
# Análisis de Comentarios de Pacientes en una Clínica en Maryland utilizando Técnicas de Aprendizaje No Supervisado

**Mauricio Gonzalez Caro**

**Eva Karina Diaz Gavalo**

**Juan Felipe Padilla Sepúlveda**

**Andrés Eduardo Quiñones Ortiz**

**Universidad de los Andes**

**Maestría en Inteligencia Analítica de Datos**

**Aprendizaje No Supervisado**

**2024**


## Resumen

El objetivo de este proyecto es mejorar la evaluación de las encuestas de satisfacción que una clínica envía a sus pacientes después de cada atención. Con 12 ubicaciones en Maryland, Estados Unidos, y aproximadamente 30.000 pacientes por año, este hospital comunitario brinda servicios médicos, dentales, pediátricos y de salud mental. Mediante una breve encuesta anónima, la clínica recopila datos demográficos de los pacientes, así como niveles de satisfacción y otros comentarios, que suelen ser extensos y están disponibles en idiomas como el inglés y el español.

Actualmente, las respuestas de las encuestas son recogidas en medios digitales, sin embargo su revisión, análisis y derivación a las áreas especializadas de atención se hace de manera manual, lo que lo convierte en un proceso lento y laborioso. Este enfoque impide la identificación de posibles acciones o resultados a partir de la retroalimentación y limita la capacidad de la clínica para implementar mejoras más profundas y específicas. Si bien es fácil abordar las respuestas estructuradas, las preguntas de respuesta abierta están infrautilizadas a pesar de ser una fuente de retroalimentación valiosa.

Es debido a esto que el proyecto propone el uso de herramientas de procesamiento del lenguaje natural (NLP) y técnicas de aprendizaje no supervisado, como algoritmos de agrupamiento (Clustering), para clasificar las respuestas, identificar temas recurrentes y captar emociones. Al utilizar estas técnicas, se espera poder transformar los datos en información procesable que ayude a mejorar la práctica clínica y la experiencia del paciente.

## Introducción

En el ámbito de la salud, comprender la satisfacción y las inquietudes de los pacientes es esencial para elevar la calidad del servicio y la atención brindada. Este enfoque permite identificar áreas de oportunidad y mejorar aspectos específicos del cuidado que se ofrece en una de las más grandes clínicas comunitarias en el estado de Maryland. Por razones de privacidad, el nombre de la clínica se reserva.

Cada vez que un paciente es atendido, se le envía una breve encuesta diseñada para medir la satisfacción con el servicio recibido. Esta encuesta, que es anónima, permite identificar la ubicación del servicio, el proveedor que lo prestó, y recopila información demográfica básica de los pacientes, como sexo, etnia, raza y lenguaje. Además, incluye preguntas de tipo sí/no para evaluar la satisfacción general, así como una pregunta abierta que invita al paciente a compartir cualquier comentario adicional que considere relevante. El propósito de esta encuesta es tomar decisiones informadas basadas en la retroalimentación obtenida, con el fin de mejorar la experiencia del paciente y fortalecer la relación entre ellos y la clínica, asegurándoles que sus opiniones son valoradas y consideradas.

Actualmente, el análisis de estas encuestas se enfoca en las respuestas estructuradas, realizar el análisis de los comentarios es un trabajo laborioso, ya que demanda mucho tiempo la lectura de cada comentario para poder determinar cual requiere de una acción específica. Además, no se puede establecer tendencias ni encontrar patrones de comportamiento con este análisis manual. Las respuestas varían en longitud y pueden estar escritas tanto en inglés como en español.

Este proyecto se enfoca en el análisis de estos comentarios de texto libre con el objetivo de clasificarlos en temas y comprender el sentimiento expresado en ellos, utilizando técnicas de aprendizaje no supervisado como un algoritmo de clustering para agrupar los comentarios para identificar patrones y tendencias que permitan responder a la pregunta: ***¿Qué sentimientos y temas recurrentes se expresan en los comentarios de los pacientes, y cómo se pueden utilizar estos para mejorar los servicios de la clínica?***

Realizando una revisión inicial de literatura, artículos como Text Mining In Healthcare y Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care | BMJ Quality & Safety presentan técnicas de minería de datos, con un enfoque en la clasificación de minería de texto en

salud. Se analizan en detalle los avances y mejoras propuestas en este campo, junto con los desafíos relacionados con el tipo de datos en el sector.

