



GOBIERNO DEL
ESTADO DE MÉXICO



TECNOLÓGICO DE ESTUDIOS SUPERIORES DE TIANGUISTENCO

OPCIÓN IX

TESIS

**“SISTEMA DE APRENDIZAJE SUPERVISADO PARA ANÁLISIS DE OPINIONES
EN EL CORPUS MUCHOCINE”**

QUÉ PARA OBTENER EL TÍTULO DE:

INGENIERO EN SISTEMAS COMPUTACIONALES

PRESENTA:

MARICARMEN FLORES SOLER

DIRECTOR DE TESIS:

D. EN C. OMAR GONZÁLEZ BRITO

TIANGUISTENCO ESTADO DE MÉXICO, JULIO 2023

Dedicado con amor:

A mi madre: Araceli Soler Ruiz

A mi padre: José Luis Flores Segura

*A mis abuelitas: Amelia Martina Segura Aranda †
y Macrina Rogelia Ruiz Escandón †*

A Dios

Agradecimientos

A mi **director de tesis** el Dr. Omar González Brito que me guio, me demostró su apoyo y confianza, a su vez me aconsejo y me brindo sus conocimientos.

A mis **asesoras** la Dra. Laura Cleofas Sánchez y la Mtra. Silvia Salas Hernández que me dieron sus consejos y apoyo, a su vez me brindaron sus conocimientos.

A mi señora **Madre Araceli Soler** que me ha brindado su apoyo en todos los aspectos de mí vida, paciencia y su amor incondicional.

A mi señor **Padre José Luis Flores** por apoyarme en todos los aspectos, darme aliento, por darme la confianza y creer en mí.

A mis **amigas** Lore y Aide por su amistad, apoyo incondicional a lo largo de nuestra carrera, sus consejos y por los momentos tan divertidos y lindos que compartimos juntas.

Resumen

La minería de opiniones es una rama que se desprende de la minería de textos, que, a su vez, es una subdisciplina de la minería de datos. Su enfoque principal se centra en detectar el sentimiento, emoción o polaridad que se comunican a través de un conjunto de palabras. La metodología que utiliza la minería de opiniones se utiliza principalmente para evaluar las opiniones, emociones y actitudes expresadas (Robles Domínguez, 2019).

En la presente investigación se realizó un estudio para desarrollar un método de análisis de opiniones en críticas de cine en español. Se utilizó el corpus "Muchocine" creado por Cruz (Cruz et al., 2008) constituido por 3878 críticas de películas en idioma español.

Se ha desarrollado un método con base a la metodología de clasificación de textos, la cual se compone de cinco fases. En la primera fase, se hace la adquisición de datos. En la segunda fase se realiza el análisis y etiquetado de los datos, la cual consiste en aplicar el preprocesamiento, del texto del corpus, que incluye la eliminación de *stopwords*, la eliminación de caracteres especiales, signos de puntuación y números. En la tercera fase se aplica la construcción y ponderación de características, la cual se realiza mediante conjuntos de vectores binarios utilizando el enfoque de bolsa de palabras y ponderación booleana. En la cuarta fase, las críticas se someten a un proceso de clasificación con diferentes técnicas de aprendizaje supervisado como Máquina de Soporte Vectorial (*Support Vector Machine, SVM por sus siglas en inglés*), Regresión Logística (*Logistic Regression, LR por sus siglas en inglés*) y Naive Bayes. Por último, se validó el rendimiento del modelo de clasificación mediante validación cruzada a 10 pliegues utilizando la métrica de evaluación exactitud (*accuracy*). Los resultados más significativos de los experimentos que

se realizaron con las técnicas SVM, NB y LR superan el 80% de exactitud siendo resultados competitivos en comparación con el estado del arte.

Contenido

CAPÍTULO 1 INTRODUCCIÓN.....	3
1.1 Antecedentes	6
1.2 Planteamiento del problema	10
1.3 Objetivos	12
1.4 Alcances y Limitaciones.....	14
1.5 Justificación.....	16
1.6 Hipótesis.....	19
CAPÍTULO 2 ESTADO DEL ARTE	20
3.1 Análisis de Sentimientos a nivel de documento en críticas de cine en español.	21
3.2 Técnicas de clasificación de opiniones aplicadas a un corpus en español.	23
3.3 Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español.....	25
3.4 Una revisión a la minería de opiniones y los retos del <i>PNL</i>	26
3.5 Minería de opinión, acceso abierto y gobernanza.....	28
3.6 Predicción electoral usando un modelo híbrido basado en análisis sentimental: elecciones presidenciales de Colombia.	30
3.7 Minería de opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles.	32
CAPÍTULO 3 MARCO TEÓRICO.....	35
3.1 Opinión.....	36
3.1.1 Opinión pública	36
3.1.2 Comentarios.....	37
3.1.3 Análisis de comentarios.....	37
3.2 Inteligencia Artificial.....	38
3.2.1 Aprendizaje supervisado.....	39
3.2.1.1 Regresión Logística	40
3.2.1.2 Máquina de Soporte vectorial	42
3.2.1.3 Naïve Bayes.....	44
3.2.2 Aprendizaje no supervisado.....	46
3.2.2.1 Vecinos más cercanos.....	47
3.2.2.2 Redes neuronales	48
3.3 Procesamiento de lenguaje natural	50
3.3.1 Características textuales	52
3.3.1.1 Analizador léxico.....	53

3.3.1.2 Analizador sintáctico	53
3.3.1.3 Analizador semántico	54
3.3.2 Técnicas del procesamiento del lenguaje natural.....	54
3.4 Minería de opiniones	56
3.4.1 Clasificación de la minería de opiniones	57
3.5 Clasificación de textos.....	58
3.5.1. Adquisición de datos	59
3.5.2 Análisis y etiquetado de datos	60
3.5.3 Construcción y ponderación de características	61
3.5.3.1 Ponderación Booleana	61
3.5.3.2 Ponderación IDF.....	62
3.5.3.3 Ponderación TF	62
3.5.3.4 Ponderación TF-IDF.....	63
3.5.4 Selección y proyección de características.....	64
3.5.5 Entrenamiento del modelo.....	64
3.5.6 Evaluación de solución	65
CAPÍTULO 4 METODOLOGÍA.....	68
4.1 Clasificación de textos.....	68
4.1.1 Adquisición de los datos.....	70
4.1.2 Análisis y etiquetado de datos.	70
4.1.3 Construcción y ponderación de características.	71
4.1.4 Entrenamiento del modelo.....	71
4.1.5 Evaluación de la solución	72
CAPÍTULO 5 EXPERIMENTACIÓN Y RESULTADOS.....	73
CONCLUSIONES	81
REFERENCIAS	83
ANEXOS 92	

Lista de figuras

Figura 3.0 Máquina de Soporte Vectorial	42
Figura 3.1 Red Neuronal.....	49
Figura 3.2 Etapas de la clasificación de textos.....	58
Figura 3.3 (Mironczuk & Protasiewicz, 2018). Pasos básicos de la evaluación de modelos.....	65
Figura 4.0 Método propuesto.	69
Figura 5.1 Comparación de los resultados con respecto al estado del arte.	80

Lista de tablas

Tabla 4.0 Estructura del corpus “Muchocine”	70
Tabla 4.1 Representación vectorial de la ponderación booleana.....	71
Tabla 5.0 Estructura del corpus “Muchocine”	74
Tabla 5.1 Clases para el experimento 1.....	75
Tabla 5.2 Resultados del modelo con 2 clases.	75
Tabla 5.3 Resultados de la validación cruzada a 10 pliegues con el modelo de 2 clases.....	76
Tabla 5.4 Clases para el experimento 2.....	76
Tabla 5.6 Resultados de la validación cruzada a 10 pliegues con el modelo de 3 clases.....	77
Tabla 5.7 Resultados del modelo con 5 clases.	78
Tabla 5.8 Resultados de la validación cruzada a 10 pliegues con el modelo de 5 clases.....	78



CAPÍTULO

1

Introducción

La toma de decisiones es una parte esencial de nuestra vida diaria, y comprender las perspectivas de otras personas desempeña un papel fundamental en este proceso. A lo largo de la historia, las personas han buscado diversas formas de obtener información y orientación para tomar decisiones importantes (Flores Domínguez & Tapia Fabela, 2020).

Con la aparición de la web 2.0, hemos observado cómo nuestras vidas están cada vez más vinculadas al uso de las redes sociales, foros, aplicaciones web,

blogs y otros medios digitales (Insua Yañez, 2019), (Carrillo de Albornoz, 2011). En estos espacios, las personas expresan sus opiniones, comentarios, sentimientos, deseos y frustraciones. Esto ha cambiado la forma en que accedemos a la información, ya que miles de usuarios generan información de manera exponencial, haciendo imposible leer todas sus opiniones una por una. Esto plantea la pregunta fundamental de "¿Qué piensan las personas?". En la actualidad, esta pregunta puede responderse mediante la extracción, clasificación y análisis de lo que las personas escriben en el mundo digital. Por esta razón, muchas empresas consideran esta investigación como un proceso clave en la toma de decisiones. Poder comprender lo que la gente opina y "siente" sobre un producto o campaña es fundamental para tomar medidas adecuadas (Pérez Vera, 2017).

Debido a esto, existen múltiples esfuerzos para desarrollar modelos, técnicas y herramientas que permitan el análisis automático de textos. Una de las tareas importantes en el procesamiento del lenguaje natural (PLN) es la clasificación de textos, especialmente en lo que respecta a la opinión de los usuarios. En este contexto, ha surgido un tema de interés en los últimos años: la minería de opiniones (Henríquez Miranda et al., 2016), (Liu, 2012). La minería de opiniones se centra en analizar las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia diversas entidades, como productos, servicios, organizaciones, individuos, problemas, eventos, temas y sus características. Empresas y organizaciones en general desean conocer la reputación que tienen entre sus usuarios en las redes sociales, *blogs*, *wikis* y otros sitios web. Además, la mayoría de las investigaciones previas se han centrado en sistemas de minería de opiniones en inglés (Vilares, Alonso & Gómez Rodríguez, 2013), (Jiménez Zafra et al., 2014). Por lo tanto, existe un interés creciente en investigaciones en otros idiomas.

En el contexto del idioma español, para llevar a cabo la clasificación basada en opiniones, el primer paso implica la búsqueda de un recurso adecuado. Se debe verificar que la clasificación de las críticas de cine sea la más apropiada para la tarea de minería de opiniones. En este sentido, es esencial encontrar un corpus que cumpla con ciertos requisitos para ser considerado un recurso viable. Entre estos requisitos se incluye que cada crítica esté acompañada de la puntuación otorgada por su autor, lo que facilitará la distinción entre opiniones positivas, negativas o neutrales. Además, es fundamental que los textos generados por los usuarios mantengan un nivel mínimo de calidad (Cruz et al., 2008).

El corpus "Muchocine," desarrollado por Cruz, (Cruz et al., 2008), se ajusta a las características necesarias para ser considerado un recurso adecuado para la aplicación de la minería de opiniones. Este corpus se compone de un total de 3,878 críticas de películas en español. Cada crítica incluye información sobre el nombre del autor, el título de la película, la puntuación otorgada por el usuario, la calificación máxima posible, un breve resumen y la crítica en sí. Las críticas están valoradas en una escala del 1 al 5, donde 1 representa una crítica muy negativa y 5 indica una crítica muy positiva.

El sistema que estamos desarrollando integrará varios métodos para analizar datos con el objetivo de poder clasificar las opiniones en 3 clases como positivas, negativas y neutras, en 2 clases, positiva y negativa y en 5 clases muy mala, mala, neutra, buena y muy buena a partir del corpus "Muchocine" esto mediante aprendizaje supervisado.



1.1 Antecedentes

A lo largo de la historia de la humanidad, la comunicación ha sido una característica fundamental que nos distingue como especie. Nuestra capacidad de razonar y relacionarnos con otros seres humanos ha dado origen al desarrollo de la escritura como un medio para plasmar conocimientos, sentimientos y emociones en forma de documentos, poesía y obras maestras. A pesar de la diversidad temática de estas expresiones, todas comparten un elemento común: la información (Pérez Vera, 2017).

La comprensión de lo que piensan los demás desempeña un papel vital en el proceso de toma de decisiones (Pang & Lee, 2008). A lo largo de la historia, las personas han recurrido a diversas formas de búsqueda de orientación para tomar decisiones importantes. En tiempos antiguos, esto implicaba interpretar las entrañas de animales, observar señales de humo o analizar sueños. Culturas como la china empleaban la sabiduría poética y el I Ching, mientras que los griegos consultaban el oráculo de Delfos (Buchanan & O'Connell, 2006). En la década de 1990, las personas recurrían a amigos, conocidos, vecinos o familiares para obtener información sobre productos o servicios que deseaban adquirir (Flores Domínguez & Tapia Fabela, 2020).

Sin embargo, con la aparición de la web 2.0, la forma en que accedemos a información ha experimentado una transformación significativa. La proliferación de las redes sociales, foros, aplicaciones web, blogs y otras plataformas (Carrillo de Albornoz, 2011), (Flores Domínguez & Tapia Fabela, 2020) ha generado un crecimiento exponencial en la cantidad de opiniones y comentarios de los usuarios en línea, lo que hace imposible revisar manualmente cada una de estas expresiones.

Este abundante conjunto de datos es de gran interés para una variedad de sectores, incluyendo empresas, industrias y ámbitos académicos. No obstante, la extracción y procesamiento de esta información se ha vuelto una tarea compleja y desafiante cuando se aborda de forma manual (Henríquez Miranda et al., 2016).

En este contexto, las tendencias actuales de inteligencia artificial, en particular las técnicas de procesamiento de lenguaje natural (*PLN*), se presentan como una valiosa área de investigación. Dentro del campo de *PLN*, la minería de opiniones (*MO*) ha ganado especial atención en años recientes. La *MO* se enfoca en analizar las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia diversas entidades, que incluyen productos, servicios, organizaciones, individuos, temas y sus características (Liu, 2012).

La minería de opiniones ha emergido como un tema de investigación relevante debido a la vasta cantidad de opiniones y comentarios generados por usuarios activos en Internet. Las empresas y organizaciones, en general, buscan evaluar la percepción que los usuarios tienen de sus productos en plataformas digitales como redes sociales, *blogs*, *wikis* y otros sitios web (Henríquez Miranda et al., 2016), (Carrillo de Albornoz, 2011).

Aunque la mayoría de los estudios en esta área se han centrado en textos en inglés, también se han realizado investigaciones en otros idiomas. Por ejemplo, Cruz y sus colegas (Cruz et al., 2008) describen la creación de un corpus en español y realizan experimentos utilizando orientación semántica mediante clasificadores supervisados y no supervisados. Denecke (Denecke, 2008) utiliza un corpus en alemán para entrenar un clasificador que evalúa la polaridad de las opiniones. Zhang y su equipo (Zhang et al., 2009) aplican algoritmos basados en reglas y aprendizaje automático sobre corpus en chino. Agić, Ljubešić y Tadić (Agić et al., 2010) presentan un corpus en croata relacionado con los mercados financieros. Ghorbel, (Ghorbel & Jacot, 2011) emplean una aproximación basada en aprendizaje automático para analizar comentarios

sobre películas en francés, traduciendo el corpus al inglés para su integración con *SentiWordNet*.



1.2 Planteamiento del problema

En la sociedad actual, las opiniones desempeñan un papel fundamental, ya que se utilizan para evaluar una amplia gama de áreas y son de gran utilidad en investigaciones de mercado. El auge de las redes sociales y las plataformas que permiten a la población expresar sus opiniones ha llevado a un aumento en la necesidad de utilizar clasificadores de opiniones. Esto se debe a que estas plataformas albergan una enorme cantidad de datos, y las empresas y organizaciones buscan acceder únicamente a la información relevante (De León Camacho & Rosas Sánchez, 2017), (Henríquez Miranda et al., 2016).

Dada la abrumadora cantidad de opiniones de los usuarios, es necesario recurrir a técnicas de inteligencia artificial para procesar esta información. Estas técnicas involucran diversos modelos y metodologías desarrollados específicamente en el campo de la Minería de Datos, (Henríquez Miranda et al., 2016), (Cruz et al., 2008), (Martínez Cámara et al., 2011).

