Análisis y agrupación de quejas realizadas por usuarios de una aerolínea latinoamericana

Resumen

El análisis, priorización y solución oportuna de quejas de clientes es una tarea importante en la prestación del servicio en el transporte aéreo, donde cada día surgen nuevas aerolíneas y por lo tanto la competencia es cada vez mayor; es ahí donde la velocidad de solución toma importancia para maximizar la satisfacción de los usuarios.

El presente proyecto tiene como objetivo segmentar y analizar un conjunto de quejas interpuestas por clientes de una aerolínea latinoamericana a través de diferentes medios de contacto, con el propósito de conocer la naturaleza de los reclamos y el nivel de insatisfacción del cliente. Esto permitirá dar solución a los reclamos según su nivel de urgencia. Para lograr lo anterior, se hizo uso de técnicas de procesamiento de lenguaje natural con el fin de transformar los textos escritos por los usuarios en datos utilizables por el modelo de K-means dando como resultado seis agrupaciones. Según los resultados arrojados por cada clúster se puede evidenciar, en general, la temática de cada uno de ellos, sin embargo, para futuros análisis es importante probar nuevas técnicas de NLP para mejorar el proceso de limpieza y tener como resultado agrupaciones disjuntas entre sí.

Introducción

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 latinoamericana para posteriormente hallar diferentes grados de severidad de esta, con el fin 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.

En la literatura existe diversa información con respecto a este tema. Aldunate, Á (2022) en su artículo presenta la importancia de comprender los factores que influyen en la satisfacción del cliente a través del análisis de encuestas por medio del procesamiento de lenguaje natural y a partir de allí crear un modelo supervisado. Por su parte, Gavval, R., & Ravi, V. (2020) enuncia la necesidad de resolver oportunamente las reclamaciones para evitar pérdidas económicas, es por eso que en su artículo sobre agrupación de quejas de clientes bancarios por medio de redes sociales evidencia como el algoritmo de K-medias y variantes de este funcionan muy bien a la hora de realizar las agrupaciones.

Materiales y Métodos.

Datos utilizados.

La base de datos a utilizar es un lote de quejas históricas realizadas por clientes de una aerolínea latinoamericana, la cual fue preparada y compartida específicamente para este ejercicio académico. La información para explotar son los mensajes colectados que contienen el detalle de la queja o reclamo. La base cuenta con 10 mil observaciones.

Proceso de limpieza de los datos.

Para preparar la data a partir de los mensajes de texto, que nos permitirá construir las variables que van a alimentar nuestro modelo de aprendizaje no supervisado, se utilizaron las siguientes técnicas de Lenguaje de Procesamiento Natural (NLP): 1. Inferencia del idioma para tomar en cuenta sólo los mensajes en español, 2. Exclusión de palabras y patrones en el texto usando expresiones regulares, 3. Tokenización, 4. Eliminación de stopwords, 5. Lematización, 6. Exclusión de nombres propios usando Part of Speech (POS), por mencionar las más relevantes.

Estadísticas descriptivas de los datos.

Para realizar la exploración de los mensajes contenidos en la base de datos, se llevó a cabo la siguiente estrategia: 1. Lectura de datos crudos, 2. Cálculo de estadísticas de los mensajes sin y con preprocesamiento, 3. Comparativa de estadísticas y análisis de resultados. A continuación, se muestra un resumen de algunas de las estadísticas descriptivas calculadas sobre el texto:

Estadística	Resultados sin preprocesamiento	Resultados con preprocesamiento	Análisis de resultados
Cantidad de palabras promedio por mensaje	88 palabras	53 palabras	Al eliminar stopwords y lematizar se reducen significativamente la cantidad de palabras por mensaje
Cantidad de caracteres promedio por mensaje	526 caracteres	380 caracteres	Al eliminar stopwords y lematizar se reducen la cantidad de caracteres por mensaje
Palabras que más se repiten	De, que, el	Vuelo, pasajero, aerolínea, equipaje, compensación	Se observa el ruido que hacen las stopwords sobre la data
Trigramas que más se repiten	En el aeropuerto, en la ruta, en el vuelo	Solicitar información equipaje, información equipaje demorado	Después del preprocesamiento los grupos de tres palabras que más se repiten comienzan a describir una queja

Tabla 1. Listado de estadísticas para ilustrar la diferencia entre los datos originales y los preprocesados

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).