Por otra parte, en Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer generated reviews se describe cómo técnicas de agrupamiento se pueden usar para detectar patrones y tendencias en grandes volúmenes de datos no estructurados, como los comentarios de los pacientes utilizando reseñas escritas por pacientes sobre medicamentos para entrenar y probar modelos de aprendizaje supervisado que clasifiquen las reseñas en una escala de cinco puntos. Si bien no es nuestro objetivo utilizar modelos supervisados, en este paper se destaca a Random Forest como el mejor en la predicción de satisfacción de los usuarios; esto podría ser útil en caso de querer estimar ratings o calificaciones en temas de satisfacción. De igual manera, en Patient Satisfaction with Healthcare Services and the Techniques Used for its Assessment: A Systematic Literature Review and a Bibliometric Analysis se ofrece una visión integral de los métodos utilizados en la recolección de datos de satisfacción y cómo el análisis de texto no estructurado puede mejorar la evaluación de los servicios de salud a través de una amplia revisión bibliográfica, explorando múltiples metodologías y cuáles son las variables que más afectan lo resultados de estimación de satisfacción.

En el campo del análisis de sentimientos en salud, estudios como Sentiment Analysis in Healthcare: A Brief Review subrayan la importancia de ajustar las técnicas lingüísticas según el idioma. Mientras que métodos como la lematización y el etiquetado de partes del discurso (POS tagging) han demostrado ser efectivos en inglés, su aplicación en otros idiomas, como el árabe (o el español en nuestro caso), requiere ajustes personalizados. Este enfoque refleja un desafío común: los idiomas tienen estructuras gramaticales y léxicas únicas que impactan directamente en la precisión del análisis de sentimientos. En nuestro trabajo, esta perspectiva es muy relevante, ya que observamos que la adaptación de técnicas estándar del inglés a los comentarios en español no siempre garantiza los mismos resultados.

## Materiales y métodos

La base de datos analizada en este proyecto comprende un total de 24,324 observaciones y 27 variables, las cuales recogen información de las encuestas realizadas a pacientes tras recibir servicios en la clínica desde el 4 de diciembre de 2019 hasta el día que se descargaron los datos 21 de agosto de 2024. Estas encuestas son enviadas a los pacientes después de cada cita médica y están diseñadas para captar varias respuestas que incluyen tanto información estructurada como comentarios abiertos. Con el objetivo de centrar el análisis en aspectos directamente relacionados con los comentarios de los pacientes y su contexto, se seleccionaron las siguientes variables: ***Respondent ID, End Date, Language, y Do you have any other comments or concerns?.*** Para ver la metodología implementada ver anexo 0

Ahora bien, iniciemos con el preprocesamiento de la información. El *Respondent ID* es un identificador único en formato *float* que permite individualizar cada respuesta, la variable *End Date* recoge la fecha en la que la encuesta fue completada, y se convertirá a formato de fecha para facilitar su análisis temporal, *Language* indica el idioma en el que se completó la encuesta, y la variable *Do you have any other comments or concerns?* contiene los comentarios abiertos proporcionados por los pacientes. (ver anexo 1)

Debido a la naturaleza bilingüe de la base de datos, se consideró esencial dividirla en dos subconjuntos: uno para los comentarios en inglés y otro para los comentarios en español. Esta división es relevante, ya que las diferencias pueden influir en los resultados del análisis.

De las 11,638 filas presentes en la base de datos surveys_english, se identificaron 6,201 registros con comentarios nulos, lo que representa aproximadamente el 53.27% del total de las observaciones en inglés. En cuanto a la base de datos surveys_spanish, que contiene 12,680 filas, se detectaron 7,805 registros con comentarios nulos, lo que equivale a un 61.54% de las observaciones en español. Recordemos que esta pregunta es opcional, preguntando a los pacientes si tienen algún otro comentario adicional a las preguntas realizadas.