Las redes sociales se han convertido en un espacio de expresión libre para los usuarios, y es aquí donde se generan opiniones, sentimientos, anhelos y frustraciones las empresas tienen un interés legítimo en conocer las opiniones de sus clientes en potencia. De manera similar, en el ámbito político, es esencial detectar las opiniones de posibles votantes. En la industria hotelera, por ejemplo, plataformas como Tripadvisor se utilizan para detallar aspectos importantes para los usuarios, como la ubicación, limpieza, tamaño y precio, entre otros. Del mismo modo, en la industria cinematográfica, resulta fundamental comprender las opiniones de los espectadores sobre una película, lo que ayuda a determinar si vale la pena verla o no. Estas opiniones son de gran relevancia para la toma de decisiones y la creación de nuevos proyectos. Por lo tanto, es esencial clasificar estas opiniones para extraer información valiosa que pueda utilizarse para mejorar la producción de contenido de alta calidad (Flores Domínguez & Tapia Fabela, 2020), (Pérez Vera, 2017).



1.3 Objetivos

Objetivo general

Desarrollar un sistema para el análisis de opiniones mediante la clasificación automática de textos que utilice métodos de aprendizaje automático para identificar opiniones en el corpus “Muchocine”.

Objetivos específicos

1. Identificar las etapas para realizar el análisis de datos.
2. Identificación de corpus para el análisis de opiniones.
3. Desarrollar un método para la clasificación automática de textos
 - a. Seleccionar características mediante un método de filtro.
 - b. Preprocesamientos
 - c. Modelo de representación vectorial
 - d. Ponderación booleana
 - e. Establecer un modelo de aprendizaje.
4. Aplicar técnicas de aprendizaje supervisado, como SVM, NV y LR
5. Planificar la estrategia experimental.
6. Proceder con el entrenamiento del modelo propuesto basándose en los patrones textuales.
7. Evaluar el rendimiento del modelo propuesto con la métrica de exactitud.



1.4 Alcances y Limitaciones

Alcances

- Clasificación automática de opiniones en 2 clases (positivo, negativo), 3 clases (positivo, negativo y neutral), 5 clases (muy mala, mala, neutra, buena, muy buena).
- Emplear técnicas utilizadas en la clasificación de opiniones: *SVM*, *NB* y *RL*

Limitaciones

- El corpus en español: La cantidad de datos disponibles en español puede ser restringida en comparación con otros idiomas que son más comunes en la investigación de la clasificación automática de texto.
- El corpus en español es muy grande, por lo que hay una gran limitación en cuanto al tiempo de la experimentación.
- Falta de contexto: El sistema podría enfrentar desafíos al interpretar el entorno en el cual se expone una opinión, lo que podría causar errores en la clasificación de las opiniones.
- Exploración léxica: La exploración léxica se enfoca en el manejo de las propiedades léxicas de los textos, tales como palabras clave, frecuencia de términos y evaluación de sentimientos. No obstante, este enfoque podría descuidar otros elementos esenciales del lenguaje, tales como la estructura gramatical, el significado y la coherencia del discurso.



1.5 Justificación

Con el paso de los años, los seres humanos han buscado constantemente formas de comunicarse. A través de ella, las personas han plasmado conocimientos, emociones y pensamientos en diversos tipos de documentos, poesía y obras maestras. A pesar de las diferencias en su contenido, todos estos

elementos comparten un componente común: la información (Pérez Vera, 2017).

El conocimiento de las opiniones de otras personas desempeña un papel fundamental en el proceso de toma de decisiones. Con la llegada de la Web 2.0, la forma en que accedemos a información experimentó un cambio significativo (Flores Domínguez & Tapia Fabela, 2020). Los medios sociales, en particular, han revolucionado la manera en que accedemos a información nueva (Amaral, 2020). Los usuarios participan activamente en las redes sociales compartiendo fotos, comentando publicaciones de otros, expresando sus opiniones sobre productos adquiridos y reflexionando sobre asuntos de actualidad, entre otras cosas. Dado este panorama, el interés de las empresas por conocer las opiniones de los usuarios sobre sus productos o su propia entidad se ha vuelto especialmente relevante como por ejemplo en el sector hotelero, plataformas como TripAdvisor proporcionan detalles valorados por los usuarios, como la ubicación, higiene, espacio y costo mientras que, en empresas dedicadas a la cinematografía, es relevante conocer las percepciones de la audiencia sobre una película para determinar si merece la pena verla. (Insua Yañez, 2019), (De León Camacho & Rosas Sánchez, 2017).

Las empresas y organizaciones en general sienten un gran interés por evaluar la reputación de sus productos en plataformas digitales como las redes sociales y los blogs, esto con el fin de mejorar la marca, producto o servicio con base a las opiniones expresadas por parte de los usuarios por lo que el internet se ha convertido en un espacio donde los usuarios comparten una amplia variedad de opiniones y comentarios sobre una diversidad de temas. (Amaral, 2020).

No obstante, cada vez que alguien intenta explorar lo que otros piensan sobre algo, se enfrenta a una gran cantidad de datos que a menudo dificultan la búsqueda de información relevante para un propósito específico (Amaral,

2020). Aunque estas opiniones son valiosas desde una perspectiva comercial, realizar este proceso de manera manual resulta complicado y laborioso. El análisis de datos surge como una solución para procesar grandes volúmenes de información, identificando datos significativos que respalden una toma de decisiones eficiente y efectiva. En este contexto, se propone la creación de un sistema que integre diversos métodos de análisis de datos para automatizar la clasificación de opiniones.

Esta iniciativa busca mejorar la toma de decisiones empresariales al permitir que las empresas ajusten sus estrategias, productos o servicios según la información proporcionada por las opiniones de los usuarios. Este enfoque tiene un impacto positivo en la toma de decisiones y en el rendimiento general de las empresas.



1.6 Hipótesis

Si se desarrolla un sistema para el análisis de opiniones mediante la clasificación automática de textos se podrán clasificar las opiniones en el corpus "Muchocine" con una exactitud superior al 80%.



CAPÍTULO

2

Estado del arte

En este capítulo, se realiza una revisión exhaustiva de los estudios y desarrollo en el ámbito de la clasificación automática de textos. Se exploran enfoques contemporáneos que incluyen métodos basados en reglas, aprendizaje supervisado y técnicas de aprendizaje profundo, con el propósito de obtener una comprensión completa del estado actual de investigación en este campo.

3.1 Análisis de Sentimientos a nivel de documento en críticas de cine en español.

Flores Domínguez, (Flores Domínguez & Tapia Fabela, 2020) discute la importancia de conocer las opiniones de otras personas durante el proceso de toma de decisiones. A lo largo de la historia, la forma en que se toman las decisiones ha evolucionado, y con la aparición de la web 2.0 y la proliferación de las redes sociales, la cantidad de información disponible ha crecido de manera exponencial. Esto hace que resulte prácticamente imposible revisar manualmente todas las opiniones expresadas en estos medios.

Por lo tanto, se planteo el desarrollo de un modelo de análisis de sentimientos a nivel de documentos que utilice las opiniones recopiladas en el corpus muchocine. Este modelo emplea una red neuronal de retro propagación (*back-propagation*) para clasificar el corpus en cinco categorías distintas.

Este método consta de tres etapas distintas. En la primera etapa, se realiza un preprocesamiento del corpus, que implica la eliminación de acentos en las palabras, la eliminación de *stopwords* y signos de puntuación. En la segunda etapa, las críticas que componen el corpus muchocine se representan como un conjunto de vectores binarios utilizando la técnica de bolsa de palabras y el pesado booleano. La tercera etapa implica la clasificación del corpus en cinco categorías utilizando una red neuronal de retro propagación (*back-propagation*), con la aplicación de una validación cruzada de 10 particiones. Luego, se procede a evaluar la precisión (*accuracy*) obtenida por esta red.

Para evaluar la precisión de la red neuronal de retro propagación, se crearon cinco conjuntos de entrenamiento diferentes, cada uno con características específicas. El primer conjunto de entrenamiento no experimenta ningún tipo de preprocesamiento. En el segundo conjunto, se realizó la sustitución de palabras acentuadas. En el tercero, se eliminaron las *stopwords*. En el cuarto conjunto, se eliminaron los signos de puntuación. El quinto conjunto

experimentó tanto la sustitución de acentos como la eliminación de *stopwords*. Por último, el sexto conjunto se sometió a la sustitución de acentos, la eliminación de signos de puntuación y *stopwords*.

Es esencial destacar que en esta investigación se emplearon las 3878 críticas incluidas en el corpus muchocine del año 2008. El Análisis de Sentimientos se llevó a cabo a nivel de documento. Además, no se consideraron emoticonos en el análisis debido a la estructura del corpus.

El enfoque propuesto busca clasificar las críticas en cinco categorías distintas: críticas con puntuación 1, que son altamente negativas; críticas con puntuación 2, que son negativas; críticas con puntuación 3, que son neutras; críticas con puntuación 4, que son positivas; y críticas con puntuación 5, que son altamente positivas.

La métrica utilizada para evaluar el rendimiento del modelo de entrenamiento fue *accuracy* y se observaron variaciones en los resultados en cada una de las muestras de entrenamiento.

En particular, la sexta muestra de entrenamiento arrojó el resultado más sobresaliente en el experimento, con una precisión del 0.8499. En esta muestra, se aplicaron tres procesos de preprocesamiento diferentes: la sustitución de acentos, la eliminación de *stopwords* y la eliminación de signos de puntuación. El preprocesamiento desempeñó un papel crucial al ayudar al clasificador a interpretar de manera precisa las críticas. La eliminación de *stopwords*, debido a su alta frecuencia semántica, pero bajo valor informativo, contribuyó significativamente a la mejora de la comprensión del texto.

La quinta muestra de entrenamiento obtuvo un nivel de precisión del 0.8383, lo que indica resultados aceptables. En esta muestra, se aplicaron dos tipos de preprocesamiento al corpus, lo que contribuyó a mejorar la calidad general de los textos. En contraste, la segunda muestra de entrenamiento, que se

centró en la sustitución de acentos en el corpus, logró un nivel de precisión de 0.7891.

En la tercera muestra de entrenamiento, donde se eliminaron las stopwords del corpus, se obtuvo un nivel de precisión del 0.7829. Este resultado se atribuye a la eliminación de palabras que, debido a su alta frecuencia semántica, carecían de valor en el análisis.

Por otro lado, la primera muestra de entrenamiento, que no incluyó ningún proceso de preprocesamiento, alcanzó un nivel de precisión de 0.6877. Esto se debió a que el texto sin preprocesamiento fue interpretado de manera incorrecta por el clasificador.

Finalmente, la cuarta muestra de entrenamiento, que se caracterizó por la eliminación de signos de puntuación en el corpus, registró un nivel de precisión de 0.6091. Este resultado más bajo se atribuye a la importancia de los signos de puntuación en la determinación de las funciones gramaticales y las relaciones sintácticas en el texto, lo que dificultó la interpretación adecuada del mismo.

3.2 Técnicas de clasificación de opiniones aplicadas a un corpus en español.

Martínez Cámara, (Martínez Cámara et al., 2011), presenta un experimento realizado en un conjunto de opiniones en español sobre películas con el objetivo de evaluar el desempeño de varios clasificadores entrenados para determinar la polaridad de estas opiniones.

Para llevar a cabo los experimentos, se utilizó un corpus en español compuesto por críticas de cine. Este conjunto consta de 3,878 críticas obtenidas del sitio web "muchocine", donde las críticas se califican en una escala del 1 al 5, donde 1 representa una película de muy baja calidad y 5 indica una película excelente. Las películas calificadas con un 3 se consideran neutrales o que el

crítico no las considera ni buenas ni malas, y no se incluyeron en el estudio. De todos los campos disponibles para cada documento, como el nombre del autor, el título de la película, la puntuación asignada, etc., en la experimentación se enfocaron exclusivamente en la puntuación de la película, el resumen de la crítica (*Re*) y la crítica desarrollada (*Cr*).

Debido a esto, se crearon tres conjuntos de datos: uno que incluye solo los resúmenes (*Re*), otro solo con las críticas (*Cr*), y el último que consiste en una combinación de resúmenes y críticas (*Re_Cr*).

Se aplicaron dos algoritmos de aprendizaje automático, *SVM* y el clasificador Naive Bayes, a cada uno de estos conjuntos de datos con diferentes técnicas de preprocesamiento. Se experimentó con la inclusión o exclusión de *stopwords* y *stemmers*, además de utilizar distintos esquemas de pesado como *TF-IDF*, frecuencia de ocurrencia (*TF*), número de ocurrencias (*TO*) y ocurrencia binaria (*BTO*).

Las métricas de evaluación empleadas son las convencionales en recuperación de información: precisión, *recall* y *F1*. Estas evaluaciones se llevaron a cabo utilizando una validación cruzada de 10 particiones.

Los resultados de los experimentos revelaron una variedad de datos. Para el algoritmo *SVM*, los mejores resultados se obtuvieron en el conjunto de datos *Re_Cr* utilizando el esquema de pesado *TF-IDF* sin aplicar ningún mecanismo de reducción de características, con una precisión del 87.73%, *recall* del 87.69%, y *F1* del 87.71%. Por otro lado, el algoritmo Naive Bayes no superó los resultados obtenidos por *SVM*, aunque mostró un comportamiento similar. Los mejores resultados en Naive Bayes se lograron en el conjunto de datos *Re_Cr* después de aplicar *stopwords* y *TF-IDF*, con una precisión del 83.62%, *recall* del 84.01%, y *F1* del 84.04%.

En resumen, los resultados sugieren que la combinación de *stopwords*, *stemmers*, *TF-IDF* y *SVM* proporciona el mejor rendimiento en este contexto de análisis de opiniones en español.

3.3 Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español.

Cruz, (Cruz et al., 2008), aborda el creciente interés en el procesamiento automatizado de opiniones contenidas en documentos de texto. Este interés se ha intensificado debido al aumento exponencial de los contenidos generados por los usuarios en la web, así como al interés de empresas y gobiernos en analizar, filtrar y detectar automáticamente las opiniones expresadas por sus clientes o ciudadanos.

Para comenzar, se emprendió la tarea de encontrar un recurso adecuado para este estudio. Se decidió construir un corpus utilizando críticas de películas extraídas de las páginas *HTML* de "Muchocine2" con fecha de febrero de 2008. Estas páginas *HTML* se transformaron en archivos *XML*, uno por cada crítica. Cada archivo *XML* contenía el texto de la crítica, el nombre del autor, el nombre de la película comentada, la puntuación otorgada (en una escala del 1 al 5) y un breve resumen de la crítica escrito por el autor.

El corpus resultante comprende un total de 3,878 críticas y aproximadamente 2 millones de palabras, con un promedio de 546 palabras por crítica.

El algoritmo de clasificación propuesto se basó en la extracción de bigramas a partir de patrones morfosintácticos simples. Se asumió que este conjunto de bigramas contenía al menos algunos que expresaban opiniones, junto con otros que no lo hacían. Luego, para cada bigrama extraído, se calculó su orientación semántica (positiva o negativa) utilizando el algoritmo *PMI-IR* (*Pointwise Mutual Information - Information Retrieval*). La clasificación de la crítica se basó en la suma de las orientaciones semánticas obtenidas: se

consideró positiva si el valor calculado era mayor o igual a cero y negativa en caso contrario.

Se realizaron experimentos de clasificación utilizando una muestra aleatoria de 400 críticas del corpus. De estas, 200 tenían una puntuación de 1 o 2 y se clasificaron como críticas negativas, mientras que las otras 200 tenían una puntuación de 4 o 5 y se clasificaron como críticas positivas. No se utilizaron todas las críticas del corpus debido a limitaciones de tiempo.

Los resultados en el cálculo de la orientación semántica se basaron en la extracción de algunos de los bigramas del corpus y la orientación semántica obtenida utilizando una sola semilla por clase (*SOv1*) o varias semillas por clase (*SOv2*).

Al concluir los experimentos, se determinó que los mejores resultados se obtuvieron con el clasificador *SOv1-S*, que logró una precisión del 77.5%, y el *SOv2-NS*, con un 69.5%

3.4 Una revisión a la minería de opiniones y los retos del *PNL*

Nigro, (Nigro, 2020) se enfoca en los desafíos relacionados con el desarrollo de herramientas de minería de opiniones, que son esenciales para el análisis de redes sociales. Estas herramientas están siendo adaptadas para su uso en medios sociales, lo que plantea desafíos particulares para el procesamiento de lenguaje natural (*PLN*).

Los medios sociales presentan una serie de desafíos inherentes al aplicar técnicas convencionales de minería y análisis de sentimientos. Los *microposts*, como los *tweets*, son especialmente desafiantes para las herramientas de minería de textos y, en particular, para la minería de opiniones. Estos mensajes suelen carecer de un contexto extenso y a menudo implican conocimiento implícito. La ambigüedad es un problema particular, ya que no se puede recurrir fácilmente a una referencia central de información. A diferencia de las

publicaciones de *blogs* y los comentarios, los *tweets* generalmente no siguen un hilo de conversación y se presentan de manera aislada de otros *tweets*.

