

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP



Desarrollo de modelos de procesamiento de lenguaje natural mediante técnicas de aprendizaje automático para el análisis de sentimientos en las PQRS de clientes industriales de Vanti S.A. ESP

Juan Sebastián Lozano Forero

Universidad Ean

Facultad de ingeniería

Maestría en ciencia de datos

Ciudad, Colombia

27/mayo/2026

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Juan Sebastián Lozano Forero

Trabajo de grado presentado como requisito para optar al título de:

Magister en ciencias de datos

Director:

Luis Armando Cobo Campo

Modalidad:

Trabajo Dirigido

Universidad Ean

Facultad de ingeniería

Maestría en ciencias de datos

Ciudad, Colombia

27/mayo/2026

**Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de
sentimientos de PQRS de clientes industriales de Vanti S.A. ESP**

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, 27/mayo/2026

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Frase:

Ora como si todo dependiera de Dios,
trabaja como si todo dependiera de ti.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Agradecimientos

A Dios, por darme la vida, a mis padres, por su amor incondicional. A mi esposa, por su comprensión, compañía y respaldo incondicional. A mi hijo, por convertirse en mi mayor inspiración y en la razón para seguir creciendo y siendo mejor cada día.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Resumen

El presente trabajo de grado aborda el análisis de sentimientos en las PQRS de clientes industriales de Vanti S.A. E.S.P., ante la necesidad de transformar información textual no estructurada en insumos útiles para la toma de decisiones. Como punto de partida, se identifica que la gestión de estas comunicaciones se realiza de forma manual, lo que dificulta su análisis oportuno y limita la generación de valor a partir de los datos.

El objetivo del estudio fue desarrollar un modelo de procesamiento de lenguaje natural capaz de clasificar los textos en categorías de sentimiento (positivo, negativo y neutro). Para ello, se implementó una metodología cuantitativa que incluyó exploración de datos, limpieza y preprocesamiento, construcción de la variable objetivo mediante un enfoque híbrido (VADER, diccionarios y reglas heurísticas), vectorización TF-IDF y entrenamiento de modelos de machine learning.

Los resultados evidenciaron que el modelo Linear SVM presentó el mejor desempeño, destacándose en métricas como F1-score macro y estabilidad en validación cruzada. En conclusión, el modelo desarrollado constituye una herramienta viable para la clasificación de PQRS, aportando valor en la identificación de casos críticos y en la generación de información estratégica para la organización.

Palabras clave: Machine learning, análisis de sentimientos, PQRS, servicio al cliente, decisiones estratégicas.

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Abstract

This master's thesis addresses sentiment analysis in PQRS (Petitions, Complaints, Claims, and Suggestions) submitted by industrial clients of Vanti S.A. E.S.P., in response to the need to transform unstructured textual information into useful inputs for decision-making. As a starting point, it was identified that the management of these communications is carried out manually, which hinders timely analysis and limits the generation of value from the data.

The objective of the study was to develop a natural language processing model capable of classifying texts into sentiment categories (positive, negative, and neutral). To achieve this, a quantitative methodology was implemented, including data exploration, cleaning and preprocessing, construction of the target variable through a hybrid approach (VADER, dictionaries, and heuristic rules), TF-IDF vectorization, and machine learning model training.

The results showed that the Linear SVM model achieved the best performance, standing out in metrics such as macro F1-score and cross-validation stability. In conclusion, the developed model represents a viable tool for PQRS classification, providing value in the identification of critical cases and in the generation of strategic information for the organization.

Keywords: Machine learning, sentiment analysis, PQRS, customer service, strategic decision-making

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Contenido

	Pág.
Introducción	16
Objetivos	20
<i>Objetivo general.....</i>	<i>20</i>
<i>Objetivos específicos</i>	<i>20</i>
Justificación	21
Marco Institucional.....	27
Marco de Referencia.....	33
<i>Análisis bibliométrico</i>	<i>33</i>
<i>Fundamentación teórica</i>	<i>38</i>
<i>Métricas de evaluación de modelos.....</i>	<i>44</i>
Diseño Metodológico.....	46
<i>Población y muestra.....</i>	<i>47</i>
<i>Análisis interno y externo</i>	<i>49</i>
<i>Fases metodológicas</i>	<i>51</i>
<i>Herramientas utilizadas</i>	<i>55</i>

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

<i>Consideraciones éticas</i>	<i>56</i>
Diagnóstico Organizacional.....	58
<i>Resultados del diagnóstico interno</i>	<i>67</i>
<i>Análisis DOFA</i>	<i>68</i>
<i>Aporte del diagnóstico organizacional.....</i>	<i>69</i>
Plan de Intervención.....	71
<i>Procesamiento estadístico de datos.....</i>	<i>71</i>
<i>Selección del modelo.....</i>	<i>97</i>
<i>Implementación del modelo.....</i>	<i>107</i>
<i>Operación del modelo</i>	<i>112</i>
<i>Plan de intervención.....</i>	<i>113</i>
<i>Cronograma de implementación</i>	<i>115</i>
<i>Recursos requeridos</i>	<i>118</i>
<i>Parámetros para evaluar el éxito del modelo en producción.....</i>	<i>119</i>
<i>Análisis de Riesgos y medidas de mitigación</i>	<i>121</i>
<i>Análisis DOFA de la solución propuesta</i>	<i>124</i>
<i>Roles, responsabilidades y gobernanza del sistema</i>	<i>126</i>
Conclusiones y Recomendaciones.....	129

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

<i>Conclusiones.....</i>	<i>129</i>
<i>Recomendaciones</i>	<i>132</i>
Referencias	134
A. Anexo. Encuesta sobre manejo de PQRS	139

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Lista de figuras

Figura 1. Variación mensual en el número de contactos	22
Figura 2. Porcentaje de participación por tipo de PQRS (2022-2025).....	23
Figura 3. Porcentaje de participación mensual por tipo de contactabilidad (2022-2025).....	25
Figura 4. Cobertura de Vanti en el territorio nacional	29
Figura 5. Cantidad de documentos sobre PNL y ML	34
Figura 6. Distribución de los tipos de documentos publicados sobre PNL y ML	35
Figura 7. Red de concurrencia de palabras clave en investigación sobre PNL y ML	36
Figura 8. Canales de recepción de comunicaciones de clientes industriales	59
Figura 9. Existencia de lineamientos para el registro de información de clientes	60
Figura 10. Procedimiento de registro de comunicaciones telefónicas	61
Figura 11. Tratamiento de comunicaciones digitales de clientes	62
Figura 12. Uso de análisis del contenido y tono emocional en las PQRS	57
Figura 13. Percepción sobre la utilidad de sistemas automáticos de análisis de sentimientos	64
Figura 14. Percepción sobre la disponibilidad del procesamiento de lenguaje natural en la atención al cliente	65
Figura 15. Principales oportunidades de mejora en proceso de atención de PQRS	66

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Figura 16. Estructura y características del conjunto de datos	73
Figura 17. Análisis descriptivo de la variable objetivo	74
Figura 18. Estadística descriptiva de la longitud de texto	75
Figura 19. Distribución de longitud de texto (caracteres)	76
Figura 20. Detección de valores atípicos en longitud de texto	77
Figura 21. Top palabras más frecuentes (sin limpiar)	78
Figura 22. Ejemplos de registros con baja longitud y contenido no significativo	79
Figura 23. Comparación de longitud de texto (tokens)	80
Figura 24. Reducción promedio de tokens	81
Figura 25. Top palabras después de limpieza (dominio)	82
Figura 26. Nube de palabras del texto original (antes del proceso de limpieza)	83
Figura 27. Nube de palabras del texto procesado (después del proceso de limpieza)	83
Figura 28. Comparación de estadísticas de tokens antes y después del proceso de limpieza	84
Figura 29. Reducción de tokens por registro tras el proceso de limpieza del texto	85
Figura 30. Bigramas más frecuentes en el conjunto de datos	87
Figura 31. Trigramas más frecuentes en el conjunto de datos	88
Figura 32. Diccionario de polaridad de términos para la clasificación de sentimientos en PQRS	90
Figura 33. Implementación del modelo VADER	91

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Figura 34. Algoritmo de etiquetado basado en reglas léxicas y modelo VADER.....	92
Figura 35. Distribución de clases de sentimiento y ejemplos del etiquetado.....	92
Figura 36. Preparación del conjunto de datos para el entrenamiento del modelo.....	94
Figura 37. Representación de distribución de textos en espacio semántico (t-SNE).....	95
Figura 38. Representación de distribución de textos por categoría de sentimiento (t-SNE).....	96
Figura 39. Comparación del desempeño de modelos.....	98
Figura 40. Comparación del desempeño de modelos mediante F1-score macro.....	99
Figura 41. Matriz de confusión normalizada del modelo Linear SVM.....	101
Figura 42. Comparación entre distribución real y predicha de las clases de sentimiento.....	102
Figura 43. Términos más relevantes asociados a la clase negativa según el modelo SVM.....	103
Figura 44. Términos más relevantes asociados a la clase neutro según el modelo SVM.....	104
Figura 45. Términos más relevantes asociados a la clase positiva según el modelo SVM.....	105
Figura 46. Curvas ROC y valores AUC del modelo SVM para clasificación de sentimientos.....	106
Figura 47. Estructura de código para predicción de sentimientos mediante SVM y TF-IDF.....	107
Figura 48. Construcción del resultado e interpretación de la predicción del modelo.....	108
Figura 49. Ejemplo de clasificación de PQRS con sentimiento negativo.....	109
Figura 50. Ejemplo de clasificación de PQRS con sentimiento positivo.....	110
Figura 51. Ejemplo de clasificación de PQRS con sentimiento neutro.....	111

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Lista de Tablas

Tabla 1. Ficha técnica población participante en encuesta.....	48
Tabla 2. Análisis DOFA.....	68
Tabla 3. Plan de intervención.....	114
Tabla 4. Cronograma de actividades.....	116
Tabla 5. Recursos requeridos.....	118
Tabla 6. Parámetros de evaluación para el seguimiento del modelo.....	120
Tabla 7. Matriz de riesgos.....	122
Tabla 8. Análisis DOFA.....	124
Tabla 9. Definición de roles.....	127

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Introducción

Debido al contexto social, político y económico por el cual pasa Colombia actualmente, las empresas en el sector de servicios públicos se ven expuestas a un número creciente de desafíos relacionados con la gestión de clientes. Estas empresas están reguladas por normas que exigen responder con oportunidad a las peticiones, quejas, reclamos y sugerencias (PQRS) de los usuarios (Congreso de la República, 2015). En este entorno, la calidad del servicio al cliente ya no se define sólo por el cumplimiento técnico del suministro, sino también por la capacidad institucional para procesar la retroalimentación ciudadana, detectar insatisfacciones y reaccionar proactivamente.

Vanti S.A. ESP es una de estas empresas, en este momento es referente y líder en el sector energético de Colombia, se dedica a la distribución y comercialización de gas natural, tiene una cobertura que se extiende a 105 municipios del país, sus redes se encuentran concentradas en Bogotá, Cundinamarca, Boyacá, Santander y una parte del departamento del Cesar. Al prestar un servicio público esencial, está obligada a atender los derechos de los usuarios, pero también a gestionar la percepción de sus clientes para evitar pérdidas reputacionales y mejorar la eficiencia interna. Lo anterior como parte de una estrategia interna basada en la importancia de la relación con los clientes como un factor estratégico clave, en el que la principal fuente de información se recibe en las PQRS ya que de ellas se obtienen datos sobre la calidad percibida del servicio, las deficiencias operativas y las expectativas no atendidas.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

La empresa al atender en promedio dos mil de PQRS mensuales se enfrenta principalmente a tres retos, el primero es la cantidad elevada de mensajes y solicitudes, el segundo la heterogeneidad del lenguaje (no todas las personas se expresan ni se hacen entender de igual manera) y por último la subjetividad inherente al análisis manual que depende del criterio del analista. En consecuencia, gran parte del esfuerzo se centra en el análisis cuantitativo, por ejemplo, al intentar resolver preguntas sobre el número de incidencias que se tienen con los clientes o el porcentaje de cuantas de ellas proceden y cuales no, entre otras. Otro reto importante se presenta con el análisis cualitativo, en el cual la revisión de los comentarios se genera de manera manual, lo que representa una carga operativa para los analistas encargados de la gestión de PQRS. En estos momentos, la lectura, interpretación y clasificación de una muestra representativa de solicitudes requiere en promedio, dos horas de trabajo por día, lo que dificulta realizar análisis oportunos y limita la capacidad de respuesta frente a situaciones críticas que podrían afectar la experiencia del cliente.

La problemática que se trabajará durante el desarrollo del presente proyecto de intervención empresarial surge a raíz de la sobrecarga de información en los sistemas de gestión de PQRS de Vanti, lo cual, estratégicamente hablando, genera una pérdida de oportunidades de competitividad dada la relación directa que tiene la satisfacción del cliente con el cierre y prolongación en el tiempo de contratos de suministro, distribución y comercialización de gas, especialmente para clientes industriales.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Para atender retos como los mencionados anteriormente, en un contexto de desarrollo informático lleno de avances, ha sido posible generar herramientas de clasificación por medio de modelos de aprendizaje automático (machine Learning), con estos ya es posible hacer análisis semánticos y reconocimiento de sentimientos en grandes volúmenes de texto no estructurado. Estas técnicas han sido aplicadas en sectores como redes sociales, atención al cliente, salud e instituciones públicas (Uribe, 2023).

Actualmente en Vanti el manejo del análisis cualitativo de las PQRS de clientes industriales depende en gran medida de la revisión manual por parte de analistas, con lo cual en ocasiones se presenta baja escalabilidad, limitando la capacidad estratégica para dirigir mejoras de servicio. Lo anterior deriva la siguiente pregunta de investigación ¿Por medio de la aplicación de técnicas de procesamiento de lenguaje natural se puede hacer un sistema de análisis de sentimientos aplicable a la base de PQRS recibida en la vicepresidencia Grandes Clientes? lo anterior con el objetivo de apoyar la gestión y mejora del servicio al cliente con esto el proceso de análisis pasaría de dos horas a realizarse inmediatamente al recibir las PQRS.

Ahora bien, es importante hacer la claridad que la implementación de modelos de análisis de sentimientos, para este proyecto se plantea como un complemento al trabajo humano, orientado a mejorar la eficiencia y capacidad de respuesta del área, su intención no es reemplazar el criterio ni la experiencia de los analistas encargados de la gestión de PQRS. Por el contrario, busca funcionar como una herramienta de apoyo para facilitar la

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

priorización de casos y fortalecer la toma de decisiones a partir de los datos generados por los clientes.

Para abordar la pregunta de investigación planteada, el presente documento está diseñado con una serie de capítulos, los cuales inician con un marco institucional que incluye la descripción de la empresa en la que se desarrolla el proyecto, seguido del marco de referencia en el cual se exponen los fundamentos teóricos relacionados con el procesamiento de lenguaje natural, aprendizaje automático, análisis de sentimientos, así como casos de aplicación en gestión de PQRS o atención al cliente. Adicional, se trabaja un diseño metodológico el cual describe el enfoque de la investigación y las fases que permiten alcanzar los objetivos planteados, metodología de la investigación.

Posteriormente se abordará en el diagnostico organizacional en el que se presenta el análisis exploratorio y estadístico de los datos, incluyendo la caracterización del conjunto de información, la calidad de los datos y los principales hallazgos relevantes para el problema. En consecuencia, se continua con el plan de intervención, en el que se desarrolla la construcción y evaluación de modelos de clasificación de sentimientos, la selección del modelo óptimo y su implementación. Finalmente se presentan las principales conclusiones del estudio y recomendaciones para su implementación y futuras líneas de investigación.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Objetivos

Objetivo general

Desarrollar y evaluar un modelo de Procesamiento de Lenguaje Natural aplicado a la base de PQRS de los clientes industriales de la vicepresidencia de Grandes Clientes de Vanti S.A. E.S.P., con el fin de generar análisis de sentimientos que apoyen la gestión y la toma de decisiones estratégicas.

Objetivos específicos

1. Construir un marco teórico en el área de Ciencia de Datos, con énfasis en modelos de Procesamiento de Lenguaje Natural y análisis de sentimientos aplicados a la gestión de PQRS.
2. Realizar un diagnóstico del proceso de gestión de PQRS en la vicepresidencia de grandes clientes, evaluando los procedimientos actuales de análisis de información y las oportunidades de incorporación de modelos de Procesamiento de Lenguaje Natural.
3. Diseñar, construir y evaluar un modelo de Procesamiento de Lenguaje Natural para el análisis automatizado de sentimientos en las PQRS, utilizando métricas de desempeño que permitan validar su efectividad.
4. Elaborar un plan para la implementación del modelo de análisis de sentimientos seleccionado, definiendo los lineamientos técnicos, operativos y tecnológicos necesarios para su integración en el proceso de gestión de PQRS.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Justificación

En la era digital, las organizaciones enfrentan el desafío de gestionar grandes volúmenes de datos generados a través de múltiples canales de comunicación. Particularmente, las empresas de servicios públicos como Vanti S.A. ESP reciben diariamente numerosas Peticiones, Quejas, Reclamos y Sugerencias (PQRS) de sus clientes industriales, las cuales contienen información valiosa sobre la percepción y satisfacción del cliente, lo que las convierten en una fuente invaluable de información para comprender las percepciones, identificar deficiencias en el servicio y orientar estrategias de mejora continua.

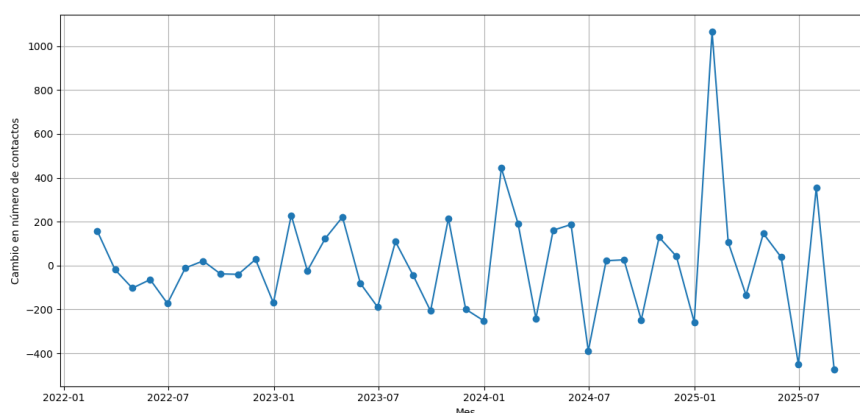
Al mirarlo bajo una óptica normativa, en Colombia, la Ley 142 de 1994 y la Ley 1755 de 2015 establecen la obligación de garantizar un servicio eficiente y transparente en la atención al usuario, reconociendo este proceso como un derecho fundamental y un mecanismo de participación ciudadana (Congreso de la República, 1994, 2015). Bajo este contexto, Vanti S.A. ESP al ser una de las empresas líderes en la distribución y comercialización de gas natural en Colombia, enfrenta el desafío de procesar y responder de manera eficiente un alto volumen de PQRS, especialmente en el segmento de clientes industriales debido a que ellos al tener procesos productivos dependientes del suministro energético, demandan respuestas rápidas, consistentes y con altos estándares de calidad.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

En la Figura 1 se presenta la evolución del número de PQRS recibidas por Vanti en el segmento industrial entre 2022 y julio de 2025. En donde se evidencia que la atención de PQRS ha adquirido una relevancia creciente en los últimos años.

Figura 1

Variación mensual en el número de contactos



Nota: Adaptada de datos internos de Vanti S.A. ESP (2025). Información confidencial.

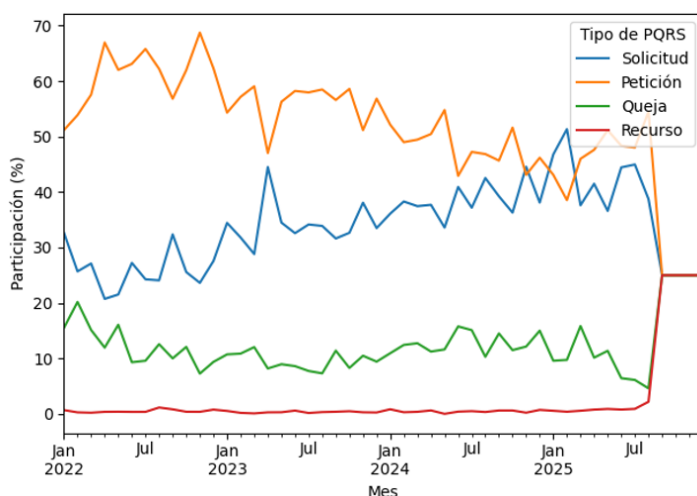
Como se observa, la tendencia muestra un incremento sostenido en el volumen de contactos, alcanzando en febrero de 2025 el mayor registro del período analizado. Este aumento refleja, por un lado, una mayor participación y exigencia de los clientes, y por otro, el desafío de la compañía para gestionar de manera eficiente grandes volúmenes de información textual. La gestión manual, aunque necesaria, resulta insuficiente para analizar de manera ágil y objetiva los sentimientos y percepciones expresadas en estos comentarios.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Ahora bien, teniendo en cuenta que se está presentando este incremento, es importante entender el porcentaje de participación por tipo de PQRS, en la figura 2 se presenta la evolución mensual en un periodo comprendido entre enero del 2022 y agosto de 2025.

Figura 2

Porcentaje de participación por tipo de PQRS (2022-2025)



Nota: Adaptada de datos internos de Vanti S.A. ESP (2025). Información confidencial.

Como se observa en la figura 2 las peticiones fueron el tipo de contacto predominante durante gran parte del periodo analizado, con valores entre el 50% y el 65%. Sin embargo, a partir del 2024 se observa un aumento progresivo en las solicitudes, tanto que sobre el segundo semestre casi las igualan y a comienzos del 2025 estuvieron por encima unos meses. En contraste, las quejas mantienen una participación menor pero

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

estable, mientras que los recursos registran un crecimiento puntual al final del periodo, reflejando un número reducido de casos que escalan a instancias de revisión formal.

La relación entre peticiones y solicitudes puede reflejarse como cambio en la relación con los clientes y sus necesidades, ya que ahora no solo piden información o aclaraciones, sino que inician gestiones más concretas y operativas. Por su parte, la estabilidad de las quejas indica que, aunque persisten manifestaciones de inconformidad, la empresa ha logrado contener su crecimiento pese al aumento general del volumen de contactos. El repunte de los recursos hacia el 2025 se relaciona al seguimiento particular de algunos casos que han requerido revisión más puntual.