Por otra parte, teniendo en cuenta que algunos pacientes no deseaban incluir comentarios adicionales, no obstante escribieron en la respuesta “No” o “Not” o alguna de sus variaciones en mayúscula y minúscula para

ambos idiomas, se hizo necesario evaluar cuántas de estas respuestas correspondían a estos casos, dando como resultado que en inglés 916 respuestas eran solo “no” o “not” y en español un total de 1,343.

De acuerdo a lo anterior y para evitar ruido al hacer el modelo, se determinó eliminar estos casos que solo contienen “no” y “not” al igual que los comentarios nulos. Una vez eliminados de las bases de datos en inglés y español, se procedió graficar la distribución temporal de las encuestas, en el que se observa un pico considerable a mediados de 2020 y 2021. Este aumento podría estar relacionado a la pandemia de COVID. (ver anexo 2)

Aunque durante la pandemia las encuestas en español fueron significativamente numerosas, en la mayoría de los casos las encuestas en inglés parecen ser más frecuentes. En 2024, la cantidad de encuestas en español y en inglés se muestra más equilibrada, aunque las encuestas en inglés siguen dominando en cantidad.

Se procedió a analizar la longitud promedio de los comentarios en la base de datos en inglés, encontrando que, de un total de 4,521 comentarios, el promedio de palabras por comentario es de 20.00, con una desviación estándar de 38.32. Esto indica una considerable variabilidad en la cantidad de texto proporcionado por los pacientes. El comentario más corto consta de una sola palabra, mientras que el más largo alcanza las 925 palabras. Los percentiles muestran que el 25% de los comentarios contienen 5 palabras o menos, el 50% tienen 11 palabras o menos, y el 75% tienen 23 palabras o menos. A pesar de la presencia de algunos comentarios extensos, la mayoría son relativamente breves, lo que sugiere una distribución desequilibrada en la longitud de los comentarios. Este desequilibrio podría introducir sesgos en el análisis, ya que los comentarios extremadamente cortos, como aquellos de 2 palabras, podrían estar duplicados o no aportar valor significativo al análisis. Por tanto, en la siguiente etapa se aplicarán técnicas de preprocesamiento de lenguaje natural para limpiar y normalizar estos datos.

Por otro lado, en la base de datos en español, se analizaron 3,532 comentarios. La longitud promedio es de 13.00 palabras, con una desviación estándar de 19.56 palabras. Al igual que en los comentarios en inglés, existe una amplia variabilidad, pero en menor grado. El comentario más largo tiene 311 palabras. Los percentiles indican que el 25% de los comentarios tienen 3 palabras o menos, el 50% tienen 7 palabras o menos, y el 75% tienen 15 palabras o menos. Estos resultados muestran que los comentarios en español tienden a ser más cortos que los comentarios en inglés. (ver anexo 3)

Después de realizar el preprocesamiento inicial, nos enfocamos en preparar los datos textuales para el análisis. La primera etapa clave fue la tokenización, que consistió en dividir los comentarios en palabras individuales o tokens. En este proceso, preservamos algunas palabras específicas como "no", "nunca", "nadie" y "pero", ya que son fundamentales para el análisis de sentimiento. Seguido de esto, procedimos con la lematización, permitiendo simplificar el texto. Esto nos permitió tratar con una representación semántica más limpia y estandarizada de los comentarios. La etapa siguiente consistió en convertir los textos lematizados a vectores numéricos para su análisis. Esta fase de vectorización fue crítica para poder aplicar algoritmos de clustering y análisis de temas.

- Para los comentarios en inglés, utilizamos Universal Sentence Encoder (USE) desarrollado por Google. Este modelo fue elegido debido a su capacidad de capturar el significado semántico completo de las oraciones, lo que lo hace adecuado para comentarios de diversas longitudes.

