrados%20se%20utiliza%20para%20calcular,de%20regresi%C3%B3n%20con%20este%20m%C3%A9todo.>
- Amat Rodrigo. Regresión logística simple y múltiple. https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple
- Roman V. Aprendizaje No Supervisado en Machine Learning: Agrupación. <https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc>
- Roman V. Aprendizaje No Supervisado en Machine Learning: Agrupación. <https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc>
- Roman V. Aprendizaje No Supervisado en Machine Learning: Agrupación <https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupaci%C3%B3n-bb8f25813edc>

	Fecha: 30/03/2022
Solicitud de Registro de Tema para Titulación	

- Wakefield K. A guide to the types of machine learning algorithms and their applications
https://www.sas.com/en_ie/insights/articles/analytics/machine-learning-algorithms.html