Revista Peruana de Computación y Sistemas 2023 5(2):29-40
https://doi.org/10.15381/rpcs.v5i2.27134

ISSN Online: 2617-2003
Facultad de Ingeniería de Sistemas e Informática
Universidad Nacional Mayor de San Marcos

Artículo de Revisión

Clasificación de quejas de los clientes empleando procesamiento de lenguaje natural: revisión sistemática de la literatura

Customer complaint classification using natural language processing: systematic literature review

José Luis Flores Poma [1,a]

[1] Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática. Lima, Perú

[a] Autor de correspondencia: jose.flores26@unmsm.edu.pe, ORCID: https://orcid.org/0000-0001-5815-1977

Resumen

Un cliente insatisfecho por algún producto y/o servicio se encuentra motivado a expresar una queja. Clasificar las quejas de forma manual es un proceso que representa elevados costos en recursos humanos y materiales. La Inteligencia Artificial (IA) permite el uso de diversos algoritmos para realizar tareas que pueden simular la inteligencia humana, una rama de esta es el Procesamiento de Lenguaje Natural (NLP por sus siglas en inglés), tiene por objetivo que las máquinas tengan la capacidad de entender lenguaje humano, permitiendo, por ejemplo clasificar y categorizar datos de forma automática. Este artículo proporciona una revisión sistemática de la literatura abordando desafíos en la clasificación de textos de queja, tales como la falta de balance de clases, la presencia de datos sin etiquetar y la interpretación de los resultados de los modelos. Se exploran técnicas de pre procesamiento, como la tokenización, la remoción de stopwords y la lematización, que influyen en el rendimiento de los modelos. Adicionalmente, se discuten las métricas de rendimiento como precision, recall y F1-score. Se muestran las tendencias actuales y futuras líneas de investigación. Para tal fin se analizaron 24 artículos publicados entre 2018 y 2023 extraídas de las bases de datos de Web of Science y Scopus.

Palabras clave: Procesamiento de lenguaje natural, machine learning, queja de cliente, satisfacción de cliente.

Abstract

A dissatisfied customer with a product and/or service is motivated to express a complaint. Classifying complaints manually is a process that represents high costs in human and material resources. Artificial Intelligence (AI) allows the use of various algorithms to perform tasks that can simulate human intelligence, a branch of this is Natural Language Processing (NLP), its objective is that machines have the capacity to understand human language, allowing, for example, to classify and categorize data automatically. This article provides a systematic review of the literature addressing challenges in the classification of complaint texts, such as the lack of class balance, the presence of unlabeled data, and the interpretation of model results. Preprocessing techniques are explored, such as tokenization, stopword removal, and lemmatization, which influence model performance. Additionally, performance metrics such as precision, recall and F1-score are discussed. Current trends and future lines of research are shown. For this purpose, 24 articles published between 2018 and 2023 extracted from the Web of Science and Scopus databases were analyzed.

Keywords: Natural language processing, machine learning, customer complaint, customer satisfaction.

Recibido: 14/08/2023 - Aceptado: 18/12/2023 - Publicado: 20/12/2023

Citar como:
Flores Poma, José Luis (2023). Clasificación de quejas de los clientes empleando procesamiento de lenguaje natural: revisión sistemática de la literatura. Revista Peruana de Computación y Sistemas, 5(2):29-40. https://doi.org/10.15381/rpcs.v5i2.27134

1. Introducción

Actualmente los usuarios están cada vez más informados sobre productos y/o servicios, además, las regulaciones son más estrictas, generando que la competencia entre empresas sea agresiva. En este contexto la retención de clientes es crucial [1]. Múltiples investigaciones concluyen que la correcta gestión de quejas incrementa la satisfacción de los clientes produciendo a su vez la lealtad de los mismos [2]

La gestión de quejas se ha explorado desde un ámbito general. Por ejemplo [3] analizaron datos de los usuarios que fueron registrados por la American Customer Satisfaction Index1 (ACSI) en un periodo de 10 años. Sus conclusiones sugieren que las empresas con mejor gestión de quejas tienen mayor lealtad de los clientes. Por otra parte [4] realizaron una investigación mediante encuestas evidenciando una relación entre la prontitud de atención y la satisfacción del cliente con el proceso de gestión de quejas. Finalmente, [5] tras un estudio empírico concluyeron que las respuestas de satisfacción y fidelidad dependen del tipo de queja y la forma en que fueron atendidos.

Un aspecto relevante a considerar sobre el tratamiento de quejas es su clasificación ya que es especialmente útil para las organizaciones, permitiendo abordar problemas relevantes y gestionarlos de manera eficiente [6]. Un estudio realizado por [7] sobre el proceso de gobernanza, concluyó que clasificar con precisión los informes de queja por parte de los ciudadanos promueve la modernización del gobierno. Otro ejemplo es el de las autoridades de la ciudad de Malang en Indonesia implementaron una herramienta en línea para el registro de quejas, posteriormente se envían al administrador del sistema para su clasificación [8]. En relación a las ciudades inteligentes se ha podido observar que mediante el uso de la clasificación de quejas se pueden averiguar de inmediato los principales problemas de los ciudadanos para brindar mejores servicios [9]. Sin embargo es muy común que esta tarea se realice de forma manual ocasionando costos elevados de recursos humanos y materiales [10].

En este contexto, la inteligencia artificial ha demostrado ser de utilidad, ya que incluye diversos algoritmos para extraer datos, aprender arreglos complejos, principalmente en datos grandes y multifacéticos, para respaldar los procedimientos de predicción, clasificación y toma de decisiones [11]. Una rama de esta área es el Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) que utiliza técnicas computacionales para aprender, comprender y producir contenido escrito en lenguaje humano [12].

Si bien es cierto, los usuarios cuentan con múltiples herramientas que permiten registrar sus quejas, el procesamiento se realiza de forma manual [13]. Además [6] refiere que las investigaciones acerca de análisis de textos de queja son escasos en la actualidad. Revisar la aplicación del Procesamiento de Lenguaje Natural para la clasificación de quejas cobra relevancia.

El objetivo de este artículo es proporcionar una visión panorámica de la literatura sobre clasificación automática de quejas mediante el uso de Procesamiento de Lenguaje Natural. Se responderán preguntas referentes a los modelos que se han utilizado para la clasificación automática de quejas, los resultados obtenidos tras su ejecución, además qué características reúnen los dataset empleados y cuáles métricas validan los modelos. Las respuestas obtenidas ayudarán a determinar cuál es la tendencia actual en cuanto a clasificación de quejas empleando procesamiento de lenguaje natural.

Este paper está organizado de la siguiente forma: Sección 2 describe la metodología usada en la investigación, Sección 3 se realiza el análisis dando respuesta a las preguntas planteadas por la investigación Sección 4 conclusiones.