- Para los comentarios en español, inicialmente intentamos usar el modelo USE multilingüe, pero experimentamos problemas técnicos que impidieron su uso efectivo. Por ello, decidimos utilizar FastText de Facebook, un modelo para el análisis de texto en español. FastText es especialmente eficiente en la generación de vectores que capturan relaciones semánticas, y fue la mejor alternativa disponible para nuestros comentarios en español.

Una vez convertidos los comentarios en vectores, no reducimos la dimensionalidad de los vectores. La razón detrás de esta decisión es que tanto USE como FastText ya generan representaciones vectoriales optimizadas y suficientemente densas, diseñadas para capturar la semántica completa de las oraciones sin necesidad de reducir la información. Aplicar técnicas de reducción de dimensionalidad, como PCA, habría eliminado información relevante que estos modelos preservan.

Para encontrar el número adecuado de clusters, empleamos el algoritmo K-Means apoyándonos en el Método de la Silueta para evaluar distintos valores de K. Tras evaluar K en un rango de 2 a 10, seleccionamos el valor de K que maximizó el coeficiente de silueta y para facilitar la visualización de los clusters, aplicamos una reducción dimensional con PCA a dos componentes principales donde coloreamos los clusters óptimos.

Con los clusters ya definidos, aplicamos el modelo Latent Dirichlet Allocation (LDA) a los comentarios de cada cluster. Este modelo nos permitió identificar temas dentro de cada grupo, extrayendo palabras clave que describen los tópicos dominantes. Para cada cluster, generamos cuatro temas principales que reflejan las inquietudes o áreas de interés de los pacientes. Este análisis temático nos proporcionó una comprensión más profunda de lo que caracteriza cada grupo de comentarios.

Finalmente, para una representación visual clara de los temas identificados, creamos nubes de palabras para cada cluster. Estas nubes de palabras ofrecieron una visualización intuitiva de los términos más frecuentes en cada grupo, facilitando la interpretación de los temas. La visualización permitió identificar rápidamente las palabras clave que dominan en los comentarios de cada cluster, lo cual fue fundamental para la interpretación de los resultados.

## Resultados y Discusión

El análisis de las nubes de palabras (word clouds) en ambos idiomas revela patrones interesantes en los comentarios de los pacientes:

- Inglés: en este idioma se encuentran palabras como time, go, moment thank, love, proffesional, service, great, entre otras que resaltan aspectos positivos y negativos de la clínica. (ver anexo 4). En este idioma el modelo propuesto presentó un mejor agrupamiento de los temas y palabras clave. Se establece un sentimiento dominante en cada uno de los clusters.

- Español: se resaltan términos como *excelente, gracias* y *amable;* esto sugiere a primera vista una recepción positiva de los servicios recibidos por parte del personal de la clínica, estando presentes palabras como doctor y enfermera. (ver anexo 5) Sin embargo, estas palabras se repiten en varios clusters, lo que las convierte en el sentimiento dominante de la experiencia del paciente y dificulta la diferenciación entre los clusters.

Estos resultados subrayan la valoración de la calidad humana y profesional del servicio, siendo consistentes con las expectativas de la atención médica de calidad.

El uso de Truncated SVD para reducir la dimensionalidad de los conjuntos de datos TF-IDF en ambos idiomas mostró que se puede preservar una gran parte de la información original (95.31% en inglés y 84.18% en español) con un número reducido de componentes (2500 en español y 1000 en inglés). Esto valida la eficacia de esta técnica en la conservación de información relevante, permitiendo un análisis más eficiente y manejable de grandes volúmenes de datos textuales.

El análisis de los resultados del método de la silueta en inglés (K=3) reveló un rendimiento significativamente mejor en comparación con el idioma español (K=8).

