EXTRACCIÓN DE TÓPICOS EN COMENTARIOS DE APLICACIONES MÓVILES DEL SECTOR BANCARIO

SEMINARIO DE INVESTIGACIÓN II

DULCE MARIA REYES LUCAS

AGENDA

- Descripción del problema
- Objetivo
- Proyectos relacionados
- Metodología
- Cronograma de actividades
- Recursos (fuente de datos)
- Resultados
- Conclusión
- Referencias

DESCRIPCIÓN DEL PROBLEMA

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.

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 estás necesidades y que deje satisfecho al usuario.

Por ello, iimplementaré un análisis de sentimiento y modelado 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

1) Realización de web scrapping a 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 Levenstein 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) Categorización de los tópicos generados por funcionalidad.

5) Visualización de los tópicos generados a partir de la calificación otorgada por el usuario (5 estrellas) mapeada como puntuación NPS (Detractor 1-3, Neutro 4, Promotor 5)

PROYECTOS RELACIONADOS

Con respecto al análisis y modelado de tópicos encontré algunos acercamientos y publicaciones, entre ellos los siguientes:

- Does the NPS® reflect consumer sentiment? A qualitative examination of the NPS using a sentiment analysis approach
- Sentiment analysis and topic extraction of the twitter network of #prayforparis
- Modeling topic extraction-based sentiment analysis based on user reviews
- How can i improve my app? Classifying user reviews for software maintenance and evolution
- Covid-19 vaccine infodemic: sentiment analysis of the twitter content

METODOLOGÍA

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

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 Levenstein para la corrección de las malas escrituras, dejar palabras con 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.

CRONOGRAMA DE ACTIVIDADES

	Actividad	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6
	Obtención fuente de datos						
Exploración:	Comprensión de los datos						
	Preparación de los datos						
*	Limpieza y procesamiento de los datos						
Modelación:	Topic Modeling (LDA)						
(reiterativo)	Topic Modeling (NMF)						
	Topic Modeling (SVD)						
	Interpretación y descripción tópicos finales						
	Clasificación tópicos por funcionalidades						
	Split tópicos por clasificación de puntuación (5 estrellas->NPS)						
	Visualización resultados en Power BI						

RECURSOS

- **Fuente de datos:**

  Los datos que usaré en este proyecto se encuentran en las tiendas de aplicaciones de Google y Apple Store, estos son las reviews (comentarios) de la aplicación Móvil de Santander México “SuperMóvil”, la forma de extracción será mediante la técnica de web scraping, en Python se apunta a la tienda de Google y de Apple Store a los ids mx.bancosantander.supermovil y id498944221 respectivamente que pertenecen a la aplicación SuperMóvil.

- **Software:**
  - Python
  - R
  - PowerBi
  - Datos tiendas de aplicaciones

AVANCES

FUENTE DE DATOS: RECOPILACIÓN 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, continuacion_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
```

FUENTE DE DATOS: RECOPILACIÓN DE LOS DATOS

WEB SCRAPPING : GOOGLE PLAY STORE -- RESULTADOS

reviewId	userName	userImage	content	mbsUpCo	riewCreatedVersi	at	eplyConte	repliedAt	fecha	hora	aniomes
gp:AOqpT	Antonio Martinez	https://play	Muy buena app	0	5.62.3	2022-03-27 14:23:23			2022-03-2	14:23:23	2022-03
gp:AOqpT	David Orozco	https://play	Lenta y con muchos errores	0	5.62.3	2022-03-27 14:14:56			2022-03-2	14:14:56	2022-03
gp:AOqpT	Diana lizbeth Mejia Hernández	https://play	Muy buena	0	5.62.3	2022-03-27 14:11:42			2022-03-2	14:11:42	2022-03
gp:AOqpT	Hector Hernandez Cisneros	https://play	Excelente	0	5.62.3	2022-03-27 13:48:36			2022-03-2	13:48:36	2022-03
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			2022-03-2	13:47:11	2022-03
gp:AOqpT	Jose María Ponce Becerril	https://play	Buena	0	5.62.3	2022-03-27 13:44:32			2022-03-2	13:44:32	2022-03
gp:AOqpT	ISRAEL ZAMORANO	https://play	Pesi servicio en sucursal la app esta peor todo los di	0	5.62.3	2022-03-27 13:41:15			2022-03-2	13:41:15	2022-03
gp:AOqpT	Cesar Octavio Loza Saucedo	https://play	muy mala, pesima	0	5.62.3	2022-03-27 13:38:55			2022-03-2	13:38:55	2022-03
gp:AOqpT	Daniel Ramirez Gama	https://play	Exelente App	0		2022-03-27 13:35:01			2022-03-2	13:35:01	2022-03
gp:AOqpT	Brandon Cruz	https://play	Muy buena y rapida	0	5.62.3	2022-03-27 13:31:57			2022-03-2	13:31:57	2022-03
gp:AOqpT	Hec Urizar	https://play	Todo bien	0		2022-03-27 13:30:44			2022-03-2	13:30:44	2022-03
gp:AOqpT	Naytai Torreblanca Hernández	https://play	Es una app muy buena y muy confiable	0	5.62.3	2022-03-27 13:29:11			2022-03-2	13:29:11	2022-03
gp:AOqpT	Eduardo Lira	https://play	Una magnífica herramienta	0	5.62.3	2022-03-27 13:22:39			2022-03-2	13:22:39	2022-03
gp:AOqpT	Iveth García	https://play	Buena	0		2022-03-27 13:22:35			2022-03-2	13:22:35	2022-03
gp:AOqpT	Nancy Martinez melquiades	https://play	Excelenteapp	0	5.62.3	2022-03-27 13:15:52			2022-03-2	13:15:52	2022-03
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			2022-03-2	13:14:11	2022-03
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			2022-03-2	13:08:26	2022-03
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			2022-03-2	12:52:32	2022-03
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			2022-03-2	12:50:49	2022-03
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			2022-03-2	12:50:32	2022-03

FUENTE DE DATOS: RECOPILACIÓN DE LOS DATOS

WEB SCRAPPING : APPLE STORE -- RESULTADOS

