# C:\Users\gkrios\Downloads\isntitucional vertical azul.jpgCarátula

UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA  
*La Universidad Católica de Loja*

**ÁREA TÉCNICA**

INGENIERO EN SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

TRABAJO DE TITULACIÓN.

Minería de texto para el análisis de sentimientos, en procesos de enseñanza aprendizaje al utilizar la red social Facebook

**AUTOR:** Astudillo Bustamante, Ivo Andrés  
**DIRECTORA:** Jara Roa, Dunia Inés

Loja – Ecuador  
2021

**Aprobación del director del Trabajo de Titulación**

Loja, día, de mes, de año

Título académico.

Nombres y Apellidos completos del coordinador (a) de Titulación

**Coordinador(a) de Titulación**

Ciudad. -

De mi consideración:

El presente Trabajo de Titulación denominado: (nombre del trabajo) realizado por Nombres y Apellidos completos del autor (a) o autores (as), ha sido orientado y revisado durante su ejecución, por cuanto se aprueba la presentación del mismo. Así mismo, doy fe que dicho Trabajo de Titulación ha sido revisado por la herramienta antiplagio institucional.

Particular que comunico para los fines pertinentes.

Atentamente,

Firma del Director del Trabajo de Titulación

Nombres y Apellidos completos del Director del Trabajo de Titulación.

C.I.:

**Declaración de autoría y cesión de derechos**

“Yo, Nombres y Apellidos completos, declaro y acepto en forma expresa lo siguiente:

* Ser autor(a) del Trabajo de Titulación denominado: Nombre del trabajo, de la Titulación …………….…., específicamente de los contenidos comprendidos en: se debe colocar los nombres de los capítulos elaborados en el Trabajo de Titulación, por ejemplo. Introducción, Capítulo 1. Marco teórico de la pobreza y las políticas públicas, Capítulo 2. Evidencia empírica. Metodología de la investigación, Capítulo 3. Descripción de la población seleccionada, Capítulo 4. Relato del hogar, Conclusiones y Recomendaciones, siendo nombres y apellidos completos, director (a) del presente trabajo; y, en tal virtud, eximo expresamente a la Universidad Técnica Particular de Loja y a sus representantes legales de posibles reclamos o acciones judiciales o administrativas, en relación a la propiedad intelectual. Además, ratifico que las ideas, conceptos, procedimientos y resultados vertidos en el presente trabajo investigativo son de mi exclusiva responsabilidad.
* Que mi obra, producto de mis actividades académicas y de investigación, forma parte del patrimonio de la Universidad Técnica Particular de Loja, de conformidad con el artículo 20, literal j), de la Ley Orgánica de Educación Superior; y, artículo 91 del Estatuto Orgánico de la UTPL, que establece: “Forman parte del patrimonio de la Universidad la propiedad intelectual de investigaciones, trabajos científicos o técnicos y tesis de grado que se realicen a través, o con el apoyo financiero, académico o institucional (operativo) de la Universidad”.
* Autorizo a la Universidad Técnica Particular de Loja para que pueda hacer uso de mi obra con fines netamente académicos, ya sea de forma impresa, digital y/o electrónica o por cualquier medio conocido o por conocerse, sirviendo el presente instrumento como la fe de mi completo consentimiento; y, para que sea ingresada al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública, en cumplimiento del artículo 144 de la Ley Orgánica de Educación Superior.

Firma: ...............................................................

Autor: Nombres y Apellidos completos

C.I.: ..............................................................

