Primera entrega del proyecto: Propuesta inicial

Integrantes:

Andrés Parra Garzón

Carlos Cañas Díaz

Shirley Sánchez Sedano

Resumen.

El análisis, priorización y solución oportuna de quejas de los clientes es una tarea importante en la prestación de un servicio, y más aún en un sector como el de aerolíneas, donde cada día la competencia en precios es más agresiva; es ahí donde la velocidad de solución de problemas inherentes a la operación cobra importancia para maximizar la satisfacción del cliente. El presente proyecto busca, haciendo uso de las quejas puestas por los clientes de una aerolínea a través de diferentes medios de contacto, generar segmentación de estas, que permita conocer el nivel de insatisfacción y el sentimiento que tuvo el cliente al escribirla. Esto permitirá clasificar las quejas en las cuales el cliente está inconforme pero calmado y su solución puede esperar un tiempo prudente , hasta las quejas donde el cliente se encuentra muy molesto y necesita atención inmediata para su problema, asociado al proceso donde se presenta la queja: no es lo mismo un cliente que después de su vuelo tuvo un problema con equipaje dañado, el cual puede estar molesto pero se puede solucionar en cierto tiempo, que un cliente que se encuentra haciendo check-in en el aeropuerto y su reserva no aparece, el cual necesita atención inmediata y su nivel de inconformidad es mayor.

La complejidad de esta tarea reside en dos pasos importantes: el primero es ser capaces de utilizar una queja escrita por una persona, la cual es una línea de texto o un párrafo sin estructura, y convertirla en información que sea utilizable y comparable con otras quejas recibidas. Para esto, haciendo uso de técnicas de procesamiento del lenguaje natural y de análisis de sentimientos, se transformarán estas quejas de manera que tengan información útil del contenido de la queja, el sentimiento del cliente cuando la escribió y el nivel de inconformidad que tuvo al escribirla. El segundo paso importante es, con esta información transformada, cómo poder agrupar miles de quejas recibidas en grupos homogéneos que permita darle a la empresa información valiosa de cuáles son las más urgentes por atender y resolver, y cuáles pueden ser resueltas más adelante. Para esto se utilizarán técnicas de segmentación de datos, las cuales buscan patrones internos en la información y que tan parecidas son las quejas, independientemente de las palabras usadas, el largo del texto de la queja y del lenguaje utilizado por la persona. Si se contaran con pocas quejas esto sería una tarea sencilla para una persona, pero al contar con miles de quejas por día, se hace necesario hacer uso de técnicas más avanzadas.

Introducción.

Análisis y agrupación de quejas realizadas por usuarios de una aerolínea a través del análisis de sentimientos.

La satisfacción de los clientes es el indicador directo de crecimiento tanto en ingresos como en imagen de una organización. Por lo tanto, no dar solución adecuada a las reclamaciones de los usuarios puede afectar de manera importante la reputación de la entidad, generando pérdida de clientes y dinero. Por lo que, el propósito de este proyecto es realizar agrupación

de las quejas realizadas a una aerolínea para posteriormente hallar diferentes grados de severidad de esta, con el propósito de identificar cuáles son las quejas más recurrentes, cuales se deben resolver con mayor prontitud y además aprovechar las inconformidades para implementar medidas que ayuden a mejorar los procesos de la aerolínea.

La solución de este problema es de gran importancia para la aerolínea que recibe las quejas, sobre todo para las áreas relacionadas con la solución de estas como el área de marketing, servicio al cliente y gestión de PQRS.

Revisión preliminar de la literatura.

Detección de sugerencias en tuits de usuarios y seguidores de aerolíneas con minería de sugerencias: Este trabajo de grado se centra en la recolección de tuits creados por usuarios de aerolíneas mexicanas: Aeroméxico, Interjet, TAR, Viva Aerobús y Volaris. En total se recolectaron 21.161 comentarios que se escribieron entre el 9 de agosto y el 27 de octubre del 2019.
El análisis consiste inicialmente en realizar procesamiento de lenguaje natural donde se pasan todos los comentarios a minúsculas, se eliminan acentos, emoticones, comillas, guiones bajos, stopwords y caracteres numéricos. Este procedimiento es similar al realizado en nuestro proyecto, además de que trata del mismo tema. Sin embargo, en esta tesis se hace uso de personas externas para etiquetar los comentarios según una clasificación de sugerencias, y hacer uso finalmente modelos supervisados, este es nuestra gran diferencia ya que nuestro problema es de carácter no supervisado.