```
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
```

...

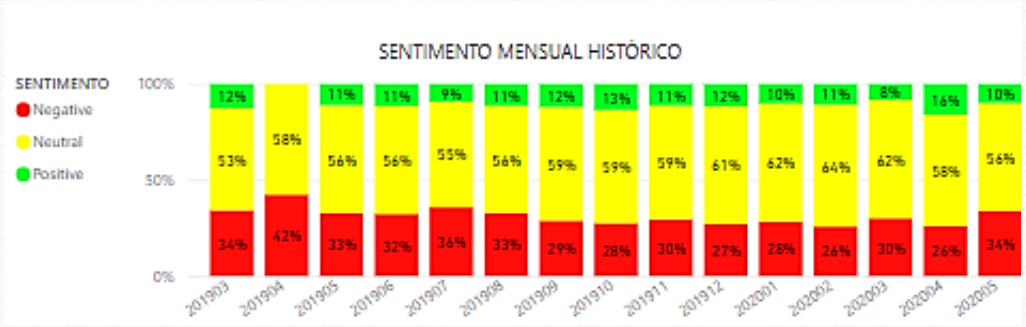
```
df=pd.DataFrame(np.array(sant.reviews),columns=['review'])
df2=df.join(pd.DataFrame(df.pop('review').tolist()))
```

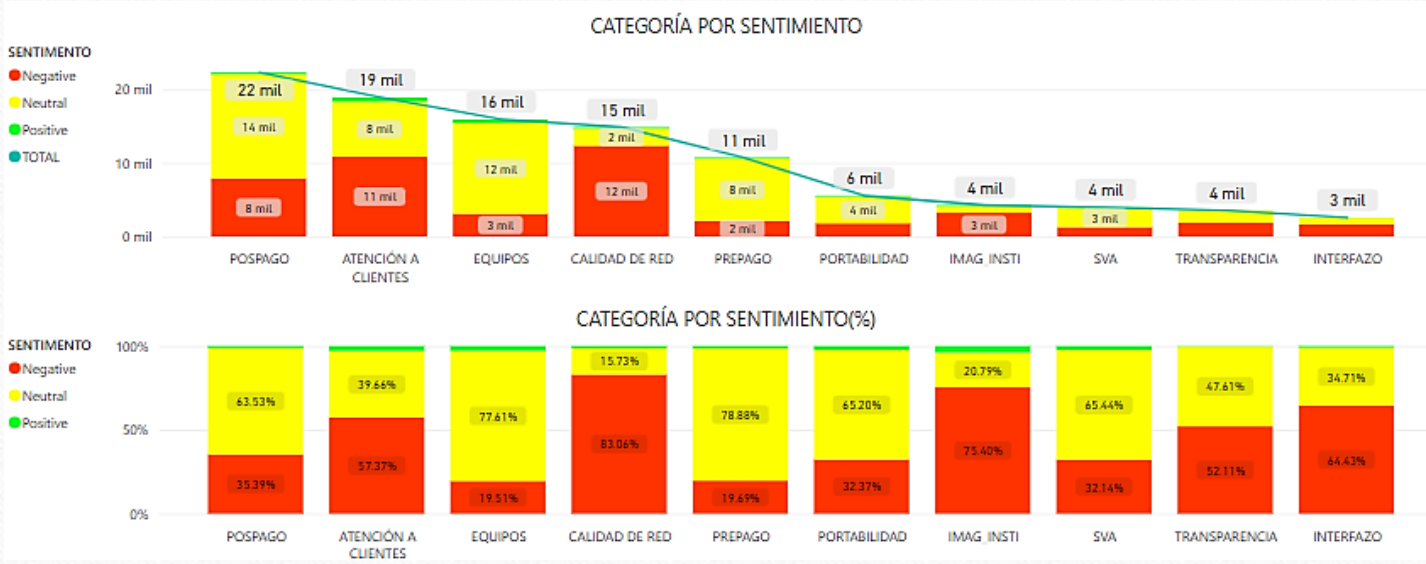
```
df2.head(3)
```

	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

RESULTADOS

TABLERO PROPUESTO PARA EL RESUMEN POR FUNCIONALIDAD HERRAMIENTA: POWER BI





RESULTADOS

TABLERO PROPUESTO PARA EL RESUMEN POR FUNCIONALIDAD HERRAMIENTA: POWER BI

REFERENCIAS

- 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
- I.V. (2019, 6 noviembre). Tipos de gráficos y diagramas para la visualización de datos. ingeniovirtual.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