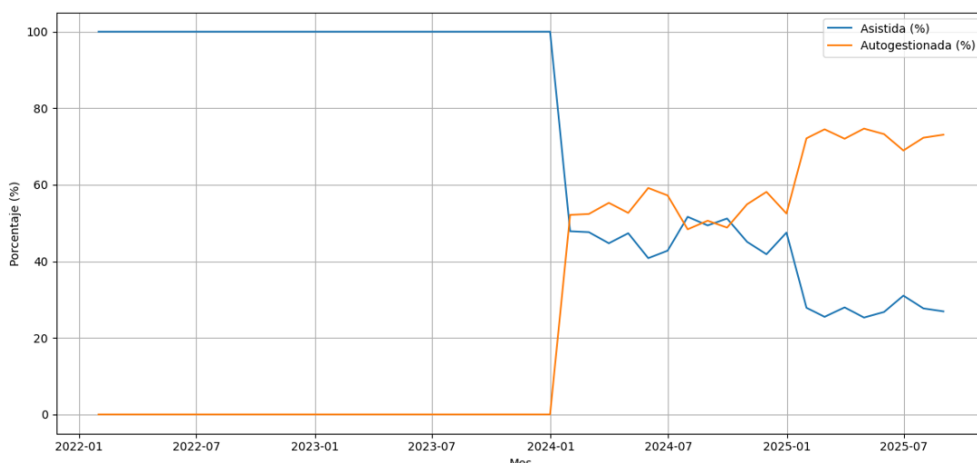
Otro factor clave de análisis corresponde a la forma en que históricamente las personas se han comunicado con la empresa, en la figura 3 se observa la evolución mensual entre enero de 2022 y agosto de 2025 en donde se evidencia el paso de una atención totalmente asistida, en la que todos los contactos con la empresa requerían la intervención directa del personal, a una autogestionada digitalmente, en la cual se dio la oportunidad a los usuarios de resolver sus solicitudes de manera independiente.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 3

Porcentaje de participación mensual por tipo de contactabilidad (2022-2025)



Nota: Adaptada de datos internos de Vanti S.A. ESP (2025). Información confidencial.

El comportamiento mostrado en la gráfica refleja una transformación digital significativa en la gestión del servicio al cliente en la vicepresidencia de Grandes Clientes de Vanti, orientada hacia la eficiencia y autonomía del usuario. Mientras que en 2022 y 2023 la atención dependía completamente de canales asistidos, en 2024 emerge un modelo orientado hacia la autogestión. Este cambio es un reflejo de los esfuerzos realizados para la adopción de nuevas plataformas digitales y como la preferencia por parte de los clientes hacia el uso de herramientas más ágiles y disponibles en cualquier momento. No obstante, el incremento de interacciones digitales también trae consigo nuevos retos como el de comprender y analizar automáticamente el sentimiento y la percepción del usuario en estos canales masivos.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Frente a este contexto, la presente investigación y plan de acción propone el uso del procesamiento de lenguaje natural mediante las técnicas de machine learning para la automatización del análisis de sentimientos en las PQRS de clientes industriales.

Entiéndase análisis de sentimientos, como una técnica del procesamiento de lenguaje natural (PLN), que permite interpretar y clasificar las emociones expresadas en textos, facilitando la comprensión de las necesidades y expectativas de los usuarios (IBM, s.f.).

Para autores como Xander Gelbuchs, la implementación de modelos de análisis en la gestión de PQRS puede mejorar significativamente la calidad del servicio al cliente, al permitir respuestas más empáticas y personalizadas (e-Contact, 2025). Además, el uso de PLN en la automatización del análisis de PQRS contribuye a una gestión más eficiente y oportuna de las solicitudes, optimizando los recursos humanos y tecnológicos de la empresa (Verhoef et al., 2021). Esto es especialmente relevante en el contexto actual, donde la rapidez y precisión en la atención al cliente son factores clave para la competitividad. Por lo tanto, desarrollar modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS en Vanti S.A. ESP no solo responde a una necesidad organizacional, sino que también contribuye al avance del conocimiento en la aplicación de tecnologías inteligentes en la gestión empresarial.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Marco Institucional

Para comprender el contexto en el cual el presente proyecto se desarrollará es importante presentar y entender el marco institucional de la empresa en donde se va a llevar a cabo. Por lo cual, el siguiente apartado tiene como propósito ofrecer una visión estructurada de la empresa Vanti S.A. ESP, destacando su enfoque estratégico, su misión y visión organizacional, así como los principales aspectos operativos que la caracterizan en el sector energético. Asimismo, se expondrán los segmentos de mercado a los que atiende y las líneas de negocio emergentes que actualmente desarrolla.

Posteriormente, se va a profundizar específicamente en la vicepresidencia de grandes clientes ya que es allí donde tendrá lugar la ejecución del proyecto, detallando su estructura, funciones y alcance en términos de cobertura y gestión de negocios, con el fin de evidenciar la pertinencia y aplicabilidad del proyecto.

Valores corporativos

Quienes somos

Somos líderes en el negocio de distribución y comercialización de gas natural en Colombia para los segmentos residencial, comercial, institucional, industrial y vehicular. Igualmente desarrollamos productos y servicios de valor agregado para nuestros clientes (Grupo Vanti, s.f.).

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Misión

Distribuimos y comercializamos gas natural y otros servicios que generan avance, experiencias valiosas, impulsan el desarrollo de los sectores y mejoran la calidad de vida de las comunidades en las que operamos (Grupo Vanti, s.f.).

Visión

Ser líderes en Colombia en ofrecer diversas soluciones energéticas y servicios sostenibles de fácil acceso distinguiéndonos por nuestro actuar responsable, excelencia operacional, servicio y pasión por nuestros clientes (Grupo Vanti, s.f.).

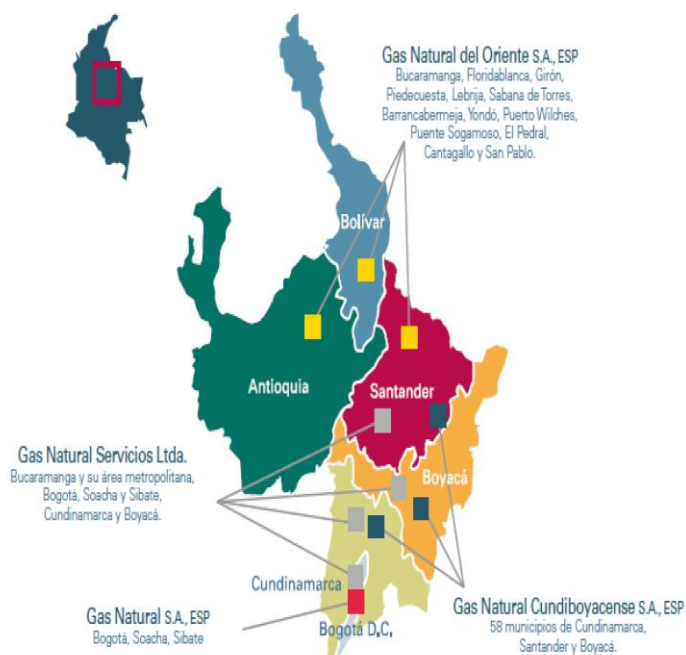
Cobertura en el país

Vanti S.A. ESP es una empresa líder en el sector energético de Colombia dedicada a la distribución y comercialización de gas natural. Está ubicada en 105 municipios del país, es importante destacar que su presencia geográfica está limitada a ciertas regiones de Colombia, debido a que sus redes de distribución de gas no están extendidas a todo el territorio nacional. Actualmente, estas redes se encuentran concentradas en Bogotá, Cundinamarca, Boyacá, Santander y una parte del departamento del Cesar. En la siguiente figura se muestra en detalle los lugares de cobertura de Vanti.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Figura 4

Cobertura de Vanti en el territorio nacional



Nota: Adaptada de datos internos de Vanti S.A. ESP (2025). Información confidencial.

En la figura se observan los departamentos y los municipios en detalle en donde tiene presencia en el territorio nacional. La empresa está organizada en filiales según la distribución de municipios en donde opera, como muestra la imagen, esta Gas Natural Sa esp que trabaja en Bogotá y sus alrededores; Gas natural cundiboyacense que está ubicada en 58 municipios de Cundinamarca y Boyacá; Gas Oriente, que está ubicada en Santander y atiende ciudades como Bucaramanga y Barrancabermeja; Gas Nacer, la cual está ubicada en el departamento del Cesar (Vanti S.A. E.S.P., 2022).

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Estructura Organizacional

Dentro de su estructura organizacional, Vanti se divide en varias vicepresidencias estratégicas que gestionan las diferentes áreas del negocio, entre ellas: Hogar y Comercio, Grandes Clientes, Estrategia y Finanzas, Tecnología, Asuntos Legales y la Vicepresidencia Técnica. Para esta ocasión, el enfoque estará en la Vicepresidencia de Grandes Clientes, ya que es precisamente en esta área donde se desarrollará el proyecto que busca analizar el impacto y las oportunidades en el servicio a los clientes industriales que consumen grandes cantidades de gas para realizar sus operaciones en sectores como construcción, manufactura, vidrio, alimentos y bebidas, entre otras (Vanti S.A. E.S.P., 2022).

La vicepresidencia grandes clientes está conformada por tres principales direcciones, las cuales son: industrias, movilidad y soluciones energéticas. Y dos de apoyo que son servicio y experiencia y Cas. A continuación, se muestra una caracterización de los clientes que atiende cada unidad de negocio.

Caracterización de clientes

Mercado grandes comercios e instituciones

Son clientes de carácter público o privado de la mediana y gran empresa en el sector terciario de nuestra economía, que tienen varios puntos de suministro o cuentas de servicio con Vanti. Sus consumos de gas natural no superan los 74.500 m³ / mes (100 KPCD). Entre las actividades económicas encontramos entidades públicas y oficiales,

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

clínicas, colegios públicos y privados, universidades, hoteles, plazas de mercado y clubes de recreación y grandes superficies (Vanti S.A. E.S.P., 2022).

Mercado clientes regulados

Son clientes de carácter privado en los sectores primario y secundario de nuestra economía, los principales sectores donde más se ubican estos clientes son agrícola, alimentos y bebidas, automotriz, plásticos, cerámica, vidrios, ladrilleras, entre otros. Estos clientes cuentan con varios puntos de suministro o cuentas de servicio con Vanti sus consumos de gas natural 3.000 a 74.500 m³ / mes (100.000 KPCD) (Vanti S.A. E.S.P., 2022).

Mercado clientes no regulados

Son clientes de carácter público o privado en los sectores primario y secundario de nuestra economía, que tienen un régimen tarifario especial y negociado por volumen de consumo con Vanti. Su actividad comercial o manufacturera tiene influencia a nivel nacional o de ciudad y sus consumos de gas natural son superiores a los 74.500 m³ / mes (100 KPCD). Entre los sectores que más clientes concentran son petróleos, químicos, ladrilleras, construcción, automotriz, manufactura y alimentos y bebidas (Vanti S.A. E.S.P., 2022).

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Mercado de movilidad

Son estaciones de servicio que prestan el servicio de tanqueo con gas natural vehicular a automóviles o camiones. También se trabaja con la conversión a gas de vehículos de gasolina y la negociación con marcas de camiones para la fabricación de vehículos dedicados al gas natural.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Marco de Referencia

Análisis bibliométrico

Para comprender de manera más profunda la evolución del conocimiento y las tendencias de investigación asociadas al machine learning y al procesamiento de lenguaje natural, se realizó un análisis bibliométrico. Con él se buscó examinar de forma cuantitativa, la producción científica y/o académica existente y evolución temporal de las publicaciones relevantes.

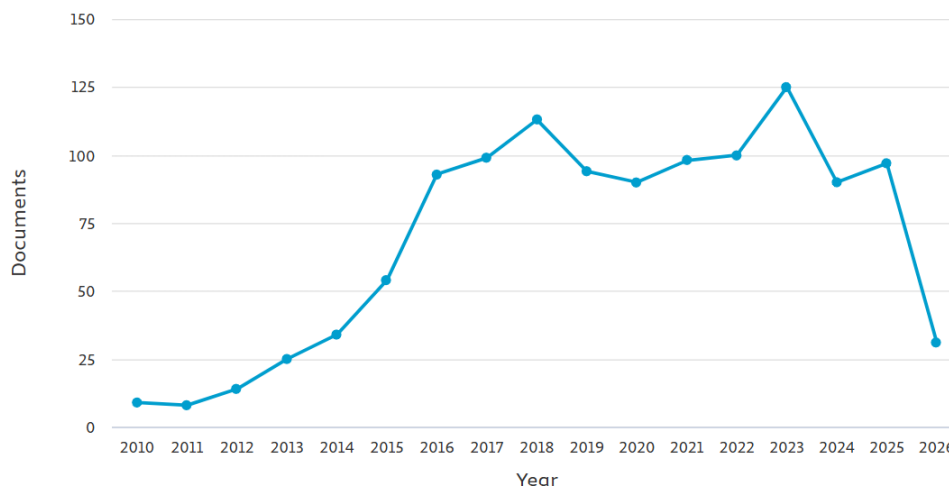
Para este ejercicio, se utilizó como fuente de investigación la base de datos de Scopus, en ella, la cadena de búsqueda se realizó con las palabras clave en inglés ya que en esta biblioteca la mayoría de producción científica fuerte está indexada en este idioma, estas palabras utilizadas fueron “Natural Language Processing” o “Sentiment Analysis” o “Opinion Mining”, publicadas entre el 2009 y el 2027. En cuanto al tipo de documento no se discriminó en la búsqueda, es decir, se buscó en artículos, conferencias, revistas y libros, entre otros.

En el análisis de la extracción de la data, lo primero que se observa es el desarrollo histórico que han tenido las publicaciones sobre estos temas, en la figura 5 se muestra este comportamiento desde el 2019 al 2026.

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Figura 5

Cantidad de documentos sobre PNL y ML por año



Nota. Elaboración propia a partir de datos recuperados de Scopus (2026).

La figura 5 muestra un crecimiento exponencial a partir del 2011 en el desarrollo de documentos y publicaciones sobre el procesamiento de lenguaje natural y machine learning, este aumento podría estar asociado al desarrollo de nuevas metodologías, avances tecnológicos o a un interés creciente de la comunidad científica por el tema. El crecimiento alcanza el punto máximo en el 2023 con cerca de 125 documentos.

Posteriormente, se identifica una tendencia a la baja en el número de documentos publicados en el 2024 el cual alcanza cifras cercanas a las manejadas en el 2019 y 2020, esta disminución puede interpretarse como una señal de saturación del campo, cambios en las prioridades de investigación, o como el efecto de proyecciones incompletas para años futuros. En conclusión, el análisis temporal revela una dinámica maduración de un campo

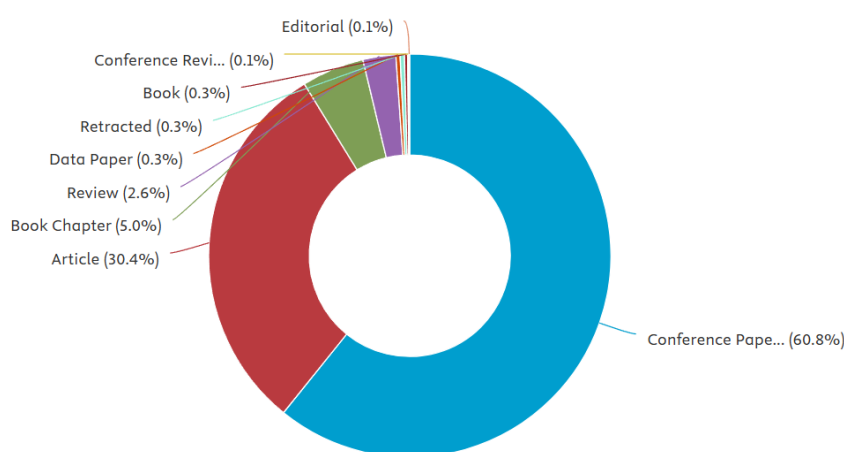
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

de investigación que pasó de ser incipiente a experimentar un notable auge, para luego mostrar una moderación en su crecimiento.

Ahora bien, es importante entender cuáles son los tipos documentos que a lo largo del transcurrir temporal que ya se analizó previamente, han tenido más concurrencia de publicación, en la figura 6 se muestra una grafica de torta en donde se evidencia este comportamiento.

Figura 6

Distribución de los tipos de documentos publicados sobre PLN y ML



Nota. Elaboración propia a partir de datos recuperados de Scopus (2026).

En la figura 6 se observa que la mayor parte de la producción científica corresponde a conferencias, los cuales representan el 61% del total de publicaciones. Este comportamiento es un reflejo del desarrollo histórico en donde en donde gran parte de los avances y desarrollos se difunden inicialmente en congresos y eventos científicos

A continuación, la figura 7, realizada con la herramienta VOSviewer, muestra una red de coocurrencia de palabras clave en investigaciones sobre procesamiento de lenguaje natural, análisis de sentimientos y minería de opiniones, a partir del análisis bibliométrico realizado en Scopus.

[illegible]

Nota. Figura generada en VOSviewer con datos exportados de Scopus (2026).

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

En la figura 7 se observa que los términos con mayor relevancia y frecuencia dentro de la red corresponden a “sentiment analysis”, “natural language processing”, “opinion mining” y “data mining”, los cuales se ubican en la zona central del mapa. La proximidad entre los nodos y la cantidad de conexiones refleja una fuerte relación entre el análisis de sentimientos y las técnicas de procesamiento de texto. En la imagen también se puede observar la concentración de clústeres, se evidencia uno relacionado con temáticas de machine learning, en este aparecen términos como “support vector machines”, “decision trees”, “naive bayes” y “logistic regression”. Otro clúster se orienta hacia enfoques más recientes asociados con “deep learning”, “neural networks”, “transformer models” y “embeddings”.

También se identifican aplicaciones frecuentes del análisis de sentimientos en contextos como redes sociales, experiencia del cliente, minería de opiniones y análisis de productos y servicios. Esto demuestra que la investigación en procesamiento de lenguaje natural posee un carácter interdisciplinario y una amplia aplicabilidad en escenarios organizacionales y de análisis de comportamiento.

En conjunto, este análisis bibliométrico permitió identificar que el análisis de sentimientos y el procesamiento de lenguaje natural representan líneas de investigación con crecimiento sostenido en los últimos años, especialmente en contextos relacionados con atención al cliente y análisis de texto. Asimismo, permitió reconocer las técnicas y

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

modelos más utilizados en la literatura, lo cual sirvió como apoyo para la selección metodológica implementada en el presente proyecto.

Luego de los hallazgos encontrados en el análisis bibliográfico, se procederá a exponer la fundamentación teórica que respalda este trabajo. En esta sección se presentarán los conceptos clave, modelos y teorías que explican y sostienen la investigación, permitiendo enmarcar el proyecto dentro de un cuerpo teórico sólido y pertinente.

Fundamentación teórica

La comunidad científica en los últimos años ha generado un sistema de reglas de procesamiento de algoritmos con el fin de ayudar a los computadores a comprender y manipular el lenguaje humano para cerrar la brecha entre la comunicación humana y el entendimiento de estas máquinas (Pineda, 2021), uno de estos esfuerzos se centra en la clasificación de textos, ubicándolos como positivos, negativos o neutros para tratar de entender el grado de favorabilidad con el cual las personas buscan calificar cualquier tema que estén tratando, a esto se le conoce como “análisis de sentimientos”. Este hace parte de una rama de la inteligencia artificial que propende por la interacción lingüística entre los humanos y las computadoras mediante procesos que permiten interpretar, procesar y generar lenguaje como si se tratara de una persona, esta rama se conoce como Procesamiento del lenguaje natural (Stryker & Holdsworth, 2024).

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Representación textual y transformación de datos

El objetivo de la investigación del lenguaje natural es crear una máquina que sea capaz de dominar el lenguaje escrito y hablado, sin embargo, al ser un objetivo demasiado amplio, los esfuerzos de los investigadores se centran en tareas más específicas como la clasificación de textos, la traducción, la creación de resúmenes, las respuestas a preguntas, entre otras. (Géron, 2023).

Es necesario resaltar que para lograr el cumplimiento de estas tareas específicas el PLN se estructura en etapas complementarias. En primer lugar, se requiere una representación adecuada del texto, ya que los algoritmos o modelos de aprendizaje automático no pueden procesar directamente datos en lenguaje natural, estos requieren de técnicas con las cuales se generan las transformaciones de palabras para permitir capturar relaciones semánticas.

Una de las principales técnicas de vectorización para transformar texto en representaciones numéricas que puedan ser interpretadas por modelos de machine learning es la TF-IDF (Term Frequency – Inverse Document Frequency). Su funcionamiento se basa en la asignación de peso a cada término según su frecuencia dentro de un documento y su nivel de relevancia respecto al conjunto total de documentos analizados. De esta manera, las palabras que aparecen con frecuencia en un texto específico, pero que no son comunes en todos los documentos, reciben una mayor ponderación, permitiendo identificar términos con mayor capacidad discriminativa (Ramos, 2003).

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Redes neuronales aplicadas al procesamiento de lenguaje natural

Otro de los algoritmos más representativos en esta fase previa a los modelos de clasificación son las redes neuronales, existen varios tipos de ellas, pero en general todas tratan de conjuntos de neuronas conectadas entre sí para conseguir un resultado, se asimila al comportamiento de redes que hacen parte del cerebro humano las cuales cuentan con entradas y salidas de estímulos y respuestas, para el caso de los modelos aprendizaje automático la entrada corresponde con la cantidad de características del conjunto de datos, y el número de neuronas de salida se corresponderá con las distintas clases que tiene el problema (Pineda, 2021).

Una de las redes más utilizadas son redes neuronales recurrentes (RNN), su característica principal incluye realimentación entre sus neuronas, con lo cual pueden recordar las salidas en momentos anteriores y utilizarlas como insumo para alimentar nuevas entradas, con lo cual resuelve problemas como series temporales, reconocimientos de voz, traducción de lenguajes y análisis de sentimientos, entre otros donde el elemento que se desea predecir depende de elementos anteriores los cuales mantienen una secuencia (Pineda, 2021).

Otro tipo de red neuronal importante para el PLN se trata de una arquitectura llamada LSTM (memoria a larga a corto plazo), que permite guardar información por mucho más tiempo ya que cuenta con celdas donde se almacena datos tanto a corto como a largo plazo. Las celdas LSTM al almacenar información permiten realizar operaciones

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

como leer datos o borrarlos dependiendo de la importancia y la relevancia que estos tengan (Minaee et al., 2021).

Ahora bien, es importante aclarar que las dos arquitecturas mencionadas anteriormente no fueron implementadas en el presente proyecto, sin embargo, su inclusión en el marco teórico tiene como propósito presentar referentes relevantes para contextualizar la evolución de este tipo de soluciones.

Modelos supervisados para clasificación de texto

Una vez que el texto ha sido transformado en representaciones numéricas, se emplean diversos algoritmos de aprendizaje supervisado tales como Naive Bayes, regresión logística, máquinas de soporte vectorial lineales (Linear SVM), clasificadores basados en descenso de gradiente estocástico (SGD), Random Forest y XGBoost. Con estos se es posible identificar patrones lingüísticos a partir de datos previamente etiquetados con las técnicas mencionadas previamente. La característica estos modelos es que son de clasificación lineal, probabilísticos y de ensamble, los cuales han demostrado un desempeño robusto en tareas de procesamiento de lenguaje natural (Jurafsky & Martin, 2023).

A continuación, se examinará desde un enfoque teórico estos modelos mencionados, los cuales dentro de la presente investigación se consideraron para abordar la problemática planteada.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Naïve Bayes

El primero es el clasificador Naive Bayes el cual es un modelo probabilístico que asume independencia condicional entre las características, lo cual, ha posicionado su uso ampliamente para la clasificación de texto, debido a que modela la frecuencia de ocurrencia de palabras dentro de los textos especialmente cuando se utilizan representaciones como Bag of Words o TF-IDF (Manning et al., 2008).