**Dedicatoria**

**Agradecimiento**

# **Índice de Contenido**

[CarátulaI](#_top)

[Aprobación del director del trabajo de fin de titulaciónII](#Aprobación)

[Declaración de autoría y cesión de derechosIII](#Declaración)

[DedicatoriaIV](#Dedicatoria)

[AgradecimientoV](#Agradecimiento)

[Índice de contenidosVI](#Índice)

[Resumen1](#Resumen)

[Abstract2](#Abstract)

[Introducción3](#Introducción)

[Capítulo uno5](#Capítulo_1_Estado_del_arte)

[Contexto de la investigación5](#Capítulo_1_Estado_del_arte)

* 1. Tema.......................................................................................................................13
  2. Planteamiento del problema...................................................................................13
  3. Alcance...................................................................................................................13
  4. Objetivos.................................................................................................................14
     1. *General*................................................................................................................14
     2. *Específicos*..........................................................................................................14
  5. Metodología y estrategia de desarrollo...................................................................16
  6. Estructura del documento.......................................................................................16

Capítulo dos...........................................................................................................................15

Marco teórico.........................................................................................................................15

* 1. Minería de datos15
  2. Descubrimiento de conocimiento16
  3. Datos estructurados vs datos no estructurados18
  4. Minería de texto20
     1. Preprocesamiento22
     2. Representación o forma intermedia23
  5. Técnicas de minería de texto25
     1. Naïve Bayes27
     2. Vecino más cercano KNN28
     3. Arboles de decisión30
     4. Máquinas de vectores de soporte31
     5. Redes neuronales artificiales33
  6. Herramientas para la minería de textos..................................................................36
  7. Aplicaciones de la minería de textos43
     1. Análisis de sentimientos44
  8. Trabajos relacionados46
     1. Primer estudio46
     2. Segundo estudio47
     3. Tercer estudio48
     4. Cuarto estudio49

Escribir el título del capítulo (nivel 1)1

* 1. Escribir el título del capítulo (nivel 2)2
     1. Escribir el título del capítulo (nivel 3)3

Escribir el título del capítulo (nivel 1)4

* 1. Escribir el título del capítulo (nivel 2)5

Escribir el título del capítulo (nivel 3)6

## **Índice de Figuras**

[**Figura 1.** Pipeline del procesamiento de datos 19](#_Toc61912196)

[**Figura 2.** El proceso de descubrimiento del conocimiento 20](#_Toc61912197)

[**Figura 3.**Plataformas sociales más usadas en el mundo 22](#_Toc61912198)

[**Figura 4.** Esquema general de la minería de texto 24](https://utpl-my.sharepoint.com/personal/iaastudillo_utpl_edu_ec/Documents/Documents/Trabajo%20de%20Titulación/Redacción%20TT/Documento/git/TT_IvoAstudillo_revision_v4.docx#_Toc61912199)

[**Figura 5.** Framework tradicional para el análisis de texto 25](https://utpl-my.sharepoint.com/personal/iaastudillo_utpl_edu_ec/Documents/Documents/Trabajo%20de%20Titulación/Redacción%20TT/Documento/git/TT_IvoAstudillo_revision_v4.docx#_Toc61912200)

[**Figura 6.** Marco general para el aprendizaje automático supervisado aplicado a la clasificación 29](#_Toc61912201)

[**Figura 7.** Propuesta de solución 31](#_Toc61912202)

[**Figura 8.** Tendencia de precisión de la clasificación en K de 1 a 50 33](https://utpl-my.sharepoint.com/personal/iaastudillo_utpl_edu_ec/Documents/Documents/Trabajo%20de%20Titulación/Redacción%20TT/Documento/git/TT_IvoAstudillo_revision_v4.docx#_Toc61912203)

[**Figura 9.** Árbol de decisión para decidir si esperar por una mesa 34](https://utpl-my.sharepoint.com/personal/iaastudillo_utpl_edu_ec/Documents/Documents/Trabajo%20de%20Titulación/Redacción%20TT/Documento/git/TT_IvoAstudillo_revision_v4.docx#_Toc61912204)

[**Figura 10.** Precisión de los clasificadores en términos de precisión y tiempo 36](https://utpl-my.sharepoint.com/personal/iaastudillo_utpl_edu_ec/Documents/Documents/Trabajo%20de%20Titulación/Redacción%20TT/Documento/git/TT_IvoAstudillo_revision_v4.docx#_Toc61912205)

[**Figura 11.** Estructura de una neurona artificial. 37](#_Toc61912206)