En la mayoría de las técnicas de minería de opiniones, se emplea el aprendizaje automático como enfoque principal. Sin embargo, esto presenta limitaciones cuando se trata de aplicaciones que involucran diversos dominios, idiomas y tipos de texto, ya que se requiere entrenar modelos específicos para cada uno de ellos, lo que demanda grandes cantidades de datos de entrenamiento para lograr resultados satisfactorios.

Casi todos los estudios relacionados con la minería de opiniones en Twitter han utilizado técnicas de aprendizaje automático. Su objetivo principal ha sido clasificar los *tweets* de manera automática en categorías de sentimientos positivos, negativos y neutros. Esto se logra mediante la construcción de un clasificador binario sencillo que utiliza características como n-gramas y el etiquetado de partes del discurso (*POS*). Estos clasificadores se entrenan utilizando ejemplos previamente anotados según la presencia de sentimientos positivos y negativos (Pak & Paroubek, 2010).

Además, hay una amplia variedad de herramientas comerciales que utilizan enfoques basados en búsquedas para llevar a cabo análisis de sentimientos en *tweets*. En términos generales, el usuario introduce un término de búsqueda y estas herramientas recuperan todos los *tweets* que contienen el término, categorizándolos como positivos o negativos (y ocasionalmente como neutros). A menudo, estas herramientas también proporcionan representaciones gráficas, como gráficos circulares u otros tipos de gráficos.

Por lo que al finalizar la revisión de la literatura se llegó a la conclusión de que el avance en el desarrollo de herramientas de minería de opinión está en curso, y los resultados iniciales son alentadores. Se tiene confianza en que las estrategias de retroceso que forman parte de la metodología incremental conducirán a un sistema exitoso. Se defiende la idea de utilizar en gran medida

técnicas superficiales en el procesamiento del lenguaje, como la fragmentación en lugar de un análisis completo.

3.5 Minería de opinión, acceso abierto y gobernanza.

Nigro, (Nigro, 2020b) plantea un nuevo enfoque en las necesidades de investigación relacionadas con la minería de opiniones, el acceso abierto a la información y la gobernanza. Menciona que el concepto de minería de opiniones se está volviendo más complejo y completo debido a las técnicas y tecnologías relacionadas. Sin embargo, a medida que avanza la tecnología de la información, observa una disminución en la formulación de políticas, leyes, normativas y regulaciones. También señala que el impacto del uso de la información en un mundo globalizado y polarizado está disminuyendo.

Las redes sociales han abierto oportunidades sin precedentes para que las personas expresen públicamente sus opiniones, lo que ha dado lugar a numerosas situaciones en ámbitos como lo público, religioso, político y económico. Por lo tanto, plantea la necesidad de controlar las opiniones en las redes sociales para comprenderlas y monitorear los comentarios e intereses relacionados.

En el ámbito de la investigación sobre deliberación en línea, se han desarrollado herramientas como *Compendium*, *Debatepedia*, *Cohere* y *Debategraph*. Estas herramientas están diseñadas para proporcionar una estructura lógica a una serie de declaraciones y para vincular argumentos con la evidencia que los respalda. En el contexto político, esto resulta muy útil ya que permite a los votantes comprender cuáles partidos políticos o individuos tienen posiciones más cercanas a las suyas. Por ejemplo, *SmartVote.ch* solicita a los votantes que indiquen su nivel de acuerdo con una serie de declaraciones políticas y luego compara sus posiciones con las de los partidos políticos.

Las aplicaciones de minería de opiniones desempeñan un papel fundamental como base para la creación colaborativa de políticas a gran escala. Estas aplicaciones tienen la capacidad de dar coherencia a una gran cantidad de contribuciones y son valiosas en la detección temprana de posibles problemas al capturar la retroalimentación inicial de los ciudadanos. A menudo, para recopilar comentarios de manera estructurada, se han utilizado encuestas específicas, pero este enfoque es costoso, ya que requiere una inversión significativa en diseño y recopilación de datos, y puede ser complicado, ya que las personas a menudo no están dispuestas a participar en encuestas.

La minería de opiniones resulta beneficiosa al detectar problemas sin la necesidad de hacer preguntas directas, lo que permite reflejar de manera más precisa la realidad. Así mismo, el *software* de mapeo de argumentos se convierte en una herramienta valiosa para asegurar que los debates políticos sean racionales y basados en pruebas sólidas, evitando repeticiones constantes de los mismos argumentos. El interés gubernamental en la minería de opiniones no es algo nuevo, sino que se remonta a los primeros días de la sociedad de la información. En consecuencia, los métodos automatizados para el análisis de contenido han experimentado un crecimiento exponencial.

El mercado de herramientas para la minería de opiniones está repleto de proveedores que ofrecen soluciones. La mayoría de estas aplicaciones se centran en analizar las opiniones de los clientes sobre productos y servicios, priorizando el análisis de sentimientos que identifica emociones positivas o negativas mediante la comprensión del lenguaje natural. Además, hay herramientas de libre acceso, muchas de las cuales se enfocan en proporcionar asesoramiento al votante y ofrecen un mapeo avanzado de argumentos. Estas herramientas de acceso gratuito suelen ser desarrolladas en gran medida por la comunidad académica y organizaciones no gubernamentales.

Por lo que se concluyó que las actuales soluciones para la minería de opiniones y el análisis de sentimientos están avanzando rápidamente, lo que suele llevar

a una disminución en la necesidad de intervención humana para clasificar los comentarios. Sin embargo, se han identificado varios desafíos, como la detección de spam y revisiones falsas, que a menudo se abordan mediante la identificación de duplicados. Además, se busca mejorar la comparación entre revisiones cualitativas y resúmenes de reseñas, así como la detección de valores atípicos y la evaluación de la reputación de los revisores.

3.6 Predicción electoral usando un modelo híbrido basado en análisis sentimental: elecciones presidenciales de Colombia.

Callejas Cuervo, (Callejas Cuervo & Vélez Guerrero, 2019) propusieron un modelo híbrido con el propósito de predecir el resultado de la primera vuelta de las elecciones presidenciales de Colombia en 2018. El objetivo principal de este modelo es minimizar el error absoluto y mejorar la calidad de las predicciones finales. Para lograr esto, se registraron y analizaron las actividades de los usuarios en Twitter y Facebook mediante el uso de algoritmos de inteligencia artificial. Los resultados obtenidos demostraron que este enfoque proporciona predicciones precisas que concuerdan coherentemente con la realidad.

Esto se debe a que las redes sociales, en particular Twitter y Facebook, están experimentando un aumento significativo en su uso debido a su amplia penetración y popularidad. Estas plataformas albergan una cantidad creciente de información que incluye opiniones e interacciones, lo que proporciona una descripción detallada de los comportamientos sociales. Dado el volumen masivo de datos disponible, ha surgido un interés creciente en la predicción de fenómenos relacionados con los usuarios, incluyendo la capacidad de prever eventos de alcance poblacional.

La metodología empleada en esta investigación se divide en tres etapas fundamentales. En primer lugar, se llevó a cabo un monitoreo de la actividad en las redes sociales relacionada con las campañas presidenciales, incluyendo la recolección de opiniones y la determinación de la polaridad sentimental hacia los candidatos y sus propuestas. En segundo lugar, se realizó un seguimiento y recopilación de datos numéricos de encuestas llevadas a cabo por empresas privadas para los medios de comunicación en Colombia. Estas encuestas se difundieron al público durante el período de la campaña electoral. Por último, se fusionaron los datos obtenidos de las redes sociales con los resultados de las encuestas para obtener los resultados finales.

Es importante destacar que el seguimiento y la obtención de los resultados finales se llevaron a cabo de manera anticipada, es decir, antes de la celebración de las elecciones presidenciales en la primera vuelta.

La base de datos utilizada en este estudio consta de tres conjuntos de datos independientes, todos ellos obtenidos de las redes sociales, específicamente de Twitter. Estos conjuntos de datos se recopilaron durante un período de 120 días, abarcando tanto el período previo como el transcurso de la campaña electoral.

En la primera etapa de análisis de datos de redes sociales, se aplican técnicas de *Procesamiento del Lenguaje Natural (PLN)* y *Análisis Sentimental (AS)* de manera general para evaluar la favorabilidad y la imagen de los candidatos. Estos candidatos se clasifican según la polaridad de los sentimientos, que puede ser Positiva, Negativa o Neutral. En este estudio, se emplea específicamente un modelo de *Máquinas de Soporte Vectorial Lineales (LSVM)*. Los conjuntos de datos *TWCo18* y *FBCo18* se dividen en tres grupos: un conjunto de entrenamiento (30%), un conjunto de validación (10%) y un conjunto de predicción (60%).

En la segunda etapa, que aborda el análisis de datos procedentes de encuestas y sondeos de opinión, se utiliza *Facebook Prophet*. Este modelo tiene

la capacidad de realizar predicciones numéricas precisas mediante la inferencia de datos de entrada. Durante esta fase, los datos contenidos en *OPCo18* se analizan como observaciones en series de tiempo en el marco del modelo propuesto.

El modelo híbrido, que corresponde a la etapa tres, fusiona la información para predecir los resultados electorales utilizando un modelo de aprendizaje basado en celdas *Long Short-Term Memory (LSTMs)*.

En las tres etapas se calcula la función de pérdida *RMSE* y se obtienen los siguientes resultados: en la etapa con *LSVM* se alcanza un resultado del 5.36%, en la etapa de pronóstico con *FB Prophet* se registra un 6.37%, y en el modelo híbrido con *LSTMs* se logra el mejor porcentaje con un 2.47%.

En conclusión, se ha observado que las predicciones basadas en los sentimientos expresados por la gente a través de plataformas digitales pueden generar resultados precisos para anticipar los resultados electorales. El hallazgo principal indica que la fiabilidad aumenta cuando se combina la información de las interacciones sociales con los datos numéricos de encuestas.

Al comparar los resultados obtenidos mediante la metodología empleada (Callejas Cuervo & Vélez Guerrero, 2019) con el margen de error de algunas de las firmas encuestadoras en la tarea de predecir los resultados finales, se evidencia una clara ventaja al utilizar el modelo híbrido (con un 2.19% menos de error que *Invamer*, un 4.98% menos de error que *Cifras y Conceptos*, y un 6.99% menos de error que *Guarumo*).

3.7 Minería de opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles.

Henríquez Miranda, (Henríquez Miranda et al., 2016) se enfoca en el interés de las empresas y organizaciones por evaluar la reputación que sus usuarios tienen

en plataformas como redes sociales, blogs, wikis y otros sitios web. El estudio se centra en la creación de un sistema de minería de opiniones en español que analiza comentarios de clientes sobre diferentes hoteles. Esta minería de opiniones busca comprender las opiniones, sentimientos, valoraciones, actitudes y emociones de las personas hacia diversas entidades, como productos, servicios, organizaciones, individuos, problemas, sucesos, temas y sus atributos.

El recurso utilizado en esta investigación se basa en *LEX*, que es una adaptación al español de *ANEW*. Esta adaptación implicó la traducción al español de las 1,034 palabras originales de *ANEW* (Bradley & Lang, 1999) y fue realizada por un filólogo profesional especializado en el idioma inglés.

Para poner a prueba el sistema, se eligió el dominio del turismo y se recopilaron 1,816 opiniones de usuarios sobre hoteles. Estas opiniones se catalogaron en una escala de cinco niveles (1 para negativo y 5 para positivo).

En cuanto a la metodología, el modelo consta de cuatro fases: extracción, preprocesamiento, identificación de sentimientos y clasificación de la opinión. En la primera fase, se extraen los textos que contienen las opiniones de diversas fuentes en línea, como la web y redes sociales. La siguiente fase implica la aplicación de varias técnicas para limpiar el texto, como la eliminación de palabras y signos irrelevantes, corrección ortográfica, normalización y lematización, entre otras.

En la fase de identificación de sentimientos, se trabaja con el texto normalizado para identificar las características esenciales que revelen los sentimientos asociados a las opiniones. La última fase asigna polaridades a las palabras identificadas previamente, es decir, determina los sentimientos asociados. Para esto, se utiliza *ANEW*, centrándose en la dimensión de valencia.

Para evaluar el sistema, se llevaron a cabo cuatro experimentos, utilizando la precisión como métrica de validación. En el primer experimento, se analizaron

palabras utilizando solo *ANEW* sin considerar negaciones. En el segundo experimento, se analizaron palabras utilizando *ANEW* y se manejaron negaciones. En el tercer experimento, se utilizaron palabras de *ANEW* junto con un diccionario de sentimientos creado manualmente, sin tratar negaciones. Finalmente, en el cuarto experimento, se usaron palabras de *ANEW* junto con un diccionario de sentimientos creado manualmente y se manejaron negaciones.

Los resultados variaron para cada experimento, y el experimento 4 obtuvo el mejor rendimiento, con una precisión del 94%. Esto se debió a la inclusión de un diccionario manual que contenía palabras representativas de sentimientos positivos y negativos que no estaban presentes en *ANEW*. En cuanto al manejo de negaciones, el algoritmo utilizado para reconocer este tipo de opiniones demostró eficacia al aumentar los resultados en un 3% (Henríquez Miranda et al., 2016).



CAPÍTULO

3

Marco teórico

El presente capítulo del marco teórico tiene como propósito ofrecer una visión panorámica y minuciosa de los aspectos teóricos y metodológicos relacionados con la minería de opiniones. Al explorar enfoques y técnicas fundamentales, se establecerán los cimientos para desarrollar la metodología de investigación e implementar modelos y algoritmos con el objetivo de alcanzar los objetivos planteados en este estudio.

3.1 Opinión

En el campo de la informática, las opiniones pueden ser representadas como una "quíntupla" que consta de cinco elementos: el objeto que está siendo evaluado, la característica específica de ese objeto que se está comentando, el sentimiento o actitud hacia esa característica, la persona que emite la opinión y el momento en el que se realiza la opinión (Liu, 2012).

Las opiniones proporcionan una forma de evaluar cómo las personas ven un tema o producto específico (Guzmán Cabrera, 2019). Además, estas opiniones pueden detallar aspectos o atributos específicos de una entidad, lo que significa que, en una revisión de un producto, el autor puede compartir tanto aspectos positivos como negativos sobre las características individuales del producto (López Ramos & Arco García, 2019).

3.1.1 Opinión pública

El término "opinión pública" abarca diversas definiciones. Por ejemplo, según el diccionario de la lengua española, se refiere al "sentir o estimación en que coincide la generalidad de las personas acerca de asuntos determinados". Aunque esta definición incluye términos comúnmente asociados con la opinión pública, también señala que el concepto implica una realidad más compleja. En su origen griego y latino, "opinión pública" tendría dos significados: uno relacionado con aspectos sociopolíticos y la noción de polis, y otro vinculado a un aspecto visual-intelectual, conectado con la idea de exhibición (Rueda & Alamán, 2020).

3.1.2 Comentarios

Los comentarios se refieren a expresiones verbales o escritas que contienen opiniones, evaluaciones o críticas sobre diversos temas, como productos, servicios o eventos. Estos comentarios pueden ser generados por usuarios en línea en varias plataformas, como redes sociales, sitios web de reseñas, *blogs* o foros de discusión. En entornos en línea, los comentarios se caracterizan por ser parte de conversaciones colectivas en las que la interacción entre múltiples usuarios trasciende las limitaciones geográficas y físicas. Además, el análisis de comentarios implica trabajar con el lenguaje natural, lo que lo convierte en una técnica esencial para comprender el contenido de los comentarios y extraer información relevante de ellos (Romero, 2016).

3.1.3 Análisis de comentarios

El análisis de comentarios se refiere al proceso de examinar y extraer información valiosa de las opiniones y comentarios expresados por individuos en línea sobre una variedad de temas, como productos, servicios o eventos. Este análisis se lleva a cabo utilizando técnicas de procesamiento de lenguaje natural y minería de datos con el propósito de obtener una comprensión más profunda de la percepción pública y obtener información que pueda ser útil para la toma de decisiones.

Es esencial destacar que el análisis de comentarios implica trabajar con el lenguaje natural, lo que lo convierte en una técnica fundamental para este tipo de análisis, ya que los comentarios se expresan en palabras y frases en el lenguaje cotidiano (Romero, 2016).

3.2 Inteligencia Artificial

La inteligencia artificial ha sido influenciada por diversas disciplinas a lo largo de la historia. La filosofía, desde hace más de 2000 años, ha explorado teorías sobre el razonamiento y el aprendizaje, considerando la mente como una máquina codificada y el pensamiento como el determinante de la acción correcta. Las matemáticas proporcionaron las herramientas necesarias para la lógica y la incertidumbre estadística, mientras que la psicología brindó herramientas para investigar la mente humana. La lingüística se volvió esencial para representar y manejar el conocimiento, y las ciencias de la computación han permitido que la IA se convierta en una metodología aplicada.