Regresión logística

El siguiente modelo en abordar se trata de la regresión logística, la cual estima la probabilidad de pertenencia a una clase mediante una función logística. Es ampliamente utilizada en problemas de clasificación de texto debido a su interpretabilidad y buen desempeño en espacios de alta dimensionalidad. Específicamente hablando del análisis de sentimientos, mediante la asignación de pesos a las palabras permite identificar la polaridad de los comentarios, lo cual es de gran utilidad para interpretación de modelos más complejos (James et al., 2021).

Linear support Vector Machine (Linear SVM)

El modelo que sigue es el linear SVM (máquinas de soporte vectorial) el cual se concentra en encontrar un hiperplano óptimo que maximiza el margen entre las clases. Se considera altamente efectivo en tareas de clasificación de texto debido a su capacidad de manejar datos de alta dimensionalidad (Cortes & Vapnik, 1995).

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Stochastic Gradient Descent (SGD Classifier)

El siguiente propiamente no es un modelo como tal, sino un método de optimización utilizado para entrenar modelos lineales se trata del clasificador SGD. Este trabaja minimizando una función de error ajustando los parámetros del modelo. Para el análisis de sentimientos es bastante útil ya que permite procesar grandes volúmenes de información textual, ajustando los parámetros del modelo de manera iterativa y eficiente, lo cual es clave para trabajar con conjuntos de datos de gran tamaño, como los utilizados en procesamiento de lenguaje natural (Bottou, 2010).

Random Forest

A continuación, se encuentra el Random Forest el cual combina múltiples árboles de decisión, lo que mejora la precisión y reduce el sobreajuste. Resulta útil en problemas de clasificación ya que permite capturar relaciones no lineales entre las variables (Breiman, 2001), sin embargo, vale la pena aclarar que debido a la alta dimensionalidad y naturaleza de los datos textuales dentro del lenguaje natural su desempeño en tareas de procesamiento puede verse limitado.

XGBoost

El último modelo que se va a tratar es el XGBoost, el cual es un algoritmo de boosting basado en árboles de decisión que construye modelos de manera secuencial, optimizando los errores de iteraciones anteriores. Se considera uno de los más utilizados

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

en tareas de clasificación y predicción debido a su escalabilidad y alto rendimiento (Chen & Guestrin, 2016).

Todos los modelos mencionados no solo han demostrado su eficacia en la representación semántica del lenguaje, además han facilitado avances significativos en tareas de clasificación de sentimientos, lo cual es de gran importancia para el presente proyecto.

Métricas de evaluación de modelos

Con el fin de comparar el desempeño de los diferentes modelos de machine learning implementados en la investigación, se utilizaron diversas métricas de evaluación empleadas en tareas de clasificación supervisada y procesamiento de lenguaje natural. Estas permitieron analizar no solo la capacidad predictiva de los modelos, sino su estabilidad y eficiencia computacional (Géron, 2023).

Una de las métricas utilizadas fue el Accuracy, la cual mide la proporción de predicciones correctas realizadas por el modelo sobre el total de observaciones evaluadas. Pese a que esta métrica proporciona una visión general del desempeño, diversos autores señalan que puede resultar insuficiente en escenarios con desbalance entre clases, debido a que tiende a favorecer las categorías con mayor cantidad de registros (Géron, 2023). Por esta razón, también se empleó el F1-score macro, ya que con ella se calcula el promedio del desempeño obtenido en cada clase sin considerar el tamaño. Su uso resulta especialmente importante en problemas de análisis de sentimientos, donde las categorías

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

pueden presentar distribuciones desiguales. Complementariamente, se utilizó el F1-score weighted, el cual pondera el desempeño de cada clase según la cantidad de observaciones presentes, permitiendo obtener una evaluación más ajustada al comportamiento global del conjunto de datos (Raschka et al., 2022).

Adicionalmente, se incorporaron las métricas Cross Validation Mean (CV mean) y Cross Validation Standard Deviation (CV std), estas están relacionadas con validación y permiten evaluar la estabilidad y capacidad de generalización de los modelos frente a diferentes particiones de los datos, reduciendo el riesgo de sobreajuste y proporcionando mayor confiabilidad en los resultados obtenidos (Kuhn & Silge, 2022).

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Diseño Metodológico

El presente proyecto se realizará bajo un enfoque cuantitativo, de tipo descriptivo y aplicado mediante la medición de características que derivan en un marco conceptual sobre la aplicación de modelos de lenguaje natural usando técnicas de machine learning. De este enfoque se obtienen conclusiones objetivas a partir de la medición de características observables (Bernal, 2022), lo cual busca resultados objetivos y generalizables a partir de los datos analizados, ya que al medir características específicas de las PQRS y relacionarlas con los sentimientos expresados por los clientes, se asegura una interpretación basada en evidencia práctica, reduciendo el riesgo de subjetividad. Además, el enfoque ayuda a identificar patrones sobre las opiniones de los clientes, lo que es útil para la empresa a la hora de tomar decisiones informadas con miras a mejorar sus servicios. Es importante aclarar que no se va a trabajar con una metodología experimental, la cual busca establecer relaciones de causa y efecto mediante la manipulación de variables (Monje, 2011) debido a que se va a trabajar con una base real de comentarios de clientes, no se podrán alterar variables por temas éticos y de integridad en la información proporcionada por los clientes.

Para entender cómo se va a desarrollar el presente proyecto, primero se abordará la población y muestra, posteriormente se expondrá como va a ser el diseño metodológico, luego se va a explorar las herramientas que serán utilizadas, posteriormente se analizará el instrumento complementario y finalmente se tendrán en cuenta las consideraciones éticas.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Población y muestra

En el contexto de la intervención empresarial realizada con este proyecto, metodológicamente se identifican dos tipos de población según su participación, el primero, de carácter analítico, son los registros históricos de PQRS utilizados para el desarrollo del modelo de lenguaje natural; el segundo, con un componente más diagnóstico, se trata de los colaboradores del área de servicio y experiencia al cliente que participaron en una encuesta de análisis organizacional aplicada como herramienta de validación.

La población del componente analítico está conformada por la totalidad de los registros de PQRS recibidas de los clientes industriales de la vicepresidencia de Grandes Clientes de Vanti S.A. ESP dentro del periodo comprendido entre enero del 2022 a febrero del 2026. Para la muestra se adoptó un enfoque censal, la cual es reconocida por ser una investigación exhaustiva en la que se consigue información del total de las unidades de información que componen el universo a investigar (Instituto Nacional de Estadística – Uruguay, s. f.) lo anterior aplicando criterios de calidad y completitud tras el proceso de depuración y preprocesamiento de datos.

La segunda población, del componente diagnóstico, está conformada por los colaboradores del área encargada de la recepción, gestión y análisis de las PQRS. La muestra fue de tipo no probabilístico por conveniencia, integrada por los funcionarios directamente involucrados en el proceso, quienes aceptaron participar voluntariamente en

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

la aplicación del instrumento de tipo encuesta (Ver en anexos) conformado por 10 preguntas de tipo cerrado y de percepción, estructuradas mediante escala Likert y preguntas de selección múltiple. La encuesta tuvo como objetivo identificar aspectos relacionados con el proceso actual de gestión de PQRS, tiempos de análisis, dificultades operativas, percepción frente al manejo manual de la información, y oportunidades de mejora mediante herramientas analíticas. Con el fin de caracterizar el proceso de aplicación del instrumento, a continuación, se presenta la ficha técnica correspondiente a la encuesta.

Tabla 1

Ficha técnica población participante en encuesta

Aspecto	Descripción
Población objetivo	Colaboradores del área de Servicio y Experiencia al Cliente involucrados en la recepción, gestión y análisis de PQRS
Total, de colaboradores del área	26
Número de participantes	18
Tipo de muestra	No probabilística por conveniencia
Tipo de instrumento	Encuesta
Medio de aplicación	Microsoft Forms
Fecha de aplicación	01/02/2026
Participación	Voluntaria
Objetivo del instrumento	Identificar percepciones, prácticas y necesidades relacionadas con la gestión de PQRS

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Nota. Elaboración propia.

Con el fin de garantizar la claridad y pertinencia de las preguntas frente a los objetivos planteados en la investigación, el instrumento fue sometido a un proceso de validación de contenido con el jefe del área de Servicio y Experiencia al Cliente de Vanti S.A. E.S.P., considerando su conocimiento del proceso de gestión de PQRS y su experiencia en el análisis de las solicitudes recibidas por los clientes.

Durante la validación se evaluaron aspectos relacionados con la capacidad del instrumento para identificar problemáticas, necesidades y oportunidades de mejora dentro del proceso actual. Como resultado, se realizaron ajustes en la redacción y organización de algunos ítems, buscando facilitar la interpretación de las preguntas por parte de los participantes y mejorar la calidad de la información recolectada.

Este proceso permitió fortalecer la validez del instrumento y asegurar que las preguntas estuvieran alineadas con el contexto organizacional y con los objetivos del componente diagnóstico desarrollado en la presente investigación.

Análisis interno y externo

Con el propósito de comprender el contexto organizacional en el cual se desarrolló el presente proyecto de intervención empresarial, se realizó un análisis interno y externo enfocado en el proceso de gestión de PQRS del área de Servicio y Experiencia al Cliente. Este análisis permitió identificar factores organizacionales, operativos y tecnológicos que

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

influyen directamente en la gestión de la información y en la atención de los clientes industriales.

El análisis interno se orientó a identificar las principales fortalezas y debilidades presentes dentro del proceso actual de recepción, gestión y análisis de las PQRS. Para ello, se consideraron elementos como los procedimientos operativos existentes, el manejo manual de la información, los tiempos requeridos para la revisión de solicitudes, la disponibilidad de datos históricos y el nivel de apoyo tecnológico disponible para el procesamiento de información textual. Este análisis se desarrolló a partir de la información recolectada mediante la encuesta aplicada a los colaboradores del área, así como de la revisión del proceso organizacional actualmente implementado.

Por su parte, el análisis externo permitió identificar oportunidades y amenazas provenientes del entorno tecnológico, competitivo y regulatorio que pueden impactar el proceso de gestión de PQRS. En esta etapa se consideraron aspectos como el crecimiento de herramientas basadas en inteligencia artificial y procesamiento de lenguaje natural, el aumento en las expectativas de los clientes frente a la calidad del servicio y la presión regulatoria relacionada con la atención oportuna de las solicitudes ciudadanas. Asimismo, se analizaron las tendencias actuales en automatización y analítica de datos aplicadas a procesos de experiencia del cliente.

A partir de ambos análisis fue posible consolidar un diagnóstico organizacional que permitió identificar oportunidades de mejora y justificar la pertinencia de implementar una

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

herramienta de análisis de sentimientos basada en técnicas de machine learning y procesamiento de lenguaje natural para apoyar la gestión de las PQRS dentro de la organización.

Fases metodológicas

Para abordar la problemática identificada, se definió una metodología estructurada compuesta por fases de exploración, limpieza, transformación y modelado de datos. Este enfoque permitió garantizar la calidad de la información y establecer una base sólida para construir y evaluar un modelo de análisis de sentimientos alineado con las necesidades de la organización

Fase 1 Comprensión del problema y contexto

En esta fase se analizaron las características del proceso de atención al cliente, identificando que las observaciones contenidas en las PQRS constituyen una fuente clave de información no estructurada cargada de elementos emocionales que no están siendo adecuadamente identificados. Con lo cual se pudo establecer que mediante la automatización de sentimientos en las PQRS se puede mejorar la gestión de la información y apoyar la toma de decisiones.

Fase 2 Exploración de datos

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Análisis exploratorio del conjunto de datos para comprender la naturaleza de la información y definir estrategias de preprocesamiento, esto mediante la identificación de la estructura del dataset, la observación de la calidad de los datos (valores nulos, inconsistencias), la distribución de la variable objetivo, la longitud de los textos, la frecuencia de palabras, bigramas y trigramas, y los registros con contenido no significativo.

Fase 3 Preparación y limpieza de los datos

Con base en los hallazgos del análisis exploratorio, se implementarán procesos de limpieza y transformación, como eliminación de registros vacíos o irrelevantes, normalización de texto (minúsculas, eliminación de caracteres especiales), eliminación de valores atípicos y reducción de longitud de los textos.

Fase 4 Construcción de la variable objetivo (etiquetado de sentimientos)

Esta fase se desarrolla por medio de un enfoque híbrido para el etiquetado de sentimientos. Este se compone de tres capas, en la primera se desarrolla un diccionario con comentarios reales tanto positivos como negativos que ha recibido la empresa. Luego pasa al modelo VADER el cual realiza los análisis de sentimientos, es importante aclarar que esta herramienta fue diseñada para el idioma inglés, por lo tanto, no se plantea como el etiquetador principal, por tanto, metodológicamente para este proyecto, se usa únicamente

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

como capa de desempate cuando el texto no contiene ninguna keyword del dominio, es decir, para casos ambiguos donde incluso un lexicón nativo en español probablemente tampoco daría una señal clara. Finalmente, en el tercer paso se hace un análisis de urgencia basado en criterios del negocio. Este enfoque permitió asignar etiquetas de sentimiento (positivo, negativo, neutro) a cada registro, generando una variable objetivo robusta y coherente con el contexto de las PQRS que recibe Vanti.

Fase 5 Transformación del texto a variables numéricas

Con el objetivo de preparar los datos para el entrenamiento de los modelos, se aplica la técnica TF-IDF para convertir los textos en representaciones numéricas, generando un espacio de características de alta dimensionalidad. De igual manera, se realiza la división del dataset en entrenamiento y prueba mediante muestreo estratificado.

Fase 6 Análisis de la estructura del espacio vectorial

Por medio de la técnica de reducción de dimensionalidad t-SNE, se visualiza la distribución de los textos en el espacio vectorial, lo cual permite identificar patrones y agrupaciones, analizar la superposición entre clases y comprender la complejidad del problema de clasificación.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Fase 7 Entrenamiento y evaluación de modelos

Los modelos de clasificación que se implementaron y evaluaron, fueron: Naive Bayes, Regresión Logística, Linear SVM, XGBoost, SGD Classifier, Random Forest.

Se evaluaron por medio de métricas como: Precision, Recall, F1-score y validación cruzada.

Fase 8 Selección del modelo óptimo

Para la elección del modelo óptimo, en el análisis de las métricas antes dichas, se tuvo en cuenta el mejor desempeño en F1-score, la capacidad para manejar datos de alta dimensionalidad y la estabilidad entre el entrenamiento y la prueba.

Fase 9 Análisis e interpretación del modelo

Esta fase permite interpretar el comportamiento del modelo y validar su coherencia con el contexto del negocio, se analizaron los resultados del modelo mediante la matriz de confusión, identificación de términos más relevantes por clase y comparación de distribución real vs predicha.

Fase 10 Propuesta de implementación

En esta fase se describe la construcción, evaluación y selección del modelo de clasificación, así como su implementación en un entorno práctico.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Herramientas utilizadas

Para el desarrollo del presente proyecto el lenguaje de programación escogido como base para todo el desarrollo fue Python debido a su amplio uso para ciencia de datos y a la disponibilidad de librerías especializadas en análisis de texto. El proceso de desarrollo y análisis se llevó a cabo en Google Colab y entornos locales mediante entornos interactivos como Jupyter Notebook, lo cual permitió ejecutar código de forma incremental, facilitando la exploración de datos, la validación de resultados y la construcción del modelo. Para la manipulación y transformación de los datos se empleó la librería Pandas, con la cual se gestionó el conjunto de datos, se realizó el proceso de limpieza y estructuración de la información. En el tratamiento del texto se utilizaron herramientas de procesamiento de lenguaje natural apoyadas principalmente en técnicas de limpieza, normalización y análisis léxico, complementadas con librerías como NLTK y SpaCy, según la necesidad de cada etapa del proceso.

La construcción de los modelos de clasificación se realizó mediante la librería Scikit-Learn, utilizando algoritmos como Naive Bayes, Regresión Logística, Support Vector Machine (SVM), entre otros, así como técnicas de vectorización como TF-IDF, fundamentales para la representación numérica del texto. Finalmente, para el análisis e interpretación de resultados se utilizaron librerías de visualización como Matplotlib y Seaborn, con las cuales se hicieron gráficos descriptivos, comparativos y de evaluación del modelo, facilitando la comprensión de los resultados obtenidos.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Consideraciones éticas

Para terminar la parte metodológica del proyecto, es importante destacar que en cumplimiento con la Ley 1581 de 2012 sobre protección de datos personales en Colombia (Congreso de la República de Colombia, 2012), se van a anonimizar datos personales presentes en las PQRS ya que el tratamiento de la información es estrictamente para uso académico y cuenta con la autorización de Vanti S.A. E.S.P. Con esto se adoptan medidas orientadas a la protección de la integridad y seguridad de los datos, evitando su divulgación no autorizada y garantizando un uso responsable de la información durante todas las etapas del proyecto.

Finalmente, desde la perspectiva ética y de gobernanza de inteligencia artificial, el presente proyecto reconoce que las decisiones relacionadas con la gestión de PQRS no deben delegarse completamente a sistemas automatizados debido a que los sistemas de inteligencia artificial aplicados a contextos de alto impacto social deben mantener mecanismos de supervisión humana, transparencia y control sobre las decisiones automatizadas (Organisation for Economic Co-operation and Development [OECD], 2019; ISO, 2023). En consecuencia, la propuesta plantea el uso del modelo como una herramienta de apoyo para la toma de decisiones, manteniendo supervisión humana continua sobre los resultados generados por la inteligencia artificial, lo anterior busca garantizar criterios de responsabilidad, transparencia y validación humana en aquellos casos que puedan tener impacto sobre la atención y experiencia de los clientes, en

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

concordancia con principios de IA responsable y buenas prácticas de gobernanza tecnológica.

**Desarrollo de modelos de procesamiento de lenguaje natural para el
análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP**

Diagnóstico Organizacional

Con el propósito de complementar el análisis del proceso actual de atención a PQRS en la empresa y obtener información de primera mano sobre la forma en que los colaboradores gestionan estas solicitudes, se aplicó la encuesta expuesta en el diseño metodológico, como instrumento de recolección de datos para identificar temas como la percepción del volumen y complejidad del proceso y el nivel de conocimiento y disposición frente a soluciones basadas en procesamiento de lenguaje natural.

La encuesta fue aplicada a una población clave dentro del proceso de gestión de las PQRS, conformada por colaboradores del área de servicio y experiencia al cliente, entre ellos, los responsables por recibirlas y quienes participan en las actividades administrativas relacionadas con su gestión y análisis. La encuesta fue diseñada con preguntas de selección múltiple y percepción, orientadas a identificar aspectos relacionados con los canales de recepción de las PQRS, mecanismos de registro de la información, nivel de estandarización de los procesos, uso actual de herramientas analíticas y percepción frente a la incorporación de soluciones basadas en procesamiento de lenguaje natural. La aplicación se realizó de manera interna y los resultados obtenidos fueron posteriormente consolidados y analizados mediante estadística descriptiva y representaciones gráficas, con el propósito de identificar patrones, fortalezas, limitaciones y oportunidades de mejora dentro del proceso actual de gestión de PQRS.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

A continuación, se analizará en detalle cada una de las preguntas con sus respectivas respuestas con en fin de exponer los hallazgos que se identificaron, los cuales influyeron en el diseño de la propuesta presentada en el presente proyecto.

La primera de ellas indaga sobre los canales por los cuales se recibe las comunicaciones por parte de los clientes, como se observa en la figura 8, la mayoría de las comunicaciones con los clientes se presenta por correo electrónico, atención vía telefónica y por medio de WhatsApp.

Figura 8

Canales de recepción de comunicaciones de clientes industriales

1. ¿Por cuáles canales recibe usted con mayor frecuencia las comunicaciones o PQRS de los clientes industriales? (*puede marcar más de una opción*)



Nota. Elaboración propia.

El hecho de que la mayoría de las comunicaciones provengan de correos electrónicos puede representar una ventaja para la implementación de técnicas de procesamiento de lenguaje natural ya que este medio elimina etapas como transcripción o digitalización, con lo cual la recepción se realiza de forma directa, mejorando la eficiencia

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

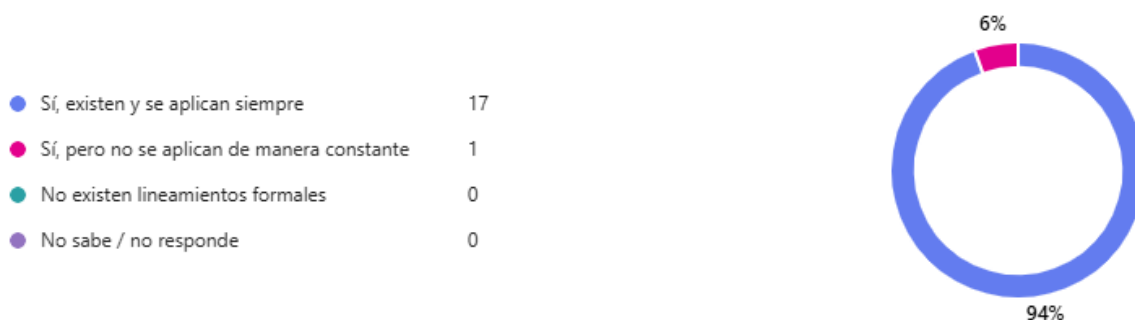
analítica. también al trabajarse con el texto original (sin transcripciones de terceros) se reduce la pérdida de información, lo cual puede reducir el sesgo por síntesis de agentes.

De manera complementaria, el segundo canal de comunicación más usado son las llamadas telefónicas en el cual, generalmente se trata de situaciones que requieren interacción inmediata. A diferencia del correo electrónico, la comunicación por este medio presenta mayores desafíos para el análisis automatizado, con lo cual se deben buscar mecanismos de transcripción estandarizada que permitan integrar las llamadas al modelo de análisis de sentimientos. A propósito, en la pregunta número 8 se indaga para saber si hay lineamientos para redactar o registrar la información de un cliente. En la figura 9 se observa los resultados de dicha pregunta.

Figura 9

Existencia de lineamientos para el registro de información de clientes

8. ¿Existen lineamientos o protocolos que indiquen cómo redactar o registrar la información del cliente?



Nota. Elaboración propia.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

En la gráfica anterior se observa que todos los encuestados tienen claro que existen lineamientos y la mayoría de ellos los aplican de manera consistente, mientras que una persona afirma que pese a que existen estos no son aplicados de manera constante. Lo anterior indica que en la vicepresidencia existe una base normativa sólida, lo cual mitiga de manera significativa el sesgo por síntesis en la transcripción que realiza el personal que atiende las llamadas.

En las preguntas 3 y 4 se indaga por el canal de recepción vía telefónica. Primero, la pregunta 3 tiene como objetivo conocer el número de personas que recibe por este medio la comunicación, contestando más de la mitad afirmativamente (el 56% contra el 44% que no recibe comunicaciones por este medio). Posteriormente la pregunta 4, para quienes, si utilizan este medio, busca entender cómo se registra lo que se expresa durante la comunicación. En la figura 9 se observan los resultados.