[**Figura 12.** Estructura de una red neuronal con una capa oculta 37](https://utpl-my.sharepoint.com/personal/iaastudillo_utpl_edu_ec/Documents/Documents/Trabajo%20de%20Titulación/Redacción%20TT/Documento/git/TT_IvoAstudillo_revision_v4.docx#_Toc61912207)

[**Figura 13.** Procesamiento de textos 40](#_Toc61912208)

[**Figura 14.** Desarrollo del proceso metodológico 51](#_Toc61912209)

## **Índice de Tablas**

[**Tabla 1.** Ejemplo de una hoja de cálculo de datos médicos 21](#_Toc61920248)

[**Tabla 2.** Resultados de la ejecución del algoritmo Naïve Bayes 31](#_Toc61920249)

[**Tabla 3.** Exactitud y error promedio por ejecución 39](#_Toc61920250)

[**Tabla 4.** Aplicaciones de la minería de texto. 42](#_Toc61920251)

[**Tabla 5.** Herramientas para la minería de texto 47](#_Toc61920252)

[**Tabla 6.** Tamaño de los corpus utilizados 50](#_Toc61920253)

[**Tabla 7.** Resultados sobre el corpus de validación 50](#_Toc61920254)

[**Tabla 8**. Resultados de la clasificación 52](#_Toc61920255)

[**Tabla 9.** Polaridad en cuentas infantiles en español 53](#_Toc61920256)

[**Tabla 10.** Trabajos relacionados 54](#_Toc61920257)

[**Tabla 11.** Distribución de los comentarios 59](#_Toc61920258)

**Resumen**

**Abstract**

**Introducción**

**Capítulo uno**

**Contexto de la investigación**

La finalidad de este capítulo es presentar una introducción al presente trabajo de titulación. Se empieza con la sección 1.1. en la cual se describe el tema de la investigación. Seguido se tiene el planteamiento del problema, en la sección 1.2. En la sección 1.3. se describe el alcance de la investigación. En la sección 1.4. se plantean los objetivos tanto general como específicos. En la sección 1.5 se analiza la metodología y estrategia de desarrollo y se finaliza este capítulo con una explicación de la estructura del documento en la sección 1.6.

## **Tema**

Minería de texto para el análisis de sentimientos, en procesos de enseñanza aprendizaje al utilizar la red social Facebook.

## **Planteamiento del problema**

En la mayoría de las universidades del mundo, los entornos virtuales de aprendizaje y sistemas de gestión del aprendizaje se han convertido en las principales herramientas informáticas en procesos de enseñanza – aprendizaje. Estos sistemas permiten que el docente realice un seguimiento exhaustivo de distintas actividades como: tareas, evaluaciones, foros, chats de discusión, etcétera. Además, sirven como una especie de red social educativa, en donde el docente y los alumnos interactúan constantemente, pero... ¿qué hay de las redes sociales tradicionales?, ¿es posible llevar a cabo procesos de enseñanza – aprendizaje en alguna red social, por ejemplo Facebook?

De ser así, ¿qué actitudes muestran los estudiantes al utilizarla? La presente investigación pretende responder a estas interrogantes, partiendo de la base que ya se tiene grupos de Facebook que están dirigidos por un docente, en los cuales se desarrollan actividades académicas de una materia en concreto. Particularmente las actividades planteadas por la docente son: chats de discusión de algún tema en específico y el desarrollo colaborativo de tareas.

De esta manera, la presente investigación está orientada al análisis computacional de los datos generados en los grupos de estudio de Facebook con el fin de determinar si los estudiantes están conformes al utilizar esta red social como herramienta de apoyo en procesos de enseñanza – aprendizaje.

## **Alcance**

La investigación se centra en el análisis de sentimientos de estudiantes de la carrera de Ingeniería Informática de la Universidad Técnica Particular de Loja, específicamente de la modalidad Abierta y a Distancia, que interactúan en grupos de estudio de la materia de Inteligencia Artificial en la red social Facebook. El dataset será generado con textos de aproximadamente treinta grupos conformados por cinco estudiantes cada grupo. El texto será pre - procesado con técnicas de () para posteriormente aplicar técnicas de minería de textos y machine learning. Para llevar a cabo la experimentación se va a trabajar con el lenguaje de programación Python.