En inglés, el uso de Lemmatized_Comments fue altamente eficiente para capturar tokens y palabras clave, lo que permitió una segmentación clara en tres clusters bien definidos. A cada uno de estos clusters se le pudo asignar un sentimiento dominante de los beneficiarios que accedieron a los servicios del hospital: Cluster 1 reflejó "Descontento / Negativo", Cluster 2 representó "Inquietud / Necesidad", y Cluster 3 mostró "Gratitud / Satisfacción". Esto sugiere que, en inglés, la estructura del lenguaje y el análisis lematizado facilitaron una clasificación precisa de las emociones de los usuarios. (ver anexo 6)

En cambio, en español, a pesar de aplicar el mismo proceso, no se logró asignar un sentimiento distintivo para los ocho clusters resultantes. Este desafío probablemente se deba a la mayor complejidad del idioma español, caracterizado por su diversidad léxica y morfológica, que hace más difícil la identificación clara de patrones

emocionales en los datos. La variedad de formas verbales, sinónimos y matices semánticos en español complica la segmentación precisa, para este idioma, sería necesario un refinamiento adicional del proceso de lematización o la implementación de modelos más eficientes para capturar el verdadero sentido emocional de los comentarios de los beneficiarios. (ver anexo 7)

## Conclusiones

Con base en lo expuesto previamente, podemos concluir que el uso de técnicas como Truncated SVD y lematización indudablemente ayuda de manera significativa en el manejo de grandes volúmenes de datos textuales, preservando información relevante y permitiendo una segmentación efectiva. En particular, el análisis de comentarios en inglés, permitió clasificar adecuadamente sentimientos y temas recurrentes, lo que brinda una valiosa herramienta para identificar áreas de mejora y también las fortalezas, permitiendo dar respuesta a la pregunta problema planteada al inicio. No obstante, en el caso de respuestas en español presentó un mayor desafío debido a la complejidad del idioma, lo que precisa un mayor refinamiento en los métodos implementados para poder obtener resultados coherentes que puedan ser de utilidad.

Adicional a lo anterior, se recomienda revisar y modificar la pregunta final de la encuesta, específicamente la que solicita comentarios adicionales. En lugar de dejar el campo abierto directamente, se sugiere implementar una pregunta condicional que primero pregunte al paciente si desea proporcionar comentarios adicionales con las opciones "Sí" o "No". Solo en caso de seleccionar "Sí", se desplegaría un campo de texto para que el paciente ingrese sus observaciones. Este ajuste no solo optimizaría la recopilación de datos, sino que también ayudaría a filtrar respuestas irrelevantes o breves, como "no", "no gracias", "no thanks" y similares, las cuales introducen ruido en el análisis. Al reducir estas respuestas ineficaces, se facilitaría el proceso de extracción de insights valiosos y se mejoraría la calidad de la retroalimentación obtenida. De esta manera, la clínica podría enfocarse en comentarios más detallados y constructivos, permitiendo implementar mejoras más precisas y relevantes en sus servicios.

En cuanto a los pasos clave a seguir, se podría realizar un análisis más detallado y segmentado, dado que hasta el momento el análisis ha sido general, sin embargo, existe una oportunidad significativa en el desglose de los resultados, dado que la clínica cuenta con 12 sedes y el resultado de cada una puede diferir con las otras, esta profundización permitiría mayor precisión al identificar oportunidades de mejora dentro de las sedes, servicios e incluso los profesionales.

Finalmente, en estudios futuros se podría automatizar el análisis de estos resultados, incluyendo también una forma de derivar a las áreas responsables para que se puedan realizar intervenciones oportunas y de una manera más eficiente. La implementación de estas recomendaciones podría ayudar a la clínica en la gestión de la calidad del servicio brindado a sus pacientes, garantizando que las experiencias de los pacientes puedan ser tomadas en cuenta como insumos para el mejoramiento continuo.