2. Satisfacción del cliente (SC) y Queja del cliente (QC)

Este apartado se concentra en realizar una revisión acerca de la satisfacción del cliente (SC) y queja del cliente (QC) que serán analizados posteriormente en esta investigación.

2.1. Satisfacción del cliente (SC)

Se entiende por satisfacción del cliente al estado de agrado o decepción en comparación entre la percepción y la expectativa por un producto o servicio [14]; esta métrica se obtiene de variados y complejos factores tales como precio, servicio, calidad [15]. La satisfacción del cliente puede traer la lealtad del cliente beneficiando a la empresa con recompras o recomendaciones además de reducir los costos de atraer nuevos clientes [16]; por el contrario un cliente insatisfecho podría no regresar o sentirse motivado en difundir comentarios negativos dañado la imagen de la empresa [17].

2.2. Queja del cliente (QC)

La queja de los clientes tiene como objetivo expresar insatisfacción hacia un producto o servicio como consecuencia de experiencias desagradables o fallas [18]

3. Metodología

Con la finalidad de presentar un alto nivel de descripción se emplea el procedimiento propuesto por [19], el cual contempla las siguientes fases:

- Planificación de la revisión: Es el proceso más importante antes de comenzar la revisión, aquí

1 American Customer Satisfaction Index (ACSI) es la única medida nacional de satisfacción del cliente entre industrias en los Estados Unidos. El Índice mide la satisfacción de los consumidores domésticos de EE. UU. con la calidad de los productos y servicios ofrecidos por empresas nacionales y extranjeras con una participación significativa en los mercados de EE. UU

se plantean las preguntas a responder, se establece el protocolo base de la revisión.

- Realización de la revisión: Se ejecuta la obtención de la data, acorde a responder las preguntas formuladas por la investigación; se establecen los criterios de selección, exclusión.
- Reportando la revisión: Es la fase final del proceso de revisión, aquí los resultados hallados se documentan y difunden a los potenciales interesados.

3.1. Planificación de la revisión

Esta investigación se direcciona a responder preguntas acerca de la clasificación de quejas empleando Procesamiento de Lenguaje Natural:

P1: ¿Qué modelos se han utilizado para la clasificación de quejas empleando procesamiento de lenguaje natural?

P2: ¿Cuál fue el tipo de salida esperada en los modelos de clasificación de quejas empleando procesamiento de lenguaje natural?

P3: ¿Qué métricas se han utilizado para evaluar la precisión de los modelos clasificación de quejas empleando procesamiento de lenguaje natural?

P4: ¿Cuáles son las características de los dataset empleados para el entrenamiento, evaluación de los modelos de clasificación de quejas empleando procesamiento de lenguaje natural?

Para ubicar los artículos apropiados que respondan las preguntas planteadas se realizó la búsqueda de información en las bases de datos Scopus y Web of Science, limitándose al período transcurrido entre 2018 y mayo de 2023; los filtros se aplicaron al abstract, keywords y tittle. Empleando los operadores lógicos AND y OR, se combinó la búsqueda, resultando la cadena que se muestra en la Tabla 1. Determinadas las investigaciones se establecieron los criterios de selección, exclusión tal como se aprecia en la Tabla 2.

Tabla 1. Cadena de búsqueda

Base de datos	Cadena de búsqueda
Scopus	TITLE-ABS-KEY("complaint classification" AND ("NLP" OR "natural language processing" OR "machine learning" OR approach OR method OR strategy OR model OR technique))
WoS	TS=("complaint classification" AND ("NLP" OR "natural language processing" OR "machine learning" OR approach OR method OR strategy OR model OR technique))

Tabla 2. Criterios de selección, exclusión

Criterio de selección	Criterio de exclusión
CS1: El periodo debe estar dentro de los años 2018 a mayo 2023.	CE1: El artículo se encuentra fuera del ámbito de la investigación (no presentan modelos, algoritmos, componentes, métricas, procesamiento de lenguaje natural).
CS2: El artículo debe ser relevante para responder al menos 1 de las preguntas de la investigación.	CE2: Libros, reportes técnicos, revisiones, handbooks.
CS3: Estar publicado en algún journal o conferencia	CE3: Artículos en proceso de revisión.
CS4: El idioma debe ser inglés	CE4: Artículo duplicado

3.2. Realización de la revisión:

Se obtuvo un conjunto primario de papers con las estrategias de búsqueda, posteriormente se aplicaron los criterios de selección, exclusión. Una revisión inicial se hizo evidente, permitiendo establecer si los documentos eran relevantes para contestar las preguntas planteadas en la investigación, el flujo se aprecia en la Figura 1

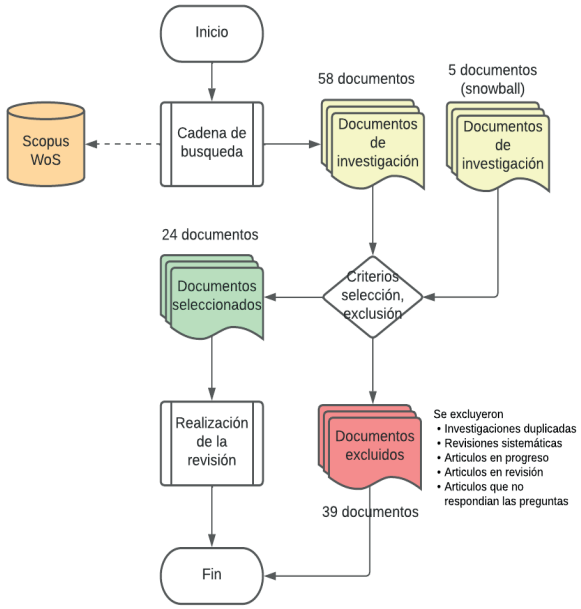
3.2.1. Aseguramiento de la calidad

Con la finalidad de determinar si los documentos seleccionados cumplían los parámetros necesarios para ser usados en la investigación se ejecutaron los criterios de selección y exclusión quedando como se muestra en la Tabla 3. Por otro lado la Figura 2 muestra la cantidad de publicaciones por años. Nótese que el 2022 fue el año con mayor producción con siete investigaciones.

Tabla 3. Número de investigaciones halladas con la cadena de búsqueda

Descripción	Scopus	Web of Science	Total
Resultados elegibles	45	13	58
Año 2018 - 2022	31	7	38
Artículo, artículo de conferencia	28	7	35
Artículo estado final	28	6	34
Publicado journal, conferencia	28	6	34
Idioma inglés	28	6	34
Revisión título, abstract	22	6	28
Artículos duplicados	21	0	21

Figura 1. *Flujo del proceso realizado*



3.3. Reportando la revisión

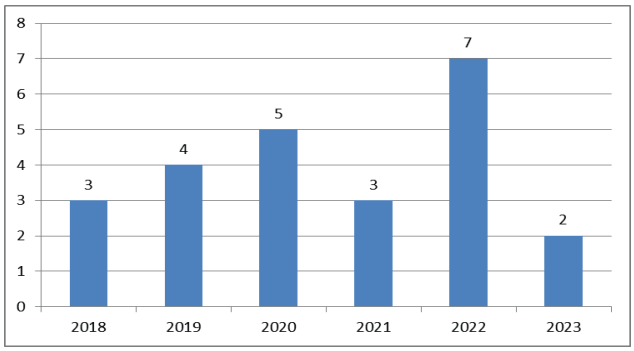
3.3.1. Fuente de datos:

De los resultados obtenidos se aprecia que la tendencia es dispersa, la Tabla 4 muestra el detalle del proceso.