Sin embargo, la idea de desarrollar máquinas autónomas surge en el siglo XVII con filósofos como John Locke y Gottfried Leibniz planteando preguntas sobre la posibilidad de que las máquinas pudieran tener conocimiento. En el siglo XX, en 1936, Alan Turing diseñó una máquina que pensaba y realizaba tareas como los seres humanos, marcando un punto crucial en la evolución del concepto de la IA y en el desarrollo de los procesos y metodologías para su implementación (Martinsanz y Peñas, 2005).

La inteligencia artificial se define como el esfuerzo por emular la inteligencia humana en las máquinas, y según Russell (Russell, 2019), la inteligencia artificial recibe información del entorno y toma acciones mediante algoritmos de Procesamiento del Lenguaje Natural (PLN). Se pueden identificar cuatro tipos de inteligencia artificial, según Maldonado Ramones (2022):

1. Máquinas reactivas: Estos sistemas se basan en principios básicos de Inteligencia artificial y responden a situaciones específicas sin almacenar información previa.

2. Memoria limitada: La inteligencia artificial con memoria limitada guarda información y predicciones anteriores al recopilar datos, empleando tres modelos distintos: aprendizaje por refuerzo, memoria a corto plazo y redes antagónicas.
3. Teoría de la mente: Este enfoque de IA busca comprender cómo se sienten los humanos y se basa en inferir estados mentales y emociones.
4. Conciencia de sí mismo: Se prevé que en algún momento la IA pueda desarrollar una conciencia de sí misma.

La inteligencia artificial se enfoca en aprovechar técnicas de cálculo con instalaciones avanzadas de investigación y pronóstico para procesar diversos tipos de datos, lo que permite la toma de decisiones (Morán López, 2022).

3.2.1 Aprendizaje supervisado

Los algoritmos supervisados son aquellos en los que el sistema es capaz de aprender a partir de datos de entrenamiento previamente etiquetados (Escortell Pérez, 2017). Su funcionamiento se basa en descubrir la relación entre valores de entrada, conocidos como características, y una o más variables de salida, denominadas etiquetas. Para lograr este aprendizaje, se enseña al algoritmo cuál es el resultado deseado para un valor específico, y mediante el entrenamiento con múltiples ejemplos, el algoritmo es capaz de combinar estos valores y realizar predicciones, incluso con datos que no ha visto previamente (Herrera Contreras, 2020).

El objetivo principal de este enfoque de aprendizaje es encontrar una función que pueda generar el valor de salida correcto para un patrón de entrada dado (Siurob Palomero, 2019). Estos algoritmos son considerados los más

sencillos de implementar y comprender. Se basan en un conocimiento previo, donde los datos de entrenamiento consisten en tuplas (X, Y) , donde X representa las variables que predicen una salida determinada Y . La variable por predecir, Y , puede ser de naturaleza cuantitativa en problemas de regresión o de tipo cualitativo en problemas de clasificación dentro de un modelo de aprendizaje supervisado (Cedeno Moreno & Vargas Lombardo, 2020).

En el ámbito de los algoritmos supervisados, podemos identificar dos enfoques principales: los algoritmos de regresión y los de clasificación (Escortell Pérez, 2017):

1. Regresión: En este tipo de algoritmos, los valores de salida corresponden a una o más variables continuas.
2. Clasificación: Por otro lado, en los algoritmos de clasificación, las muestras son asignadas a dos o más clases distintas, y el objetivo es aprender de los ejemplos existentes para poder clasificar nuevas muestras en el futuro.

3.2.1.1 Regresión Logística

La Regresión Logística, también conocida como *Logistic Regression* en inglés, es uno de los algoritmos más sencillos y ampliamente utilizados para la clasificación de variables. Su implementación es fácil, lo que lo convierte en una elección común para predecir el resultado de una variable definida. En la Regresión Logística, solo se trabajan con dos categorías, donde un evento se clasifica como 1 si ocurre y como 0 si no ocurre, lo que la hace una técnica binomial. Al aplicar esta técnica, se modela la probabilidad de un resultado

en función de las características individuales de las variables ingresadas (Morán López, 2022).

La estimación se realiza mediante el uso de la función logística, también conocida como función sigmoide (ecuación 1) y la ecuación de regresión logística (ecuación 2).

$$p = \frac{1}{1 + e^{-y}} \quad (1)$$

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n \quad (2)$$

Los coeficientes β_i , también conocidos como coeficientes beta (Lunt, 2015), representan la influencia o importancia de las variables predictoras. Al aplicar la función sigmoide a la regresión logística, se obtiene la ecuación 3.

$$p = \frac{1}{1 + e^{-\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}} \quad (3)$$

Según Lunt (2015), los modelos basados en regresión logística generan predictores para un resultado distribuido normalmente y se fundamentan en la suposición de que los datos son linealmente separables (Lovera & Cardinale, 2023).

El objetivo es calcular la probabilidad de que una de las opciones de la variable dicotómica dependiente ocurra en función de la puntuación en una serie de variables independientes que pueden estar en diferentes escalas de medida (Pauli, 2019).

3.2.1.2 Máquina de Soporte vectorial

En este modelo, los datos a clasificar se representan como un vector n -dimensional de características, ubicando cada entrada del conjunto de datos como un punto en un espacio. El algoritmo busca determinar un hiperplano que mejor separe los puntos etiquetados en el conjunto de entrenamiento (Lovera & Cardinale, 2023).

El objetivo de SVM es encontrar un hiperplano en un espacio n -dimensional, donde n es el número de características (*features*), que permita clasificar claramente los datos, como se muestra en la figura 3.0

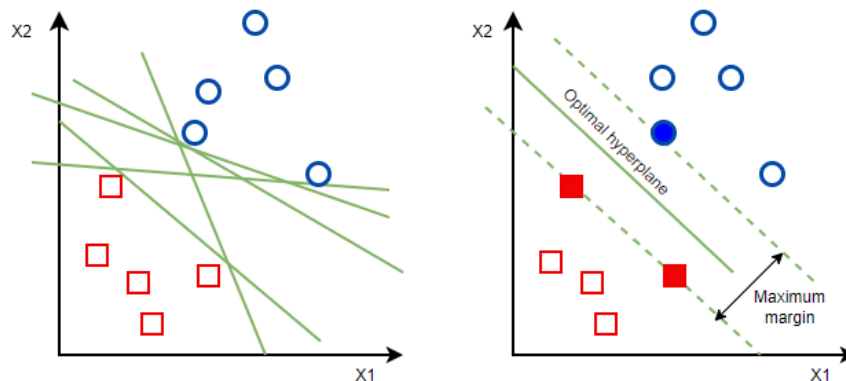


Figura 3.0 Máquina de Soporte Vectorial

El mejor hiperplano será aquel que tenga la mayor distancia a las clases. Las dimensiones del hiperplano están determinadas por el número de características del conjunto de datos (Alarcón Ramos, 2022).

El objetivo es encontrar un hiperplano que pueda clasificar los vectores de palabras maximizando su distancia con los ejemplos más cercanos de cada clase. Las máquinas de soporte vectorial cumplen con esta condición y

denominan al hiperplano que logra esta maximización como el "hiperplano de margen" o "separación óptima" (HSO). Expresado como: $(w \cdot x) + b = 0$ donde $w, x \in \mathbb{R}^n$, $b \in \mathbb{R}$. En este contexto, la tarea consiste en encontrar el vector de características, al que se pueden asociar pesos w . Además, la función de distancia que mide la cercanía del ejemplo más cercano al hiperplano separador viene dada por la ecuación (4).

$$d(x, w, b) = \frac{|(w \cdot x) + b|}{\|w\|} \text{ con } \|w\| = \sqrt{(w \cdot w)} \quad (4)$$

En el caso de los clasificadores no lineales, cuando no existe una superficie de decisión lineal, el vector de entrada X se mapea a uno de mayor dimensión denominado espacio de características $F = \{\phi(x) \mid x \in X\}$ utilizando la ecuación (5).

$$x = \{x_1, x_2, \dots, x_n\} \rightarrow \phi(x) = \{\phi(x)_1, \phi(x)_2, \dots, \phi(x)_n\} \quad (5)$$

Con el fin de plantear una función Kernel K tal que, para todo $x, z \in X$, se expresa como lo indica la ecuación (6).

$$k(x, z) = \langle \phi(x) \cdot \phi(z) \rangle \quad (6)$$

En la ecuación (6), $\phi(x) \cdot \phi(z)$ representa el producto punto dentro del espacio de características de una dimensión arbitraria. Una vez encontrado este espacio de mapeo apropiado, se resuelve el problema de la misma manera que en el caso lineal (Suarez Mosquera, 2021).

En ciertos problemas, nos encontramos con la necesidad de tratar con más de dos variables predictoras o conjuntos de datos que no pueden ser completamente separados, así como clasificaciones en más de dos categorías. Las limitaciones computacionales de las máquinas de aprendizaje lineal hacen que no sean adecuadas para muchas aplicaciones en estas situaciones. Sin embargo, la representación mediante funciones *Kernel* proporciona una solución a este problema al proyectar la información a un espacio de características de mayor dimensión. Esto aumenta la capacidad computacional de las máquinas de aprendizaje lineal ya que mapearemos el espacio de entradas a un nuevo espacio de características de mayor dimensionalidad (Escortell Pérez, 2017).

3.2.1.3 Naive Bayes

Este tipo de clasificador forma parte de un grupo de modelos probabilísticos que se fundamentan en la premisa de que todos los eventos son independientes de una variable determinada (Xu, 2018). Esto implica que el impacto del valor de un predictor (B) en una clase (A) no está relacionado con los valores de otros predictores, lo cual se conoce como "independencia condicional de clase".

El teorema de Bayes establece la siguiente relación, considerando la variable de clase "y" y el vector de características dependientes x_1 hasta x_n como lo indica la ecuación (7).

$$P(y|x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n|y)}{P(x_1, \dots, x_n)} \quad (7)$$

Aplicando la suposición ingenua de independencia condicional entre cada par de características dada la variable de clase, la relación se puede expresar de la siguiente manera (8).

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i|y) \quad (8)$$

Para toda i, esta relación es simplificada como lo indica la ecuación (9).

$$P(y|x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^n P(x_i|y)}{P(x_1, \dots, x_n)} \quad (9)$$

Como $P(x_1, \dots, x_n)$ es constante dada la entrada, podemos usar la siguiente regla de clasificación (10):

$$\begin{aligned} P(y|x_1, \dots, x_n) &\propto P(y) \prod_{i=1}^n P(x_i|y) \\ &\Downarrow \\ \hat{y} &= \arg \max_y P(y) \prod_{i=1}^n P(x_i|y) \end{aligned} \quad (10)$$

y podemos usar la estimación *Maximum A Posteriori* (MAP) para estimar $P(y)$ y $P(x_i|y)$; el primero es entonces la frecuencia relativa de la clase y en el conjunto de entrenamiento (Siurob Palomero, 2019).

Existen distintos tipos de modelos de Naive Bayes según las suposiciones que se hacen sobre la distribución de las características (Alarcón Ramos, 2022):

- Naive Bayes Gaussiano: Este modelo asume que las características siguen una distribución normal (gaussiana). Es apropiado cuando las características numéricas del conjunto de datos se distribuyen normalmente.
- Naive Bayes Multinomial: Se utiliza cuando los datos tienen una distribución multinomial, lo que es común en problemas de clasificación de documentos. En este caso, se utilizan las frecuencias de palabras como predictores. Es ampliamente empleado en tareas de análisis de texto y clasificación de documentos.
- Naive Bayes Bernoulli: Similar al modelo multinomial, pero en este caso, las variables predictoras son variables booleanas independientes (binarias), es decir, toman solo dos valores (verdadero o falso). Este tipo de modelo también se utiliza en clasificación de documentos y para problemas donde se busca determinar si una característica está presente o no.

Cada uno de estos modelos de Naive Bayes tiene sus ventajas y desventajas, y la elección del más adecuado dependerá del tipo de datos y la naturaleza del problema que se esté abordando.

3.2.2 Aprendizaje no supervisado

El aprendizaje no supervisado es un enfoque que permite generar conocimiento utilizando únicamente los datos de entrada, a diferencia del aprendizaje supervisado, donde los datos están etiquetados. En este método, los algoritmos buscan patrones de similitud entre los datos para agruparlos en conjuntos afines (Herrera Contreras, 2020).

En el aprendizaje no supervisado, los datos se organizan de manera que se pueda describir su estructura. Esto puede implicar la agrupación de datos en *clústeres* o la búsqueda de diferentes formas de examinar datos complejos para simplificarlos u organizarlos. A diferencia del aprendizaje supervisado, en el aprendizaje no supervisado no se cuenta con conocimiento previo ni se tienen tuplas (X, Y) etiquetadas. En su lugar, solo se dispone de la variable X para trabajar (Cedeno Moreno & Vargas Lombardo, 2020).

Por ende, el aprendizaje no supervisado se enfoca en describir la estructura de los datos y buscar patrones que permitan simplificar el análisis. Su enfoque es exploratorio, ya que no se busca una respuesta específica ni se tiene conocimiento previo, sino más bien descubrir posibles relaciones o agrupaciones dentro de los datos (González Guerra, 2020).

3.2.2.1 Vecinos más cercanos

El método de clasificación basado en vecinos (*KNN*) es un enfoque de aprendizaje no generalizador o basado en instancias, lo que significa que no construye un modelo explícito, sino que se basa en recordar sus datos de entrenamiento. Para clasificar una nueva instancia *KNN* utiliza votación, donde los vecinos más cercanos según la distancia seleccionada votan, y la clase se asigna a la que obtiene la mayoría de los votos.

El parámetro K en *KNN* representa el número de vecinos cercanos que se tendrán en cuenta para la votación. Para evitar empates, K debe ser un número impar. La selección de K se realiza probando diferentes valores en el algoritmo *KNN* y eligiendo aquel que minimice los errores sin comprometer la precisión. Es importante tener en cuenta que, si K es igual a 1, la clasificación dependerá únicamente de un vecino cercano, lo que puede llevar a asignarle

una etiqueta basada en la mayoría de los puntos cercanos. A medida que K aumenta, la votación se vuelve más estable, pero si K es demasiado grande, es posible que se generen más errores en la clasificación. Por lo tanto, es fundamental encontrar un valor de K que proporcione un equilibrio adecuado entre estabilidad y precisión (Alarcón Ramos, 2022).

3.2.2.2 Redes neuronales

La red neuronal es un tipo de estructura de datos que, a través de iteraciones, aprende qué variables son más importantes para la clasificación (Goldberg, 2017), (Pedregosa et al., 2011). Se puede visualizar la red neuronal como una tupla (X, W, Σ, f, Y) , donde:

1. X representa el conjunto de entradas.
2. W es el conjunto de pesos. $w_{i,j,k}$ donde i representa el número de aristas entrantes de la neurona j en la capa k
3. Σ es la función de agregación.
4. f simboliza la función de activación.
5. Y es la salida.

En este caso, las entradas representan el estímulo al que se somete la red neuronal, mientras que las salidas corresponden a la respuesta que la red proporciona ante dicho estímulo. La red neuronal puede visualizarse como un grafo en el que cada nodo de una capa está conectado con todos los nodos de la capa siguiente y la anterior. El conjunto de entradas $(X_1 \dots X_n)$ se refiere a los parámetros que podemos extraer de la muestra o conjunto de datos que se utiliza para entrenar o evaluar la red neuronal.

Asimismo, el conjunto de pesos de la neurona $(w_{1,1} \dots w_{1,n} \dots w_{m,n})$ determinará la importancia de la salida de una neurona de la capa anterior como entrada

para la neurona de la capa siguiente. Cada arista del grafo tendrá un peso "w" asociado. Cada neurona generará una salida a través de la cual transmitirá información a las neuronas de la capa siguiente, o directamente a la salida de la red neuronal, en caso de ser la última capa de neuronas. De esta manera, los pesos juegan un papel crucial en la transmisión y procesamiento de la información en la red neuronal (Carrión García et al., 2021).

La red neuronal que se muestra en la figura 3.1 ajustará los sesgos y los pesos de cada una de sus neuronas, así como las conexiones entre ellas, mediante la aplicación de una función de coste que se reducirá mediante el algoritmo del descenso del gradiente estocástico. Es importante destacar que este clasificador es un modelo de caja negra, lo que significa que resulta difícil explicar el motivo detrás de las decisiones tomadas por la red neuronal. En otras palabras, a pesar de su eficacia, la interpretación y explicación de los resultados obtenidos puede ser compleja debido a la complejidad interna de la red y su proceso de aprendizaje.