Algoritmo de clustering seleccionado.

El algoritmo de K-Medias forma parte de los algoritmos de clustering basados en centroides, el cual agrupa las observaciones en un número predefinido de K clústeres de forma que, la suma de las varianzas internas de los clústeres sea lo menor posible. Esto se logra en cada clúster a través de la suma de todos los pares de distancias euclidianas cuadráticas de las observaciones, dividida por el número total de observaciones en el mismo. Este proceso se repite hasta que las asignaciones llegan al punto en que no cambian.

Para aplicar el algoritmo K-Medias se utilizó la función KMeans de Sklearn, cuyos parámetros principales son los siguientes:

- **n_clusters:** Número de clústeres predefinido
- **init:** Método de inicialización para asignar los centroides iniciales, puede ser alejándolos lo más posible o asignándolos aleatoriamente
- **n_init:** Número de veces que se va a repetir el proceso, cada vez con una asignación aleatoria inicial distinta
- **max_iter:** Máximo de iteraciones permitidas

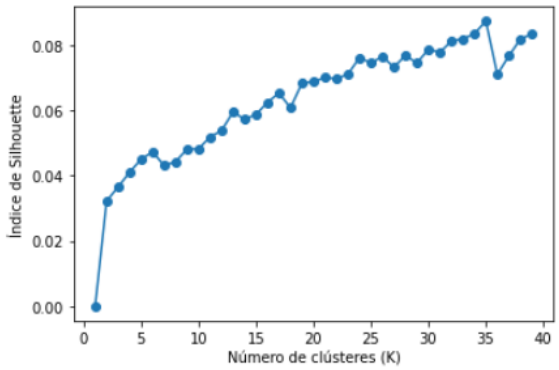
Algoritmos utilizados

Luego de realizar la transformación de los datos mediante el uso de TF-IDF se realizaron pruebas utilizando los siguientes algoritmos: K-medias, Jerárquico aglomerativo, DBSCAN y detección de comunidades. Analizando los resultados de cada uno de ellos obtenemos que:

- Haciendo uso del algoritmo jerárquico aglomerativo se obtiene un clúster de mayor tamaño frente a los demás clústeres, y no se logran encontrar un diferenciador contundente entre los diferentes grupos. Al analizar las palabras dentro de los clústeres se encuentra que vuelo, maleta, esperar, compensación aparecen en cada uno de ellos, con diferente frecuencia. Se encontraron 5 clústeres usando esta técnica
- El algoritmo DBSCAN logró identificar algunos outliers de la información, sin embargo, la calibración del épsilon mediante las técnicas estudiadas no ayudó a este fin, por lo que mediante prueba y error se intentó identificar un valor, pero no fue posible, se obtuvo al igual que en al caso anterior un clúster más grande que los demás, sin ninguna diferenciación importante en cuanto frecuencia de los datos
- Con detección de comunidades se lograron encontrar 15 clústeres, con información satisfactoria que da noción de la temática de la queja y de la severidad de esta, sin embargo, la técnica no permite seleccionar la cantidad de grupos deseados, en cambio solo permite seleccionar la cantidad mínima de elementos para conformar un clúster, es por esto que algunos de los clústeres conformados fueron muy similares a otros, y el análisis de 15 clústeres se hizo complejo
- K-medias también logró resultados satisfactorios y relevantes, siendo capaz de clasificar la temática de la queja, el origen de estas y la severidad de acuerdo con las palabras únicas que presentaron alta frecuencia. Como fue posible seleccionar la cantidad de clústeres y los clústeres generados son parecidos a los de detección de comunidades en cuanto a las palabras únicas que dan sentido al grupo, se selecciona esta técnica como la adecuada para el problema de este proyecto