Identificación y análisis de quejas en Twitter de los principales bancos en México de 2018 a 2019 mediante técnicas de minería de datos y recuperación de información: En esta tesis como dice su título, se centra en las quejas que se les realizan a los principales bancos en México. Con el fin de buscar tendencias y detectar comportamientos similares, el procedimiento en este caso consistió en realizar el procesamiento de lenguaje natural donde después de cambiar el texto a minúsculas, eliminar acentos, caracteres especiales, URL, signos de puntuación y stopwords, usar unigramas y ponderación con TF-IDF; se realiza el proceso de clusterización por medio de K-medias en donde se pueden identificar los temas correspondientes a cada grupo de queja, posteriormente a cada clúster se crea una subdivisión por medio de clustering jerárquico divisivo con el criterio Ward. Aunque el tema no es similar, este trabajo nos da ideas para que una vez empleemos NLP no quedarnos únicamente con la clusterización inicial, sino analizar cada grupo con el objetivo de realizar subdivisiones dentro de cada agrupación de quejas.

Humanizing Customer Complaints with NLP Algorithms: En este artículo se muestra la importancia de prestar atención a las emociones humanas en cada una de las quejas para realizar un mejor análisis, y es por esto que en uno de los casos del artículo se presenta el “análisis de sentimientos”. Inicialmente se realiza el procesamiento de lenguaje natural, el cual es explicado de manera teórica el detalle de los pasos a seguir, para posteriormente realizar la agregación del sentimiento asociado a cada queja y realizar las respectivas agrupaciones de esta. Este artículo es de interés en nuestro proyecto, ya que nos muestra la importancia de realizar el análisis de sentimientos para obtener mejores resultados.

Using Text Mining and Cluster Analysis to Improve Customers Complaints System: En este trabajo el objetivo es utilizar métodos de minería de texto y agrupamiento para mejorar el sistema de quejas de los clientes de un sistema de transporte. Para realizar la minería de texto se utilizó el software Rapid Miner y posteriormente se realizó clusterización

con K-medias. Aunque la temática del trabajo no es igual ni el procesamiento inicial de las quejas, ya que en nuestro caso se realizará procesamiento de lenguaje natural, la finalidad es la misma al realizar el agrupamiento por medio de técnicas de aprendizaje no supervisado.

Descripción de los datos.

La base de datos a utilizar para el presente ejercicio académico se trata de data histórica de quejas realizadas por clientes de una aerolínea. Dicha información se encuentra como un mensaje desestructurado en la variable “descripcion". Utilizando técnicas de NLP se buscará dar una respuesta a la pregunta planteada.

Originalmente la base de datos propuesta contenía más de 76 mil observaciones, sin embargo, durante el presente ejercicio exploratorio identificamos la necesidad de reducir significativamente el universo de datos ya que, cada observación puede contener mensajes lo suficientemente grandes como para afectar el rendimiento del modelo, considerando que las ejecuciones se harán en computadores personales. Finalmente, se tomó la decisión de trabajar con una base de datos de 10 mil observaciones con mensajes de 250 palabras en promedio.

Estrategia. Para realizar la exploración de los mensajes contenidos en la base de datos, se llevó a cabo la siguiente estrategia: 1. Extracción de datos, 2. Cálculo de estadísticas sin preprocesamiento, 3. Cálculo de estadísticas con preprocesamiento, 4. Comparativa de estadísticas, 5 Observaciones y conclusiones para cada comparación.

Limitaciones. Consideramos que este ejercicio está limitado por la madurez de las librerías de NLP en español, con respecto a su contraparte en inglés, ya que los mensajes a procesar son en su mayoría en idioma español. Si bien, existen estrategias para hacer uso de las librerías más maduras, como traducir los textos al inglés, por el momento optamos por mantenerlos en su idioma original.

Preprocesamiento. Para preparar la data que nos permitirá construir las variables que nos van a alimentar nuestro modelo de aprendizaje no supervisado, se utilizaron las siguientes técnicas: 1. Tokenizacion, 2. Conversión de texto a minúsculas, 3. Eliminación de signos de puntuación, 4. Eliminación de stopwords, 5. Lematización.