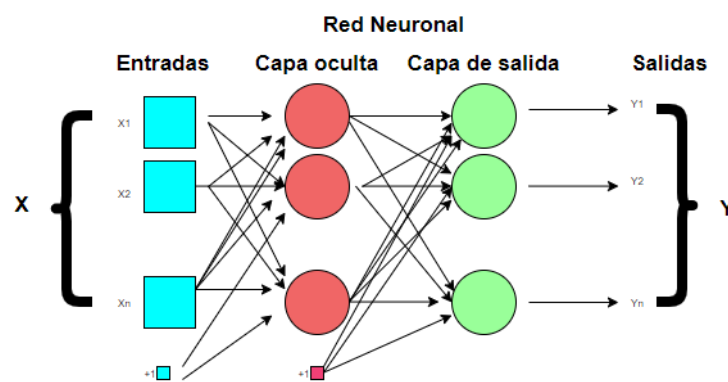


Figura 3.1 Red Neuronal.

Las redes neuronales son consideradas también como un sistema capaz de emular algunas de las actividades que realizan los seres humanos, como

memorizar y asociar hechos. Su unidad básica es la neurona artificial, que consta de tres elementos esenciales: función de entrada, función de activación y función de salida. Estas funciones permiten que las neuronas procesen la información de entrada, generen una activación y produzcan una salida que será transmitida a otras neuronas o al resultado final de la red (Suarez Mosquera, 2021).

3.3 Procesamiento de lenguaje natural

El Procesamiento de Lenguaje Natural (PLN), también conocido como NLP por sus siglas en inglés, es una de las tareas fundamentales de la Inteligencia Artificial que implica la manipulación de lenguajes naturales mediante herramientas computacionales. El PLN no se limita simplemente a transcribir la información comunicada, sino que se refiere a la habilidad de las máquinas para procesar y comprender el mensaje comunicado en lenguaje natural. En otras palabras, el PLN consiste en utilizar el lenguaje natural para comunicarnos con el computador, y este último debe ser capaz de entender el mensaje que se le proporciona (Davia García, 2021).

El Procesamiento de Lenguaje Natural (PLN) es un área de investigación y producción científica muy activa que tiene numerosas aplicaciones, como la búsqueda de respuestas automáticas, corrección ortográfica, reconocimiento de voz, generación de resúmenes, traducción automática y análisis de sentimientos, entre otras. Estas aplicaciones se basan en el procesamiento de textos publicados en Internet y, especialmente, en textos de redes sociales.

Las tareas básicas del PLN implican dividir los textos en piezas elementales, como palabras u oraciones, y comprender las relaciones entre estas piezas, así como su significado cuando se combinan de diversas formas. Estas tareas

permiten "normalizar" los textos, es decir, obtener una representación estructurada de ellos para facilitar su procesamiento y análisis (Cardoso et al., 2019).

Tuvo sus inicios en la década de 1950, cuando surgió como una nueva disciplina en el campo de las ciencias de la computación. Su propósito era crear sistemas con la suficiente inteligencia para que la interacción entre humanos y máquinas pudiera realizarse a través del lenguaje natural. Desde entonces, el PLN ha experimentado un constante desarrollo y avance, convirtiéndose en un área de investigación clave dentro de la Inteligencia Artificial. Su objetivo principal sigue siendo facilitar la comunicación y comprensión entre humanos y computadoras mediante el uso del lenguaje cotidiano (Pauli, 2019).

Esta área de conocimiento tiene diversas aplicaciones, mencionadas por Hernández y Gómez (Hernández & Gómez, 2013), que incluyen:

1. Recuperación y extracción de información: Se enfoca en obtener información relevante de textos no estructurados, aplicando procesos de transformación matemáticos o basados en propiedades para obtener una fuente de datos estructurada con las secciones de interés de los textos.
2. Minería de datos: Su objetivo es descubrir relaciones o patrones ocultos dentro de los textos, para ello se requiere que los datos estén estructurados, lo cual se logra mediante un proceso de normalización.
3. Traducción automática: Busca convertir textos de un idioma a otro manteniendo su significado. Para lograrlo, se emplea una representación intermedia de los textos y luego se ajusta de acuerdo con las reglas gramaticales del idioma destino.

4. Sistemas de búsqueda de respuestas: Utilizados principalmente en motores de búsqueda, estos sistemas buscan proporcionar respuestas precisas a las búsquedas realizadas por usuarios. Se emplean técnicas de extracción y recuperación de información, y el reto principal radica en la generación de respuestas, por lo que es necesario calificar los resultados obtenidos.
5. Generación de resúmenes automáticos: Puede realizarse a nivel de documento o documentos, y puede ser de tipo abstractivo o extractivo. En el abstractivo, se basa en técnicas de parafraseo para producir síntesis y aún se encuentra en desarrollo.

3.3.1 Características textuales

El sistema de Procesamiento de Lenguaje Natural (PLN) tiene como objetivo facilitar la comunicación entre el usuario y el propio sistema mediante una serie de tareas. Estas tareas conducen a una arquitectura de niveles, donde las oraciones que se analizarán serán comprendidas por el sistema PLN. En total, se contemplan 4 niveles de análisis, sin embargo, no todos los niveles serán implementados en cada caso (Cortez Vásquez et al. 2009), (Sobrinho Sande 2018), ya que su desarrollo dependerá de las funciones específicas que deba desempeñar un sistema particular.

Las características textuales del Procesamiento de Lenguaje Natural (PLN), organizados de menor a mayor complejidad, son los siguientes (Pauli, 2019):

1. Nivel de análisis morfológico.
2. Nivel de análisis sintáctico.
3. Nivel de análisis semántico.
4. Nivel de análisis pragmático.

3.3.1.1 Analizador léxico

Un analizador léxico, también conocido como *lexer* o *scanner*, es una parte esencial del procesamiento del lenguaje natural (NLP). Forma parte de la fase de análisis sintáctico y se encarga de descomponer un flujo de texto (una secuencia de caracteres) en unidades más pequeñas llamadas "tokens".

La lexicología es una disciplina que se dedica a estudiar el significado de las unidades léxicas de una lengua, así como las relaciones sistemáticas y las conexiones entre ellas. Los sistemas que operan a nivel léxico se encargan de interpretar el significado de las palabras de manera individual. El analizador léxico se encarga de combinar las unidades léxicas más elementales para formar estructuras más complejas, que también tienen significado (Mancera, 2020).

3.3.1.2 Analizador sintáctico

Un analizador sintáctico, también conocido como *parser*, es una parte fundamental del procesamiento del lenguaje natural (NLP). Su objetivo es analizar la estructura gramatical de un texto y determinar cómo las palabras y frases se combinan para formar oraciones y frases bien formadas dentro de un lenguaje.

La sintaxis se ocupa del estudio de las reglas y el orden en que las palabras se organizan dentro de una oración, con el propósito de determinar la función que cada palabra cumple en dicha estructura (Mancera, 2020).

3.3.1.3 Analizador semántico

El analizador semántico, también conocido como análisis semántico, es una etapa crucial en el procesamiento del lenguaje natural (PLN) que tiene como objetivo comprender el significado de un texto más allá de su estructura gramatical. Mientras que el analizador léxico y el analizador sintáctico se centran en la identificación de palabras y su estructura gramatical, respectivamente, el analizador semántico va más allá y busca capturar el significado y la interpretación de las oraciones y textos.

Los sistemas que procesan a un nivel semántico determinan los significados posibles de las palabras. Este tipo de procesamiento requiere de técnicas de separación de ambigüedad semántica para palabras que tengan varios significados (Mancera, 2020).

3.3.2 Técnicas del procesamiento del lenguaje natural

1. Se mostrarán a continuación de forma breve los distintos conceptos que se utilizan en el procesamiento del lenguaje natural (Alarcón Ramos, 2022), (Calvo Madurga, 2020), (Cardoso et al., 2020), (Valladares Cedillo, 2022), (Carrión García et al., 2021):
2. *Token*: Se refiere a una instancia de una secuencia de caracteres que se agrupa como una unidad semántica en un documento, y se procesa como tal.
3. *Tokenización*: Es el proceso de asignar un *token* a una secuencia de caracteres que se desea tratar como una entidad única. Al tratar el texto de esta manera, se pueden contar los *tokens*, lo que a menudo se utiliza como características en análisis posteriores.

4. Normalización: Implica la eliminación de signos de puntuación como comas, puntos, tildes, entre otros, para estandarizar el texto y facilitar su procesamiento.
5. Bolsas de palabras: Es una forma de representar el vocabulario de un corpus mediante la creación de una matriz en la que cada columna corresponde a un *token* y las filas a las frases del corpus. En la matriz se registra cuántas veces aparece cada *token* en cada oración.
6. Preprocesado de texto: Es un paso fundamental en el procesamiento de lenguaje natural que implica dar forma al corpus según las necesidades específicas, como decidir si se eliminan o conservan signos de puntuación, si se otorga importancia a las mayúsculas y si se filtran palabras vacías como artículos, pronombres y preposiciones.
7. Filtrado de palabras vacías (*stopwords*): Consiste en la eliminación de palabras que carecen de significado relevante, como artículos, pronombres y preposiciones.
8. Lematización: Es un proceso que cuenta las palabras que comparten una raíz lingüística en un diccionario de datos, devolviéndolas a su forma base. Por ejemplo, convierte diferentes conjugaciones de un verbo en su forma infinitiva.
9. *Stemming*: Implica eliminar prefijos o sufijos de las palabras, dejando solo el "núcleo" o *stem*. Por ejemplo, en las palabras "niño" y "niñera", el *stem* sería "niñ". Este proceso no siempre produce cortes precisos en las palabras y, por lo tanto, tiene limitaciones.

3.4 Minería de opiniones

La minería de opiniones es una rama que se deriva de la minería de textos, la cual, a su vez, forma parte de la minería de datos. Su enfoque principal radica en identificar el sentimiento, emoción o polaridad transmitida por un conjunto de palabras. Esta técnica se emplea principalmente para analizar las opiniones, emociones y actitudes expresadas por individuos capaces de comunicar sus conocimientos (Robles Domínguez, 2019).

Esta tecnología opera a través del procesamiento del lenguaje natural, la lingüística computacional y la minería de textos. Utilizando algoritmos, es posible evaluar el estado de ánimo de un comentario o conjunto de comentarios, lo que permite identificar tendencias y comprender las reacciones del público frente a determinados contenidos (Buesa, 2020).

Hablar de minería de opiniones implica adentrarse en un campo en constante expansión, relacionado con la evaluación de los elementos subjetivos presentes en los contenidos creados por los usuarios. Dentro de este ámbito, se encuentran aplicaciones que efectúan un análisis superficial de los textos, adaptándose a la tarea o el problema en cuestión. En términos generales, existen dos categorías principales de tareas asociadas con la minería de opiniones (Robles Domínguez, 2019):

1. Detección de polaridad: Esto se refiere a la habilidad de determinar si una opinión es positiva o negativa. Más allá de una evaluación básica de la polaridad, a menudo se busca obtener un valor numérico dentro de un rango específico, lo que se podría considerar como un "rating" objetivo relacionado con la opinión en cuestión.

2. Análisis del sentimiento basado en características: Este proceso implica la capacidad de identificar las diversas características de un producto mencionadas en la opinión de un usuario y, para cada una de estas características, determinar su polaridad. Estos enfoques son mucho más complejos y detallados que la simple detección de polaridad.

3.4.1 Clasificación de la minería de opiniones

La técnica de Minería de Opiniones abarca una variedad de actividades orientadas al procesamiento de texto sin estructura. Su objetivo principal es asignar una valoración numérica a expresiones subjetivas relacionadas con opiniones y sentimientos, al mismo tiempo que busca determinar el nivel de polaridad, ya sea positiva, negativa o neutral, asociado a diversas "entidades". Las tareas definidas incluyen las siguientes (Romero, 2016):

La Clasificación de la Subjetividad: Es una tarea cuyo propósito es identificar secciones de texto que contienen un significado o carga subjetiva, expresados por el autor del texto, que pueden incluir opiniones, expresiones de sentimientos, y más. Esta tarea permite diferenciar entre fragmentos de texto que expresan una opinión sobre un tema en particular y aquellos que simplemente relatan un hecho sin ofrecer una perspectiva. A menudo, se considera que la clasificación de la subjetividad es un paso inicial antes de realizar otras actividades de Minería de Opiniones, como la determinación de la polaridad.

La Clasificación de la Intensidad: Es una tarea que tiene como objetivo categorizar los textos en función de la intensidad emocional que expresan. En la mayoría de los enfoques que abordan este problema, se realiza una clasificación en tres categorías de intensidad: positiva, neutra y negativa.

3.5 Clasificación de textos

La clasificación de textos se refiere al procedimiento de asignar categorías a un documento. Este proceso brinda una mayor organización, lo que facilita el entendimiento e interpretación de los datos. Para llevar a cabo la clasificación de textos, se han desarrollado diversos algoritmos que permiten, de forma automática, predecir la clase de nuevos documentos no vistos previamente, utilizando un conjunto de ejemplos creados por especialistas (Coello Guilarte, 2019). La figura 3.2 muestra la metodología de clasificación de textos propuesta por Mironczuk, (Mironczuk & Protasiewicz, 2018).

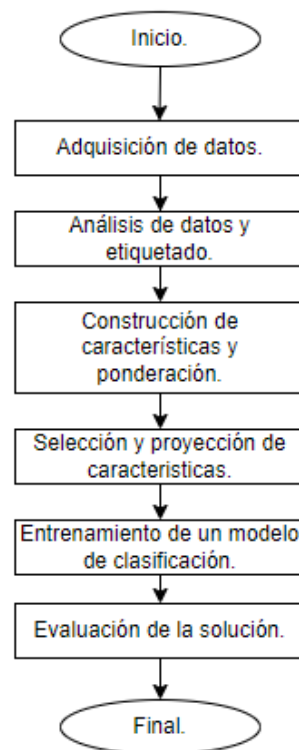


Figura 1.2 Etapas de la clasificación de textos.

3.5.1. Adquisición de datos

El proceso inicial en la realización de la clasificación de textos es la fase de adquisición. En esta etapa, se lleva a cabo la obtención del conjunto de documentos digitales que serán objeto de análisis.

La adquisición de datos, que engloba el proceso de capturar y registrar información proveniente de diversas fuentes para su posterior análisis en aplicaciones científicas, tecnológicas o de investigación (Mironczuk & Protasiewicz, 2018).

Corpus

De acuerdo con la Real Academia Española (RAE), un corpus se refiere a un conjunto exhaustivo y ordenado de datos o textos de carácter científico, literario, u otros, que pueden servir como base en investigaciones (Corpus, s. f.).

Otra definición sostiene que el corpus de texto, o simplemente corpus, constituye la base de todas las actividades en el Procesamiento del Lenguaje Natural (PLN). Se trata de un extenso conjunto de datos en formato de texto, en diversos idiomas como inglés, francés, entre otros. Este corpus puede estar compuesto por un solo documento o varios, y su principal objetivo es ser ampliamente utilizado para análisis lingüísticos y estadísticos, así como para la creación de herramientas de PLN. Las fuentes para un corpus pueden incluir redes sociales como Twitter, sitios de blogs, foros de discusión abierta, libros y más. Según la tarea en cuestión, el corpus podría incluso incluir archivos de formatos distintos, como sonido, imágenes o videos (Siurob Palomero, 2019).

Corpus Muchocine

Muchocine es un sitio web en español que se dedica a albergar reseñas y opiniones sobre películas. Las críticas en el sitio web se organizan en tres categorías principales: por género cinematográfico, por el título de la película y por décadas en la historia del cine (Flores Domínguez & Tapia Fabela, 2020).

El corpus "Muchocine en español" creado por Cruz (Cruz et al., 2008) se obtuvo extrayendo las páginas HTML de las críticas de películas en el sitio web Muchocine en febrero de 2008. Este corpus contiene un total de 3878 críticas, cada una de las cuales ha sido calificada con valores que van de 1 a 5. En esta escala, el valor 1 indica que una película es muy mala, el 2 indica que es mala, el 3 es neutral, el 4 representa una película buena y el 5 indica que es muy buena.

Es importante destacar que las críticas en este corpus son escritas por usuarios comunes, lo que agrega complejidad a la tarea de análisis. Los textos pueden contener errores ortográficos, incoherencias y problemas de puntuación, además de variar en longitud.

3.5.2 Análisis y etiquetado de datos

En la etapa de análisis y etiquetado de datos, se concentra principalmente en la creación de conjuntos de datos con etiquetas. Hay dos enfoques para llevar a cabo el etiquetado de datos: uno implica asignar etiquetas a grupos de textos y el otro consiste en asignar etiquetas o múltiples etiquetas a segmentos específicos de un texto (Mironczuk & Protasiewicz, 2018).