Resultados del algoritmo seleccionado

Para el algoritmo de K-medias se hizo uso del coeficiente de Silhouette con el fin de encontrar la cantidad de clústeres adecuados. Los resultados se pueden observar a continuación:



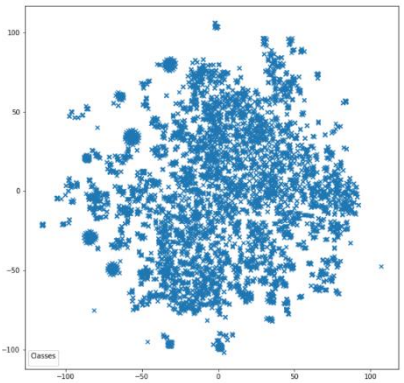
Gráfica 1. Coeficiente de Silhouette

Podemos observar que el valor del índice no es muy variable para los primeros 40 clústeres, teniendo un máximo en 35. Para la primera iteración se utiliza esta cantidad de clústeres y se analiza el resultado. Para el análisis de resultado se utilizará la técnica de reducción de dimensionalidad y visualización t-SNE propuesta por Van der Maaten y Hinton (2008), la cual se ajusta a problemas de alta dimensionalidad como es este caso, embebiendo los datos en un espacio de dos o tres dimensiones para su visualización, donde elementos con alta probabilidad de ser similares se modelan como puntos cercanos, y elementos con baja probabilidad de ser similares se modelan como puntos lejanos.

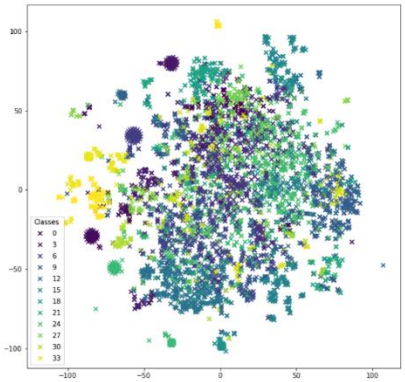
Inicialmente utilizando esta técnica, sin ninguna etiqueta de clúster, los datos se visualizan como la gráfica 2. Se observa una nube de puntos muy homogénea, con algunos pequeños grupos focalizados en la periferia de la gráfica.

Ahora bien, haciendo uso de K-medias con 35 clústeres, el resultado se muestra en la gráfica 3. El algoritmo fue capaz de agrupar y separar los pequeños grupos observados previamente, sin embargo, los clústeres se encuentran superpuestos unos con otros, esto afianza la suposición inicial de que la información suministrada es muy homogénea en cuanto a los elementos y palabras que contienen las quejas.

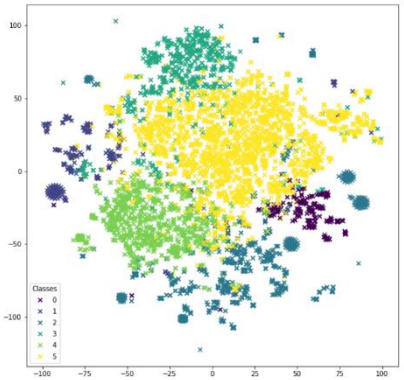
Debido a este resultado, y a la complejidad de análisis de 35 clústeres, se decide iterar mediante ensayo y error disminuyendo la cantidad de clústeres hasta encontrar una solución satisfactoria, que pueda clasificar la nube central, pero que nos permita identificar esos pequeños grupos de las periferias. Dicho número de clústeres propuesto es 6. Con 6 clústeres y volviendo a usar la técnica t-SNE, podemos ver el resultado final en la gráfica 4.