Tabla 4. *Número de investigaciones seleccionadas según publicador*

Publicador	Nro. Investigaciones
Association for Computational Linguistics (ACL)	4
Elsevier B.V.	1
Elsevier Ltd	3
European Alliance for Innovation	1
IEEE Computer Society	1
Institute of Electrical and Electronics Engineers Inc.	3
Institute of Physics Publishing	1
Little Lion Scientific	1
Science and Information Organization	1
Science Publishing Corporation Inc	1
SPIE	1
Springer	3
Springer Science and Business Media Deutschland GmbH	1
Springer Verlag	1
Telecommunications Society and Academic Mind	1
Total	24

Figura 2. *Número de publicaciones seleccionadas por año*



Tendencias de investigación:

De acuerdo al análisis realizado con las investigaciones seleccionadas se evidencia que la coocurrencia de palabras vinculadas con *"complaint classification"* son: *classification models, deep learning, sentiment analysis, machine learning, social media, text processing, text classification, support vector machine* (Figura 3). Por otro lado los focos de atención por parte de los investigadores en cuanto refiere a *"customer satisfaction"* son: *competitive business, business organizations, complaints, praises, sales, machine learning* (Figura 4). Al hablar de *"natural language processing"*, las tendencias se centran en: *text classification, classification models, semantics, sentiment analysis, feature extraction, dataset* (Figura 5).

Figura 3. *Coocurrencia de palabras de acuerdo a complaint classification*

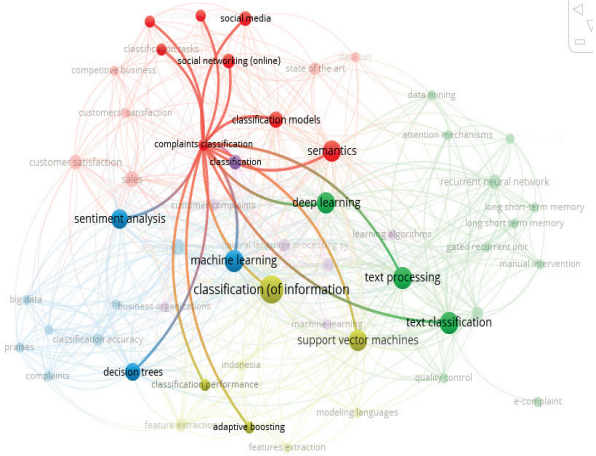


Figura 4. *Coocurrencia de palabras de acuerdo a customer satisfaction*

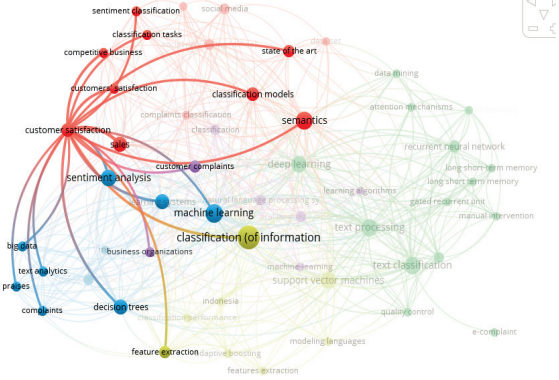
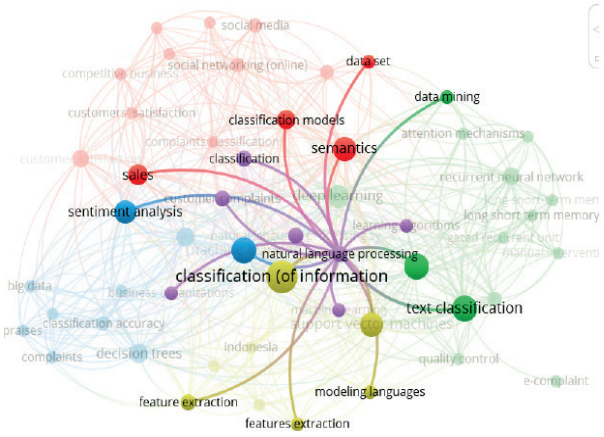


Figura 5. *Coocurrencia de palabras de acuerdo a natural language processing*



3.4. Análisis:

En este apartado la investigación dará respuesta a las preguntas planteadas, para cumplir tal fin se analizarán los documentos de referencia.

3.4.1. P1: ¿Qué modelos se han utilizado para la clasificación de quejas empleando procesamiento de lenguaje natural?:

Los modelos empleados constan de múltiples componentes que, varían de acuerdo a las necesidades de la implementación, no existe consenso a la hora de definirlos, sin embargo es común emplear: recolección de datos, pre procesamiento de datos, entrenamiento y evaluación.

Ya que, la clasificación de quejas trata con narrativas en lenguaje natural los modelos requieren de pre procesamiento como una etapa crucial antes del entrenamiento, por ejemplo [20] realizaron segmentación de palabras y remoción de stopwords, por otro lado [7] eliminaron información inútil tales como direcciones URL, nombres de compañías, adicionalmente segmentaron las narrativas con longitud de palabras limitado a 340. De esta manera, en caso de exceder este número se truncarían.

Un elemento primordial para la clasificación automática de quejas radica en el etiquetado, suele pasar que no esté disponible. Existirán casos en los que será necesario realizar esta tarea de forma manual [21]. Frecuentemente los modelos presentan problemas de sesgo, esto se debe a muchos factores, sin embargo, uno importante a tener en cuenta es la correcta distribución de los datos tal es el caso de la investigación hecha por [22], ellos pre procesaron los textos de tal forma que el género estuviese correctamente balanceado.

Los modelos recurren a la Tokenización, que consiste en representar las oraciones en unidades normalmente separadas por espacios en blanco. Por ejemplo [23] realizan este proceso para posteriormente remover los *stopwords* de tal forma que eviten impactos negativos en la tarea de clasificación, por otro lado [10] tokenizaron el corpus y aplicaron normalización, stemming y lemmatización para reducir los términos del vocabulario, removieron los stopwords y convirtieron el resultado a *Term frequency – Inverse document frequency* (TF-IDF) [24], proponen una arquitectura multinivel, como consecuencia de realizar dos tipos de procesamiento de tweets, en una primera etapa análisis de sentimientos, implementaron tres algoritmos deep learning LSTM, Bi-LSTM y CNN para identificar los tweets negativos, que, en una segunda etapa son la materia prima para la clasificación de quejas implementada con BERT *uncased*.