3.5.3 Construcción y ponderación de características

Se describe como el procedimiento de generar variables o atributos adicionales utilizando los datos ya disponibles. Estos nuevos atributos se diseñan para capturar información importante y distintiva que no se encuentra en los atributos originales (Mironczuk & Protasiewicz, 2018).

Los datos de tipo textual poseen dos formas de representación:

1. Una representación o modelo de espacio vectorial, donde un documento se convierte en un vector de ponderaciones de atributos, y las palabras (términos, frases) dentro del documento se transforman en atributos.
2. Representación visual en la que un documento se representa en forma de un gráfico, en el cual los nodos simbolizan palabras y las conexiones entre nodos representan las relaciones entre las palabras.

Después de que las características sean seleccionadas, se les puede asignar valores numéricos. Hay esquemas de ponderación de características ampliamente reconocidos y utilizados en los ámbitos de clasificación y procesamiento de textos. Además, también existen enfoques de ponderación de características más recientes que pueden ser menos conocidos (Mironczuk & Protasiewicz, 2018).

3.5.3.1 Ponderación Booleana

La forma más básica de asignar peso es la binaria o booleana; en esta, el peso asignado a una palabra en un documento es 1 si la palabra aparece en el documento y 0 si no lo hace. Esta técnica proporciona información sobre la relación entre un documento y una categoría según los atributos que contiene

(Coello Guilarte, 2019). La representación de la ponderación booleana de un término en los documentos se logra mediante vectores cuyo tamaño es igual al número de términos únicos generados a partir de la colección de documentos (Vázquez, 2008).

3.5.3.2 Ponderación IDF

Factor *IDF* de cada elemento del árbol en cada documento: Es el coeficiente que determina la capacidad discriminativa del término de un documento con respecto a la colección.

$$IDF(n) = \log_{10} \frac{N}{(DF(n) + 1)} \quad (11)$$

1. Donde N es el número total de documentos.
2. DF es el número de documentos donde aparece el término n .
3. El logaritmo se utiliza para obtener un coeficiente bajo de fácil manejo, y el +1 funciona como factor correctivo del resultado (Vuotto et. al, 2015).

3.5.3.3 Ponderación TF

Este método de ponderación, conocido como "*TF*", se fundamenta en la frecuencia con la cual una palabra aparece en un documento específico. Investigaciones previas han evidenciado que la aplicación de esta ponderación resulta provechosa para identificar términos estrechamente vinculados al ámbito temático en consideración (Nadeem et al., 2016). En la ecuación 11 se resume:

$$tf_{ij} = f_{ij} \quad (11)$$

En el contexto en el que la frecuencia f_{ij} representa la cantidad de veces que el término w_i aparece en el documento d_j , si esta frecuencia se ajusta en función de la longitud del documento, la ecuación 12 experimenta una transformación a:

$$tf_{ij} = \frac{f_{ij}}{|d_j|} \quad (12)$$

donde $|d_j|$ se refiere a la extensión del documento en relación con la cantidad de términos presentes en él.

3.5.3.4 Ponderación TF-IDF

El método *TF-IDF* es ampliamente empleado para asignar pesos a los atributos, y evalúa la importancia de cada atributo c_j en relación con un usuario específico mediante la siguiente fórmula (13):

$$W(u, c_j) = FF(u, c_j) * IUF(c_j) \quad (13)$$

La importancia del atributo c_j para el usuario u se calcula combinando dos componentes: (i) Una medida de la frecuencia del atributo c_j para el usuario u , denominada Frecuencia de Atributo (FF, por sus siglas en inglés, *Feature Frequency* en inglés), y (ii) una medida de la diferencia entre usuarios (Barranco et al, 2012).

3.5.4 Selección y proyección de características

Las técnicas de selección de atributos buscan retener únicamente los atributos o características más significativos o descriptivos, mientras que descartan los demás. En este sentido, existen tres enfoques estándar para la selección de atributos: filtro, envolvente e integrado.

1. Los métodos de filtro se basan en una función de evaluación para clasificar las características según su importancia.
2. Los métodos de envolvente son algoritmos generales que exploran subconjuntos de características en busca de los más apropiados.

Las técnicas de proyección de características, por otro lado, reorganizan las características existentes en dimensiones diferentes. La meta aquí es generar nuevos ejes de datos para que la estructura y la varianza del conjunto de datos nuevo reflejen de manera cercana la estructura del conjunto de datos original (Mironczuk & Protasiewicz, 2018).

3.5.5 Entrenamiento del modelo

Se emplean diversos enfoques de aprendizaje con el fin de instruir a una función de clasificación para identificar un concepto objetivo. Este proceso se lleva a cabo solamente cuando se cuenta con un conjunto de entrenamiento previamente elaborado. Los métodos de entrenamiento pueden categorizarse en enfoques de aprendizaje supervisado, semisupervisado, por conjunto, activo, de transferencia, entre otros. (Mironczuk & Protasiewicz, 2018).

3.5.6 Evaluación de solución

Una vez que se ha desarrollado un modelo de clasificación, es posible seleccionar y definir métricas para evaluar su rendimiento. La Figura 3.3 ilustra tres etapas para llevar a cabo la evaluación de modelos.

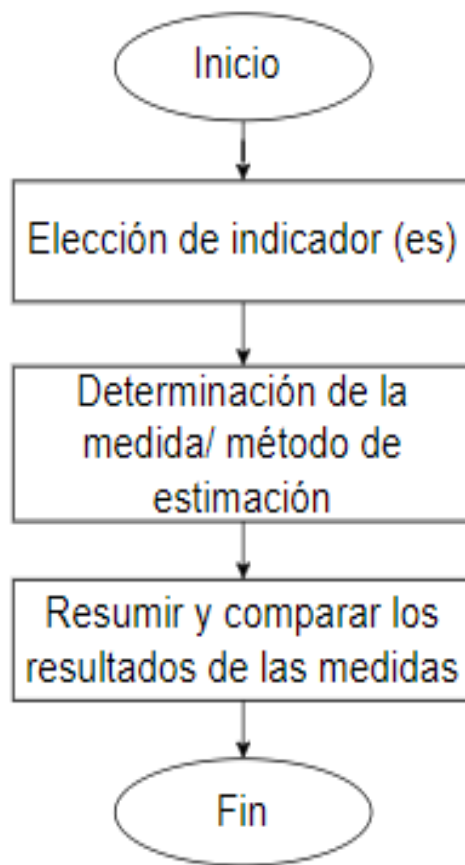


Figura 3.3 (Mironczuk & Protasiewicz, 2018). Pasos básicos de la evaluación de modelos.

En primer lugar, se evalúa un solo indicador o conjunto de indicadores (Mironczuk & Protasiewicz, 2018). Existen diversos indicadores estadísticos avanzados, que abarcan la precisión, la recuperación, la exactitud, la puntuación F, la especificidad, el área bajo la curva (AUC), la tasa de error, entre otros (Sokolova & Lapalme, 2009). El cálculo de estos indicadores está

vinculado al tipo de problema de clasificación, ya sea binario, multiclase o multietiquetado (Mironczuk & Protasiewicz, 2018).

En segundo lugar, definimos un procedimiento para calcular o estimar los valores del indicador. Luego especificamos cómo dividir los conjuntos de datos disponibles. En tercer lugar, se compilan y contrastan todos los valores obtenidos (Mironczuk & Protasiewicz, 2018).

Las métricas de evaluación son instrumentos empleados para valorar la efectividad y la exactitud de un clasificador (Powers, 2011).

Precisión: Nos permite cuantificar la proporción de casos identificados correctamente dentro del conjunto total de casos clasificados como pertenecientes a una clase. También se denomina valor predictivo positivo, y se calcula como la fracción de casos relevantes respecto a los casos recuperados (Castro et al., 2022).

$$Precisión = \frac{VP}{VP + FP}$$

Exactitud: Es un valor numérico en el rango de 0 a 1 que indica la proporción de predicciones correctas realizadas por el modelo. Por lo general, se expresa como un porcentaje (Alarcón Ramos, 2022), (Hastie, et al., 2009).

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

Donde:

1. VP son verdaderos positivos
2. VN son verdaderos negativos
3. FP son falsos positivos
4. FN son falsos negativos

Recuerdo: Se define como la proporción entre los documentos que el clasificador asignó a una clase específica y el número total de documentos que realmente pertenecen a esa clase (Coello Guilarte, 2019), (Davis, 2006).

$$\text{Recuerdo} = \frac{VP}{VP + FN}$$

F-measure: La puntuación *F-measure* es una métrica que integra tanto la precisión como la exhaustividad para ofrecer una evaluación integral del rendimiento de un clasificador (Powers, 2011). Se calcula utilizando la fórmula:

$$F - \text{measure} = 2 * \frac{\text{Precision} * \text{Recuerdo}}{\text{Precision} + \text{Recall}}$$



CAPÍTULO

Metodología

4

4.1 Clasificación de textos

El presente capítulo proporciona una panorámica general acerca del desarrollo metodológico del proceso, la recopilación de los datos, así como el análisis y la obtención de resultados. Se enfatiza la importancia de seleccionar el conjunto de datos adecuado para esta investigación, con el objetivo de garantizar su autenticación. Asimismo, se detallará la metodología empleada para el análisis de los datos.

Para la construcción del método propuesto se empleó la metodología de clasificación de textos, la cual fue propuesta por Mironczuk, (Mironczuk & Protasiewicz, 2018), la cual consta de las siguientes etapas: 1.- Adquisición de datos; 2.- Análisis de datos y etiquetado; 3.- Construcción de características y ponderación; 4.- Selección y proyección de características; 5.- Entrenamiento de un modelo de clasificación; y 6.- Evaluación de la solución, no obstante, en la presente investigación se excluyó la etapa de selección y proyección de características, lo cual se puede observar en la figura 4.0.

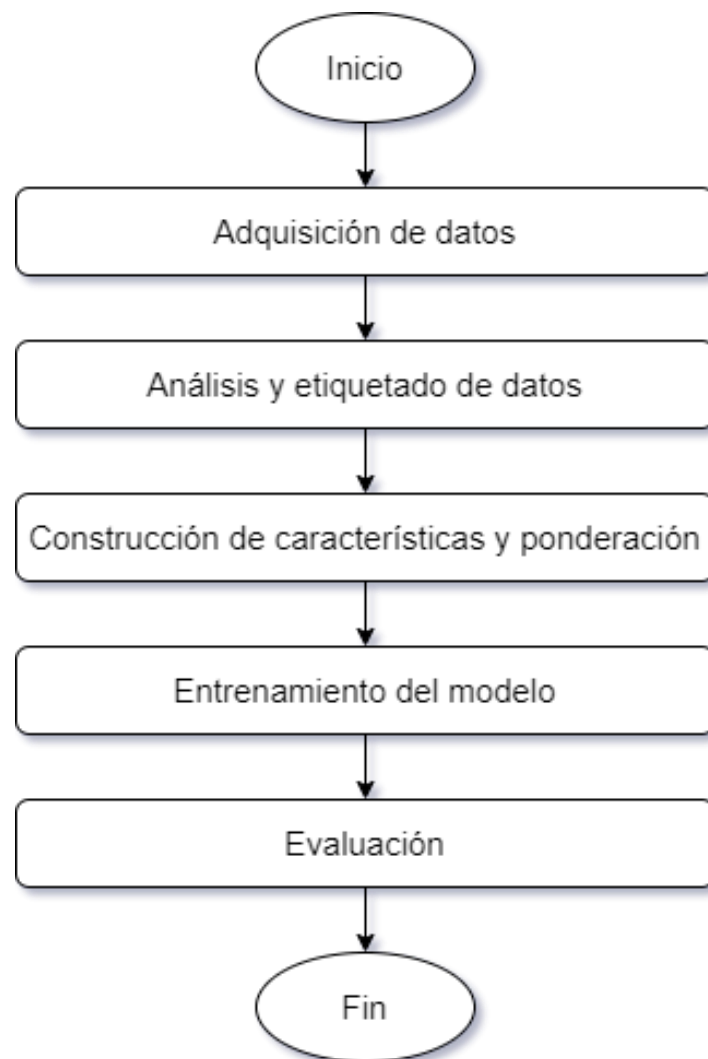


Figura 4.0 Método propuesto.

4.1.1 Adquisición de los datos.

El corpus utilizado para la presente investigación se obtuvo de la página web de TIMM (<http://timm.ujaen.es/recursos/mce-corpus/>): "Muchocine". Este corpus ha sido utilizado para la tarea de minería de opiniones. El corpus "Muchocine" creado por Cruz, (Cruz et al., 2008) se obtuvo extrayendo las páginas HTML de las críticas de películas en el sitio web Muchocine en febrero de 2008. Este corpus contiene un total de 3878 críticas, cada una de las cuales ha sido calificada con valores que van de 1 a 5. En esta escala, el valor 1 indica que una película es muy mala, el 2 indica que es mala, el 3 es neutral, el 4 representa una crítica buena y el 5 indica que es muy buena. En la tabla 4.0 se muestra de manera grafica la descripción de cada uno de los elementos del corpus utilizado para la presente investigación.

Polaridad	Escala	Número de documentos por clase
1	Muy mala	351
2	Mala	923
3	Neutra	1253
4	Buena	890
5	Muy buena	461
Total, de críticas		3878

Tabla 4.0 Estructura del corpus "Muchocine".

4.1.2 Análisis y etiquetado de datos.

El modelo de representación a utilizar es unigramas de palabras. En dicha representación se realizó un preprocesamiento, donde se eliminaron, caracteres especiales como; signos de puntuación, signos de admiración e interrogación, así como la eliminación de *stopwords* y números esto con la finalidad de dejar el texto limpio.

4.1.3 Construcción y ponderación de características.

En esta fase, se emplea el enfoque propuesto, el cual involucra la elección de un documento por crítica. Una vez que los documentos por crítica se extraen, las características son utilizadas en el proceso de ponderación booleana.

La ponderación booleana consiste en asignar un valor a una característica dentro de un documento, reflejando su relevancia. Se le asigna un valor de uno cuando la característica está presente y un valor de cero cuando no lo está. A través de esta ponderación, se logra determinar la importancia de cada una de las características.

La representación vectorial de la ponderación booleana implica asignar una columna a cada documento de cada crítica y una fila a cada característica. Posteriormente, se llena la tabla con los valores de ponderación, asignando un 1 cuando la característica está presente en el documento de la crítica y un 0 en caso contrario. Este procedimiento se repite hasta que se han tenido en cuenta todos los documentos de cada crítica, la tabla 2.2 exhibe esta representación de forma visual.

Documento/crítica	Característica 1	Característica 2	Característica n-1	Característica n
Documento 1 crítica 1	0	0	1	1
Documento n-1 crítica 1	1	1	0	1
Documento n crítica 1	0	1	1	0
Documento 1 crítica n-1	0	0	1	1
Documento n-1 crítica n-1	1	1	0	0
Documento n crítica n	1	0	0	1

Tabla 4.1 Representación vectorial de la ponderación booleana.

4.1.4 Entrenamiento del modelo

En esta investigación, se emplearon los modelos de aprendizaje supervisado, *Máquina de Soporte Vectorial*, *Naive Bayes* y *Regresión Logística*.

Para los modelos SVM y NV, se utilizaron los parámetros con un valor de C igual a uno, utilizando una clasificación de uno contra todos. En el caso del SVM, se empleó un Kernel lineal.

Los modelos se entrenaron utilizando las características de cada crítica. Una vez finalizada la etapa de entrenamiento, se procedió a generar el modelo para pasar a la etapa de evaluación.

4.1.5 Evaluación de la solución

La métrica utilizada para evaluar el desempeño fue la exactitud. La exactitud se define como la proporción de predicciones correctas realizadas por el modelo en comparación con el total de predicciones realizadas. Se establece en base a los valores de verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN). El cálculo de la exactitud se obtiene al dividir el número de predicciones correctas entre el total de predicciones realizadas, como se muestra en la siguiente ecuación:

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$



CAPÍTULO

5

Experimentación y resultados

En este capítulo se expondrán los resultados derivados de la experimentación con los clasificadores: *Máquina de Soporte Vectorial*, *Naive Bayes* y *Regresión Logística*. Estos clasificadores se aplicaron al corpus “Muchocine” creado por Cruz, (Cruz et al., 2008) el cual se utilizó para este estudio, con el propósito de evaluar su eficacia en la clasificación de críticas. Además, se detallarán los resultados obtenidos con cada uno de los clasificadores y se llevara a cabo una comparación entre ellos.