## Bibliografía

Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: Opportunities and challenges. Journal of Biomedical Semantics, 9(1), 1-14. https://doi.org/10.1186/s13326-018-0179-8

Abualigah, L., Alfar, H., Shehab, M., & Abu Hussein, A. M. (2020). Sentiment analysis in healthcare: A brief review. En Recent Advances in NLP: The Case of Arabic Language (pp. 129-141). Springer. https://doi.org/10.1007/978-3-030-34614-0_7

Allenbrand, C. (n.d.). Supervised and unsupervised learning models for pharmaceutical drug rating and classification using consumer-generated reviews. Journal of Pharmaceutical Sciences. https://doi.org/10.1016/j.jphs.2023.07.007

Ferreira, D. C., Vieira, I., Pedro, M. I., Caldas, P., & Varela, M. (2023). Patient satisfaction with healthcare services and the techniques used for its assessment: A systematic literature review and a bibliometric analysis. Journal of Healthcare Quality Research. https://doi.org/10.3390/jhqr10001171

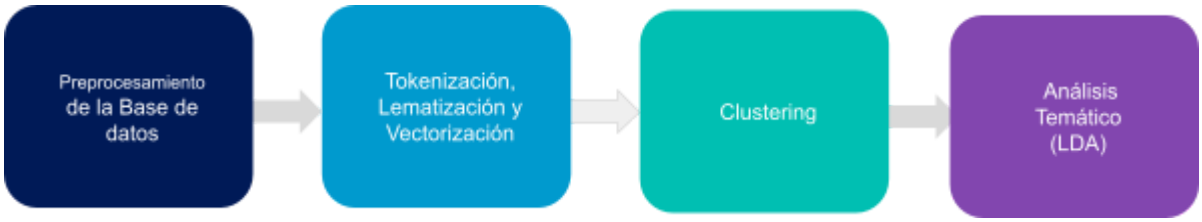
Berger, S., Saut, A., & Berssaneti, F. T. (2020). Using patient feedback to drive quality improvement in hospitals: A qualitative study. University of São Paulo. https://www.researchgate.net/publication/346380130_Using_patient_feedback_to_drive_quality_improvement_in_hospitals_a_qualitative_study

Rana, D., & Mahajan, P. Y. (2020). Text mining in healthcare. Journal of Health Informatics. https://www.researchgate.net/publication/338336460_Text_Mining_In_Healthcare

Wagland, R., Recio-Saucedo, A., Simon, M., Bracher, M., Hunt, K., Foster, C., Downing, A., Glaser, A., & Corner, J. (2016). Development and testing of a text-mining approach to analyze patients' comments on their experiences of colorectal cancer care. BMJ Quality & Safety, 25(8), 604-613. https://doi.org/10.1136/bmjqs-2015-004189