Los modelos empleados para la clasificación de quejas son diversos y multifacéticos, estos no están definidos de forma estándar, en su lugar se ajustan a las necesidades de los investigadores, la Tabla 7 da cuenta de los variados componentes de los modelos revisados.

Dos elementos clave que se repiten en todos los modelos revisados son: recopilación/adquisición de datos y pre procesamiento de datos, esto debido a que, el entrenamiento requiere datos históricos para realizar el proceso; en cuanto al pre procesamiento, por tratarse de lenguaje natural, el texto siempre contiene irregularidades que pueden distorsionar el resultado esperado, existen casos en que los corpus deben ser procesados para definirse únicamente en minúsculas.

Existe preferencia por el uso de la lemmatización, ya que es un proceso que se asemeja a la derivación reduciendo los verbos y los sustantivos a su forma singular, a diferencia del stemming que es considerado un proceso más destructivo. La remoción de stopwords es clave para que el modelo no se vea afectado por la presencia de caracteres que pueden afectar la idea que trata de expresar la narrativa.

Los cinco algoritmos más empleados en los modelos revisados son SVM representado el 15%, seguido por GRU (12%), BERT (12%), MLP (9%) y LSTM (9%), tal como se muestra en la Tabla 5. Claramente el uso de deep learning marca la tendencia en cuanto a tareas de clasificación de quejas, a través de redes neuronales, la Figura 6 detalla el recuento realizado.

Tabla 5. *Recuento de algoritmos empleados en las investigaciones revisadas*

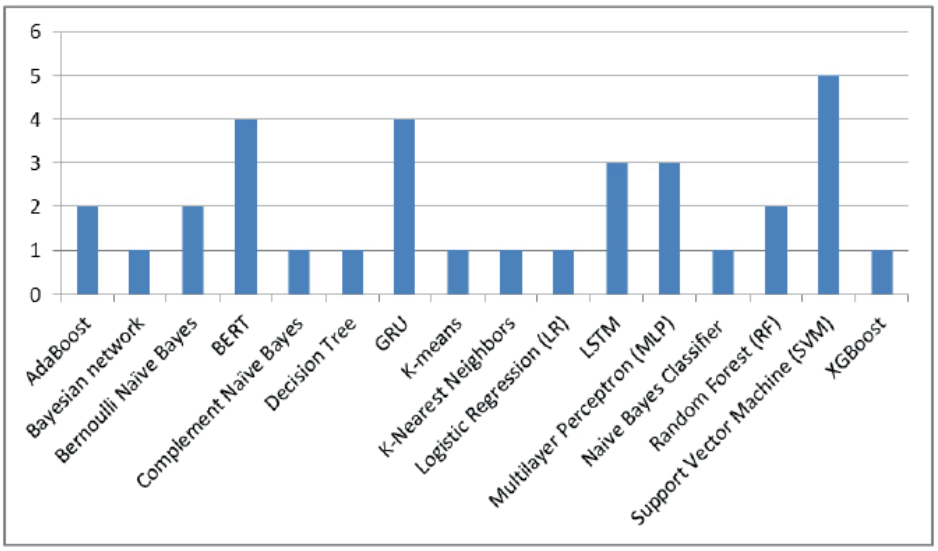
Algoritmo	Referencia	Recuento
Support Vector Machine (SVM)	[20], [10], [9], [25], [8]	5
BERT	[24], [7], [26], [27]	4
GRU	[28], [29], [6], [30]	4
LSTM	[31], [10], [29]	3
Multilayer Perceptron (MLP)	[32], [33], [30]	3
AdaBoost	[9], [23]	2
Bernoulli Naïve Bayes	[23], [8]	2
Random Forest (RF)	[23], [32]	2
Bayesian network	[34]	1
Complement Naïve Bayes	[23]	1
Decision Tree	[23]	1
K-means	[34]	1
K-Nearest Neighbors	[23]	1
Logistic Regression (LR)	[23]	1
Naive Bayes Classifier	[35]	1
XGBoost	[9]	1

3.4.2. P2: ¿Cuál fue el tipo de salida esperada en los modelos de clasificación de quejas empleando procesamiento de lenguaje natural?:

Las salidas por lo general se categorizaron de acuerdo al tipo de organización de desde donde se recolectaron los datos, por ejemplo [25] realizaron un sistema de clasificación de quejas, para tal fin emplearon datos de comentarios registrados en twitter acerca de la línea de cercanías en Jakarta – Indonesia, las seis clases jerárquicas resultantes fueron: cola y retraso, KRL Capacidad, higiene, instalación, categoría de servicio, seguridad; estas se vinculan directamente con el tipo de servicio ofrecido.

Asimismo [23] realizaron una investigación acerca del sector alimentos y economía, en este caso agruparon los resultados en cuatro categorías de acuerdo a la información de la *Economic and Food Safety Authority*

Figura 6. Recuento uso algoritmos en los modelos revisados



(ASAE), estas son: *food safety, regard economic offenses (FisEc - purchasing power parity), regard economic offenses (FisEc - business corporation), complains out food or economic fields*, es de apreciar que las salidas son de carácter jerárquico.

La investigación realizada por [27], se enfocó en categorizar la severidad de las quejas, para tal fin se basaron en un estudio previo que agrupa las quejas en cuatro clases: *no explicit reproach, disapproval, accusation, blame*.

En la industria de la construcción [33], realizaron un modelo para categorizar las quejas sobre problemas de calidad, se detalla que la definición de clases es realizada manualmente por oficiales, sin embargo su criterio de clasificación no es universalmente aceptado debido a lo subjetivo de la tarea. Aun así doce son las etiquetas consideradas para la salida del modelo, estas son: *Ash or sand on surface, Leakage, Crack of floor slab, Hollowing or cracking, Floor thickness is substandard, Structural dimensions, Construction impact, Completion acceptance, Building materials, Foundation, Decoration, Others*.

El tipo de salida esperada, por lo general depende del etiquetado previo en el dataset, es necesario contabilizar las columnas y validar el balanceo de tal forma que el modelo no se vea perjudicado. Un detalle importante a tener en cuenta, se refiere a la falta de previsión, los modelos son entrenados con clases establecidas, no se toma en cuenta la posible aparición de una nueva clase.

La Tabla 8 muestra las salidas esperadas en los modelos verificados**.**

3.4.3 P3: ¿Qué métricas se han utilizado para evaluar la precisión de los modelos clasificación de quejas empleando procesamiento de lenguaje natural?