Figura 10

Procedimiento de registro de comunicaciones telefónicas

4. Cuando un cliente se comunica por llamada telefónica, ¿cómo se registra lo que expresa durante la conversación?



Nota. Elaboración propia

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

En la figura se observa que entre quienes atienden PQRS vía telefónica hay una distribución equilibrada entre la transcripción literal del mensaje y la síntesis del contenido en palabras propias. Lo anterior puede presentar un aspecto crítico a considerar para el entrenamiento y desempeño de los modelos de análisis de sentimientos ya que puede generar variabilidad en la calidad y riqueza del dato textual.

Continuando por la misma vía de análisis, las preguntas 5 y 6 realizan la indagación por el canal de comunicación vía correo electrónico o WhatsApp. La 5 tiene como objetivo conocer el número de personas que recibe por este medio la comunicación, contestando la mitad afirmativamente, es decir, de las dieciocho personas que participaron, nueve las reciben por este medio. Posteriormente la pregunta 6, para quienes, si utilizan este medio, indaga sobre cómo se registra lo que se expresa durante la comunicación. En la figura 11 se observan los resultados.

Figura 5

Tratamiento de comunicaciones digitales de clientes

6. Cuando recibe comunicaciones por correo electrónico o WhatsApp, ¿qué tratamiento le da al mensaje?



Nota. Elaboración propia.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Lo primero que se destaca de esta pregunta, es que ninguna persona contestó que no hay un procedimiento definido, lo cual es buena señal ya que es una confirmación de que hay políticas para el manejo de las comunicaciones con los clientes. Lo segundo es que la mayoría de las respuestas se dividen entre quienes copian y pegan textualmente lo que escribe el cliente y quienes realizan una síntesis del contenido principal. Lo anterior significa que, pese a que hay un procedimiento definido, puede que no se haga un seguimiento riguroso respecto a este criterio, lo cual puede influir en la consistencia de los datos más adelante.

La pregunta número 7 es importante ya que indaga respecto de si hay un sistema empresarial definido para el registro de las comunicaciones, dando como resultado que la mayoría de los encuestados (78%) afirman que sí, hay un aplicativo empresarial en el cual se consigna la información, lo cual es bueno ya que demuestra que hay un nivel adecuado de centralización de la información lo que facilita la integración de soluciones de analítica y tableros de control. En esta vía, el resultado de la pregunta número 8 también es positivo ya que el 94% de las personas confirman que si existen lineamientos sobre como se deben redactar o registrar la información de los clientes y sobre todo, lo más importante es que estos se aplican.

A continuación, en la figura 11, se mostrará una gráfica con el resultado de la pregunta número 9, la cual indaga si actualmente hay algún tipo de análisis de contenido o tono emocional en los mensajes de los clientes.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 12

Uso de análisis del contenido y tono emocional en las PQRS

9. ¿La empresa realiza actualmente algún tipo de análisis del contenido o tono emocional de los mensajes de los clientes?



Nota. Elaboración propia.

Como se observa en la gráfica, la mayoría de los encuestados (el 61%) confirma que si hay análisis, pero se realizan de manera manual. Por su parte una proporción menor (el 22%) reporta algún nivel de automatización. Finalmente, el 17% desconoce si este análisis se lleva a cabo. Lo anterior evidencia una oportunidad e idoneidad para el desarrollo de automatizaciones mediante el uso de modelos de procesamiento de lenguaje natural, con el fin de reducir la carga operativa para beneficio de la vicepresidencia y el área de servicio y experiencia al cliente.

Ahora bien, con la intención de conocer sobre la percepción de las personas que trabajan con el manejo y gestión de las PQRS, respecto a la importancia del desarrollo de herramientas automatizadas, a continuación, se muestran las figuras 12 y 13.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 6

Percepción sobre la utilidad de sistemas automáticos de análisis de sentimientos

10. En su experiencia, ¿qué tan útil sería contar con un sistema que identifique automáticamente el sentimiento (positivo, negativo o neutro) de las PQRS?



Nota. Elaboración propia

Figura 14

Percepción sobre la disponibilidad del procesamiento de lenguaje natural en la atención al cliente

11. ¿Considera que la incorporación de herramientas de procesamiento de lenguaje natural (PLN) podría facilitar la atención al cliente?



Nota. Elaboración propia.

En las anteriores graficas se observa que el 83% considera muy útil un sistema automático que identifique los sentimientos y el 61% está totalmente de acuerdo con la

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

incorporación de herramientas de procesamiento de lenguaje natural para facilitar la atención al cliente.

Para finalizar este análisis de los resultados de la encuesta, en la figura 14 se presenta una nube de palabras en donde se muestra los principales aspectos mencionados con relación a lo que las personas consideran que se debe mejorar en el proceso actual de atención de PQRS.

Figura 7

Principales oportunidades de mejora en proceso de atención de PQRS



Nota. Elaboración propia.

Como se observa en la gráfica, el eje central por el cual gira el discurso en torno a este tema se encuentra la palabra “cliente” lo que evidencia una orientación hacia la importancia del usuario dentro del proceso. De manera complementaria, palabras como “respuesta”, “más tecnología”, “solución” y “mayor celeridad” reflejan una inclinación hacia la oportunidad en la atención, buscando mejorar factores claves como el uso de tecnologías y la necesidad de reforzar en los tiempos de respuesta.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Resultados del diagnóstico interno

El diagnóstico interno realizado en el área de Servicio y Experiencia de la vicepresidencia de Grandes Clientes de Vanti S.A. E.S.P. permitió identificar diferentes aspectos relacionados con el proceso actual de gestión de PQRS. Entre los principales hallazgos se evidenció una alta dependencia de procesos manuales para la revisión, clasificación y análisis de las solicitudes recibidas por los clientes, situación que incrementa los tiempos operativos y dificulta el análisis oportuno de grandes volúmenes de información textual.

De igual manera, se permitió identificar que gran parte del conocimiento obtenido a partir de las PQRS depende de la interpretación individual de los analistas, lo cual puede generar diferencias en la priorización y seguimiento de los casos. Así mismo, se evidenciaron limitaciones para detectar patrones recurrentes de inconformidad o tendencias de insatisfacción de manera sistemática, especialmente cuando el volumen de solicitudes aumenta.

No obstante, el diagnóstico también permitió identificar fortalezas importantes dentro del área, como la disponibilidad de información histórica de PQRS, el conocimiento operativo de los colaboradores sobre el proceso y la disposición organizacional hacia el uso de herramientas analíticas que contribuyan al fortalecimiento de la experiencia del cliente.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

En conjunto, estos resultados evidenciaron la necesidad de incorporar mecanismos tecnológicos que permitan optimizar el análisis de información no estructurada y apoyar la toma de decisiones mediante herramientas basadas en procesamiento de lenguaje natural y machine learning.

Análisis DOFA

Con base en los hallazgos obtenidos en el análisis de la encuesta aplicada al personal del área de Servicio y Experiencia al Cliente, a continuación se presenta la siguiente matriz DOFA orientada al desarrollo e implementación de un modelo de procesamiento de lenguaje natural para el análisis de sentimientos en las PQRS.

Tabla 2

Análisis DOFA sobre la encuesta

Fortalezas	Oportunidades
Existencia de un sistema centralizado para registrar PQRS.	Implementación de modelos NLP para análisis automatizado.
Lineamientos definidos para registrar información de clientes.	Integración con herramientas analíticas como Microsoft Power BI.
Alto volumen de comunicaciones digitales en texto.	Optimización de tiempos de respuesta y priorización de casos.
Disposición positiva del personal hacia herramientas automáticas.	Aprovechamiento estratégico de información histórica de PQRS.
Existencia de políticas para manejo de comunicaciones.	Fortalecimiento de la toma de decisiones basada en datos.
Debilidades	Amenazas
Predominio del análisis manual de PQRS.	Resistencia al cambio tecnológico.
Variabilidad en la forma de registrar llamadas.	Inconsistencias en la calidad de los datos textuales.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Aplicación no homogénea de lineamientos.	Riesgos asociados al manejo de datos sensibles.
Falta de automatización en análisis de sentimientos.	Dependencia de la calidad de las transcripciones.
Subutilización de la información histórica.	Incremento de expectativas de los clientes en tiempos de respuesta.

Nota. Elaboración propia.

A partir de esta matriz DOFA, se evidencia que la organización cuenta con condiciones favorables para el desarrollo de la propuesta, especialmente por la existencia de información centralizada, lineamientos operativos y disposición positiva frente a la incorporación de herramientas tecnológicas. No obstante, también se identifican desafíos relacionados con la calidad y estandarización de los datos, lo cual se constituye como una oportunidad para una vía de acción por medio de la automatización con miras hacia el fortalecimiento de la eficiencia operativa.

Aporte del diagnóstico organizacional

El diagnóstico organizacional desarrollado dentro de la presente investigación permitió identificar oportunidades de mejora asociadas al proceso de gestión de PQRS en el área de Servicio y Experiencia de la vicepresidencia de Grandes Clientes de Vanti S.A. E.S.P. estas se constituyen el punto de partida para el diseño de la propuesta de intervención ya que identificaron necesidades operativas y organizacionales relevantes para la posible implementación del sistema dentro del proceso actual en el área debido a

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

que evidenciaron aspectos como la alta dependencia de procesos manuales lo que genera dificultades para establecer alertas tempranas frente a casos con alta carga negativa o potencial escalamiento, esto afecta la capacidad de respuesta del área y la priorización de recursos.

En respuesta a estas oportunidades de mejora, el diseño de la propuesta plantea el desarrollo e implementación de un modelo de procesamiento de lenguaje natural con el fin de automatizar el análisis textual, mejorar la priorización de casos críticos, optimizar los tiempos de respuesta y generar insumos analíticos que fortalezcan la toma de decisiones estratégicas.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Plan de Intervención

Tras los hallazgos identificados luego del diagnóstico organizacional, y apoyados en el marco de referencia que sustenta teóricamente los conceptos que aborda la presente investigación, se procede a la etapa de desarrollo del procesamiento estadístico de los datos, para así posteriormente, realizar la selección modelo óptimo, es aquí donde se analizan de manera comparativa los resultados obtenidos considerando métricas de desempeño como precisión, recall y F1-score, así como la capacidad de generalización de cada algoritmo frente a los datos de prueba. Cabe resaltar que la selección no obedece únicamente al análisis de criterios cuantitativos, también incorpora aspectos cualitativos y operativos, tales como la interpretabilidad del modelo, su eficiencia computacional y su viabilidad de implementación en el contexto organizacional. Con lo cual, se busca identificar la alternativa que no solo presente el mejor desempeño técnico, sino que se ajuste de manera adecuada a las necesidades del proceso de gestión de PQRS para Vanti.

Procesamiento estadístico de datos

La sección de procesamiento estadístico de los datos se desarrolló con el propósito de comprender en detalle la estructura y calidad de la base de datos seleccionada antes de iniciar la construcción de los modelos de machine learning debido a que identificar cómo estaba conformada la información, qué variables eran útiles para la investigación y cuáles requerían ajustes o depuración para garantizar resultados confiables.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

A continuación, se analizarán aspectos como la cantidad de registros disponibles, la presencia de datos vacíos, inconsistencias en los textos, duplicidad de información y diferencias en la forma en que eran registradas las observaciones de las PQRS. De igual manera, se realizarán análisis descriptivos y visualizaciones con el fin de entender el comportamiento general de la información y detectar posibles factores que pudieran afectar el desempeño de los modelos.

Población y muestra

La población objeto de estudio está conformada por las PQRS registradas por los clientes industriales de Vanti S.A. ESP en los canales de atención dispuestos para ello, como llamadas telefónicas, correos electrónicos y medios digitales, en el periodo comprendido entre 2022 y 2026. Es importante aclarar que los mensajes analizados no son textos escritos directamente por los clientes, sino transcripciones desarrolladas por trabajadores de Vanti en el momento de registrar la comunicación. Durante la atención, el funcionario anota en el sistema el resumen o descripción textual de lo expresado por el usuario; de igual manera, en los casos gestionados por correo electrónico, el empleado transcribe o adapta el contenido relevante de la solicitud o reclamo. Pese a ello, se considera que mantienen validez y pertenencia para el propósito del estudio ya que estos registros se generan de manera estandarizada según el sistema de gestión de PQRS, el cual contiene protocolos que exigen al funcionario consignar con fidelidad el motivo del

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

contacto y los elementos clave de la conversación. En consecuencia, el conjunto de datos mantiene la representatividad necesaria para extraer patrones válidos y generar información útil para la mejora de los procesos de atención al cliente.

El dataframe utilizado para el entrenamiento y pruebas del modelo, proviene de unas bases de datos suministradas por la empresa, las cuales, fueron exportadas a csv y facilitadas en una memoria usb. Esta información contiene datos históricos de contactos que se han presentado desde el 2022 hasta febrero del 2022. A continuación, se muestra una descripción de los hallazgos encontrados en esta primera exploración.

Análisis exploratorio

La base de datos consta de 39.083 registros y 44 columnas, está compuesta por datos de los clientes (por protección a los clientes, sus datos fueron anonimizados), información administrativa y la descripción textual relacionada con los casos atendidos. En la figura 16 se muestra el detalle de las columnas.

Figura 16

Estructura y características del conjunto de datos

#	Column	Non-Null Count	Dtype	#	Column	Non-Null Count	Dtype
0	Empresa Id	1 non-null	object	23	Autoriza respuesta e-mail	31246 non-null	object
1	Empresa	39072 non-null	object	24	Unidad de envío actual	39071 non-null	object
2	Cliente	1 non-null	object	25	Fecha de envío actual	39070 non-null	object
3	AÑO	39072 non-null	float64	26	Fecha finalizado	38952 non-null	object
4	MES	39072 non-null	object	27	Unidad de envío finalizado	34427 non-null	object
5	UNIDAD 1	39072 non-null	object	28	Resultado	38946 non-null	object
6	UNIDAD	39072 non-null	object	29	Motivo de cierre	14731 non-null	object
7	Provincia	37657 non-null	object	30	Días finalización	39072 non-null	float64
8	Municipio	35841 non-null	object	31	Días de gestión	39069 non-null	float64
9	Municipio contacto	1077 non-null	object	32	Fecha devolución	10905 non-null	object
10	Id ticket PQR	30745 non-null	float64	33	Marcarción	4 non-null	object
11	Origen del TCK	30745 non-null	object	34	Fecha de inicio marcarción	26 non-null	object
12	Radicación Especial	2140 non-null	object	35	Fecha de fin marcarción	17 non-null	object
13	Tipo de matriz	39072 non-null	object	36	Número días marcarción	36744 non-null	float64
14	Evento	39072 non-null	object	37	Observaciones	39033 non-null	object
15	Actividad	39072 non-null	object	38	Fecha Alta	12441 non-null	object
16	Actividad global	39072 non-null	object	39	Creado el	24225 non-null	object
17	Código causa	39072 non-null	object	40	Código	24429 non-null	object
18	Código antiguo	39072 non-null	object	41	Días 360	7728 non-null	float64
19	Causa	39072 non-null	object	42	Categoría de Cuenta	24094 non-null	object
20	Fecha radicación	39072 non-null	object	43	Días quejas	0 non-null	float64
21	Unidad Generadora	38875 non-null	object	dtypes: float64(7), object(37)			
22	Estado	39072 non-null	object				

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Nota. Elaboración propia.

La figura 16 presenta un resumen de la estructura del conjunto de datos utilizado en la investigación, incluyendo el nombre de las variables, la cantidad de valores no nulos y el tipo de dato asociado a cada una. Si se analiza con detenimiento los valores nulos de columnas como Municipio contacto, fecha de devolución, marcación o radicación especial, se encuentra un alto nivel de datos faltantes, lo cual lleva a concluir que es necesario un proceso de depuración, para no afectar el desempeño de los modelos más adelante.

Otro tema por destacar de la base es que una de las columnas con mayor valor para la presente investigación es la número 37 llamada Observaciones, ya que es la base para el desarrollo del modelo de análisis de sentimientos, debido a que contiene el texto del cual se realizarán los análisis. En la siguiente imagen se mostrará en detalle datos relevantes sobre esta variable.

Figura 17

Análisis descriptivo de la variable objetivo

```
=====
ANÁLISIS DE COLUMNAS OBJETIVO
=====

Observaciones
Registros no nulos : 39,033 (99.9%)
Valores nulos      : 50
Valores únicos     : 36,921
Ejemplo            : "https://apps.mypurecloud.com/directory/#/engage/admin/interactions/56ea668f-d047-4a40-9598-ca75836b5d1290100634BNA380Se"
Ejemplo            : "FORM_MAIL_INSCRIPFACTURA: Cliente solicita duplicado de factura al validar se evidencia que la cuenta contrato pertenec"
```

Nota. Elaboración propia.

Cómo se observa en la imagen 16 el 99,9% de los registros contienen datos, lo cual garantiza una adecuada cantidad de información para el entrenamiento. Aleatoriamente se

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

extrajeron dos ejemplos del tipo de comentarios que se pueden encontrar, llama la atención el primero, en el cual se observa la presencia de prefijos o códigos puede introducir ruido, por lo que podría requerirse una etapa adicional de limpieza (la cual más adelante se realizará).

Ahora bien, siguiendo con esta línea de análisis, dada la relevancia de esta variable para el desarrollo de los modelos, en las próximas imágenes se llevará a cabo un análisis detallado enfocado en ella. La figura 17 presenta un resumen estadístico de la longitud de los textos contenidos en la variable `texto_completo`, medida en número de caracteres, lo cual permite comprender la estructura y variabilidad de la información textual utilizada en el modelo de procesamiento de lenguaje natural.

Figura 8

Estadística descriptiva de la longitud de texto

```
Campo texto_completo creado.
Longitudes de texto (caracteres):
count    39033.0
mean      271.9
std       303.4
min        1.0
25%       165.0
50%       254.0
75%       255.0
max      12752.0
```

Nota. Elaboración propia.

Gracias a la figura 18 se observa que existe una alta variabilidad en la longitud de los textos, reflejada en la desviación estándar, esto es un reflejo de comparar el texto más

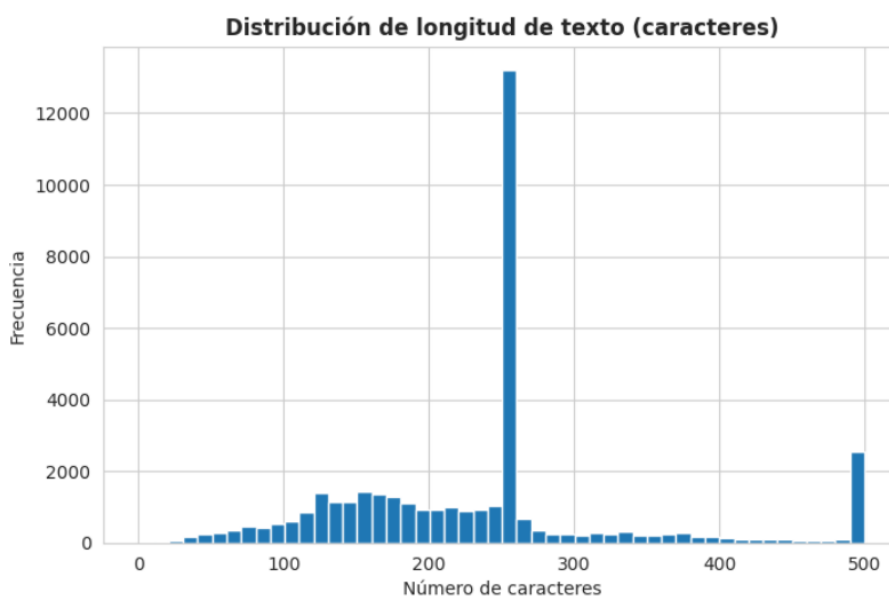
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

largo con 12.752 caracteres y textos con 1 carácter, pero general, el promedio se encuentra sobre los 271 caracteres, lo cual sugiere que la mayoría de los textos se concentran en un rango relativamente estrecho, lo que es positivo para el modelado mediante técnicas basadas en TF-IDF y n-gramas. La figura 19 busca complementar visualmente este análisis, mostrando la distribución real de las longitudes de texto.

Figura 19

Distribución de longitud de texto (caracteres)



Nota. Elaboración propia.

Como se observa en la imagen, la mayoría de los textos se concentran en un rango aproximado entre 100 y 300 caracteres, lo cual coincide con los estadísticos previamente analizados (media cercana a 272 y mediana de 254). Se identifican también los outliers

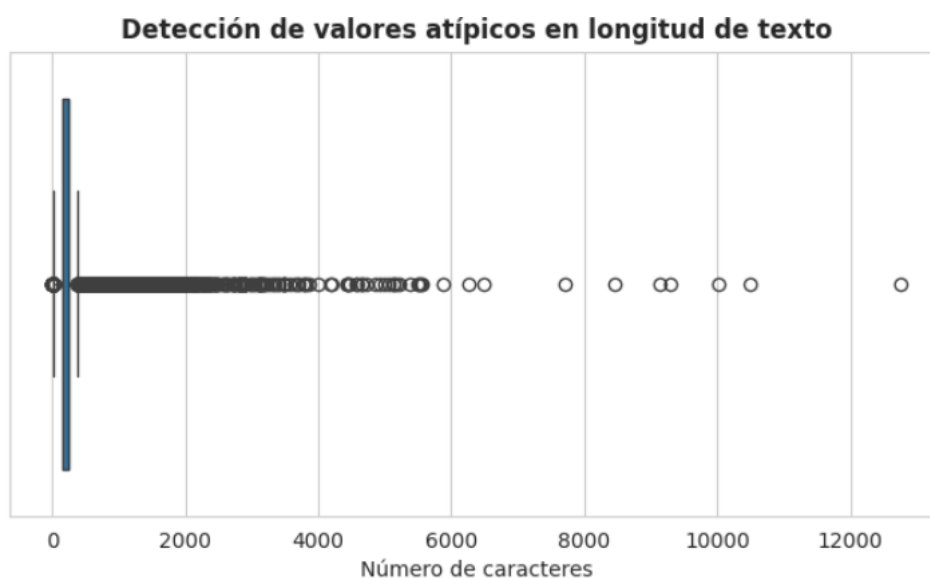
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

que están sobre los 500 caracteres o más. En la figura, se realiza para hacer mirar en detalle estos valores atípicos.

Figura 20

Detección de valores atípicos en longitud de texto



Nota. Elaboración propia.

Como se observa en la figura 20 hay una mayor concentración en un rango de caracteres relativamente bajo, sin embargo, también se evidencia la presencia de una gran cantidad de puntos dispersos hacia la derecha del gráfico, los cuales corresponden a valores atípicos o extremos que influyen en la dispersión total ya que presentan asimetría en la distribución. La importancia de la identificación de estos valores atípicos es que permite establecer criterios de depuración orientados a la mejora en la representatividad del conjunto de datos.