# Anexos

## **Anexo 0.** Metodología implementada

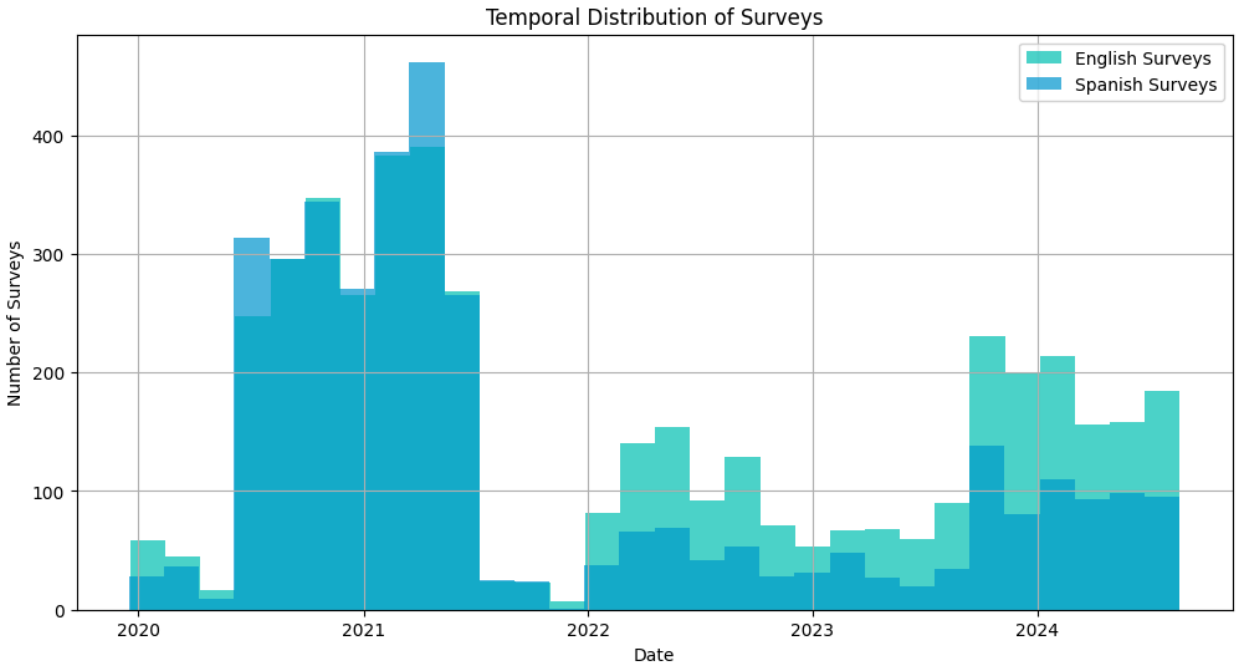



## **Anexo 1.** Datos recopilados en la encuesta de satisfacción

| | Respondent ID | Collector ID | Start Date | End Date | language | What sex were you assigned at birth? | What is your current Gender Identity? | What is your sexual orientation? | Are you a veteran? | Please check your race: | ... | Please select the Ethnicity: | Where was your appointment? | Appointment with who? | Unnamed: 20 | When you contacted this health professional's office to get an appointment for care you needed, did you get it as soon as you needed? | During this visit, did this health professional listen carefully to you? | During this visit, were clerks and receptionists at this health professional's office as helpful as you thought they should be? | Have you ever missed or re-scheduled your appointment because you were unable to pay the nominal fee? | Would you refer other family members and friends to CCI for care? |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | NaN | NaN | NaN | NaN | NaN | Response | Response | Response | Response | White | ... | Response | Response | Response | Other (please specify) | Response | Response | Response | Response | Response |
| 1 | 1.918263e+10 | 252301175.0 | 8/21/2024 0:30 | 8/21/2024 0:32 | es_US | Female | Female | NaN | NaN | NaN | ... | NaN | Greenbelt | Ajmera, Dr. | NaN | Yes | Yes | Yes | NaN | NaN |
| 2 | 1.918258e+10 | 252301175.0 | 8/20/2024 17:39 | 8/20/2024 17:43 | en | Male | Male | Straight/Heterosexual | No | NaN | ... | Non-Hispanic/Latino | Takoma Park | Other (please specify) | Cecilia Carrington | Yes | Yes | Yes | No | Yes |
| 3 | 1.918257e+10 | 252301175.0 | 8/20/2024 16:58 | 8/20/2024 17:03 | es_US | Female | Female | Straight/Heterosexual | No | NaN | ... | Hispanic or Latino | Takoma Park | Other (please specify) | Ginecóloga | Yes | Yes | Yes | No | Yes |
| 4 | 1.918256e+10 | 252301175.0 | 8/20/2024 16:31 | 8/20/2024 16:32 | es_US | Female | Female | Do not want to disclose | No | White | ... | Hispanic or Latino | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 24319 | 1.120152e+10 | 235506314.0 | 12/6/2019 10:10 | 12/6/2019 10:11 | en | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | Osinovskava, Renata | NaN | Yes | Yes | Yes | NaN | NaN |
| 24320 | 1.119934e+10 | 235506314.0 | 12/5/2019 14:45 | 12/5/2019 14:46 | en | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | Santhirasegaram, Dr. | NaN | Yes | Yes | Yes | NaN | NaN |
| 24321 | 1.119930e+10 | 235506314.0 | 12/5/2019 14:29 | 12/5/2019 14:30 | en | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | Santhirasegaram, Dr. | NaN | Yes | Yes | Yes | NaN | NaN |
| 24322 | 1.119915e+10 | 235506314.0 | 12/5/2019 13:39 | 12/5/2019 14:19 | en | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | Santhirasegaram, Dr. | NaN | Yes | Yes | Yes | NaN | NaN |
| 24323 | 1.119534e+10 | 235506314.0 | 12/4/2019 10:46 | 12/4/2019 10:47 | en | NaN | NaN | NaN | NaN | NaN | ... | NaN | NaN | Burke, Lorrie-Anne | NaN | Yes | Yes | Yes | NaN | NaN |