Todos los modelos necesitan ser evaluados para medir su desempeño, las investigaciones revisadas tratan acerca de implementaciones dirigidas a la clasificación de textos de queja, por tanto son unánimes con respecto al uso de las métricas siendo estas: *Accuracy, Precision, Recall y F1 Score*; vienen dadas por la siguiente definición:

$$Acuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score = 2 * \frac{Recall * Precision}{Recall + Precision} \quad (4)$$

Los valores representados son *True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN),* la Tabla 6 muestra las métricas empleadas en los modelos revisados

3.4.4 P4: ¿Cuáles son las características de los dataset empleados para el entrenamiento, evaluación de los modelos de clasificación de quejas empleando procesamiento de lenguaje natural?

El dataset es el elemento más importante dentro de un modelo de inteligencia artificial ya que, sin la información histórica que provee, no es posible realizar el entrenamiento. Las características de los dataset varían dependiendo de su proceso de adquisición, por ejemplo [24] reunieron alrededor de 20 mil tweets en un periodo de seis meses basados en la información geográfica y los keywords/hashtags específicos del dominio de la electricidad, se consideraron ocho atributos (text –tweet-, location, tweet_id, userid, created_at, retweet_count, hashtags, mentions).

Tabla 6. Métricas utilizadas para evaluar los modelos de clasificación de quejas mediante NLP

Descripción	Referencia	Métricas
Automating Public Complaint Classification Through JakLapor Channel: A Case Study of Jakarta, Indonesia	[9]	- Accuracy, precision, recall, F1 score
A joint attention enhancement network for text classification applied to citizen complaint reporting	[7]	- Accuracy, runtime, F1 score, precision, and recall
Automating Complaints Processing in the Food and Economic Sector: A Classification Approach	[23]	- Accuracy, precision, recall, F1-score
CitEnergy : A BERT based model to analyse Citizens' Energy-Tweets	[24]	- Accuracy, precision, recall, F-measure
Modeling for Car Quality Complaint Classification based on Machine Learning	[20]	- Accuracy, precision, recall, F1-score
Deep Learning-Based Complaint Classification for Indonesia Telecommunication Company's Call Center	[29]	- Accuracy, precision, recall, F1-score
Customer Critique Analysis System for PT. KCI's Twitter	[25]	- Accuracy
Convolutional neural network: Deep learning-based classification of building quality problems	[33]	- Accuracy, precision, recall, F1 score
Complaint Classification Using Hybrid-Attention GRU Neural Network	[6]	- Accuracy, precision, recall, F1 score
A LSTM based Tool for Consumer Complaint Classification	[31]	- Accuracy
Anonymo: Automatic Response and Analysis of Anonymous Caller Complaints	[10]	- Accuracy, precision, recall, F1 score
Automated complaints classification using modified nazief-adriani stemming algorithm and naive bayes classifier	[35]	- Accuracy, precision, recall, F1 score
Automatic Complaint Classification System Using Classifier Ensembles	[8]	- Accuracy
Complaint Classification using Word2Vec Model	[30]	- Accuracy

Por otro lado [33] recopilaron 4087 narrativas de queja sobre problemas de calidad en la construcción, estos se agruparon en 12 clases de forma manual. Una investigación desarrollada por [6] uso dos dataset de diferentes industrias, el primero en idioma chino acerca de quejas de delivery segmentados en 8 categorías, de las cuales seleccionaron las 3 con mayor número de casos, y, otro en idioma inglés con 555,958 registros que contenían 66,806 narrativas de queja, igual que con el caso anterior solo consideraron las 3 clases con mayor ocurrencia.

Dentro de la industria vehicular [20], emplearon un dataset de 2400 registros agrupados en 8 categorías de 300 narrativas cada una. La dificultad para la obtención de datos no es un problema aislado, esto debido a las empresas propietarias que, en última instancia no están dispuestas a proporcionar este insumo a las investigaciones, [34] tuvo que recurrir a un dataset público de GitHub que contenía 100 mil registros de queja, se dividió 70% para entrenamiento y 30% para pruebas.

La Tabla 9 muestra el detalle de los dataset empleados en las investigaciones seleccionadas.

Al hablar de procesamiento de lenguaje natural es inevitable encontrarnos con multiplicidad de expresiones que van más allá del entendimiento computacional, en ese sentido las investigaciones exploradas analizan diversos escenarios en la adquisición de datos; estos pueden ser en formato hablado que luego deberá ser convertido a texto.

Sin embargo un conjunto de datos textuales requiere de cierto tratamiento para ser útil como materia prima para entrenamiento de un modelo de inteligencia artificial; debido a esto el pre procesamiento es tremendamente relevante. Realizado el tratamiento, los datos deben estar etiquetados, en caso no sea así se procede con esta tarea, finalmente el análisis permite explorar las categorías y validar si están balanceadas. La tendencia de división de datos varía entre 60% y 80% para entrenamiento y 20% a 40% para pruebas.

Un problema frecuente que se presenta en las investigaciones es precisamente la fuente de recolección ya que los casos de estudio son diversos y, tratándose de narrativas de queja las entidades dueñas de los datos, no siempre están dispuestas a colaborar con los estudios

4. Conclusiones

En estos tiempos las empresas competitivas comprenden que las quejas son una excelente fuente de retroalimentación acerca de la experiencia del cliente, sin embargo por lo general su gestión se realiza de forma manual. Los estudios de clasificación de quejas empleando Procesamiento de Lenguaje Natural son escasos.

Con esto en mente, se ha realizado una revisión sistemática de la literatura acerca de la clasificación de quejas empleando procesamiento de lenguaje natural. Para tal fin se recurrió a las bases de datos indexadas Web of Science y Scopus, el resultado inicial identificó 58 artículos potenciales; adicionalmente, mediante la técnica *snowball*, se añadieron 5 documentos desde ACL Anthology por considerarse investigaciones importantes para responder las preguntas planteadas; luego del proceso de selección/exclusión se obtuvieron un total de 24 *papers*.

Los modelos presentan similitudes en cuanto se refiere a: adquisión de datos, pre procesamiento, entrenamiento y validación. Siendo el pre procesamiento la tarea principal a realizar debido a que las quejas expresan "sentimientos negativos" por lo tanto la intensidad de la narrativa puede llevar a sesgos. Múltiples estrategias tales como limpieza de datos, balanceo, remoción de *stopwords*, etc., se emplean para no obtener resultados anómalos.

El entrenamiento de los modelos requiere fundamentalmente del etiquetado de las narrativas, en caso de no existir este proceso se realiza de forma manual. Por tratarse de un procedimiento altamente subjetivo, es posible recurrir a investigaciones que hayan categorizado quejas símiles. Los algoritmos más empleados entre los investigadores son SVM, GRU, BERT, MLP y LSTM.

Ninguna de las investigaciones revisadas contemplan una alternativa al problema de las clases nuevas, un trabajo futuro propuesto es la aplicación del *zero-shot* a la clasificación de quejas.