Análisis exploratorio.

Cantidad de palabras por mensaje.

Imagen 1. Cantidad de palabras por mensaje (sin preprocesamiento)

Imagen 2. Cantidad de palabras por mensaje (con preprocesamiento)

En esta comparativa se observa que la mayoría de los mensajes conserva su densidad de palabras aún después del preprocesamiento, salvo para un pequeño grupo para el cual se reduce significativamente su cantidad de palabras a menos de mil.

Palabras que más se repiten.

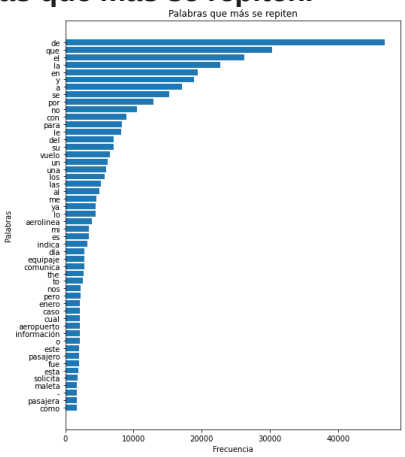

Imagen 3. Palabras que más se repiten en mensaje (sin preprocesamiento)

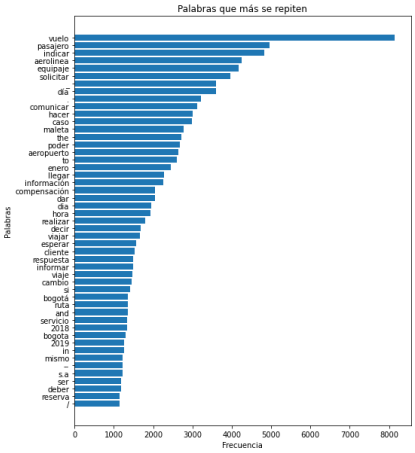

Imagen 4. Palabras que más se repiten en mensaje (con preprocesamiento)

En esta comparativa hay varios aspectos relevantes a destacar: se observa el ruido que hacen los stopwords sobre la data, el incremento en la frecuencia de algunas palabras como resultado del proceso de lematización, y que hay mensajes en otros idiomas diferentes al español que habrá que manejar.

Named Entity Recognition (NER).

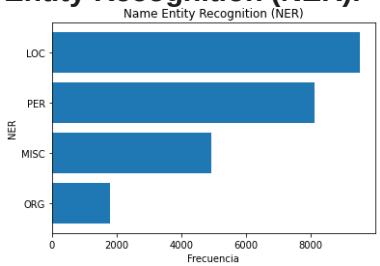

Imagen 5. Aplicación de NER (sin preprocesamiento)

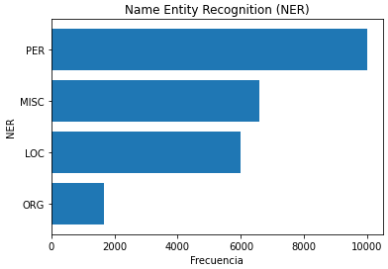

Imagen 6. Aplicación de NER (con preprocesamiento)

Al hacer el análisis de reconocimiento de entidades vemos que al principio prima la información geográfica por sobre el resto. Luego del preprocesamiento se logran identificar más personas. Asimismo, es destacable que la información miscelánea también crece.

Part of Speech (POS).

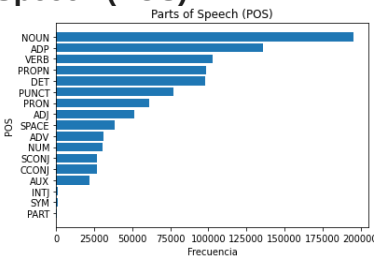

Imagen 7. Aplicación de POS (sin preprocesamiento)

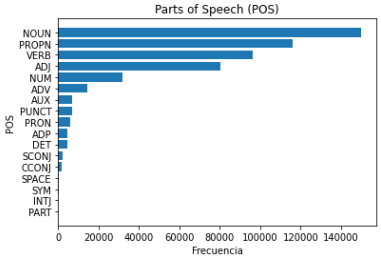

Imagen 8. Aplicación de POS (con preprocesamiento)

En el análisis POS vemos que antes del preprocesamiento la información es bastante variada, gramaticalmente hablando; sin embargo, luego de procesar la base, la información se concentra en nombres, nombres propios, verbos y adjetivos.