Para el caso de esta investigación se dispone del corpus "Muchocine" (Cruz et al., 2008) el cual está constituido por críticas de cine clasificadas en 5 clases con textos extraídos de la página web <https://muchocine.net/>, dicho corpus se encuentra desbalanceado pues no se cuenta con la misma cantidad de archivos por clase, como se muestra en la tabla 5.0

Polaridad	Escala	Número de documentos por clase
1	Muy mala	351
2	Mala	923
3	Neutra	1253
4	Buena	890
5	Muy buena	461
Total, de críticas		3878

Tabla 5.0 Estructura del corpus "Muchocine".

Para la experimentación se tomaron las 3878 críticas que conforman el corpus para aplicar el preprocesamiento, el cual consiste en la limpieza y la preparación de las críticas para su posterior clasificación. Se eliminaron los caracteres especiales como; signos de puntuación, admiración e interrogación, así como la eliminación de las *stopwords* y los números. Esto con el fin de poder reducir la dimensionalidad del problema, descartando así información poco relevante para el análisis de opiniones.

Concluido el preprocesamiento del corpus "Muchocine" se realizaron 3 muestras, cada una diferenciada por la cantidad de clases utilizadas para cada experimento, para cada muestra se utilizó un conjunto de entrenamiento (*Train*), el cual consta del 80% de los datos y un conjunto de validación (*Test*) conformado por el 20% restante de los datos.

Para evaluar la capacidad de clasificación del modelo se realizó la validación cruzada a 10 pliegues, técnica particularmente útil cuando se tiene un conjunto de datos limitado y se desea maximizar la utilización de los datos disponibles para entrenar y evaluar el modelo, ya que proporciona una

estimación más confiable del rendimiento del modelo de comparación, por lo que se tuvo que entrenar y evaluar el modelo 10 veces. Así mismo para cada muestra los datos se representaron en n-gramas de palabras de tamaño 1.

Para el experimento 1 se tomaron únicamente 2 clases; **positivo** y **negativo** como se muestra en la tabla 5.1, las cuales están conformadas por las clases muy mala, mala, buena y muy buena del corpus original, dejando de lado la clase neutra, teniendo así un total de 2625 críticas.

Clases del corpus original.	Número de documentos por clase
Muy mala	351
Mala	923
Neutra	1253
Buena	890
Muy buena	461
Total, de criticas	3878

Clases tomadas para la muestra 1.	Número de documentos por clase	Clase nueva	Número de documentos por clase
Muy mala	351	Negativa	1274
Mala	923		
Buena	890	Positivo	1351
Muy buena	461		
Total, de criticas			2625

Tabla 5.1 Clases para el experimento 1.

En dicha muestra se usó el modelo de representación de bolsa de palabras en conjunto con el peso booleano, en la tabla 5.2 se muestran los resultados obtenidos con respecto a los tres clasificadores, teniendo como mejor resultado una exactitud de 83.62591%, mientras que en la tabla 5.3 se muestra cada resultado obtenido de la validación cruzada a 10 pliegues.

No. De clases	Clases			
		SVM	NB	LR
2	Positivo	82.29003	81.9847	83.62591
	Negativo			

Tabla 5.2 Resultados del modelo con 2 clases.

No.	Número de clases.	Clases			
			SVM	NB	LR
1	2	Positivo Negativo	81.2977	83.3969	83.5877
2			83.5877	83.7786	85.3053
3			81.4885	79.3893	81.1068
4			82.8244	80.916	85.1145
5			79.9618	82.8244	80.916
6			79.7709	81.8702	81.4885
7			83.0152	82.4427	84.3511
8			84.1603	84.3511	84.7328
9			83.9694	82.4427	85.4961
10			82.8244	78.4351	84.1603

Tabla 5.3 Resultados de la validación cruzada a 10 pliegues con el modelo de 2 clases.

Con el experimento 2, se trabajó con 3 clases: **positivo**, **negativo** y **neutro**, en donde se utilizaron las 3878 críticas. Dos de estas categorías se formaron a partir de la unión de las clases; muy mala y mala para conformar la clase negativa mientras que buena y muy buena conforman la clase positiva, como se indica en la tabla 5.4, mientras que para la clase neutro se ocuparon las 1253 críticas que conformaban la categoría neutra del corpus original.

Clases del corpus original.	Número de documentos por clase
Muy mala	351
Mala	923
Neutra	1253
Buena	890
Muy buena	461
Total, de críticas	3878

Clases tomadas para la muestra 2.	Número de documentos por clase	Clase nueva	Número de documentos por clase
Muy mala	351	Negativa	1274
Mala	923		
Buena	890	Positivo	1351
Muy buena	461		
Neutra	1253	Neutro	1253
Total, de críticas			3878

Tabla 5.4 Clases para el experimento 2.

Así como para el experimento anterior se empleó el modelo de representación de bolsa de palabras junto con el peso booleano. Los resultados correspondientes a los tres clasificadores se detallan en la tabla 5.5, destacándose una exactitud de 60.29673%, Además, la tabla 5.6 presenta los resultados de la validación cruzada.

No. De clases	Clases			
		SVM	NB	LR
3	Positivo	58.18061	59.03221	60.29673
	Negativo			
	Neutro			

Tabla 5.5 Resultados del modelo con 3 clases.

No.	Número de clases.	Clases			
			SVM	NB	LR
1	3	Positivo Negativo Neutro	57.5483	60.5161	60
2			61.2903	61.6774	62.7096
3			56.6451	58.4516	59.2258
4			57.6774	57.1612	61.2903
5			59.2258	60.129	61.4193
6			57.2903	58.3225	59.0967
7			58.0645	58.5806	59.4838
8			59.0967	59.0967	59.8709
9			56.9032	55.2258	58.0645
10			58.0645	61.1612	61.8064

Tabla 5.6 Resultados de la validación cruzada a 10 pliegues con el modelo de 3 clases.

Para crear la muestra 3 del último experimento se utilizaron las clases en su forma original, por lo que en esta ocasión se hicieron uso de las 5 categorías del corpus original, muy mala, mala, neutra, buena y muy buena. Así mismo se utilizó un peso booleano y un modelo de representación de bolsa de palabras. Los resultados obtenidos de la experimentación en cuanto a cada uno de los tres clasificadores se pueden observar en la tabla 5.7, distinguiéndose como

mejor resultado un 52.07174% de exactitud. Al mismo tiempo en la tabla 5.8 se muestran los resultados obtenidos en cuanto a la validación cruzada.

No. De clases	Clases			
		SVM	NB	LR
5	Muy mala	50.3588	49.03748	52.07174
	Mala			
	Neutra			
	Buena			
	Muy buena			

Tabla 5.7 Resultados del modelo con 5 clases.

No.	Número de clases.	Clases			
			SVM	NB	LR
1	5	Muy mala Mala Neutra Buena Muy buena	50.734	47.9608	51.876
2			51.0603	49.5921	54.323
3			50.8972	46.0032	52.8548
4			51.0603	49.9184	52.2022
5			52.202	47.7977	52.3654
6			50.0815	48.7765	51.2234
7			50.0815	53.181	53.3442
8			49.1027	49.9184	50.4078
9			50.734	49.2659	52.6916
10			47.6345	47.9608	49.429

Tabla 5.8 Resultados de la validación cruzada a 10 pliegues con el modelo de 5 clases.

Como se puede apreciar en los resultados de los experimentos anteriores, los clasificadores obtienen mejores resultados entre menos clases se ocupen, como se puede observar en la figura 5.0, siendo los mejores resultados los obtenidos en el modelo a 2 clases con un porcentaje de exactitud que supera el 80%.

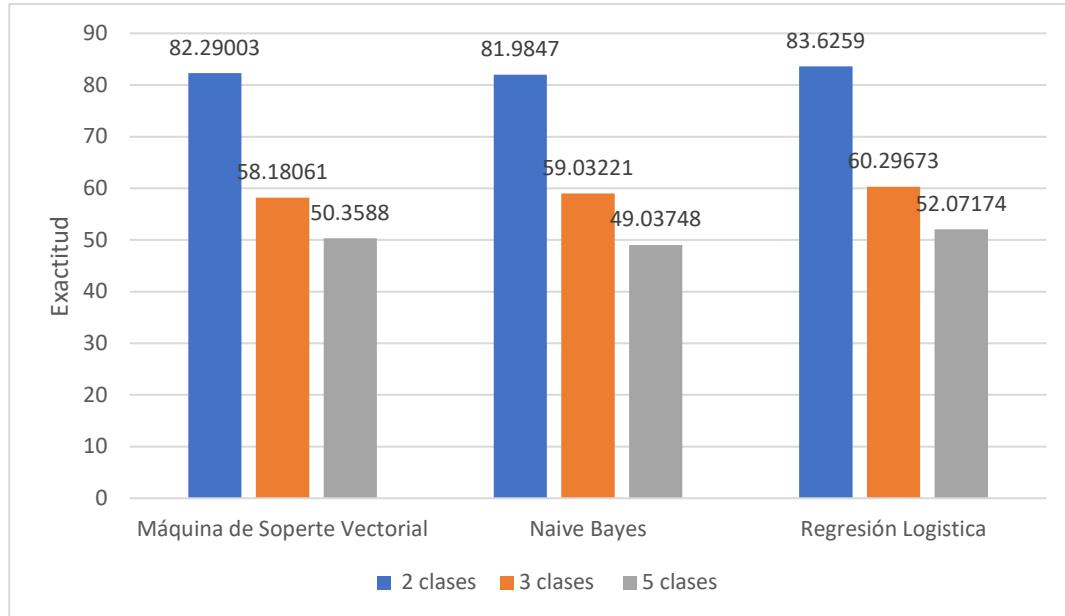


Figura 5.0 Comparación de los resultados de los tres experimentos.

En la revisión de la literatura la tarea de minería de opiniones en el corpus de críticas de cine en español "Muchocine" se han utilizado diferentes clasificadores. En este estudio, proponemos el uso de tres clasificadores, dos de los cuales ya se habían utilizado en un trabajo anterior como lo son Máquina de Soporte Vectorial y Naive Bayes, mientras que se propone el uso de un nuevo clasificador siendo este la Regresión Logística, evaluando su rendimiento mediante la métrica de exactitud.

En el trabajo de Cruz, (Cruz et al., 2008) se emplean clasificadores que calculan la orientación semántica con una semilla ($S0v1 - S$) de forma supervisada y orientación semántica no supervisada utilizando dos semillas ($S0v2 - NS$) evaluando los resultados con la métrica *accuracy*. Mientras que, en el estudio de Martínez Cámara, (Martínez Cámara et al., 2011) se hace uso de dos clasificadores los cuales son Naive Bayes y SVM, aplicando las métricas recuerdo, *F1-measure* y *accuracy* para evaluar los clasificadores. Por otro lado, Flores Domínguez, (Flores Domínguez & Tapia Fabela, 2020) utiliza una la red neuronal *Back-propagation* evaluándola con *accuracy*.

Los resultados obtenidos en este estudio son comparables con los hallados en el estado del arte. Sin embargo se debe de tomar en cuenta que dichos trabajos cuentan con diferencias significativas que pudieron alterar los resultados finales como; el número total de críticas empleado, el número de clases, el preprocesamiento utilizado, los modelos de selección de términos, el pesado de los términos así como las herramientas empleadas para cada estudio, sin embargo aun con estas discrepancias los resultados fueron competitivos con los demás

obteniendo un *accuracy* mayor al expresado en el trabajo de Cruz, (Cruz et al., 2008) y manteniéndose al a par con Martínez Cámara, (Martínez Cámara et al., 2011) y Flores Domínguez (Flores Domínguez & Tapia Fabela, 2020) , como se puede observar en la figura 5.1

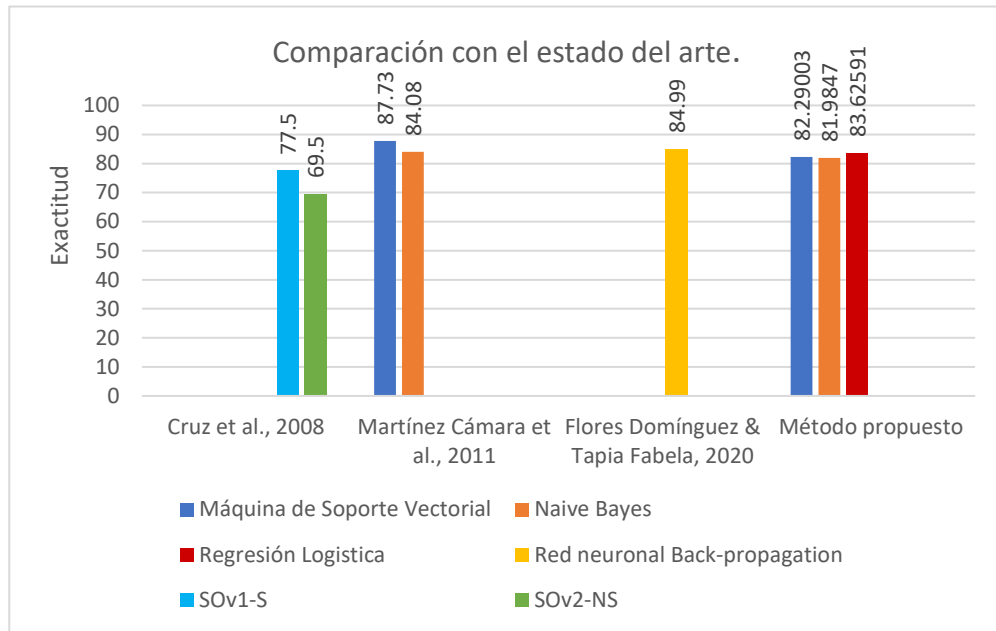


Figura 5.1 Comparación de los resultados con respecto al estado del arte.



Conclusiones

En este capítulo, se exponen las conclusiones generales de la investigación, resaltando las contribuciones clave. Además, se hace una breve referencia a los estudios previos que han empleado los métodos de aprendizaje supervisado *SVM*, *LR*, y *NB* y no supervisado, así como aquellos que han analizado el corpus “Muchocine” de críticas en español.

En la presente investigación se lograron alcanzar todos los objetivos establecidos, desde la creación del modelo de análisis de opiniones hasta la realización de las pruebas experimentales correspondientes.

Así mismo podemos destacar que la hipótesis previamente planteada la cual afirmaba que, “si se desarrolla un sistema para el análisis de opiniones mediante la clasificación automática de textos se podrán clasificar las opiniones en el corpus “Muchocine” con una exactitud superior al 80%” fue llevada a cabo con éxito, puesto que en los resultados obtenidos destacamos que el resultado con una mayor relevancia fue de 83.6259% de exactitud, esto se logró implementando diversas técnicas en las críticas de cine como un preprocesamiento en el texto, retirando caracteres especiales como; signos de puntuación, signos de admiración e interrogación, así como la eliminación de *stopwords* y números. Además, se experimentó con diferentes conjuntos de clases y se utilizaron clasificadores como SVM, NB y RL.

Así bien los resultados obtenidos del modelo de análisis de opiniones permitieron observar cómo se comportaba el corpus “Muchocine” dependiendo cuantas clases se utilizaban y si estas influían a la hora de realizar los experimentos con los clasificadores, destacando así los mejores resultados fueron cuando se hizo uso de dos clases negativa y positiva obteniendo un resultado competitivo con los logrados en el estado de arte

Finalmente, un punto a destacar en los resultados es que el clasificador *Regresión Logística* fue el que sobre salió ante los demás clasificadores en esta investigación y puesto que dicho método no se había ocupado en trabajos donde emplearan el corpus “Muchocine” podemos destacar que este clasificador puede ser útil a la hora del análisis de opiniones.