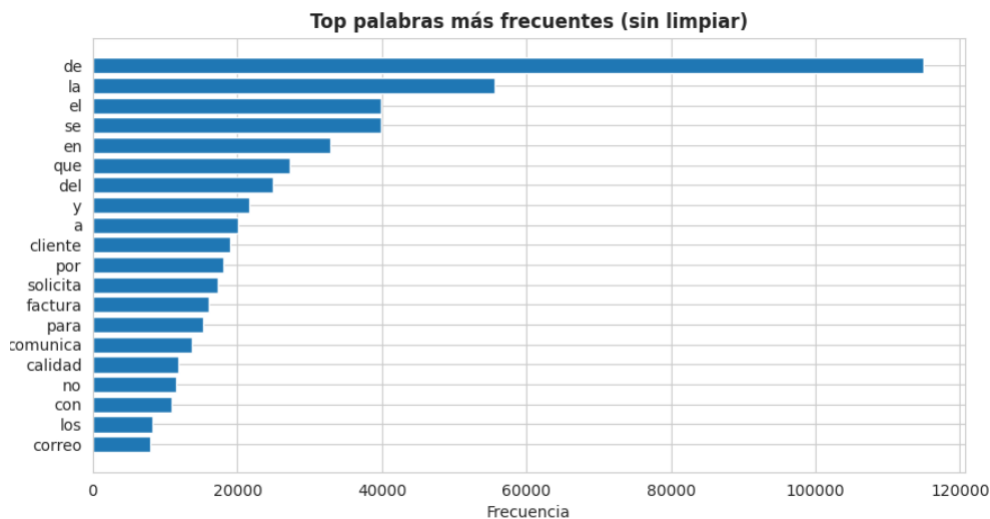
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Con el objetivo de comprender la composición inicial del texto contenido en las PQRS, a continuación, en la figura 21 se presenta una gráfica de las palabras más frecuentes sin aplicar procesos de limpieza o preprocesamiento.

Figura 9

Top palabras más frecuentes (sin limpiar)



Nota. Elaboración propia.

Como se observa en la figura 21 en este top antes de la limpieza, entre las palabras más frecuentes se encuentran stopwords como “de”, “la”, “el”, “se”, “en”, “que”, “del”, “y”, “a”, “por”, estos textos cumplen una función gramatical dentro de las oraciones, pero a la hora de modelar no contiene información significativa para el análisis de sentimientos, lo cual puede afectar negativamente el desempeño del modelo.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

En la siguiente figura, la 22 se encuentra otro ejemplo de transformación que se debe hacer dentro del preprocesamiento ya que incide en la calidad del dato que llegará al modelo. Se trata de textos que tienen una longitud máxima de 7 caracteres.

Figura 22

Ejemplos de registros con baja longitud y contenido no significativo

Observaciones	texto_completo	n_chars
31741	.	1
2729	C	1
16440	1	1
2264	0	1
1845	V	1
30383	.	1
3644	Prueba	6
21039	GFY 514	7
21112	TUN 439	7
26626	6249189	7

Nota. Elaboración propia.

Como se observa en la figura 22 hay registros con caracteres individuales como “C”, “1”, “0”, “V”, “.”. También caracteres numéricos que no aportan contenidos semánticos y códigos alfanuméricos que se asemejan a placas de vehículos.

A partir del análisis exploratorio realizado, en el cual se identificó las características del texto que dentro de la base no aportan valor al modelo y pueden afectar negativamente el desempeño, se procedió a aplicar un proceso de limpieza y depuración, orientado a eliminar esos registros irrelevantes con contenido textual insuficiente,

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

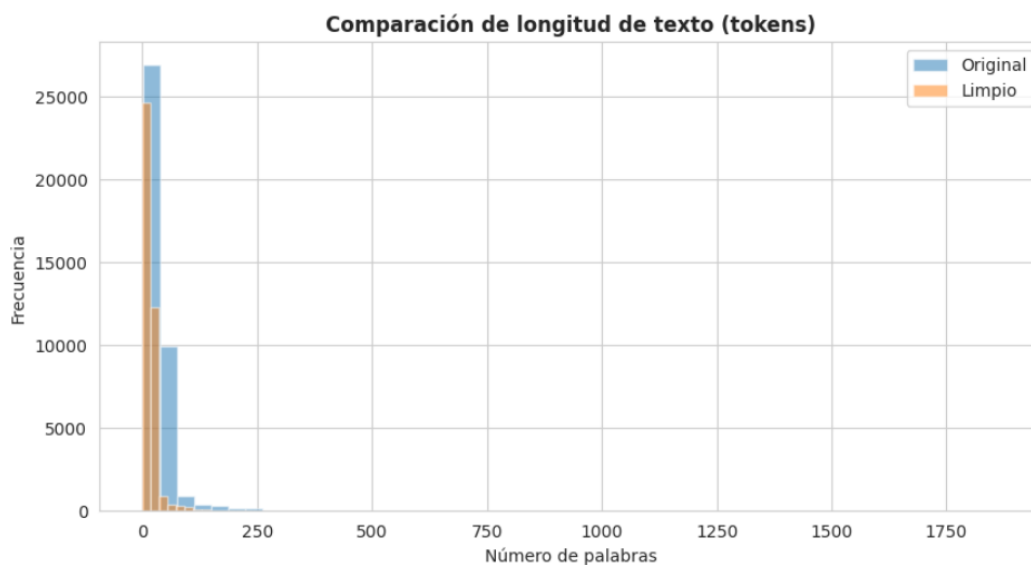
caracteres aislados, valores de prueba y expresiones sin relevancia semántica para el análisis de sentimientos.

A continuación, se presentan los resultados obtenidos tras la aplicación de estas etapas de limpieza, en los cuales se evidencia la transformación del conjunto de datos y la mejora en su estructura, permitiendo así contar con una base más consistente y adecuada para su uso en los modelos de clasificación de sentimientos.

La figura 23 es la primera muestra del cambio, en ella se presenta una comparación entre la distribución de la longitud de textos medida en número de tokens, antes y después del proceso de preprocesamiento.

Figura 10

Comparación de longitud de texto (tokens)



Nota. Elaboración propia.

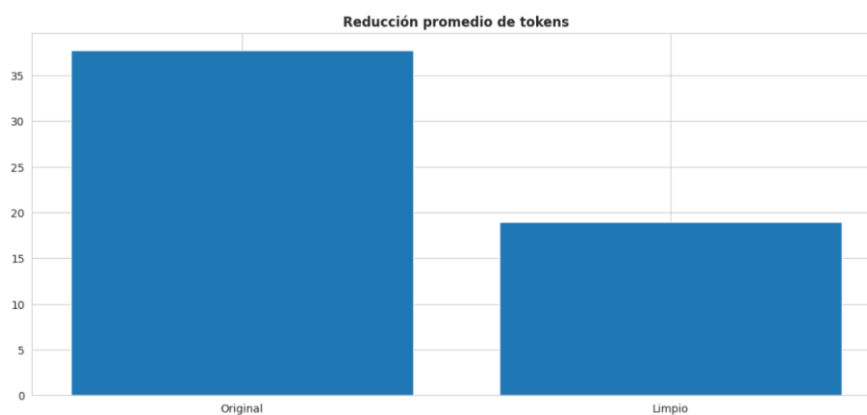
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Al analizar la gráfica 23 se observa una distribución más limpia que presenta menor dispersión, especialmente en valores altos, con lo cual se podría concluir que se redujo la presencia de textos excesivamente largos y se eliminaron elementos redundantes dentro del contenido. Como complemento a esto, la figura 24 presenta una gráfica de barras que mide el impacto antes y después de aplicar la limpieza.

Figura 24

Reducción promedio de tokens



Nota. Elaboración propia.

En la figura 24 se observa que el promedio de tokens paso de 37 palabras por registro a 19, esto representa una reducción cercana al 50% de la cantidad de palabras.

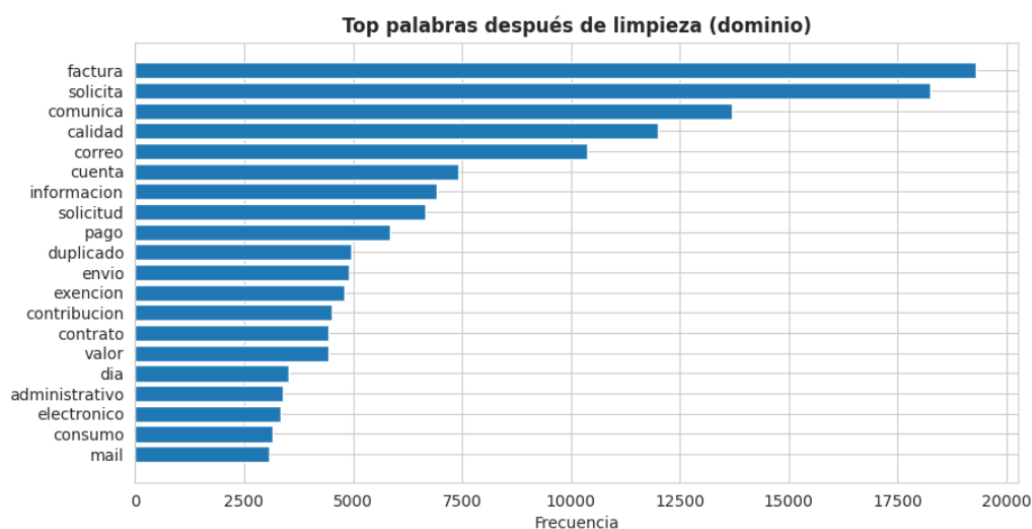
Ahora bien, en la gráfica 25 se vuelve a realizar el ejercicio del top de las palabras más usadas una vez hecha la limpieza ya antes explicada.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 25

Top palabras después de limpieza (dominio)



Nota. Elaboración propia.

En la figura 25 se observa que ahora las palabras más frecuentes están relacionadas con términos dentro de un contexto operativo de las PQRS, con palabras como “factura”, “solicita”, “comunica”, “información”, entre otras. Lo que se contrasta si al remitirse a la figura 20 en donde las palabras más predominantes eran “de”, “la”, “el”, “se”, “en”, “que”, “del”, “y”, “a”, “por”.

Siguiendo en la línea de comparación entre la base antes y después del proceso de limpieza, en las figuras 26 y 27 se observan dos nubes de palabras con el antes y el después lo cual visualmente demuestra el gran impacto luego del preprocesamiento.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Figura 26

Nube de palabras del texto original (antes del proceso de limpieza)



Nota. Elaboración propia.

Figura 27

Nube de palabras del texto procesado (después del proceso de limpieza)



Nota. Elaboración propia.

Las figuras 26 y 27 permiten identificar de manera intuitiva la relevancia y frecuencia de los términos dentro del conjunto de datos. En la imagen 26 se observa una

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

alta predominancia de palabras como “de”, “la”, “el”, “en”, “y”, “http”. Mientras que en la imagen 27 las palabras que más destacan son “solicita información”, “factura”, “correo electrónico”, “calidad administrativa”, entre otras.

A continuación, en la figura 28 se observan las estadísticas de los tokens antes y después.

Figura 11

Comparación de estadísticas de tokens antes y después del proceso de limpieza

Estadísticas de tokens ANTES y DESPUÉS de la limpieza:

	Antes (tokens)	Después (tokens)
count	39033.0	39033.0
mean	37.7	19.0
std	45.0	21.9
min	1.0	0.0
25%	21.0	11.0
50%	32.0	16.0
75%	40.0	20.0
max	1865.0	926.0

Nota. Elaboración propia.

En la figura 28 lo primero que se observa es que el número de registros no varía luego de la limpieza, esto porque en el proceso no se hizo una eliminación masiva de registros, sino una transformación del contenido textual.

Respecto a las otras medidas estadísticas se observa por ejemplo que la media pasó de 37.7 tokens a 19 tokens lo que indica que se hizo una reducción aproximada del 50% en la longitud promedio de texto (tal y como se mostraba en la figura 23) al hacer la

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

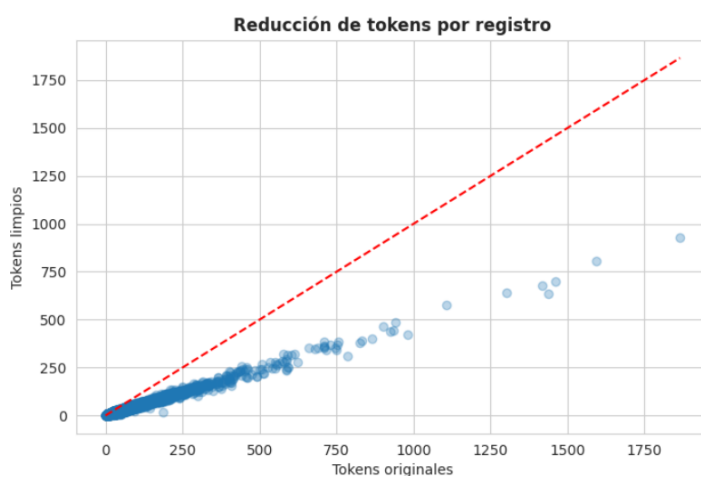
ESP

eliminación de palabras irrelevantes. La disminución en la dispersión muestra que los textos son más homogéneos en longitud debido a que pasa de 45 a 21.9. Los percentiles también disminuyeron, lo cual confirma que la limpieza impactó de manera uniforme todo el conjunto de datos, no solo casos aislados. Finalmente, en cuanto a mínimos y máximos también se presentó una reducción considerable, mientras los primeros pasaron de uno a cero, los máximos pasaron de 1865 a 926.

Para terminar esta etapa de comparativos entre el texto antes y después del proceso de limpieza, se presenta en la figura 29, en ella se presenta la relación entre el número de tokens originales y los tokens resultantes después del proceso de limpieza aplicado a los textos de las PQRS.

Figura 29

Reducción de tokens por registro tras el proceso de limpieza del texto



Nota. Elaboración propia.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

En la figura 29 se conforma por una serie de puntos que representan registros individuales y una línea diagonal punteada que indica el escenario ideal en el cual no existiría reducción en la cantidad de tokens. Se observa que la gran mayoría de los puntos se ubican por debajo de la línea de referencia, esto evidencia una disminución sistemática en la cantidad de tokens tras el proceso de limpieza.

Adicionalmente, se puede ver que las reducciones más notables son en los textos de mayor longitud, lo que sugiere que estos contienen una mayor proporción de elementos innecesarios. Sin embargo, la relación entre tokens originales y tokens limpios mantiene una tendencia aproximadamente lineal, lo que indica que el proceso de limpieza conserva la estructura general de la información sin distorsionar significativamente el contenido.

Ahora bien, una vez realizado el proceso de limpieza y depuración de los datos, se procede a desarrollar el análisis léxico del contenido textual. Esta etapa busca identificar patrones de uso, frecuencias y relevancia dentro del contexto de Vanti, mediante el examen de las palabras y expresiones presentes en las PQRS. El objetivo es comprender la composición del lenguaje utilizado por los clientes, destacando los términos más representativos y su relación con las temáticas abordadas en las comunicaciones. Esta etapa permite transformar el texto en una fuente de información estructurada, facilitando la interpretación de los datos y sirviendo como base para el desarrollo de modelos de análisis de sentimientos.

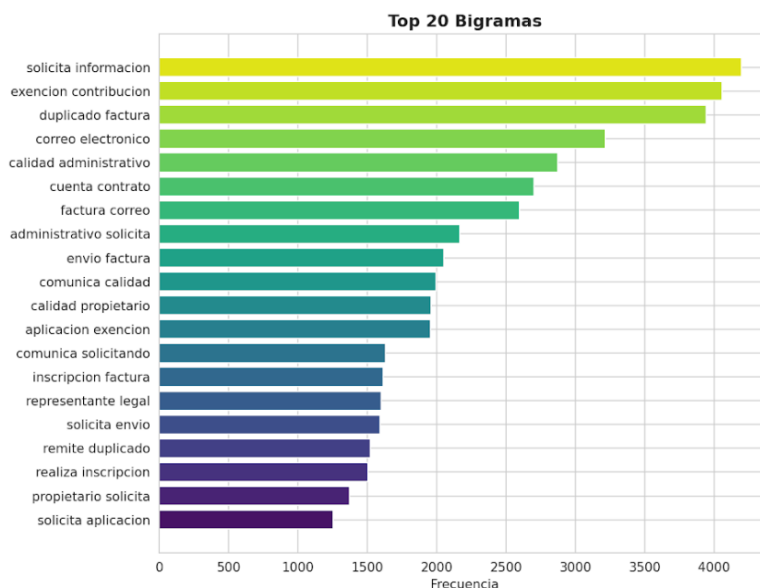
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

En primera medida, en esta etapa lo primero que se hizo fue generar n-gramas con el objetivo de capturar relaciones semánticas más específicas que en el análisis de palabras individuales, luego de eliminar el ruido de palabras que no aportan nada como números, placas, correos electrónicos, entre otros. Estos son más propicios para la generación de diccionarios con las palabras claves positivas y las palabras claves negativas, con estos, se podrá realizar la clasificación de sentimientos en el texto mediante la implementación de un modelo híbrido (más adelante se abordará este tema en detalle). En la figura 30 se presenta gráficamente el resultado del ejercicio para obtener los bigramas más frecuentes dentro de la base de PQRS.

Figura 30

Bigramas más frecuentes en el conjunto de datos



Nota. Elaboración propia.

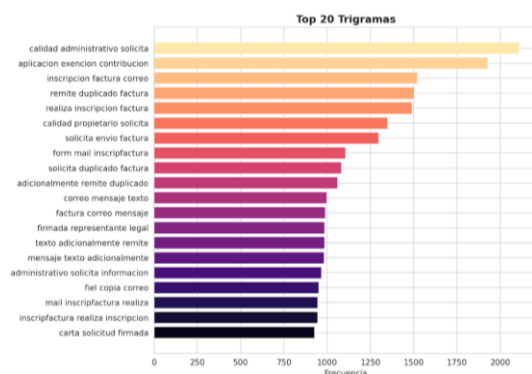
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Como se mencionaba anteriormente, la importancia de obtener estos bigramas por encima de los unigramas radica en que permiten capturar un contexto semántico más preciso, por ejemplo, referirse a “factura” por si sola es ambigua, pero “duplicado de factura” o “envío factura” aporta un significado más claro y empieza a denotar una intención, con lo cual más adelante facilitará la categorización de las PQRS. En consecuencia, la figura 29 presenta los veinte bigramas más frecuentes, los cuales incluyen palabras como “solicita información”, “exención contribución”, “duplicado factura”, entre otros. A simple vista, se observa que predomina un lenguaje tipo operativo y administrativo, con terminología relacionada con temas como facturación y regulatorios o tributarios, sin embargo, aún no se visualizan claramente el manejo emocional en la intención comunicativa. Para esto se pasó a una segunda etapa en la que se prueba el uso de trigramas, en la figura 31 se muestra el resultado del ejercicio.

Figura 31

Trigramas más frecuentes en el conjunto de datos



Nota. Elaboración propia.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

La figura 31 presenta los veinte trigramas más frecuentes encontrados en la base de datos, estos permiten capturar estructuras lingüísticas más complejas en las cuales se observan expresiones con alta carga administrativa como “calidad administrativa solicita”, “aplicación extensión contribución”, “solicita duplicado factura”, entre otras. Como se observa, estas combinaciones ya integran acción más objeto más contexto, lo que facilita la interpretación automatizada.

Una vez realizado el análisis léxico del contenido textual, que dio como resultado los n-gramas que se acaban de mostrar, la siguiente etapa es la de etiquetado de sentimientos, la cual se configura como un paso previo para el entrenamiento de los modelos supervisados de clasificación. En esta fase, cada registro es asociado a una categoría de sentimiento previamente definida, permitiendo establecer una variable objetivo sobre la cual los algoritmos de aprendizaje automático puedan entrenarse.

Es importante mencionar que, para evitar sesgos en la clasificación de sentimientos, se decidió trabajar con tres capas, la primera de ellas cada registro es asociado a una categoría de sentimiento previamente definida, permitiendo establecer una variable objetivo sobre la cual los algoritmos de aprendizaje automático puedan entrenarse. En la figura 32 se muestra como fue el proceso para la categorización.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 32

Diccionario de polaridad de términos para la clasificación de sentimientos en PQRS

```
# CAPA 1: Diccionario de polaridad por dominio PQR
KEYWORDS_NEG = {
    # Escalamiento y reclamo formal
    'comunica sra', 'calidad administrativo', 'calidad propietario',
    'desacuerdo conceptos', 'desacuerdo conceptos facturados',
    'inconformidad', 'inconform', 'reclamo', 'queja',
    # Cobros incorrectos
    'cobro indebido', 'cobro incorrecto', 'error factura',
    'factura incorrecta', 'valor elevado', 'cobro excesivo',
    'cargo incorrecto',
    # Negativa / no resolución
    'no accede', 'rechaza', 'negar', 'niega', 'negado',
    'no procede', 'improcedente',
    # Urgencia / legal
    'tutela', 'superintendencia', 'sspd', 'demanda', 'juridico',
    'suspension', 'corte', 'cortado', 'suspendido',
    # Tono negativo explícito
    'indignacion', 'molestia', 'perjuicio', 'dano',
    'incumplimiento', 'irregularidad', 'irregular',
    'exigir', 'exijo', 'exige',
}

KEYWORDS_POS = {
    'quede satisfecho', 'atención recibida', 'agradezco gestion', 'agradezco',
    # Resolución favorable
    'accede', 'aprobado', 'aprueba', 'solucionado',
    'resuelto', 'favorable', 'acepta', 'aceptado',
    'procede', 'confirmado',
    # Satisfacción del cliente
    'satisfecho', 'conforme', 'felicitaciones', 'felicitación',
    'excelente', 'muy bien', 'agradece', 'agradecido',
    'gracias atencion', 'buen servicio', 'bien atendido',
    'correcto', 'de acuerdo',
}
```

Nota. Elaboración propia.

Se observa en la figura 32 que se presentan la definición de variables de keywords tanto positivas como negativas, en ellas se manejan categorías que se etiquetan con algunos casos de comunicaciones reales que han tenido clientes y las cuales se tomaron para generar este diccionario. Para el negativo se encuentran “escalamiento y reclamo formal”, “cobros incorrectos”, “negativa/no resolución”, “urgencia/legal”, “tono negativo explícito”; para el positivo “resolución favorable”, “satisfacción del cliente”.

Luego de pasar por la primera capa, sigue a la segunda, la cual se realiza por medio de del score VADER, el cual es un modelo de análisis de sentimientos basado en un enfoque léxico, diseñado para evaluar la polaridad de textos cortos considerando

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

elementos como intensificadores, negaciones y puntuación (Hutto & Gilbert, 2014). En la figura 33 se muestra como fue la aplicación de Vader.

Figura 33

Implementación del modelo VADER

```
# CAPA 2: VADER score
def vader_label(texto):
    score = SIA.polarity_scores(str(texto))['compound']
    if score >= 0.08: return 'Positivo'
    elif score <= -0.08: return 'Negativo'
    else: return 'Neutro'

def vader_score(texto):
    return SIA.polarity_scores(str(texto))['compound']
```

Nota. Elaboración propia.

Lo que se observa en la figura 33 es la definición de la función con VADER, la cual más adelante se aplicará usando las tres capas que se están explicando.

Finalmente, la tercera capa consiste en una detección de urgencias, es decir, palabras puntuales como demanda tutela, explosión, urgente. Que para efectos prácticos fuerzan la etiqueta Negativo si las capas anteriores no han decidido. Esta capa captura PQRS de alta criticidad que podrían haber sido clasificados como Neutros por ausencia de keywords emocionales.