## **Objetivos**

Para el presente trabajo de titulación se han definido los siguientes objetivos:

### **General**

Identificar las actitudes de los estudiantes cuando se utiliza la red social Facebook en procesos de enseñanza aprendizaje, para detectar la pertinencia de uso en procesos formativos reglados.

### **Específicos**

* Creación de un corpus textual con las intervenciones de cada uno de los grupos en la red social Facebook.
* Elegir la metodología más apropiada para la realización de este tipo de minería de texto.
* Seleccionar el algoritmo de aprendizaje automático que mejor se ajuste a los datos recolectados a fin de obtener la mayor tasa de aciertos posibles.
* Realizar la evaluación e interpretación de los resultados obtenidos.

## **Metodología y estrategia de desarrollo**

El presente trabajo de titulación costa de tres etapas principales; la primera etapa corresponde a la elaboración del marco teórico acerca de la minería de texto. Así como, de las diferentes técnicas usadas en problemas de minería de texto, poniendo especial énfasis en el análisis de sentimientos. La segunda etapa corresponde al proceso de extracción de los datos, procesamiento y aplicación de los algoritmos de minería de texto. Finalmente, en la última etapa se llevará a cabo el análisis de los resultados obtenidos.

## **Estructura del documento**

El presente trabajo de titulación está estructurado en cuatro capítulos, se plantean de tal manera que están ordenados lógicamente con la finalidad de responder adecuadamente a los objetivos de la investigación.

En el **capítulo uno** se presenta el contexto de la investigación, en el cual se incluye; el tema, planteamiento del problema, alcance, objetivos, metodología y estrategia de desarrollo y estructura del documento.

El **capítulo dos** constituye el marco teórico de la investigación en cuestión. Por un lado, a manera de introducción se empieza explicando lo que es la minería de datos en general. Seguido de esto se aborda a profundidad la minería de texto y sus diferentes etapas, así como también las principales técnicas utilizadas en procesos de clasificación. También se hace alusión a los distintos usos que tiene a la minería de textos, profundizando especialmente en el análisis de sentimientos. Luego de esto se analizan las principales herramientas existentes en el mercado. Para finalizar esta sección se presenta un apartado con cuatro trabajos relacionados y un estudio comparativo de estos.

El **capítulo tres** ....

El **capítulo cuatro** .....

**Capítulo dos**

**Marco teórico**

En el presente capítulo se define el marco teórico de la investigación. Se inicia con el estudio de la minería de datos en la sección 2.1. En la sección 2.2 se analiza el descubrimiento de conocimiento. En la sección 2.3 se estudia los datos estructurados y los datos no estructurados. Seguido de esto, en la sección 2.4 se aborda la minería de texto. En la sección 2.5 se revisan las principales técnicas de la minería de texto usadas en problemas de clasificación. En la sección 2.6 se presentan distintas aplicaciones que tiene la minería de textos. En la sección 2.7 se comparan las principales herramientas existentes en el mercado para resolver problemas de minería de textos. En la sección 2.8 se estudian cuatro trabajos relacionados. Finalmente, en la sección 2.9 se realiza un estudio comparativo de estos trabajos.



## **Minería de datos**

En la actualidad la cantidad de datos que generan las aplicaciones y sistemas informáticos es abrumadora. Desde una pequeña tienda de zapatos hasta las grandes empresas y organizaciones almacenan todo tipo de información. Imágenes, videos, datos numéricos, texto, etcétera; se guardan en las conocidas bases de datos las cuales permiten almacenar cantidades muy grandes de datos a costos relativamente bajos y que su integración y administración resulte cada vez más sencilla.