La información aquí mostrada es relevante para conducir futuros estudios de clasificación de queja de los clientes.

Referencias

[1] S. Bengul y C. Yilmaz, «Effects of Customer Complaint Management Quality on Business Performance in Service Businesses», BU Journal, vol. 32, n.o 2, jul. 2018, doi: 10.21773/boun.32.2.4.

[2] S. Von Janda, A. Polthier, y S. Kuester, «Do they see the signs? Organizational response behavior to customer complaint messages», Journal of Business Research, vol. 137, pp. 116-127, dic. 2021, doi: 10.1016/j.jbusres.2021.08.017.

[3] F. V. Morgeson, G. T. M. Hult, S. Mithas, T. Keiningham, y C. Fornell, «Turning Complaining Customers into Loyal Customers: Moderators of the Complaint Handling–Customer Loyalty Relationship», Journal of Marketing, vol. 84, n.o 5, pp. 79-99, sep. 2020, doi: 10.1177/0022242920929029.

[4] G.-C. A. Ogbeide, S. Böser, R. J. Harrinton, y M. C. Ottenbacher, «Complaint management in hospitality organizations: The role of empowerment and other service recovery attributes impacting loyalty and satisfaction», Tourism and Hospitality Research, vol. 17, n.o 2, pp. 204-216, abr. 2017, doi: 10.1177/1467358415613409.

[5] A. Kuster-Boluda, N. V. Vila, y I. Kuster, «Managing international distributors' complaints: an exploratory study», JBIM, vol. 35, n.o 11, pp. 1817-1829, abr. 2020, doi: 10.1108/JBIM-11-2018-0336.

[6] S. Wang, B. Wu, B. Wang, y X. Tong, «Complaint Classification Using Hybrid-Attention GRU Neural Network», en Advances in Knowledge Discovery and Data Mining, Q. Yang, Z.-H. Zhou, Z. Gong, M.-L. Zhang, y S.-J. Huang, Eds., en Lecture Notes in Computer Science, vol. 11439. Cham: Springer International Publishing, 2019, pp. 251-262. doi: 10.1007/978-3-030-16148-4_20.

[7] Y. Wang, Y. Zhou, y Y. Mei, «A joint attention enhancement network for text classification applied to citizen complaint reporting», Appl Intell, mar. 2023, doi: 10.1007/s10489-023-04490-y.

[8] M. Ali-Fauzi, «Automatic complaint classification system using classifier ensembles», Telfor J, vol. 10, n.o 2, pp. 123-128, 2018, doi: 10.5937/telfor1802123A.

[9] S. M. Intani, B. I. Nasution, M. E. Aminanto, Y. Nugraha, N. Muchtar, y J. I. Kanggrawan, «Automating Public Complaint Classification Through JakLapor Channel: A Case Study of Jakarta, Indonesia», en 2022 IEEE International Smart Cities Conference (ISC2), Pafos, Cyprus: IEEE, sep. 2022, pp. 1-6. doi: 10.1109/ISC255366.2022.9922346.

[10] A. Azhar, S. Maweekumbura, R. Gunathilake, T. Maddumaarachchi, A. Karunasena, y M. Nadeeshani, «Anonymo: Automatic Response and Analysis of Anonymous Caller Complaints», en 2022 IEEE Symposium on Wireless Technology & Applications (ISWTA), Kuala Lumpur, Malaysia: IEEE, ago. 2022, pp. 110-115. doi: 10.1109/ISWTA55313.2022.9942736.

[11] A. Ahani et al., «Evaluating medical travelers' satisfaction through online review analysis», Journal of Hospitality and Tourism Management, vol. 48, pp. 519-537, sep. 2021, doi: 10.1016/j.jhtm.2021.08.005.

[12] D. Seong, Y. H. Choi, S.-Y. Shin, y B.-K. Yi, «Deep learning approach to detection of colonoscopic information from unstructured reports», BMC Med Inform Decis Mak, vol. 23, n.o 1, p. 28, feb. 2023, doi: 10.1186/s12911-023-02121-7.

[13] X. Tian, I. Vertommen, L. Tsiami, P. Van Thienen, y S. Paraskevopoulos, «Automated Customer Complaint Processing for Water Utilities Based on Natural Language Processing—Case Study of a Dutch Water Utility», Water, vol. 14, n.o 4, p. 674, feb. 2022, doi: 10.3390/w14040674.

[14] T. Chen, L. Peng, X. Yin, J. Rong, J. Yang, y G. Cong, «Analysis of User Satisfaction with Online Education Platforms in China during the COVID-19 Pandemic», Healthcare, vol. 8, n.o 3, p. 200, jul. 2020, doi: 10.3390/healthcare8030200.

[15] S.-H. Park, M.-Y. Kim, Y.-J. Kim, y Y.-H. Park, «A Deep Learning Approach to Analyze Airline Customer Propensities: The Case of South Korea», Applied Sciences, vol. 12, n.o 4, p. 1916, feb. 2022, doi: 10.3390/app12041916.

[16] L. T. Nguyen, M. H. DANG, T. D. TAT, y D. G. T. TRAN, «Revisiting Customer Complaint Intention: A Case Study of Mobile Service Users in Vietnam», The Journal of Asian Finance, Economics and Business, vol. 8, n.o 9, pp. 121-130, sep. 2021, doi: 10.13106/JAFEB.2021.VOL8.NO9.0121.

[17] M. Nilashi et al., «Revealing travellers' satisfaction during COVID-19 outbreak: Moderating role of service quality», Journal of Retailing and Consumer Services, vol. 64, p. 102783, ene. 2022, doi: 10.1016/j.jretconser.2021.102783.

[18] G. Chen y S. Li, «Effect of Employee–Customer Interaction Quality on Customers' Prohibitive Voice Behaviors: Mediating Roles of Customer Trust and Identification», Front. Psychol., vol. 12, p. 773354, dic. 2021, doi: 10.3389/fpsyg.2021.773354.

[19] B. Kitchenham, Guidelines for performing Systematic Literature Reviews in software engineering. EBSE Technical Report EBSE-2007-01. 2007.

[20] C. X. Yu, H. Xia, Z. X. Min, y S. Ying, «Modeling for Car Quality Complaint Classification based on Machine Learning», IJACSA, vol. 13, n.o 5, 2022, doi: 10.14569/IJACSA.2022.0130569.

[21] N. Tazeen y K. Sandhya, «A Conceptual Data Modelling Framework for Context-Aware Text Classification», IJACSA, vol. 11, n.o 11, 2020, doi: 10.14569/IJACSA.2020.0111116.

[22] P. Lertvittayakumjorn et al., «Supporting Complaints Investigation for Nursing and Midwifery Regulatory Agencies», en Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, Online: Association for Computational Linguistics, 2021, pp. 81-91. doi: 10.18653/v1/2021.acl-demo.10.