En la figura 34 se muestra el proceso de ejecución de estas tres capas y en la figura 34 el resultado de su aplicabilidad.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 34

Algoritmo de etiquetado basado en reglas léxicas y modelo VADER

```
# ETIQUETADOR PRINCIPAL
def etiquetar_por_texto(texto_obs):
    """
    Etiqueta una Observación en Positivo / Negativo / Neutro
    usando 3 capas de señal sobre el texto libre.
    """
    if not isinstance(texto_obs, str) or not texto_obs.strip():
        return 'Neutro'

    t_lower = texto_obs.lower()

    # Capa 1: keywords - mayor peso, evaluación primero
    hits_neg = sum(1 for kw in KEYWORDS_NEG if kw in t_lower)
    hits_pos = sum(1 for kw in KEYWORDS_POS if kw in t_lower)

    if hits_neg >= 2 or (hits_neg == 1 and hits_pos == 0):
        return 'Negativo'
    if hits_pos >= 2 or (hits_pos == 1 and hits_neg == 0):
        return 'Positivo'

    #Capa 2: VADER como desempate
    v_label = vader_label(texto_obs)
    if v_label != 'Neutro':
        return v_label

    #Capa 3: urgencia -> fuerza Negativo si VADER no decidió
    if tiene_urgencia(texto_obs):
        return 'Negativo'

    return 'Neutro'

# APLICAR
df['vader_score'] = df['Observaciones'].apply(vader_score)
df['etiqueta_texto'] = df['Observaciones'].apply(etiquetar_por_texto)

# Etiqueta final = la misma (renombrada para compatibilidad)
df['sentimiento'] = df['etiqueta_texto']

print('Distribución etiqueta_texto:')
print(df['sentimiento'].value_counts())
print(f'\nEjemplos por clase:')
for sent in ['Negativo', 'Positivo', 'Neutro']:
    muestra = df[df['sentimiento']==sent]['Observaciones'].dropna().sample(
        min(2, (df[df['sentimiento']==sent]).sum()), random_state=42
    ).tolist()
    print(f'\n{sent}:')
    for m in muestra: print(f' {str(m)[:120]}')
```

Nota. Elaboración propia.

Figura 35

Distribución de clases de sentimiento y ejemplos del etiquetado

```
Distribución etiqueta_texto:
sentimiento
Neutro      23632
Negativo    9084
Positivo     6299
Name: count, dtype: int64

Ejemplos por clase:

Negativo:
90104712 Cliente manifiesta que no a generado ninguna conversión, no tiene vehículo y solicita se le reintegre los valor
Señora Diana Zuleidy Rocha Rodriguez, da respuesta al Ticket 9497773, manifestando, que no fue una falsificación de docu

Positivo:
**LEY 2300 de 2023**Señor.GASNACER HOY VANTRef usuario ALIMENTO FRESCO DEL CARIBE Cuenta 62753566 - factura F94153448209R
CLIENTE:Adjunto se encuentra el documento correspondiente a la "Objeción Parcial de las Facturas de Gas Natural remitida

Neutro:
Se realiza comunicación con el cliente, donde se realiza confirmación de los correos electrónicos para envío y recepción
90101059
FORM_MAIL_INSCRIPFACTURA
Se realiza la inscripción de factura por correo y mensaje de texto, adicionalmente se
```

Nota. Elaboración propia.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Como se mencionó con anterioridad figura 35, muestra los resultados de la aplicación de las tres capas para la identificación de los sentimientos, se observa entonces que la categoría neutra con un total de 23.632 observaciones es la que concentra la mayor cantidad de registros, seguida por la categoría negativa con 9.084 y la positiva con 6.299 registros. Esta distribución evidencia un claro desbalance de clases donde predominan los textos de carácter informativo o administrativo sobre aquellos con carga emocional explícita.

Asociando lo anterior a la identificación de sentimientos, se observa que en los de clase negativa se identifican expresiones como de inconformidad, reclamos y posibles acciones legales. En la clase positiva los textos presentan elementos asociados a solución efectiva de problemas y satisfacción con la atención recibida. Por su parte, en la clase neutra predominan mensajes informativos o descriptivos, sin presencia de términos que indiquen una valoración emocional clara.

Ahora bien, luego de la clasificación de sentimientos en el texto, se procede a hacer la vectorización con TF-IDF, con el fin de transformar el texto en representaciones numéricas que capturen la relevancia de los términos dentro del corpus. A continuación, en la figura 35 se muestra el resultado de la vectorización.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 36

Preparación del conjunto de datos para el entrenamiento del modelo

Distribución actual: {'Neutro': 23632, 'Negativo': 9084, 'Positivo': 6299}

Clases codificadas: {'Negativo': 0, 'Neutro': 1, 'Positivo': 2}

 Split estratificado:
 Train: 31,212 | Test: 7,803
 Negativo : 7,267 train | 1,817 test
 Neutro : 18,906 train | 4,726 test
 Positivo : 5,039 train | 1,260 test

 TF-IDF: 30,000 features

Nota. Elaboración propia.

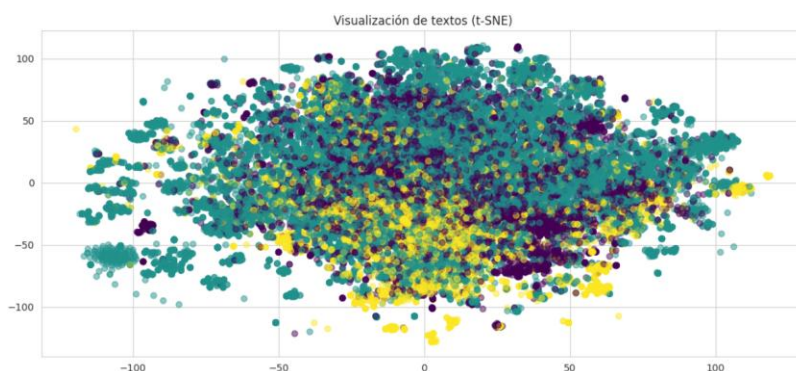
La figura 36 proporciona una cantidad relevante de datos para analizar. Primero, se observa una distribución de las clases en la cual predomina el neutro, seguido del negativo y posteriormente el positivo, dada la magnitud de las cifras, se observa que se trabaja con un dataset desbalanceado, lo cual es común en el lenguaje natural. Lo siguiente que hace es codificar las clases para volverlas numéricas y hablar el mismo idioma del modelo que se vaya a aplicar. A continuación, se dividen los datos en 80% de entrenamiento y 20% de prueba. Así mismo se mantiene la proporción de las clases para dividirlos en datos de entrenamiento y prueba. Posteriormente, se hace la vectorización TF-IDF con el fin de transformar el texto en representaciones numéricas basada en la frecuencia e importancia de los términos, generando así un espacio de características de aproximadamente 30.000 dimensiones.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Una vez realizada la transformación del texto mediante la técnica TF-IDF y definida la división del conjunto de datos entre entrenamiento y prueba se procede a explorar la estructura de los datos en el espacio vectorial. Para esto se emplea la técnica t-SNE, la cual permite la visualización de patrones y posibles separaciones entre las clases de sentimiento. A continuación, en la figura 37 se presenta dicha estructura interna de los datos.

Figura 37

Representación de distribución de textos en espacio semántico (t-SNE)



Nota. Elaboración propia.

La figura 37 se observa una representación bidimensional de los textos procesados, cada punto representa una PQRS, mientras que los colores corresponden a las diferentes clases de sentimiento. A simple vista, se puede identificar que hay una alta concentración de puntos en la región central y un poco de dispersión hacia los lados, los cuales podrían considerarse outliers debido posiblemente a textos atípicos. En contraste, esa

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

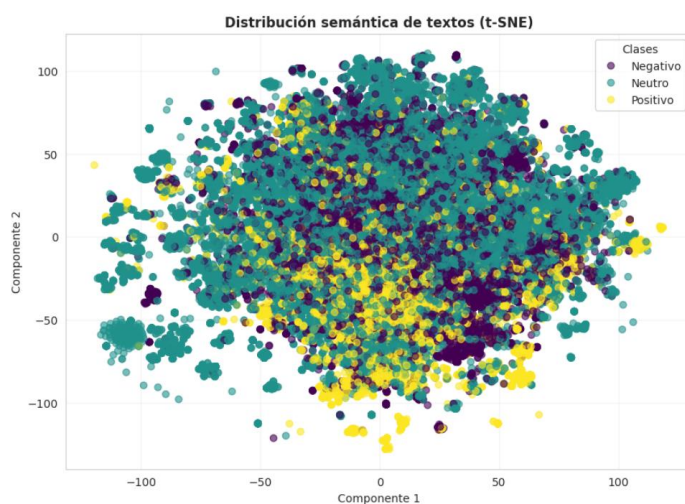
ESP

concentración en el medio puede significar similitudes semánticas entre las categorías, especialmente entre negativo y neutro.

A continuación, en la figura 38, realizada con la misma técnica t-SNE, se introduce un enfoque semántico, al asociar cada punto con su respectiva etiqueta a saber, positivo amarillo; negativo morado y neutro azul.

Figura 38

Representación de distribución de textos por categoría de sentimiento (t-SNE)



Nota. Elaboración propia.

En la figura 37 se observa, al igual que en la figura 36, una gran concentración de puntos en la región central del gráfico, adicional, se evidencia una considerable superposición entre las clases de sentimiento, lo cual sugiere que las diferencias semánticas entre categorías no son completamente separables en un espacio de baja dimensionalidad.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Al analizar la concentración de colores asociados a los sentimientos, se observa que los textos clasificados como positivos (amarillo) tienden a concentrarse con mayor frecuencia en la parte inferior del gráfico, mientras que los negativos (morado) aparecen más dispersos en diferentes regiones, y los neutros (verde) dominan la zona central. Se podría concluir que este comportamiento es consistente con la naturaleza de las PQRS, donde los textos neutros suelen ser más homogéneos (informativos), mientras que los positivos y negativos presentan mayor variabilidad en su expresión.

Selección del modelo

Con el fin de sustentar la selección del modelo más adecuado, se presenta a continuación un análisis comparativo del desempeño de los diferentes algoritmos evaluados. En la figura 38 se muestra el desempeño de diferentes modelos, evaluados mediante métricas como accuracy, F1-score macro, F1-score ponderado, así como medidas de validación cruzada (CV Mean y CV Std) y tiempo de ejecución.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 39

Comparación del desempeño de modelos

	Modelo	Accuracy	F1-Macro	F1-Weighted	CV Mean	CV Std	Tiempo(s)
1	Linear SVM	0.8624	0.8382	0.8620	0.8274	0.0023	7.4
2	XGBoost	0.8517	0.8173	0.8443	0.8078	0.0055	1240.8
3	Logistic Reg.	0.8351	0.8139	0.8383	0.8091	0.0033	12.7
4	SGD Classifier	0.8370	0.8076	0.8370	0.8025	0.0058	1.8
5	Naive Bayes (MNB)	0.7794	0.7320	0.7798	0.7253	0.0104	3.5
6	Naive Bayes (CNB)	0.7628	0.7302	0.7690	0.7283	0.0073	0.3
7	Random Forest	0.7032	0.7006	0.7192	0.6878	0.0117	115.7

Nota. Elaboración propia.

En términos generales, de acuerdo con lo que muestra la figura 39, el modelo Linear SVM es el de mejor desempeño, con un accuracy de 0.8624 y un F1-score macro de 0.8382 demuestra una alta capacidad para clasificar correctamente las tres clases de sentimiento de manera equilibrada. De igual manera, se concluye que tiene una estabilidad en su rendimiento con el valor de validación cruzada (CV Mean) de 0.8274 y una desviación estándar baja.

El que le sigue es el modelo XGBoost el cual presenta un F1-score macro de 0.8173, aunque con un tiempo de ejecución mayor, lo que puede representar una desventaja en términos de eficiencia computacional. Por su parte, la regresión logística y el SGD Classifier presentan resultados similares, con métricas ligeramente inferiores, pero tiempos de procesamiento más bajos.

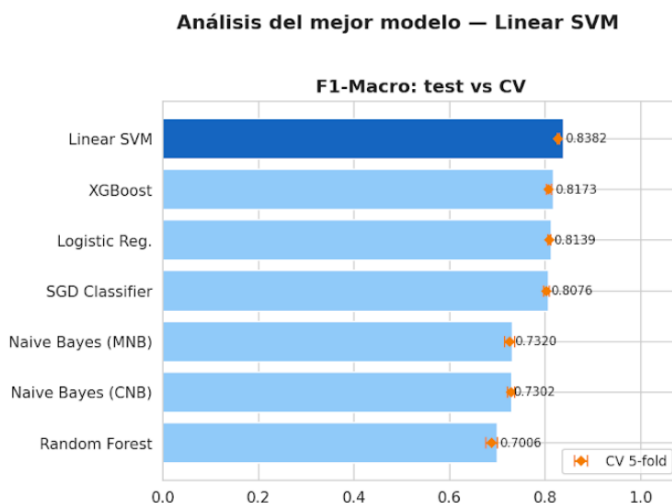
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

En contraste, los modelos de Naive Bayes y Random Forest presentan un desempeño inferior, particularmente en el F1-score macro, lo cual sugiere una menor capacidad para manejar el desbalance de clases y la complejidad del lenguaje natural presente en las PQRS.

Como complemento a lo anterior, la figura 39 muestra gráficamente los resultados concentrados en la métrica F1-score macro, tanto en el conjunto de prueba como en validación cruzada, esto con el fin de evaluar no solo la precisión global de cada modelo, sino también su capacidad para generalizar frente a datos no vistos.

Figura 40

Comparación del desempeño de modelos mediante F1-score macro



Nota. Elaboración propia.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

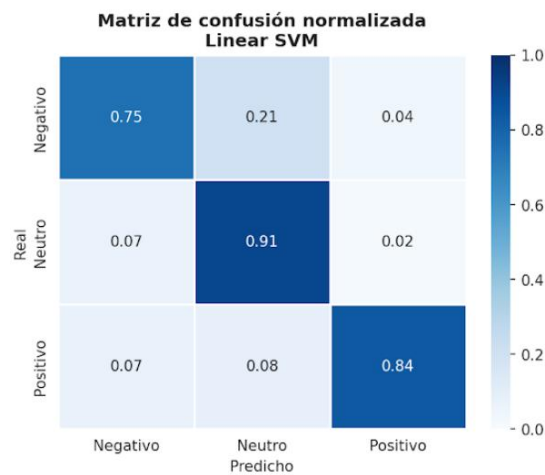
ESP

Como se observa en la figura 39 el modelo Linear SVM es el que obtiene mejor desempeño con un valor de F1-macro en el conjunto de prueba de 0.8382, este resultado denota una buena capacidad del modelo para equilibrar precisión y recall entre las diferentes clases (positivas, negativas y neutras), lo cual es especialmente relevante en problemas de análisis de sentimientos donde puede existir desbalance de clases. Los otros algoritmos que pueden capturar patrones relevantes en los datos son XGBoost (0.8173), Regresión Logística (0.8139) y SGD Classifier (0.8076). En contraparte, los modelos basados en Naive Bayes (Multinomial y Complemento) y Random Forest tienen una menor capacidad para modelar la complejidad semántica de los textos. Se podría concluir que estos últimos, basados en árboles no suelen ser tan eficientes como los modelos lineales al trabajar con representaciones como TF-IDF.

En consecuencia, el modelo seleccionado como la alternativa óptima para la clasificación de sentimientos fue el Linear Support Vector Machine (SVM), en la figura 40 se muestran las métricas globales específicas para este modelo, así como el comportamiento real del modelo en cada sentimiento que se busca clasificar dentro de las PQRS.

Figura 41

Matriz de confusión normalizada del modelo Linear SVM



Nota. Elaboración propia.

La figura 41 compara las clases reales frente a las clases predichas de los sentimientos que se buscan predecir. Se observa que clasifica correctamente el 91% de los neutros, lo cual indica una alta precisión; en cuanto a los positivos, identifica correctamente el 84%, sin embargo presenta cierta confusión con los neutros (8%) y los negativos (7%); finalmente, la clase negativa es la que presenta menor desempeño con 75% de aciertos, el problema se presenta con los neutros ya que el 21% de los negativos los clasifica como neutros, esto puede generar una subestimación de comentarios negativos, al presentarlos como neutros.

Para complementar el análisis de la figura anterior, a continuación, se muestra en la figura 42 cómo se comporta el modelo a nivel agregado en términos de volumen,

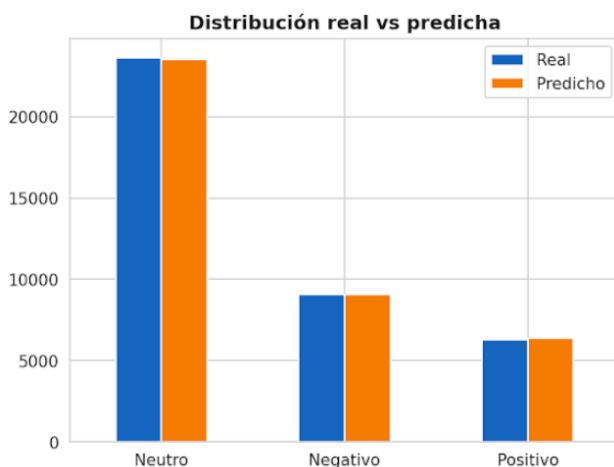
Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

comparando la distribución real entre las clases con la distribución generada por el modelo.

Figura 42

Comparación entre distribución real y predicha de las clases de sentimiento



Nota. Elaboración propia.

Como se observa en la figura 42 el modelo no presenta sesgos significativos de predicción, esto indica que hay una consistencia entre el comportamiento individual y el comportamiento agregado por tanto logra reproducir adecuadamente la proporción de cada tipo de sentimiento presente en los datos originales.

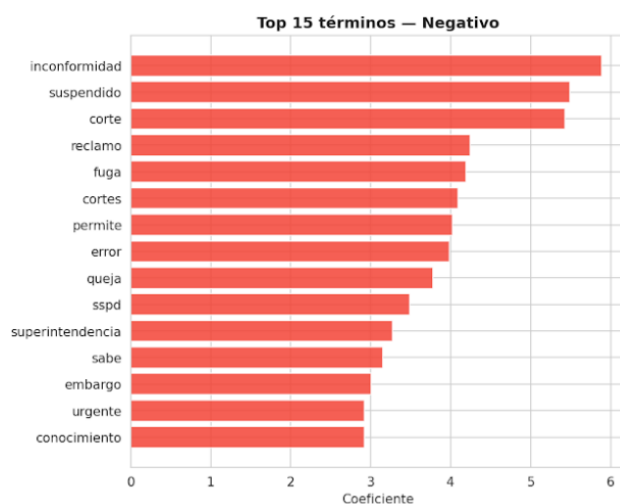
En las siguientes figuras 43, 44 y 45 se mostrará el top 15 de los principales términos con mayor peso en el modelo, estos coeficientes reflejan la importancia de cada palabra para la decisión del modelo, en este sentido, a mayor valor, mayor influencia para calificar los comentarios dentro de cada clase.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 43

Términos más relevantes asociados a la clase negativa según el modelo SVM



Nota. Elaboración propia.

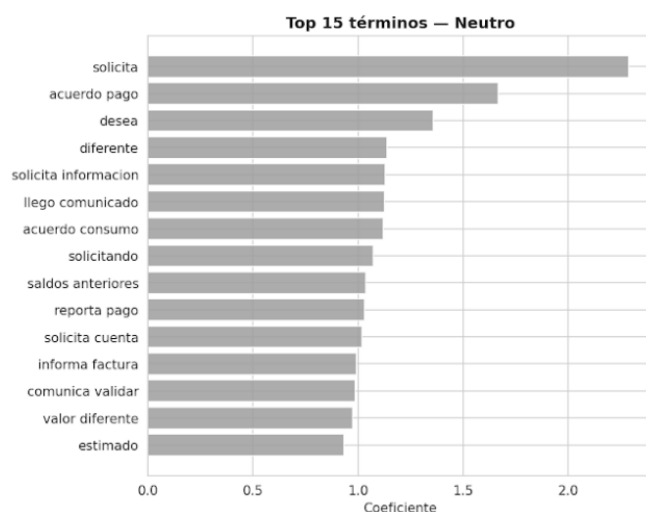
En la figura 43 se observan los términos negativos, entre ellos, el principal es “inconformidad” seguido por “suspendido” y “corte”, entre otras. En términos generales, se puede inferir que estas palabras están asociadas a situaciones problemáticas como fallas en el servicio, o insatisfacción del cliente.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 44

Términos más relevantes asociados a la clase neutro según el modelo SVM



Nota. Elaboración propia.

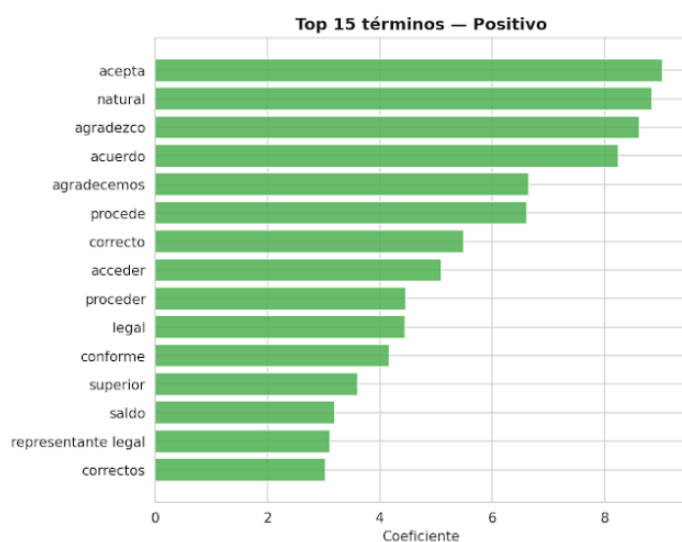
En la figura 44 se observan los términos neutros, a diferencia de los negativos, en donde predominan términos asociados a inconformidad, en este caso se encuentran palabras relacionadas con solicitudes, procesos administrativos y gestiones informativas. Entre los términos más significativos se encuentran “solicita”, “acuerdo pago”, “desea”, “diferente”, entre otros.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 45

Términos más relevantes asociados a la clase positiva según el modelo SVM



Nota. Elaboración propia.

En la imagen 45 se observan los términos positivos, A diferencia de las clases negativa y neutra, en este caso predominan términos asociados a satisfacción, conformidad y validación de procesos. Entre los términos más significativos se encuentran “acepta”, “natural”, “agradezco”, “acuerdo”, entre otros.

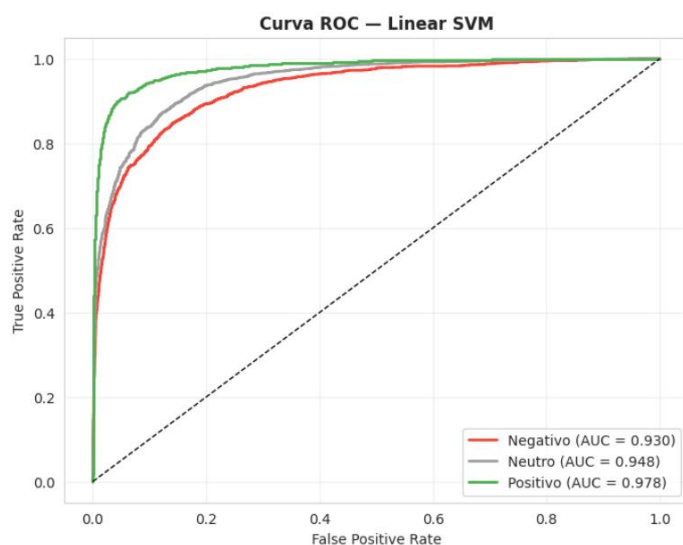
Para finalizar esta etapa de análisis del modelo escogido, la figura 44 presenta la curva ROC y del modelo Linear SVM para cada clase de sentimiento junto con sus respectivos valores de área bajo la curva (AUC), las cuales permiten evaluar la capacidad del modelo para discriminar entre las diferentes clases, considerando distintos umbrales de decisión.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 46

Curvas ROC y valores AUC del modelo SVM para clasificación de sentimientos



Nota. Elaboración propia.