Gráfica 2. t-SNE sobre datos sin agrupar



Gráfica 3. t-SNE con K-Medias (k=35)



Gráfica 4. t-SNE con K-Medias (k=6)

A pesar de que se sigue presentando superposición de clústeres, los grupos se encuentran más definidos y los valores lejanos del clúster son menores, aunque presentes. Vemos como las periferias se encuentran clasificadas. Para entender que tipo de quejas se encuentran en cada clúster, se hace uso de nubes de palabras que permita visualizar las palabras con mayor frecuencia dentro del clúster:

Gráfica 5. Comparativo de clústeres obtenidos a través de nubes de palabras

Vemos como existen palabras dominantes en cada clúster, sin embargo, en algunos de ellos no son palabras que permitan independizar el grupo de los demás, como es el caso del clúster 5 donde pasajero e indicar prevalecen. El clúster cero se caracteriza por las llamadas hechas por call center donde se crea el caso por daño en el equipaje, donde el cliente expresa ser compensado. El clúster 1 se caracteriza por inconformidad en las demoras y en los problemas operativos de las rutas. El clúster 2 lo conforman las quejas en el cual el cliente solicita información por equipaje perdido. El clúster 3 es en el que la palabra compensación prevalece, pero no es claro la temática de la queja, pero si el sentimiento de urgencia del cliente. El clúster 4 y el clúster 5 son los de mayor cantidad de elementos, donde maleta, equipaje, pasajero prevalecen, pero no es claro ni la temática ni el sentimiento. Los clústeres se pueden resumir en la siguiente tabla:

Clúster	Cantidad de elementos (%)	Temática definida	Sentimiento definido
0	4.50%	Si	Si
1	5.50%	Si	No
2	16.95%	Si	Si
3	11.71%	No	Si
4	17.07%	Si	No
5	44.27%	No	No

Tabla 2. Resumen de clústeres obtenidos usando K-Medias (k=6)

Conclusión

Con respecto a este proyecto se puede concluir que los resultados dependen fundamentalmente de la calidad de las variables seleccionadas a ingresar al modelo, es por esta razón que gran parte del tiempo se implementó en realizar una correcta limpieza de los datos, la cual consistió principalmente en elegir únicamente las quejas en idioma español, eliminar caracteres especiales, espacios, URL, números, emojis y stopwords. Además, se eliminó de manera manual un conjunto de palabras vacías que no aportaban al análisis, estas de detectaron una vez se implementó TF-IFD.

Fue importante escoger un buen paquete en Python para realizar la lematización, ya que la mayoría de las funciones están hechas para realizar un buen análisis en el idioma inglés. Por lo anterior se eligió Stanza, un paquete creado por un grupo de la universidad de Stanford, donde su característica principal es que se trata de modelos neuronales pre-entrenados que admiten 70 idiomas, entre ellos el español.

Una vez ingresadas las variables al modelo de K-medias se obtienen 6 clúster donde 4 de ellos se pueden observar claramente su temática, sin embargo, se evidencia que estas segmentaciones se traslapan entre sí en cuanto a sus términos más frecuentes.

Se recomienda para análisis futuros implementar otras técnicas y lematizadores en el idioma español para lograr una mejor limpieza de los datos, y explorar otros algoritmos de aprendizaje no supervisado que por cuestiones de tiempo no se pudieron probar. Además, realizar análisis dentro de cada clúster, es decir, es nuestro caso de observa que aún las características para cada segmentación son muy generales, por lo tanto, sería interesante aplicar análisis no supervisado dentro de cada clúster.

Bibliografía

Aldunate, Á., Maldonado, S., Vairetti, C., & Armelini, G. (2022). Understanding customer satisfaction via deep learning and natural language processing. *Expert Systems with Applications*, *209*, 118309.

Gavval, R., & Ravi, V. (2020). Clustering bank customer complaints on social media for analytical CRM via multi-objective particle swarm optimization. In *Nature Inspired Computing for Data Science* (pp. 213-239). Springer, Cham.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).