[23] G. Magalhães, B. M. Faria, L. P. Reis, H. L. Cardoso, C. Caldeira, y A. Oliveira, «Automating Complaints Processing in the Food and Economic Sector: A Classification Approach», en Trends and Innovations in Information Systems and Technologies, Á. Rocha, H. Adeli, L. P. Reis, S. Costanzo, I. Orovic, y F. Moreira, Eds., en Advances in Intelligent Systems and Computing, vol. 1160. Cham: Springer International Publishing, 2020, pp. 445-456. doi: 10.1007/978-3-030-45691-7_41.

[24] J. Bedi y D. Toshniwal, «CitEnergy : A BERT based model to analyse Citizens' Energy-Tweets», Sustainable Cities and Society, vol. 80, p. 103706, may 2022, doi: 10.1016/j.scs.2022.103706.

[25] A. Husen, S. W. Sihwi, y E. Suryani, «Customer Critique Analysis System for PT. KCI's Twitter», J. Phys.: Conf. Ser., vol. 1201, n.o 1, p. 012006, may 2019, doi: 10.1088/1742-6596/1201/1/012006.

[26] A. Singh, P. Jha, R. Bhatia, y S. Saha, «What Is Your Cause for Concern? Towards Interpretable Complaint Cause Analysis», en Advances in Information Retrieval, J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, y A. Caputo, Eds., en Lecture Notes in Computer Science, vol. 13981. Cham: Springer Nature Switzerland, 2023, pp. 141-155. doi: 10.1007/978-3-031-28238-6_10.

[27] M. Jin y N. Aletras, «Modeling the Severity of Complaints in Social Media», en Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online: Association for Computational Linguistics, 2021, pp. 2264-2274. doi: 10.18653/v1/2021.naacl-main.180.

[28] A. Singh, S. Saha, Md. Hasanuzzaman, y K. Dey, «Multitask Learning for Complaint Identification and Sentiment Analysis», Cogn Comput, vol. 14, n.o 1, pp. 212-227, ene. 2022, doi: 10.1007/s12559-021-09844-7.

[29] S. Lukitasari y F. Hidayat, «Deep Learning-Based Complaint Classification for Indonesia Telecommunication Company's Call Center», en Proceedings of the Proceedings of the 7th Mathematics, Science, and Computer Science Education International Seminar, MSCEIS 2019, 12 October 2019, Bandung, West Java, Indonesia, Bandung, Indonesia: EAI, 2020. doi: 10.4108/eai.12-10-2019.2296518.

[30] M. Rathore, D. Gupta, y D. Bhandari, «Complaint Classification using Word2Vec Model», IJET, vol. 7, n.o 4.5, p. 402, sep. 2018, doi: 10.14419/ijet.v7i4.5.20192.

[31] N. T. Thomas, «A LSTM based Tool for Consumer Complaint Classification», en 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore: IEEE, sep. 2018, pp. 2349-2351. doi: 10.1109/ICACCI.2018.8554857.

[32] S. Khedkar y S. Shinde, «Deep Learning and Ensemble Approach for Praise or Complaint Classification», Procedia Computer Science, vol. 167, pp. 449-458, 2020, doi: 10.1016/j.procs.2020.03.254.

[33] B. Zhong, X. Xing, P. Love, X. Wang, y H. Luo, «Convolutional neural network: Deep learning-based classification of building quality problems», Advanced Engineering Informatics, vol. 40, pp. 46-57, abr. 2019, doi: 10.1016/j.aei.2019.02.009.

[34] Z. Rao y Y. Zhang, «Research on Content of User Complaint Classification Based on Data Mining», en 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China: IEEE, jun. 2020, pp. 1080-1085. doi: 10.1109/ITOEC49072.2020.9141939.

[35] V. Ferdina, M. B. Kristanda, y S. Hansun, «AUTOMATED COMPLAINTS CLASSIFICATION USING MODIFIED NAZIEF-ADRIANI STEMMING ALGORITHM AND NAIVE BAYES CLASSIFIER», . Vol., n.o 5, 2019.

[36] A. Singh, S. Saha, M. Hasanuzzaman, y A. Jangra, «Identifying complaints based on semi-supervised mincuts», Expert Systems with Applications, vol. 186, p. 115668, dic. 2021, doi: 10.1016/j.eswa.2021.115668.

Apéndices

Tabla 7. *Modelos utilizados para clasificación de quejas mediante NLP*

Modelo	Referencia	Componentes
Modelo que usa tweets para clasificar sentimientos y quejas sobre el sector energía.	[24] -	– Twitter streaming API (recopilación - Twitter streaming API (recopilación de datos) – Pre procesamiento – Etiquetado de tweets – CNN, LSTM and BiLSTM – Modelo BERT uncased pre-entrenado – Pruebas
Modelo de clasificación automática de quejas sobre calidad del automóvil basado en machine learning	[20]	– Adquisición de datos – Clasificación de datos – Análisis de características de datos – Segmentación de palabras – Extracción de características (BoW) – Modelado de clasificación (SVM) – Análisis de confiabilidad
Modelo de clasificación de quejas ciudadanas	[7]	– Adquisición de datos – Preprocesamiento – BERT network (longitud, tamaño oculto) – Evaluación del modelo
Modelo de gestión de quejas mediante clasificación y enrutamiento de call center	[10]	– Agente conversacional IA – Adquisición de datos – Preprocesamiento – Lemmatization, stemming y tokenization – Vectorización TF-IDF – Entrenamiento con SVM – Pruebas
Modelo para automatizar la clasificación de quejas ciudadanas	[9]	– Adquisición de datos – Preprocesamiento – Extracción de características (Count Vectorizer, Terms Frequency-Inverse Document Frequency (TF-IDF), N-Gram, Latent Semantic Analysis (LSA) – Construcción y entrenamiento del modelo (SVM, RF, XGBoost, AdaBoost) – Evaluación del modelo
Modelo de clasificación de quejas basado en minería de datos	[34	– Adquisición de datos – Preprocesamiento – Segmentación de texto (Bayesian network) – Extracción de características. – Entrenamiento del modelo (K-means) – Evaluación del modelo
Modelo de clasificación de quejas basado en deep learning	[29]	– Adquisición de datos – Preprocesamiento – Extracción de características – Etiquetado de texto – Tokenizacion y remoción de stopwords – Construcción del modelo (RNN) – Evaluación del modelo
Modelo de automatización de clasificación de quejas en el sector alimentos	[23]	– Adquisición de datos – Preprocesamiento y normalización de palabras – Tokenizacion – Remoción stopwords – Stemming – Extracción y selección de características – Construcción de los modelos (MNB, CNB, BNB,SVM –linear-, KNN, DT, RF, AB, LR) – Evaluación y comparación de modelos
Modelo de clasificación de quejas basado en LSTM	[31]	– Adquisidor de datos – Pre procesamiento – Tokenización (Keras tokenizer API) – Remoción de stopwords – LSTM – Evaluación modelo
Modelo de clasificación automática de quejas universitarias usando Naive Bayes	[35]	– Adquisición de datos – Tokenización – Remocion stopwords – Stemming (Nazief-Adriani) – Naive Bayes Classifier – Evaluación modelo