Como se observa en la figura 46, todas las curvas se encuentran por encima de la línea diagonal de referencia. La clase positiva es la que presenta el mejor desempeño con un valor AUC de 0.978 lo que indica que tiene una alta capacidad para distinguir correctamente este sentimiento, en contraste, la clase negativa presenta un AUC de 0.930, ligeramente inferior pero aún dentro de un rango considerado alto. En la mitad de ellas, se encuentra la clase inferior con un AUC de 0.948, mostrando también un desempeño sólido. Los resultados anteriores refuerzan la conclusión de que el modelo posee una adecuada capacidad de discriminación para identificar correctamente las clases de sentimiento.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Implementación del modelo

Una vez seleccionado el modelo para la clasificación de sentimientos, a continuación, se procede a la fase de implementación, la cual lleva los resultados del proceso de modelado a un contexto práctico y aplicable dentro de Vanti. Esta parte integra todos los componentes desarrollados a lo largo del proyecto, tales como preprocesamiento de texto, vectorización mediante TF-IDF y la interpretación de resultados. A continuación, se describe la forma en que el modelo es aplicado, así como los elementos que permiten su funcionamiento dentro del flujo de análisis propuesto.

En las figuras 46, 47 se muestra como fue el desarrollo del modelo, aquí se evidencia como se llevó a cabo el proceso que recoge los elementos vistos anteriormente.

Figura 47

Estructura de código para predicción de sentimientos mediante SVM y TF-IDF

```

bundle = joblib.load('PQR_modelo_bundle.pkl')

def predecir_pqr(texto_observacion: str, verbose: bool = True) -> dict:

    advertencias = []

    #Validaciones básicas
    if not isinstance(texto_observacion, str) or not texto_observacion.strip():
        return {'error': 'Texto vacío o inválido.'}

    texto_limpio_nuevo = limpiar_texto(texto_observacion)
    n_tokens = len(texto_limpio_nuevo.split())

    if n_tokens < 3:
        advertencias.append(f'Texto muy corto ({n_tokens} tokens) - baja confianza esperada')

    # -- Vectorizar y predecir --
    X_nuevo = bundle['tfidf'].transform([texto_limpio_nuevo])
    modelo = bundle['modelo']
    le_b = bundle['label_encoder']

    pred_enc = modelo.predict(X_nuevo)[0]
    pred_lbl = le_b.inverse_transform([pred_enc])[0]

    # Probabilidades / scores de decisión
    if hasattr(modelo, 'predict_proba'):
        proba = modelo.predict_proba(X_nuevo)[0]
    elif hasattr(modelo, 'decision_function'):
        df_func = modelo.decision_function(X_nuevo)[0]
        if df_func.ndim == 0: df_func = np.array([df_func])
        exp_d = np.exp(df_func - df_func.max())
        proba = exp_d / exp_d.sum()
    else:
        proba = np.array([1.0 if i == pred_enc else 0.0
                          for i in range(len(bundle['clases']))])

    confianza = {cls: round(float(p), 4)
                  for cls, p in zip(bundle['clases'], proba)}
    max_conf = max(confianza.values())

    if max_conf < 0.50:
        advertencias.append(f'Confianza baja ({max_conf:.0%}) - resultado incierto')

    # -- Señales auxiliares --
    t_lower = texto_observacion.lower()
    hits_neg = sum(1 for kw in bundle['keywords_neg'] if kw in t_lower)
    hits_pos = sum(1 for kw in bundle['keywords_pos'] if kw in t_lower)
    v_score = SIA.polarity_scores(texto_observacion)['compound']

```

Nota. Elaboración propia.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 48

Construcción del resultado e interpretación de la predicción del modelo

```
# --- Señales auxiliares ---
t_lower = texto_observacion.lower()
hits_neg = sum(1 for kw in bundle['keywords_neg'] if kw in t_lower)
hits_pos = sum(1 for kw in bundle['keywords_pos'] if kw in t_lower)
v_score = SIA.polarity_scores(texto_observacion)['compound']

# urgencia = bool(PATRON_URGENCIA.search(texto_observacion))
urgencia = False # Temporarily set to False as PATRON_URGENCIA is undefined

resultado = {
    'sentimiento': pred_lbl,
    'confianza': confianza,
    'vader_score': round(v_score, 4),
    'hits_neg': hits_neg,
    'hits_pos': hits_pos,
    'urgencia': urgencia,
    'n_tokens': n_tokens,
    'advertencias': advertencias,
}

#Reporte visual
if verbose:
    ICONS = {'Positive': '😊', 'Negative': '😡', 'Neutro': '😐'}
    SEP = '-' * 60
    print(f'\n(SEP)')
    print(f' {ICONS.get(pred_lbl, "")} RESULTADO: {pred_lbl.upper()}')
    print(SEP)
    txt_display = texto_observacion[:180] + ('...' if len(texto_observacion) > 180 else '')
    print(f' Texto : {txt_display}')
    print(f' Tokens : {n_tokens} | VADER: {v_score:+.3f} | '
          f'KW-neg: {hits_neg} KW-pos: {hits_pos}'
          + (' 🚨 URGENTE' if urgencia else ''))
    print(f' {"-"*56}')
    print(f' Confianza por clase:')
    for cls in sorted(confianza, key=confianza.get, reverse=True):
        p = confianza[cls]
        bar = '█' * int(p * 30)
        mark = '▲' if cls == pred_lbl else ''
        print(f' {cls:10s} {p:5.1%} {bar}{mark}')
    if advertencias:
        print(f' {"-"*56}')
        for adv in advertencias:
            print(f' 🚨 {adv}')
    print(SEP)

return resultado

print('📦 Función predecir_pqr() lista')
print('👉 Uso: predecir_pqr("texto del PQR")')
```

Nota. Elaboración propia.

Lo que se observa en las dos figuras es el proceso, el cual inicia con la recepción del texto de la PQRS, pasa por una validación de que no esté vacío, se limpia el texto, se convierte a números con TF-IDF. Posteriormente se aplica el modelo con un .pkl que predice los sentimientos, una vez realizado, se calculan probabilidades (confianza), se agregan datos adicionales (Keyword, VADER, urgencia), y así se genera el resultado final.

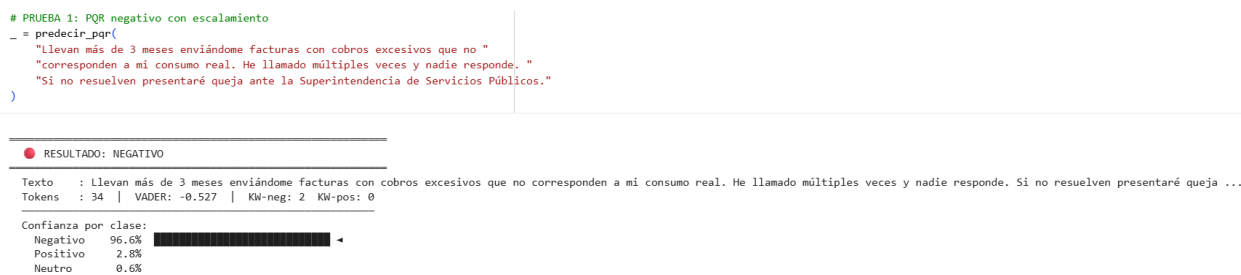
A continuación, en las figuras 48, 49 y 50 muestran la fase de pruebas para evaluar cómo es el comportamiento del modelo ante la aplicación real de comentarios que pueden llegar en las PQRS.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 49

Ejemplo de clasificación de PQRS con sentimiento negativo



Nota. Elaboración propia.

Antes de detallar los resultados, al leer el comentario que llegó a la empresa por medio de sus canales de atención, se puede inferir que es un comentario que evidencia un descontento debido a cobros excesivos que, según el cliente, no corresponden al consumo real. Ahora bien, ya viendo la aplicación del modelo en la figura 48, se observa una clasificación negativa, lo cual es coherente con el contenido del mensaje. Lo anterior se respalda con el nivel de confianza que está sobre el 97% lo cual indica la certeza sobre la categoría asignada.

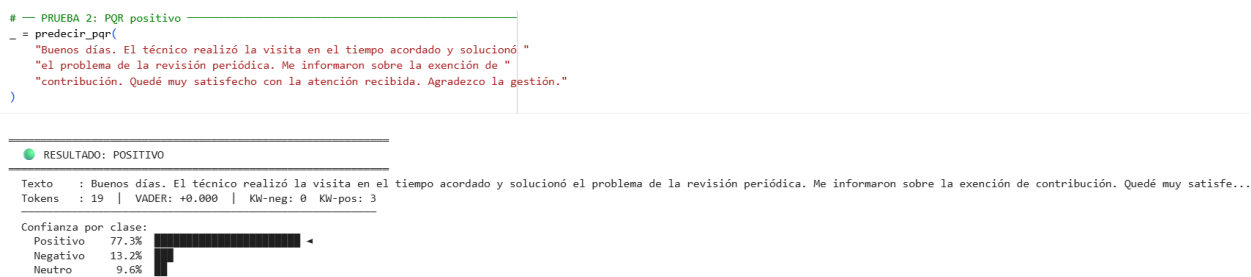
En la figura 50 se muestra otra prueba de la implementación del modelo, esta vez, como contraste con la anterior, se va a aplicar un comentario positivo.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 50

Ejemplo de clasificación de PQRS con sentimiento positivo



Nota. Elaboración propia.

En la figura 49 se observa un comentario en el cual el cliente manifiesta una satisfacción respecto a una visita para una revisión periódica. El modelo hace una clasificación de positivo con un nivel de confianza del 77%, lo cual también es consistente con el contenido del mensaje. Ahora bien, más allá del resultado que nos arroja el modelo, también es posible ver datos como el número de tokens, el cual nos muestra que tiene una longitud suficiente para permitir una calificación adecuada. También se observa la clasificación de los keywords positivos y negativos, estos corresponden al número de palabras clave negativas y positivas detectadas en el texto. En este caso, el modelo calificó en cero los negativos y tres los positivos.

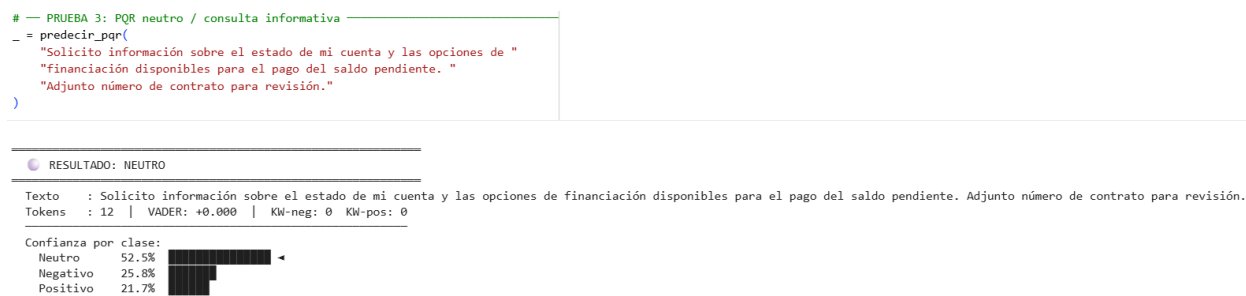
Siguiendo con las pruebas, a continuación, en la figura 50, se va a analizar una petición de un cliente respecto a una consulta informativa, en la cual no se evidencia una carga emocional explícita.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Figura 12

Ejemplo de clasificación de PQRS con sentimiento neutro



Nota. Elaboración propia.

Como se mencionaba anteriormente, en este ejemplo no se evidencia una carga emocional ya que se trata de una solicitud en donde no hay señales claras de positividad o negatividad. El modelo la identificó como “neutro” lo cual es consistente con la naturaleza del mensaje. Para este caso el nivel de confianza asociada a la predicción fue del 53%, siendo superior al de la clase negativa (25.8%) y positiva (21.7%). Estos resultados reflejan un nivel de ambigüedad en el contenido del mensaje, sin embargo, el modelo identificó adecuadamente la categoría.

En conclusión, con las pruebas expuestas anteriormente se evidencia una aplicabilidad del modelo en un entorno real, en donde se puso a prueba su capacidad para clasificar automáticamente el sentimiento de las PQRS a partir de texto libre. Es importante recalcar que, a lo largo de esta fase de implementación, se logró integrar el modelo de aprendizaje automático con los procesos de limpieza, transformación y análisis

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

del lenguaje, con lo cual se garantiza que las nuevas observaciones sean tratadas bajo las mismas condiciones que los datos de entrenamiento. De igual manera, para facilitar el uso por parte de usuarios no técnicos, se incorporaron de elementos interpretativos como niveles de confianza, identificación de palabras clave y detección de señales de urgencia que fortalecen la comprensión de los resultados.

Operación del modelo

De acuerdo con lo mencionado anteriormente en la descripción de la implementación operativa del modelo, este fue diseñado para dar alertas y permitir identificar los niveles de confianza altos, sin embargo, es importante establecer umbrales de intervención humana basados en los niveles de confianza ya mencionados, los cuales facilitan un equilibrio entre automatización y criterio profesional, entendiendo que el modelo actúa como una herramienta de apoyo a la decisión y no como un mecanismo autónomo de gestión de clientes. En este contexto, las PQRS clasificadas con niveles de confianza altos serán priorizadas automáticamente según el sentimiento identificado, mientras que aquellas predicciones con niveles de confianza bajos o moderados serán derivadas obligatoriamente a revisión humana antes de cualquier acción operativa.

El umbral que se va a manejar respecto al nivel de confianza será del 70%, las clasificaciones con niveles de confianza inferior se revisarán de manera manual por los analistas debido a la mayor probabilidad de ambigüedad semántica o interpretación

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

incorrecta del contexto. En contraparte, con las clasificadas por encima del umbral se generarán alertas prioritarias para revisión inmediata, especialmente en casos donde se detecten términos asociados con urgencia, inconformidad recurrente o posibles riesgos reputacionales para la organización.

Asimismo, el modelo permitiría identificar tendencias recurrentes de insatisfacción dentro de grupos de clientes, facilitando la generación de reportes y alertas para apoyar la toma de decisiones estratégicas y el mejoramiento continuo del servicio. En consecuencia, el modelo no se plantea como un mecanismo autónomo de decisión, sino como una herramienta de apoyo analítico orientada a fortalecer la capacidad de respuesta del área y optimizar la gestión de las PQRS mediante el uso de técnicas de procesamiento de lenguaje natural y machine learning.

Plan de intervención

La estrategia para el plan de intervención se basa en un enfoque de implementación progresiva en tres dimensiones. La tecnológica, mediante la adopción del modelo de análisis de sentimientos, integración con el sistema de PQRS; La humana, en la cual se encuentra el acompañamiento y gestión del cambio con los colaboradores del área de servicio al cliente; la organizacional, mediante la incorporación del análisis de sentimientos para el apoyo en la toma de decisiones y en los informes de gestión de calidad.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

En la tabla 3 se muestra el plan, abordando los objetivos que busca desarrollar, las actividades principales a desarrollar para conseguir el objetivo, responsables, duración estimada y los indicadores para hacer seguimiento.

Tabla 3

Plan de intervención

Objetivo específico	Actividad principal	Responsable	Duración estimada	Indicadores de logro
Fortalecer capacidades del personal de atención	Realizar talleres de capacitación sobre interpretación de resultados y uso de dashboards.	Dirección de servicio al Cliente.	2 semanas	80 % de colaboradores capacitados.
Optimizar el proceso de priorización de PQRS	Incorporar el análisis de polaridad en el flujo de clasificación de casos.	Dirección de servicio al Cliente.	1 mes	15 % de reducción en tiempo de respuesta
Monitorear el desempeño del modelo	Evaluar métricas del modelo y ajustes de precisión	Analista de datos.	Continuo	Reporte mensual de precisión y estabilidad del modelo
Promover sostenibilidad del sistema	Documentar procesos, establecer rutinas de mantenimiento y actualización	Analista de datos.	1 mes.	Manual técnico y protocolo de mantenimiento entregados

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Nota. Elaboración propia.

Como se observa en la tabla el plan de intervención se estructura basándose en una asignación de responsabilidades, estimación de tiempos y definición de indicadores que facilitan el seguimiento y la evaluación de impacto del modelo. Este plan se estructuró contemplando no solamente las actividades relacionadas al desarrollo del modelo como herramienta de apoyo en la toma de decisiones, sino a las demás actividades posteriores como el fortalecimiento del personal inmerso en el proceso de atención al cliente y gestión de las PQRS. A continuación, se explorarán los recursos requeridos para el desarrollo óptimo del plan de intervención.

Cronograma de implementación

Con el fin de establecer una ruta organizada para la implementación de la propuesta planteada, se diseñó un cronograma general de actividades que contempla las principales etapas relacionadas con la preparación de datos, entrenamiento del modelo, validación, pruebas funcionales y puesta en marcha de la herramienta dentro del área de Servicio y Experiencia al Cliente de Vanti S.A. E.S.P.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Tabla 4

Cronograma de actividades

Actividad	Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6
Recolección y consolidación de datos	X	X				
Limpieza y preprocesamiento de textos		X	X			
Construcción de variable objetivo		X	X			
Vectorización TF-IDF			X			
Entrenamiento y evaluación de modelos			X	X		
Selección del modelo óptimo				X		
Desarrollo de pruebas funcionales				X	X	
Diseño de visualización/reportes					X	

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Capacitación de usuarios	X
Implementación piloto	X
Seguimiento y monitoreo inicial	X

Nota. Elaboración propia.

Como se observa en la tabla 4 el cronograma desarrollado contempla una serie de actividades que van desde la recolección de datos al seguimiento y monitoreo del modelo ya puesto en marcha. Se estima una duración del proyecto de 6 semanas, empezando en la primera con actividades que requieren de tiempos para verificar consistencia de la información, posteriormente en la segunda semana se realizan actividades de limpieza y construcción de la variable objetivo, estas actividades duran dos semanas, este tiempo se justifica en la necesidad de validar la coherencia de las clasificaciones obtenidas antes del entrenamiento de los modelos. En la semana tres ya se empieza con las actividades de transformación de texto en representaciones numéricas, el tiempo requerido es relativamente corto debido a que corresponde a una fase automatizada dentro del flujo de procesamiento. Luego de esto, en la semana cuatro se realiza el entrenamiento de los diferentes modelos de clasificación y la evaluación de sus métricas de desempeño. El tiempo asignado responde a la necesidad de ejecutar múltiples pruebas, comparar

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

resultados y validar la estabilidad de los modelos mediante validación cruzada. Sobre la semana cinco se desarrollan mecanismos de visualización de resultados y generación de reportes para facilitar la interpretación de la información obtenida, paralelamente se realizan capacitaciones y socializaciones sobre el funcionamiento de la herramienta a los colaboradores que utilizarán el sistema. Finalmente, en la semana seis, se realiza la supervisión inicial del comportamiento del modelo en operación, incluyendo revisión de resultados, validación de clasificaciones y detección de oportunidades de mejora.

Recursos requeridos

Los recursos necesarios para que el proyecto tenga los efectos esperados, se agrupan principalmente en tres categorías. En la tabla 5 se mostrará un detalle de ellas:

Tabla 5

Recursos requeridos

Tipo de recurso	Descripción
Humanos	Equipo de analítica de datos, personal de TI, personal de servicio al cliente.
Tecnológicos	Servidor interno o entorno en la nube, licencias de Power BI, librerías de Python, almacenamiento seguro de datos.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Financieros	Inversión en horas de capacitación y soporte técnico
-------------	--

Nota. Elaboración propia.

Para efectos de simplicidad y eficiencia se encontró pertinente agrupar el requerimiento de recursos en las tres categorías mencionadas ya que la implementación del modelo va más allá del desarrollo técnico y requiere una articulación integral entre personas, procesos y tecnología. Los primeros son quienes garantizan el diseño, despliegue y mantenimiento del modelo analítico; los segundos, constituyen la base para el funcionamiento seguro, escalable y eficiente del sistema ya que permiten el procesamiento de los datos, la visualización dinámica de los resultados; y los terceros, respaldan la sostenibilidad del plan al cubrir la inversión en capacitación y soporte técnico.

Parámetros para evaluar el éxito del modelo en producción

Evaluar el desempeño del modelo una vez implementado, resulta un factor estratégicamente de alta importancia para el proyecto, es por esto que es necesario plantearse una serie de parámetros orientados a medir tanto su efectividad técnica como su aporte al proceso organizacional de gestión de PQRS. A continuación, en la tabla 6 se muestra la propuesta de actividades identificadas como parte de los parámetros de evaluación del éxito del modelo.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Tabla 6

Parámetros de evaluación para el seguimiento del modelo

Parámetro de evaluación	Objetivo
Tiempo promedio de análisis	Reducir tiempos operativos
Volumen de PQRS procesadas	Incrementar capacidad de análisis
Concordancia analista-modelo	Evaluar confiabilidad
Casos críticos detectados	Mejorar priorización
Nivel de adopción del sistema	Medir aceptación organizacional

Nota. Elaboración propia.

Como se observa en la tabla, uno de los principales indicadores corresponde a la reducción en los tiempos de análisis de las solicitudes, comparando el tiempo requerido en el proceso manual (anteriormente se mencionó que es aproximadamente de dos horas) frente al apoyo generado por el modelo automatizado, esto permitirá evaluar el impacto de la herramienta sobre la eficiencia operativa del área.

De igual manera, se propone medir la capacidad del sistema para procesar grandes volúmenes de información textual de manera consistente y oportuna, especialmente en escenarios de alta demanda de solicitudes por parte de los clientes industriales.

Otro parámetro relevante será el nivel de concordancia entre la clasificación realizada por el modelo y la validación efectuada por los analistas del área, con el fin de

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

monitorear la precisión y confiabilidad de la herramienta durante su utilización en condiciones reales.

Finalmente, se considerará el nivel de adopción y aceptación de la herramienta por parte de los colaboradores involucrados en la gestión de PQRS, entendiendo que el objetivo del sistema no es reemplazar el criterio humano, sino fortalecer la capacidad analítica y operativa del área mediante una herramienta de apoyo a la decisión.

Ahora bien, estratégicamente para el proyecto, representa de igual importancia la medición de los parámetros de evaluación el éxito del modelo en producción, como el análisis de riesgos y medidas de mitigación, a continuación, se explicará la estrategia que se desarrollará para abordar el segundo ítem.