24324 rows × 27 columns

## Anexo 2. Distribución de encuestas




## Anexo 3. Resumen Descriptivo de la Longitud de Comentarios y Tabla de Percentiles

Resumen Descriptivo de la Longitud de Comentarios:

| | Estadística | Inglés | Español |
|---|---|---|---|
| 0 | count | 4521.000000 | 3532.000000 |
| 1 | mean | 20.085379 | 13.004247 |
| 2 | std | 38.325599 | 19.560847 |
| 3 | min | 1.000000 | 1.000000 |
| 4 | 25% | 5.000000 | 3.000000 |
| 5 | 50% | 11.000000 | 7.000000 |
| 6 | 75% | 23.000000 | 15.000000 |
| 7 | max | 925.000000 | 311.000000 |

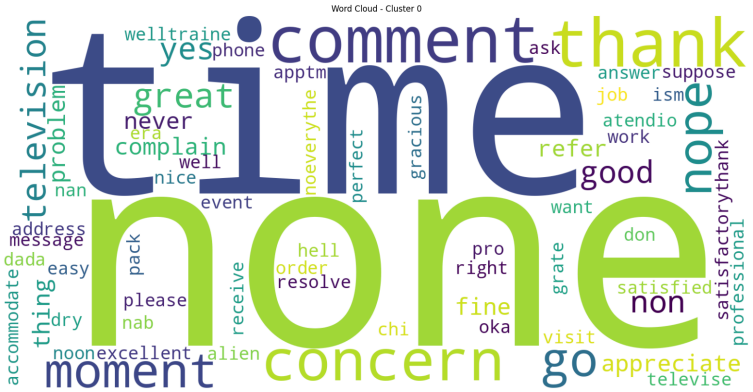
Tabla de Percentiles:

| | Percentil | Inglés | Español |
|---|---|---|---|
| 0 | 25% | 5.0 | 3.0 |
| 1 | 50% | 11.0 | 7.0 |
| 2 | 75% | 23.0 | 15.0 |

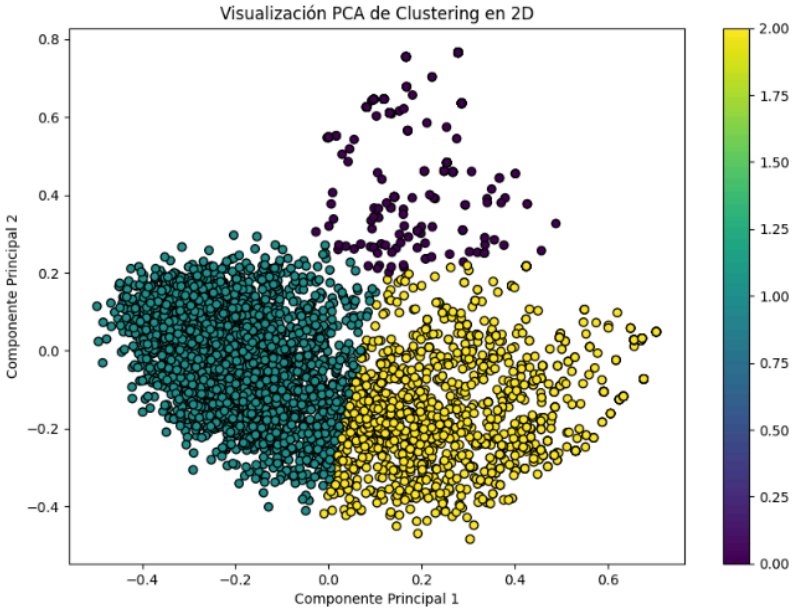
## Anexo 4. Nube de palabras: comentarios en español



## Anexo 5. Nube de palabras: comentarios en inglés

## Anexo 6: Clusters de comentarios en inglés




## Anexo 7: Clusters comentarios en Español