Referencias

- [Agić et al., 2010] Agić, Ž., Ljubešić, N., & Tadić, M. (2010). Towards sentiment analysis of financial texts in Croatian. Language Resources and Evaluation. https://bib.irb.hr/datoteka/455003.876_Paper.pdf
- [Alarcón Ramos, 2022] Alarcón Ramos, B. (2022). Análisis de sentimientos en la red social Twitter [Tesis de Grado]. Universidad Politécnica de Madrid, Madrid, España. <https://oa.upm.es/70960/>
- [Amaral, 2020] Amaral, D. (2020). Estudio de las opiniones utilizando análisis de sentimientos a nivel de aspecto. [Tesis de grado]. Universidad de la República Facultad de Ingeniería, Uruguay. <http://tinyurl.com/24j3f3dy>
- [Barranco et al., 2012] Barranco, MJ, Castro, J. & Martínez, L. (2012). Evaluación de un método de ponderación de atributos multivaluados en sistemas de recomendación basados en contenido. Universidad De Jaén. <https://t.ly/-Xfu>
- [Bradley & Lang, 1999] Bradley, M. M. & P. J. Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical Report C-1, The Center for Research in Psychophysiology, 1 - 45.
- [Buchanan & O'Connell, 2006] Buchanan Leigh & O'Connell Andrew. (2006). Una breve historia de la toma de decisiones.
- [Buesa, 2020] Buesa, C. B. (2020). La minería de opinión para el análisis del discurso de odio en las redes sociales. Un estudio de caso sobre Paulo Freire en YouTube durante el periodo 2007-2019. Commons: Revista de Comunicación y Ciudadanía Digital, 9(1), 119-159. <https://doi.org/10.25267/commons.2020.v9.i1.5>
- [Callejas Cuervo & Vélez Guerrero, 2019] Callejas Cuervo, M., & Vélez Guerrero, M. A. (2019). Predicción electoral usando un modelo híbrido basado en análisis sentimental y seguimiento a encuestas: Elecciones presidenciales de

Colombia. Revista politecnica, 15(30), 94-104.
<https://doi.org/10.33571/rpolitec.v15n30a9>

[Calvo Madurga, 2020] Calvo Madurga, A. (2020). Análisis de sentimientos y emociones en redes sociales usando ML [Trabajo de Fin de Grado]. Universidad de Valladolid, Valladolid, España.
<http://uvadoc.uva.es/handle/10324/44149>

[Castro et al., 2022] Castro, J. C. M., De Los Angeles Mejía, M. A., Armando, J., & Sanjuana, I. (2022). Clasificación de polaridad: Ciberacoso en Twitter. ResearchGate.
https://www.researchgate.net/publication/363520315_Clasificacion_de_polaridad_Ciberacoso_en_Twitter

[Cardoso et al., 2020] Cardoso, A. C., Talamé, M. L., Amor, M. N., & Monge, A. (2020). Creación de un Corpus de Opiniones con Emociones Usando Aprendizaje Automático. Tecnología y ciencia, 37, 11-23.
<https://doi.org/10.33414/rtyc.37.11-23.2020>

[Cardoso et al., 2019] Cardoso, A. C., Talame, L., Amor, M., & Neil, C. (2019). Minería de opiniones: Análisis de Sentimientos en una Red Social. XXI Workshop de Investigadores en Ciencias de la Computación (WICC 2019, Universidad Nacional de San Juan).
http://sedici.unlp.edu.ar/bitstream/handle/10915/77379/Documento_completo.pdf?sequence=1

[Carrillo de Albornoz, 2011] Carrillo de albornoz Cuadrado, J. (2011). Un Modelo Lingüístico-Semántico Basado en Emociones para la Clasificación de Textos Según su Polaridad e Intensidad. [Tesis de doctorado]. Departamento de Ingeniería del Software e Inteligencia Artificial Facultad de Informática Universidad Complutense de Madrid. <https://tinyurl.com/yqxy6u8k>

[Carrión García et al., 2021] Carrión García, J. A., Fernández Ambrona, R., Cristina Molina Gerbolés, C., & Motoso González, P. (2021). Análisis de Sentimientos de Tweets en Español Basado en Técnicas Aprendizaje Supervisado [Trabajo de fin de grado]. Departamento de Sistemas Informáticos y Computación Facultad de Informática Universidad Complutense de Madrid, Madrid, España. <https://eprints.ucm.es/id/eprint/73994/>

- [Cedeno Moreno & Vargas Lombardo, 2020] Cedeno Moreno, D., & Vargas Lombardo, M. (2020). Aprendizaje automático aplicado al análisis de sentimientos. *Revista de I+D tecnológico*, 16(2). <https://doi.org/10.33412/idt.v16.2.2833>
- [Coello Guilarte, 2019] Coello Guilarte, D. L. (2019). Clasificación translingüe para la detección de depresión en usuarios de Twitter [Tesis de Maestría]. Instituto Nacional de Astrofísica, Óptica y Electrónica. Tonantzintla, Puebla. <https://inaoe.repositorioinstitucional.mx/jspui/handle/1009/1952>
- [Corpus, s. f.] Corpus. (s. f.). «Diccionario de la lengua española». Recuperado 5 de mayo de 2023, de <https://dle.rae.es/corpus>
- [Cortez Vásquez et al. 2009] Cortez Vásquez, A., Vega huerta, H., Pariona Quispe, J. & Huayna, A. M. (2009). Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*, 6(2), 45 – 54. <https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923>
- [Cruz et al., 2008] Cruz, F. L., Troyano, J. A., Enriquez, F., & Ortega, J. (2008). Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del lenguaje Natural*, 73-80. <http://hdl.handle.net/10045/8067>
- [Davia García, 2021] Davia García, A. (2021). Estudio de la portabilidad de un sistema de análisis de sentimiento de Tweets en castellano para el catalán. [Tesis de Grado]. Escola Tècnica Superior d'Enginyeria Informàtica Universitat Politècnica de València. <https://riunet.upv.es/handle/10251/173681>
- [Davis, 2006] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233–240
- [Denecke, 2008] Denecke, K. (2008). Using SentiWordNet for multilingual sentiment analysis. 2008 IEEE 24th International Conference on Data Engineering Workshop, 507-512. <https://doi.org/10.1109/icdew.2008.4498370>
- [De León Camacho & Rosas Sánchez, 2017] De León Camacho, H. G., & Rosas Sánchez, J. A. (2017). Clasificación de opiniones en español utilizando información

- léxica y diccionarios. [Tesis de Grado]. Instituto Politécnico Nacional, Ciudad de México, México. <http://tesis.ipn.mx/handle/123456789/21447>
- [Escortell Pérez, 2017] Escortell Pérez, M. (2017). El impacto de las emociones en el análisis de la polaridad en textos con lenguaje figurado en Twitter [Proyecto Final de Máster]. Universitat Politècnica de València, Valencia España. <https://riunet.upv.es/handle/10251/86127>
- [Flores Domínguez & Tapia Fabela, 2020] Flores Domínguez, M., & Tapia Fabela, J. L. (2020). Análisis de sentimiento a nivel de documento en críticas de cine en español. Res. Comput. Sci., 149(8), 475-487. <https://tinyurl.com/2xljdsc>
- [Ghorbel & Jacot, 2011] Ghorbel, H., & Jacot, D. (2011). Sentiment analysis of French movie reviews. En Studies in computational intelligence (pp. 97-108). https://doi.org/10.1007/978-3-642-21384-7_7
- [Goldberg, 2017] Goldberg, Y. (2017), 'Neural network methods for natural language processing', Synthesis lectures on human language technologies 10(1), 1-309.
- [González Guerra, 2020] González Guerra, C. (2020). Paridad de herramientas para el análisis de sentimientos sobre un corpus en español generado desde Twitter [Tesis Magister]. Universidad Andrés Bello. Santiago de Chile. <https://repositorio.unab.cl/xmlui/handle/ria/15339>
- [Guzmán Cabrera, 2019] Guzmán Cabrera, R. (2019). Automatic Classification of Cross-Domain Opinions. Computación Y Sistemas, 23(4), 1541-1548. <https://doi.org/10.13053/cys-23-4-3307>
- [Hastie, et al., 2009] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning.
- [Henríquez Miranda et al., 2016] Henríquez Miranda, C., Guzmán, J., & Salcedo, D. (2016). Minería de Opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hoteles. Procesamiento de Lenguaje Natural, 25-32. <https://www.redalyc.org/pdf/5157/515754423002.pdf>

- [Hernández & Gómez, 2013] Hernández M., & Gómez J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. Revista Politécnica, 32(1), 87-96. <http://rua.ua.es/dspace/handle/10045/33514>
- [Herrera Contreras, 2020] Herrera Contreras, A. A. (2020). Modelo para Análisis de Sentimiento Utilizando Aprendizaje Automático [Tesis de Maestría]. Instituto Tecnológico Superior de Misantla, Misantla, Veracruz. <https://tinyurl.com/298c9uqc>
- [Insua Yañez, 2019] Insua Yañez, A. (2019). Sistema Deep Learning para el análisis de sentimientos en opiniones de productos para la ordenación de resultados de un buscador semántico. [Tesis de grado]. Facultade de Informática Universidade da Coruña. <http://tinyurl.com/2yn6zc3r>
- [Jiménez Zafra et al., 2014] Jiménez Zafra, S. M., Martínez Cámara, E., Martín Valdivia, M. T., & Ureña López, L. A. (2014). Desafíos del Análisis de Sentimientos. Actas de las V Jornadas TIMM, 15-18. <https://tinyurl.com/23prd6c4>
- [Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining. Springer. <https://doi.org/10.1007/978-3-031-02145-9>
- [López Ramos & Arco García, 2019] López Ramos, D., & Arco García, L. (2019). Aprendizaje profundo para la extracción de aspectos en opiniones textuales. Revista Cubana de Ciencias Informáticas, 13(2), 105-145. <https://tinyurl.com/2a9uss4d>
- [Lovera & Cardinale, 2023] Lovera, F. A., & Cardinale, Y. (2023). Análisis de sentimientos en Twitter: Un estudio comparativo. Revista científica de sistemas e informática, 3(1), e418. <https://doi.org/10.51252/rcsi.v3i1.418>
- [Lunt, 2015] Lunt, M. (2015). Introduction to statistical modelling: linear regression. Rheumatology, 54(7), 1137-1140. <https://doi.org/10.1093/rheumatology/ket146>
- [Maldonado Ramones, 2022] Maldonado Ramones, E. S. (2022). Análisis de sentimientos en la red social Twitter mediante el procesamiento de lenguaje natural. [Tesis de Grado]. Universidad Nacional de Chimborazo, Riobamba, Ecuador. <http://dspace.unach.edu.ec/handle/51000/10103>

- [Mancera, 2020] Mancera Barrera, J. C. (2020). Analizador semántico de emociones y sentimientos en textos contenidos en letras de canciones en español en tiempo real [Tesis de maestría]. Universidad Autónoma de Querétaro. <https://ri-ng.uaq.mx/handle/123456789/2562>
- [Martínez Cámara et al., 2011] Martínez Cámara, E., Martín Valdivia, M. T., Perea Ortega, J. M., & Ureña López, L. A. (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento Del Lenguaje Natural*, 163-170. <https://www.redalyc.org/pdf/5157/515751747017.pdf>
- [Martinsanz & Peñas, 2005] Martinsanz, G. P., & Peñas, M. S. (2005). *Inteligencia artificial e ingeniería del conocimiento*. Editorial Ra-Ma.
- [Mironczuk & Protasiewicz, 2018] Mironczuk, M. M. & Protasiewicz, J. (2018). Una visión general reciente de los elementos más avanzados de la clasificación de textos. *Sistemas expertos con aplicaciones*, 106, 36–54. <https://www.sciencedirect.com/science/article/pii/S095741741830215X>
- [Morán López, 2022] Morán López, D. M. (2022). Una comparación de las plataformas de análisis de sentimientos en Ingeniería de software: un estudio experimental. [Tesis de grado]. Pontificia Universidad Católica del Ecuador. Esmeraldas, Ecuador. <https://repositorio.pucese.edu.ec/handle/123456789/3184>
- [Nadeem et al., 2016] Nadeem, M., Horn, M., Coppersmith, G., & Sen, S. (2016). Identifying Depression on Twitter. Cornell University. <https://doi.org/10.48550/arXiv.1607.07384>
- [Nigro, 2020] Nigro, H. O. (2020). UNA REVISIÓN a LA MINERÍA DE OPINIONES y LOS RETOS DEL PNL. *Revista Ingeniería, Matemáticas y Ciencias de la Información*, 7(13), 105-110. <https://doi.org/10.21017/rimci.2020.v7.n13.a80>
- [Nigro, 2020b] Nigro, H. O. (2020b). MINERÍA DE OPINIÓN, ACCESO ABIERTO y GOBERNANZA. *Revista Ingeniería, Matemáticas y Ciencias de la Información*, 7(14), 33-37. <https://doi.org/10.21017/rimci.2020.v7.n14.a82>

- [Pang & Lee, 2008] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135. <https://doi.org/10.1561/15000000011>
- [Pak & Paroubek, 2010] Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. En *European Language Resources Association (ELRA)*. <https://tinyurl.com/k86xtuj>
- [Pauli, 2019] Pauli, P. A. (2019). Análisis de sentimiento Comparación de algoritmos predictivos y métodos utilizando un lexicón español [Tesis de grado]. Instituto Tecnológico de Buenos Aires, Buenos Aires, Argentina. <http://ri.itba.edu.ar/handle/123456789/1782>
- [Pedregosa et al., 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *SciKit-Learn: Machine Learning in Python*. HAL (Le Centre pour la Communication Scientifique Directe). <https://hal.inria.fr/hal-00650905>
- [Pérez Vera, 2017] Pérez Vera, S. A. (2017). "Análisis y Clasificación de Textos con Técnicas Semi Supervisadas Aplicado a Área Atención al Cliente. [Tesis de grado]. Pontificia Universidad católica de Valparaíso Facultad de Ingeniería. http://opac.pucv.cl/pucv_txt/txt-8500/UCD8533_01.pdf
- [Powers, 2011] Powers, D. (2011). Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *arXiv (Cornell University)*, 2(1), 37-63. <https://arxiv.org/ftp/arxiv/papers/2010/2010.16061.pdf>
- [Robles Domínguez, 2019] Robles Domínguez, M. A. (2019). Implementación del enfoque de marketing digital utilizando la minería de opiniones para la determinación de estrategias de mejora en una empresa panificadora [Tesis de maestría]. Tecnológico Nacional de México, Orizaba, Veracruz. <https://tinyurl.com/25dk4o7h>

- [Romero, 2016] Romero, M. A. (2016, 1 julio). Minería de opinión como técnica para el análisis de información en línea – IFT. Recuperado 9 de junio de 2023, de <https://tinyurl.com/23fn33>
- [Rueda & Alamán, 2020] Rueda, A. M., & Alamán, A. P. (2020). La opinión pública en la red: análisis pragmático de la voz de los ciudadanos.
- [Russell, 2019] Russell, S. (2019). DEFINICIÓN DE INTELIGENCIA ARTIFICIAL: FUNDAMENTOS DE LA IA. Obtenido de <https://tinyurl.com/yl9sspn1> [Tesis de Grado]. Universidad Nacional de Chimborazo, Riobamba, Ecuador. <http://dspace.unach.edu.ec/handle/51000/10103>
- [Siurob Palomero, 2019] Siurob Palomero, T. A. (2019). Textos de usuarios con depresión: atributos que los representan y posible detección [Proyecto Terminal]. Universidad Autónoma Metropolitana, Cuajimalpa, Ciudad de México. https://gabyrr.github.io/assets/documents/PT_TonantzinSiurob_2019.pdf
- [Sobrinho Sande 2018] Sobrinho Sande, J. C. (2018), Análisis de sentimientos en Twitter, pp. 19–22.
- [Sokolova & Lapalme, 2009] Sokolova, M. & Lapalme, G. (2009). Un análisis sistemático de medidas de desempeño para tareas de clasificación. Procesamiento y gestión de la información, 45 (4), 427–437. <https://t.ly/EYUIC>
- [Suarez Mosquera, 2021] Suarez Mosquera, J. (2021). Comparación entre algoritmos de Machine Learning en la implementación del análisis de sentimientos para textos en español [Proyecto de grado]. Universidad distrital Francisco José de Caldas, Bogotá. <https://tinyurl.com/yn9nwps>
- [Valladares Cedillo, 2022] Valladares Cedillo, J. G. (2022). Análisis de sentimientos para textos cortos en español, una revisión del estado del arte [Tesis de Grado]. Universidad Politécnica Salesiana. Quito, Ecuador. <http://dspace.ups.edu.ec/handle/123456789/22180>
- [Vázquez, 2008] Vázquez Cuchillo, J. (2008). Recuperación de Información utilizando Secuencias Frecuentes Maximales [Tesis de maestría]. Instituto Nacional de Astrofísica, Óptica y Electrónica. <https://tinyurl.com/ytdztcp>
- [Vilares, Alonso & Gómez Rodríguez, 2013] Vilares, D., M. A. Alonso, & C. Gómez Rodríguez. 2013. Clasificación de polaridad en textos con opiniones en español

mediante análisis sintáctico de dependencias. Procesamiento del lenguaje natural, 13-20.

- [Vuotto et. al, 2015] Vuotto, A., Bogetti, C., & Fernández, G. (2015). Aplicación del factor TF-IDF en el análisis semántico de una colección documental. *Biblios*, (60), 1-1
- [Xu, 2018] Xu, S. (2018). Bayesian Naive Bayes classifiers to text classification. *Journal of Information Science*, 44(1), 48-59. <https://doi.org/10.1177/0165551516677946>
- [Zhang et al., 2009] Zhang, C., Zeng, D., Li, J., Wang, F., & Zuo, W. (2009). Sentiment analysis of Chinese documents: From sentence to document level. *Journal of the Association for Information Science and Technology*, 60(12), 2474-2487. <https://doi.org/10.1002/asi.21206>

Anexos