Análisis de Riesgos y medidas de mitigación

Con el propósito de anticipar y gestionar los posibles factores que podrían afectar la ejecución del plan de acción propuesto, se plantea la implementación de una matriz de riesgos como herramienta de apoyo a la gestión del proyecto. Con esta se busca identificar, analizar y priorizar los riesgos en función de su probabilidad de ocurrencia y el impacto potencial sobre los objetivos dentro de la implementación del modelo en los procesos de gestión de PQRS de Vanti.

De acuerdo con el enfoque propuesto por el Project Management Institute los riesgos deben evaluarse considerando su probabilidad de ocurrencia y el impacto que

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

pueden generar sobre el alcance, el tiempo, los costos y la calidad del proyecto (Project Management Institute, 2021). En este sentido, el uso de matrices de impacto facilita la priorización de riesgos críticos y la definición de estrategias preventivas, lo cual resulta especialmente relevante en proyectos de analítica y transformación digital. A continuación, en la tabla 7, se presenta la matriz desarrollada en esta ocasión para evaluar los riesgos.

Tabla 7

Matriz de riesgos

Id	Riesgo Identificado	Tipo De Riesgo	Probabilidad	Impacto	Nivel De Riesgo	Estrategia De Mitigación
R1	Falta de adopción del modelo por parte del personal de servicio al cliente.	Organizacional	Media	Alta	Alto	Programas de capacitación, talleres y acompañamiento en el uso de dashboards.
R2	Baja calidad de los datos de PQRS (valores faltantes, ruido textual)	Técnico	Media	Alta	Alto	Aplicar estrategias de limpieza, imputación y validación continua de datos.
R3	Limitaciones en la infraestructura tecnológica disponible.	Tecnológico	Media	Media	Medio	Uso de entornos en la nube o servidores escalables según demanda.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

R4	Disminución del desempeño del modelo con el tiempo (model drift).	Analítico	Alta	Media	Alto	Monitoreo periódico de métricas y reentrenamiento del modelo.
R5	Insuficiencia de recursos financieros para capacitación o soporte.	Financiero	Baja	Media	Medio	Planificación presupuestal y priorización de actividades críticas.
R6	Riesgos de seguridad y confidencialidad de la información.	Legal / Seguridad	Baja	Alta	Medio	Implementar controles de acceso y almacenamiento seguro de datos.

Nota. Elaboración propia.

Con la aplicación de la matriz de impacto de riesgos se pudo identificar de manera estructurada los principales factores que podrían afectar la implementación del modelo de análisis de sentimientos en los procesos de gestión de PQRS de Vanti S.A. ESP. Los resultados evidencian que los riesgos de mayor criticidad se concentran en aspectos organizacionales, técnicos y analíticos, particularmente en la adopción del modelo por parte del personal, la calidad de los datos y la pérdida de desempeño del modelo en el tiempo. Gracias a herramientas como esta se pueden establecer de estrategias de mitigación orientadas a la capacitación, al fortalecimiento de la infraestructura tecnológica

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

y al monitoreo continuo del modelo contribuye a reducir significativamente su impacto potencial.

Análisis DOFA de la solución propuesta

A partir de los resultados obtenidos en el análisis de datos y el desempeño del modelo desarrollado, se procede a realizar un análisis DOFA con el fin de identificar las principales oportunidades, debilidades, fortalezas y amenazas asociadas a la implementación de la solución propuesta.

Tabla 8

Análisis DOFA

DEBILIDADES	OPORTUNIDADES
- Calidad y estructura de los datos: Los comentarios que se toman para el modelo pueden contener errores, ambigüedad o abreviaciones.	- Adopción de analítica avanzada en Vanti: Corporativamente la empresa está migrando hacia modelos basados en datos para mejorar la experiencia del cliente.
- Desbalance de clases: La proporción elevada de la clase neutra puede afectar la capacidad del modelo para identificar clases minoritarias.	- Disponibilidad de herramientas open source: Librerías como scikit-learn, spaCy, NLTK y TensorFlow reducen costos de implementación.
- Superposición semántica: La visualización con t-SNE evidencia que los textos no son completamente separables, lo que limita el desempeño del modelo.	- Mejora en la toma de decisiones estratégicas: El proyecto propende la identificación de patrones de insatisfacción, riesgos reputacionales y oportunidades de mejora.
FORTALEZAS	AMENAZAS
- Alto desempeño del modelo SVM	- Cambios en el lenguaje de los clientes: Nuevas expresiones, modismos o formas de comunicación pueden afectar el desempeño del modelo con el tiempo.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

- Enfoque híbrido que combina técnicas estadísticas, con reglas heurísticas y análisis léxico.

- Dependencia de la calidad de datos de entrada: Información incompleta o mal registrada puede afectar el desempeño del modelo

- Conocimiento del negocio: El entendimiento del proceso de gestión de PQRS facilita la interpretación de resultados y su aplicación práctica.

- Aspectos de privacidad y manejo de datos: Vanti al ser tan rigurosa con el uso de información de clientes, debe cumplir con normativas de protección de datos.

Nota. Elaboración propia.

En términos generales, según se observa en la tabla 8 la solución propuesta presenta condiciones favorables para su implementación dentro del proceso de gestión de PQRS de Vanti S.A. E.S.P., especialmente debido al desempeño obtenido por el modelo desarrollado, la disponibilidad de herramientas tecnológicas y la orientación organizacional hacia el uso estratégico de los datos. No obstante, también se evidencian desafíos asociados principalmente a la calidad y variabilidad de la información textual, por ejemplo, al observar la parte de la implementación del modelo, se observa la clasificación errónea de una proporción de PQRS negativas dentro de la categoría neutral.

Particularmente, la matriz de confusión normalizada del modelo Linear SVM evidencia que aproximadamente el 21% de los registros realmente negativos son clasificados como neutros. Desde una perspectiva operativa, este comportamiento representa un riesgo importante, debido a que algunos casos asociados a inconformidades o situaciones de insatisfacción podrían no ser priorizados adecuadamente dentro del proceso de atención.

En términos prácticos, esto implica que una parte de las comunicaciones que requieren

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

atención prioritaria podrían pasar inicialmente desapercibidas si el sistema fuera utilizado de manera completamente automatizada.

Para mitigar lo anterior, es importante hacer la claridad de que el modelo planteado no se concibe como un mecanismo autónomo de toma de decisiones, sino como una herramienta de apoyo para los analistas. Bajo este enfoque, los resultados generados por el sistema deben complementarse con validación humana, especialmente en aquellos casos con niveles de confianza intermedios o con presencia de indicadores asociados a urgencia o lenguaje crítico.

Roles, responsabilidades y gobernanza del sistema

Con el propósito de garantizar el adecuado funcionamiento y sostenibilidad del modelo de análisis de sentimientos dentro del proceso de gestión de PQRS, se plantea una estructura básica de roles y responsabilidades orientada al monitoreo, supervisión y uso adecuado de la herramienta dentro del área de Servicio y Experiencia al Cliente de Vanti S.A. E.S.P. A continuación, en la tabla 7 se muestra de manera estructurada la propuesta de asignación para cada agente implicado en el proceso.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.

ESP

Tabla 9

Definición de roles

Cargo	Rol dentro de la ejecución del modelo
Analista de Servicio y Experiencia al Cliente	Utilizar el modelo como herramienta de apoyo para la clasificación, revisión y priorización de PQRS, así como validar casos ambiguos o de baja confianza en las predicciones.
Jefe de Servicio y Experiencia al Cliente	Supervisar el desempeño general de la herramienta, realizar seguimiento a indicadores operativos y apoyar la toma de decisiones estratégicas derivadas del análisis de sentimientos.
Analista de Datos / Científico de Datos	Realizar monitoreo técnico del modelo, evaluar métricas de desempeño, ejecutar ajustes y apoyar procesos de actualización o reentrenamiento del sistema.
Área de Tecnología o TI	Garantizar la infraestructura tecnológica necesaria para la operación del modelo, incluyendo almacenamiento, conectividad y soporte técnico.
Líder de Transformación Digital o Innovación	Promover la adopción de herramientas analíticas dentro de la organización y apoyar iniciativas de mejora continua relacionadas con automatización y analítica de datos.
Dirección de la Vicepresidencia de Grandes Clientes	Evaluar el impacto estratégico de la herramienta y apoyar la toma de decisiones relacionadas con continuidad, escalabilidad y fortalecimiento del sistema.

Como se observa en la tabla 9, en primer lugar, los analistas encargados de la gestión de PQRS tendrían la responsabilidad de utilizar el sistema como herramienta de apoyo. Por su parte, el líder o jefe del área tendrá funciones asociadas al seguimiento y toma de decisiones estratégicas derivadas de los resultados obtenidos mediante el análisis de sentimientos. Adicionalmente, se propone que el área de analítica de datos sea responsable del mantenimiento y actualización del modelo. Ahora bien, es necesario

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

precisar que la implementación del modelo trasciende el componente analítico y requiere apoyo institucional para asegurar su sostenibilidad y apropiación dentro de la organización, es por lo que se incluyen roles asociados a tecnología, innovación y dirección estratégica.

Ahora bien, respecto a la gobernanza del sistema, lo que se busca con ella es mitigar riesgos asociados a interpretaciones erróneas y posibles sesgos en la clasificación de sentimientos. Lo primera medida es que el tratamiento de la información debe cumplir con principios de confidencialidad y protección de datos personales, garantizando el uso responsable de los registros utilizados durante la operación del modelo. De igual forma, se establece que el modelo debe operar bajo supervisión humana, entendiendo que las clasificaciones generadas corresponden a un mecanismo de apoyo a la decisión y no a un reemplazo de los colaboradores encargados de la gestión de PQRS.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Conclusiones y Recomendaciones

A continuación, se presentan las conclusiones de la intervención desarrollada en la empresa, así como las recomendaciones para la implementación del plan de intervención propuesto.

Conclusiones

El desarrollo del presente proyecto responde a la pregunta planteada al principio de la investigación, en donde se cuestionaba sobre la posibilidad de lograr por medio de machine learning con técnicas de procesamiento de lenguaje natural el análisis de sentimientos en las PQRS que gestiona el área servicio y experiencia en la vicepresidencia de Grandes Comercios. En el desarrollo del proyecto se pudo demostrar que si se puede aplicar transformando información textual no estructurada en conocimiento útil para la toma de decisiones.

En primer lugar, se comprobó que la calidad de los datos constituye un factor determinante en el desempeño de los modelos. El proceso de limpieza y preprocesamiento aplicado permitió reducir significativamente el ruido presente en los textos, mejorando la representación de la información y facilitando la identificación de patrones relevantes.

Asimismo, la construcción de la variable objetivo mediante un enfoque híbrido construido a partir de la combinación de tres capas (diccionario ajustado, análisis de

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

sentimientos con VADER, detección de urgencias) resultó fundamental para el entrenamiento de los modelos de clasificación.

En cuanto al modelado, se evaluaron diferentes algoritmos de machine learning, identificando que el modelo que ofrece mejor desempeño en términos de métricas como como F1-score macro con 0.8382 y estabilidad en validación cruzada con un rendimiento de 0.8274, es el Linear SVM. Esto confirma su idoneidad para trabajar con datos de alta dimensionalidad derivados de técnicas como TF-IDF. No obstante, se descubrió que, dada la complejidad inherente del lenguaje natural, existen diversos retos para los modelos de clasificación, como el desbalance en la distribución de las clases y la superposición semántica entre categorías de sentimiento. Lo anterior dificulta la diferenciación precisa entre textos positivos, negativos y neutros, especialmente en aquellos casos donde el contexto, la ambigüedad del lenguaje o la subjetividad de las expresiones influyen en la interpretación del mensaje.

A pesar de estas limitaciones, los resultados obtenidos evidenciaron que el modelo seleccionado logró un desempeño satisfactorio y consistente, incorporando elementos de interpretación como niveles de confianza, detección de palabras clave y análisis complementario mediante VADER, lo que facilita su comprensión y uso en contextos organizacionales.

En conjunto, los resultados obtenidos permiten concluir que el modelo desarrollado no solo cumple con el objetivo de clasificar sentimientos, sino que constituye una

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

herramienta funcional que puede aportar valor en la gestión de las PQRS, especialmente en la identificación de casos críticos y en la generación de información estratégica para la organización.

Finalmente, respecto a los objetivos que se plantearon en el desarrollo del proyecto, se evidencia un cumplimiento en cada uno de los ítems, por ejemplo, respecto al primer objetivo, se concluye que se consolidó un marco teórico con el cual se fundamentó conceptualmente la investigación desde los campos de la Ciencia de Datos. El segundo objetivo buscaba realizar un diagnóstico de la situación actual frente a la gestión de PQRS, en consecuencia, se observa que se pudo entender las fortalezas, debilidades, oportunidades y amenazas que presenta el área de servicio y experiencia. Por último, en el objetivo tres y cuatro que buscaban tanto el desarrollo del modelo como el plan de implementación al interior de la vicepresidencia, se observa, como ya se mencionó con anterioridad, que el modelo fue desarrollado satisfactoriamente y se establecieron los lineamientos para la implementación al interior de la organización. En consecuencia, se concluye que el modelo no solo cumple una función analítica, sino que se presenta como una herramienta útil para apoyar la toma de decisiones estratégicas y monitorear la experiencia del cliente, lo cual genera valor para la organización.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

Recomendaciones

En principio, antes de tratar las recomendaciones derivadas al trabajo que se realizó para Vanti, es importante mencionar una serie de recomendaciones de carácter académico, así como futuras líneas de investigación. Primero que todo, sería de gran valor la comparación entre el desempeño de modelos tradicionales de machine learning, como los que se abordaron en el presente proyecto, con arquitecturas más avanzadas como BERT o modelos generativos, con el fin de evaluar posibles mejoras en precisión de las clasificaciones. De igual forma, sería interesante validar el enfoque metodológico en otros sectores, como banca, telecomunicaciones o salud, con el fin de medir la generalización del modelo y comparar comportamientos lingüísticos entre industrias.

Ahora bien, centrados en los resultados del modelo aplicado a la empresa, a partir de los resultados obtenidos en el desarrollo del proyecto, en primer lugar, se recomienda implementar el modelo en un entorno productivo e integrarlo con herramientas de visualización como Power BI, con el fin de transformar las predicciones en indicadores estratégicos que apoyen la toma de decisiones dentro del área de servicio al cliente.

Adicionalmente, se sugiere realizar un proceso de actualización periódica del modelo mediante la incorporación de nuevos datos, lo cual permitirá mejorar su capacidad de generalización y adaptarse a cambios en el comportamiento y lenguaje de los usuarios.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

En relación con las limitaciones identificadas, se recomienda explorar técnicas para el manejo del desbalance de clases, como métodos de muestreo o ajuste de pesos en los modelos, con el objetivo de mejorar el desempeño en las categorías menos representadas.

Asimismo, se recomienda fortalecer el diccionario de palabras clave y las reglas heurísticas utilizadas, incorporando conocimiento adicional del negocio, lo cual puede contribuir a mejorar la interpretabilidad y precisión del sistema.

Finalmente, se sugiere promover el uso de soluciones analíticas dentro de la vicepresidencia, con el fin de fomentar una cultura basada en datos que permita aprovechar herramientas como la desarrollada en este proyecto para optimizar procesos y mejorar la experiencia del cliente.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Referencias

- Bernal Torres, C. (2022). *Metodología de la investigación: Administración, economía, humanidades y ciencias sociales* (1ª ed. colombiana). Pearson Educación.
- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *COMPSTAT*.
- Breiman, L. (2001). *Random Forests*. Machine Learning, 45(1), 5–32.
- Chen, T., & Guestrin, C. (2016). XGBoost: *A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.
- Congreso de la República de Colombia. (2012). Ley 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales.
<https://www.funcionpublica.gov.co/>
- Congreso de la República de Colombia. (2015). Ley 1755 de 2015. Por medio de la cual se regula el derecho fundamental de petición y se sustituye un título del Código de Procedimiento Administrativo y de lo Contencioso Administrativo.
<https://www.funcionpublica.gov.co/>.
- Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Machine Learning, 20(3), 273–297.
- DataCamp. (2024). Cross-validation in machine learning.
<https://www.datacamp.com/tutorial/cross-validation-machine-learning>.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

e-Contact. (2025). *Cómo el análisis de sentimiento optimiza la experiencia del cliente.*

Recuperado de <https://www.e-contact.cl/analisis-sentimiento-centros-experiencia/>.

Elsevier. (2026). *Scopus*. <https://www.scopus.com>.

Géron, A. (2023). *Aprende machine learning con Scikit-Learn, Keras y TensorFlow* (3.^a ed.). Anaya Multimedia.

Grupo Vanti. (s.f.). Quiénes somos. <https://www.grupovanti.com/conocenos/quienes-somos>.

Hutto, C. J., & Gilbert, E. (2014).

VADER: A parsimonious rule-based model for sentiment analysis of social media text.

Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216–225.

IBM. (s.f.). ¿Qué es el procesamiento del lenguaje natural (NLP)?

<https://www.ibm.com/mx-es/think/topics/natural-language-processing>

Instituto Nacional de Estadística (Uruguay). (s. f.). *¿Qué es un censo? Preguntas*

frecuentes. [https://www.gub.uy/instituto-nacional-](https://www.gub.uy/instituto-nacional-estadistica/comunicacion/publicaciones/preguntas-frecuentes/preguntas-frecuentes/es-censo)

[estadistica/comunicacion/publicaciones/preguntas-frecuentes/preguntas-frecuentes/es-censo](https://www.gub.uy/instituto-nacional-estadistica/comunicacion/publicaciones/preguntas-frecuentes/preguntas-frecuentes/es-censo).

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

International Organization for Standardization. (2023). *ISO/IEC 42001:2023 information technology — Artificial intelligence — Management system*. [ISO official page for ISO/IEC 42001](#)

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer.

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed. draft). Stanford University.

Kuhn, M., & Silge, J. (2022). *Tidy modeling with R*. O'Reilly Media.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Monje Álvarez, C. A. (2011). *Metodología de la investigación cuantitativa y cualitativa: Guía didáctica*. Universidad Surcolombiana.
<https://www.uv.mx/rmipe/files/2017/02/Guia-didactica-metodologia-de-la-investigacion.pdf>

Pineda, C. M. (2021). *Aprendizaje automático y profundo en Python: Una mirada hacia la inteligencia artificial*. Ediciones de la U.

Powers, D. M. W. (2011). *Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation*. Journal of Machine Learning Technologies, 2(1), 37–63.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Project Management Institute. (2021). *A guide to the project management body of knowledge (PMBOK® Guide)* (7th ed.). PMI.

Ramos, J. (2003). *Using TF-IDF to determine word relevance in document queries. Proceedings of the First Instructional Conference on Machine Learning.*

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf>

Raschka, S., Liu, Y. (Hayden), & Mirjalili, V. (2022). *Machine learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python.* Packt Publishing.

Stryker, C., & Holdsworth, J. (2024, agosto 11). *¿Qué es el PLN (procesamiento del lenguaje natural)?* IBM. <https://www.ibm.com/es-es/think/topics/natural-language-processing>

Talking with Data. (2023). Validación cruzada con R y Python: Asegura el éxito de tus modelos de machine learning. Medium.

<https://talkingwithdata.medium.com/asegura-el-%C3%A9xito-de-tus-modelos-de-machine-learning-el-poder-de-la-validaci%C3%B3n-cruzada-701e7635e299>

Uribe González, A. (2023). *Análisis de sentimientos en tiempo real para intervenciones de chat en videoconferencias por streaming* [Trabajo de grado, Universidad de Antioquia]. Repositorio Institucional Universidad de Antioquia.

<https://bibliotecadigital.udea.edu.co/>

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

Vanti S.A. E.S.P. (2022). Presentación unidad de dirección grandes clientes [Archivo PDF]. Documento interno no publicado.

Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2021). From multi-channel retailing to omni-channel customer experience management. *Journal of Retailing*, 97(2), 174–181.

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A. ESP

A. Anexo. Encuesta sobre manejo de PQRS

La presente encuesta forma parte de una investigación académica desarrollada en el marco de una tesis de maestría en Ciencia de Datos. Su propósito es recopilar información sobre el proceso de atención de PQRS en la Vicepresidencia de Grandes Clientes de Vanti S.A. ESP, con el fin de identificar oportunidades de mejora e incorporar herramientas de análisis de sentimientos.

Las respuestas recolectadas serán utilizadas exclusivamente con fines académicos y de investigación, sin ningún tipo de evaluación o repercusión sobre el desempeño de los participantes. Los datos serán tratados de forma confidencial y anonimizados, garantizando que no sea posible identificar a las personas que respondan el cuestionario. Su participación es voluntaria y contribuirá al desarrollo de soluciones basadas en analítica que fortalezcan la calidad del servicio al cliente.

¿Por cuáles canales recibe usted con mayor frecuencia las comunicaciones o PQRS de los clientes industriales? *(puede marcar más de una opción)*

- ☐ Llamadas telefónicas
- ☐ Correos electrónicos
- ☐ Mensajes de WhatsApp
- ☐ Formulario web o portal de autogestión

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

☐ Oficios o documentos físicos

☐ Otro (especifique): _____

En promedio, ¿cuántas solicitudes o comunicaciones de clientes atiende semanalmente?

☐ Menos de 10

☐ Entre 10 y 30

☐ Entre 31 y 60

☐ Más de 60

Cuando un cliente se comunica por llamada telefónica, ¿cómo se registra lo que expresa durante la conversación?

☐ Escribo un resumen en mis propias palabras

☐ Transcribo literalmente las frases del cliente

☐ Uso frases predeterminadas o plantillas

☐ Otro (especifique): _____

Cuando recibe comunicaciones por correo electrónico o WhatsApp, ¿qué tratamiento le da al mensaje?

☐ Copio y pego el texto tal como lo envía el cliente

☐ Realizo una síntesis del contenido principal

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

☐ Adapto el texto al formato de registro interno

☐ No hay un procedimiento definido

¿En qué formato o sistema se registran actualmente las PQRS?

☐ Sistema interno (CRM o aplicativo empresarial)

☐ Hoja de cálculo (Excel u otro)

☐ Documento Word / correo interno

☐ Otro (especifique): _____

¿Existen lineamientos o protocolos que indiquen cómo redactar o registrar la información del cliente?

☐ Sí, existen y se aplican siempre

☐ Sí, pero no se aplican de manera constante

☐ No existen lineamientos formales

☐ No sabe / no responde

¿La empresa realiza actualmente algún tipo de análisis del contenido o tono emocional de los mensajes de los clientes?

☐ Sí, de forma manual por los analistas

☐ Sí, mediante alguna herramienta automatizada

Desarrollo de modelos de procesamiento de lenguaje natural para el análisis de sentimientos de PQRS de clientes industriales de Vanti S.A.
ESP

☐ No, solo se clasifican por tipo de solicitud

☐ No sabe / no aplica

En su experiencia, ¿qué tan útil sería contar con un sistema que identifique automáticamente el sentimiento (positivo, negativo o neutro) de las PQRS?

☐ Muy útil

☐ Algo útil

☐ Poco útil

☐ Nada útil

¿Considera que la incorporación de herramientas de procesamiento de lenguaje natural (PLN) podría facilitar la atención al cliente?

☐ Totalmente de acuerdo

☐ De acuerdo

☐ En desacuerdo

☐ Totalmente en desacuerdo

¿Qué aspectos cree que deberían mejorarse en el proceso actual de atención de PQRS?