Adicional a los análisis anteriores, se realizaron otros que no se colocaron en el documento por restricciones de espacio. Se realizaron los análisis: cantidad de caracteres por mensaje, tamaño promedio de palabra, bigramas y trigramas que más se repiten.

Construcción de variables.

Para realizar la construcción de las variables a partir del texto preparado anteriormente, utilizamos la técnica llamada **Term Frequency-Inverse Document Frequency (TF-IDF)**, la cual se caracteriza por darle importancia a las palabras que le dan significado al mensaje por sobre las que no. Primero se calcula la frecuencia del término en función del total de términos en el mensaje (Term Frequency) y luego se calcula el peso que permite identificar fácilmente los términos que no se repiten tanto o no son tan relevantes (Inverse Document Frequency).

Originalmente obtuvimos una matriz de 2099 features, sin embargo, excluimos las variables con etiqueta numérica (0, 1, 2, 3, etc.), ya que nos parece no poseen significado per se. Al final conservamos sólo 2020 de éstas para proceder con el análisis de clustering. Por supuesto, ante la gran cantidad de variables resultantes, estamos ante un problema que requiere del uso de técnicas para reducir dicha dimensionalidad, como las que hemos aprendido hasta ahora en el curso (PCA y SVD).

	00am	abajo	abierto	able	abogado	abordaje	abordar	about	abril	abrir	...	zz062	zz064	zz07	zz08	zz09	ángel	área	él	último	único
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.245383	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.108271	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.107536	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0

5 rows × 2020 columns

Imagen 9. Vistazo a matriz de variables construida a partir de TF-IDF

Para ver más detalle sobre los análisis realizados, en el repositorio del proyecto se encuentra un cuaderno con la secuencia del proceso, técnicas y librerías utilizadas.

Propuesta metodológica.

Para el componente de Procesamiento de Lenguaje Natural se hará uso de técnicas de limpieza de datos, eliminando palabras utilizadas normalmente por las personas, pero que no agregan significado a la queja, como lo son los conectores, nombres propios, número de identificación, entre otros. Luego de esta limpieza se realizará el proceso de tokenización de las quejas con diccionarios especializados disponibles en el idioma español, que permita agrupar palabras con raíces similares, o grupos de palabra que por sí solas no tienen significado, pero con el uso en conjunto con otra palabra sí, todo esto con el fin de reducir la cantidad de palabras que se convertirán en variables del modelo de clustering.

Para el componente de clustering inicialmente se propone utilizar un modelo DBSCAN, el cual tiene la ventaja de no tener que clasificar necesariamente todos los componentes de la base de datos en un clúster, y que muestre al usuario posibles outliers y que potencialmente estos outliers sean los clientes con mayor urgencia e inconformidad de la información disponible. Como segundo posible algoritmo a utilizar se propone el modelo aglomerativo jerárquico, ya que las quejas están asociados a un proceso de la compañía. Este esquema permitirá ir agrupando quejas tanto por su nivel de incomodidad como por el proceso el cual está asociado, para así finalmente tener un único clúster.

Bibliografía.

- Jiménez Castro, R. (2021). Detección de sugerencias en tuits de usuarios y seguidores de aerolíneas con minería de sugerencias. *Instituto de Ingeniería y Tecnología*.

- Armas, D. (2020). Identificación y análisis de quejas en Twitter de los principales bancos en México de 2018 a 2019 mediante técnicas de minería de datos y recuperación de información.

- Morde, V., 2022. Humanizing Customer Complaints with NLP Algorithms. [online] Linkedin.com. Available at: <https://www.linkedin.com/pulse/humanizing-customer-complaints-nlp-algorithms-vishal-morde/> [Accessed 1 September 2022].

- HASAN, S. (2018). *Using Text Mining and Cluster Analysis to Improve Customers Complaints System* (Doctoral dissertation, The British University in Dubai (BUiD)).