Modelo de clasificación automática de quejas usando conjuntos de clasificadores	[8]	– Adquisición de datos – Tokenización, filtrado y stemming – Remoción de stopwords – BOW – Naïve Bayes, Maximum Entropy, K-Nearest Neighbors, Random Forest, Support Vector Machine – Evaluación modelo
Modelo de clasificación de quejas usando Hybrid-Attention GRU Neural Network	[6]	– Adquisición de datos – Construcción alfabeto 70 caracteres para idioma ingles – Construcción alfabeto 5000 caracteres para idioma chino – Construcción diccionario de opiniones negativas conteniendo 300 palabras – GRU bidireccional – Evaluación del modelo
Modelo de clasificación de quejas osando Word2Vec	[30	– Adquisición de datos – Pre procesamiento – Tokenización – Remoción de stopwords – Capa de incrustación (Word2Vec) – Capa GRU – Clasificador MLP – Se construyó y entreno el modelo usando la librería Keras
Modelo de clasificación de problemas de calidad de construcción	[33]	– Adquisición de datos – Etiquetado – Pre procesamiento – Segmentación de palabras y remoción de stopwords – CNN basado en deep learning – A través de la incrustación de palabras, cada texto se representa como una matriz densa de valores reales. – La extracción de características de los textos se realiza a través de núcleos de convolución – Clasificador MLP – Evaluación del modelo
Modelo de análisis de críticas de usuarios servicio de transporte público	[25]	– Adquisición de datos (twitter) – Etiquetado manual – Pre procesamiento – Remoción de stopwords – Normalización, cambio de caracteres a minúsculas – Stemming – LibSVM – Evaluación del modelo

Tabla 8. *Salida esperada de los modelos de clasificación de quejas mediante NLP*

Descripción	Referencia	Salida
Customer Critique Analysis System for PT. KCI's Twitter	[25]	– Analisis de sentimientos: (positivo, negativo, neutral) – Clasificación de quejas: 6 clases: clase -3: Cola y Retraso, clase -2 Capacidad KRL, clase -1: categoría de Higiene, clase 1: categoría de Instalaciones, clase 2: Categoría de servicio y clase 3: categoría de Seguridad
Convolutional neural network: Deep learning-based classification of building quality problems	[33]	– 12 clases (Ash or sand on surface, Leakage, Crack of floor slab, Hollowing or cracking, Floor thickness is substandard, Structural dimensions deviation or design construction problems, Construction impact, Completion acceptance, Building materials, Foundation, Decoration, Others) – (Others)
Automating Complaints Processing in the Food and Economic Sector: A Classification Approach	[23]	– 4 categorias (Category I complains that are related to food safety, categories II and III regard economic offenses and category IV relates to other complains that are neither of the food nor economic fields)
CitEnergy : A BERT based model to analyse Citizens' Energy-Tweets	[24]	– El modelo clasifica los tweets en dos clases (Clase H –alta prioridad-, clase L –baja prioridad-), estas se distinguen por la cantidad de reportes y su recurrencia.
Anonymo: Automatic Response and Analysis of Anonymous Caller Complaints	[10]	– 7 departamentos o dependencias (Debt Department, Mortgage Department, Accounts Department, Credit Reporting Department, Transaction and Foreign-Currency-Exchange Department, Loan Department, Cards Department)

Modeling for Car Quality Complaint Classification based on Machine Learning	[20]	– 8 categorias (engine/electric motor, transmission, clutch, steering system, braking system, tires, front and rear axles and suspension system, car body accessories and electrical appliances)
Modeling the Severity of Complaints in Social Media	[27]	– 4 clases: No Explicit Reproach, Disapproval, Accusation, Blame
Multitask Learning for Complaint Identification and Sentiment Analysis	[28]	– 3 clases análisis de sentimientos: positivo, negativo, neutral – 2 clases identificación quejas: positivo, negativo
Identifying complaints based on semi-supervised mincuts	[36]	– 2 clases: queja, no queja

Tabla 9. *Características de los dataset empleados para entrenar, evaluar los modelos de clasificación de quejas mediante NLP*

Descripción	Referencia	Dataset
CitEnergy : A BERT based model to analyse Citizens' Energy-Tweets	[24]	20 mil tweets de un periodo de seis meses, diferenciados geográficamente (text –tweet-, location, tweet_id, userid, created_at, retweet_count, hashtags and mentions), se divide en entrenamiento, validación y pruebas
Convolutional neural network: Deep learning-based classification of building quality problems	[33]	El conjunto de datos se obtuvo desde BCQ (Building quality complaints) que son textos registrados en la base de datos de una entidad del gobierno. El dataset se conforma de 4087 textos y se divide en tres: conjunto de datos para entrenamiento del modelo (3310), evaluación del modelo (368) y conjunto de pruebas para evaluar el desempeño del modelo (409)
Complaint Classification Using Hybrid-Attention GRU Neural Network	[6]	Se emplearon dos dataset: – -Quejas de delivery Chino categorias 73458 textos se usarotextostextosn solo los top 3 con mayor número de ejemplos – -Dataset financiero en idioma ingles 11 clases y 555,958 ejemplos solo 66,806 con narrativa, se usaron 3 clases namely credit reporting, debt collection and mortgage (procedencia kaggle
Complaint Classification using Word2Vec Model	[30]	Máximo 750 palabras en la narrativa, se usó 60% entrenamiento 40% pruebas
Modeling for Car Quality Complaint Classification based on Machine Learning	[20]	2400 registros, 300 por cada categoría, proporcionado por Beijing Car Quality Net Information Technology Limited Company
A joint attention enhancement network for text classification applied to citizen complaint reporting	[7]	73 categorías en el dataset, no se especifica el número de registros
A joint attention enhancement network for text classification applied to citizen complaint reporting	[10]	El dataset se conforma por un millón de registros en formato texto, incluye 16 columnas; se dividio 75% para entrenamiento 25% para pruebas
Automating Public Complaint Classification Through JakLapor Channel: A Case Study of Jakarta, Indonesia	[9]	El dataset contiene 56766 registros, las cinco categorías principales son: Disturbance of Peace and Order, Roads, Wild Parking, Trees, Garbage
Research on Content of User Complaint Classification Based on Data Mining	[34]	El dataset contiene 56766 registros, las cinco categorías principales son: Disturbance of Peace and Order, Roads, Wild Parking, Trees, Garbage