Ahora bien, Weiss *et al*. (2005) explican que los datos que se almacenan en tablas y que se transforman a una representación numérica se los conoce como datos estructurados. A estos datos se los puede procesar con técnicas estadísticas y obtener cierta información. El problema es que hoy en día no solo se almacenan datos estructurados, existe una gran cantidad de datos que se encuentran en un formato no estructurado por lo que si solo se trabaja con los datos que están almacenados en un formato de tablas, se estaría limitando la posibilidad de encontrar información relevante del otro conjunto de datos. Los datos estructurados y no estructurados se analizan en el apartado 2.3.

Con el paso del tiempo los investigadores se dieron cuenta que los datos contienen patrones e información que puede ser crucial para la toma de decisiones, es por eso que a partir de la década de los noventa, empezaron a aplicar nuevos métodos y técnicas que integran la estadística, la inteligencia artificial, el aprendizaje automático y los sistemas de bases de datos. Estos métodos y técnicas tienen el potencial de manejar y procesar grandes cantidades de datos con el fin de encontrar patrones ocultos y potencialmente útiles con el objetivo de que las empresas y organizaciones tomen mejores decisiones tanto para su beneficio, como para el de sus consumidores o la sociedad en general. Estos nuevos métodos y técnicas crearon un nuevo campo de investigación que se denomina minería de datos.

Para Aggarwal, como para muchos otros investigadores: “la minería de datos es el estudio de recopilar, limpiar, procesar, analizar y obtener información útil de los datos” (2015, p. 27).

Para Mourya & Gupta (2012) la minería de datos es el uso de tecnologías inteligentes que ayudan a los usuarios finales a extraer conocimiento oculto que reside en las grandes bases de datos.

La minería de datos es un proceso que contiene algunas fases, comenzando con la recopilación de los datos, seguido por la extracción de características y limpieza de los datos, el procesamiento analítico y el diseño de algoritmos (ver figura 1).

**Figura 1.**Pipeline del procesamiento de datos

Nota. Adaptado de Data Mining (p.4) [Diagrama de flujo], por Aggarwal, C. 2015. (https://doi.org/10.1007/978-3-319-14142-8). Springer International Publishing Switzerland

Colección de datos

Extracción de características

Limpieza e integración

Preprocesamiento de datos

Construcción bloque 1

Construcción bloque 2

Analítica de procesamiento

Retroalimentación (opcional)

Retroalimentación (opcional)

**Salida del Análisis**

La explicación de estas fases está fuera del alcance de esta investigación ya que como veremos posteriormente los datos que se recopilaron para la presente, son datos en formato texto no estructurados, y para el manejo de estos se utilizan técnicas de minería de texto que están bastante relacionadas con la minería de datos es por esto que es conveniente iniciar esta investigación haciendo una breve explicación de la minería de datos.

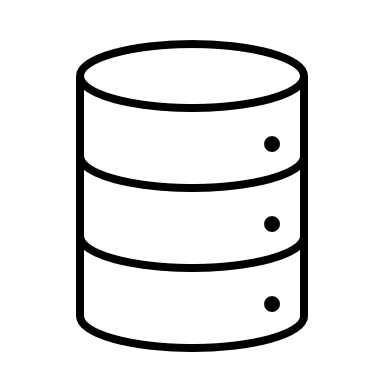
## **Descubrimiento de conocimiento**

A diferencia de Aggarwal, Bramer (2016) centra la definición de minería de datos en algoritmos que permiten descubrir patrones o información oculta de los datos. Además, señala que la minería de datos es la parte central de un proceso mayor llamado descubrimiento del conocimiento.

Bramer define al descubrimiento de conocimiento como “la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos” (2016, p. 18).

**Figura 2.**El proceso de descubrimiento del conocimiento

Nota. Adaptado de Principles of Data Mining (p. 18) [Diagrama de flujo], por Bramer, M. 2016. (https://doi.org/10.1007/978-1-4471-7307-6). Springer-Verlag London Ltd.



Origen de datos

Integración

Almacenamiento

Selección y   
pre - procesamiento

Preparación

Minería de

datos

Conocimiento

Interpretación

En la figura 2 se observa los distintos enfoques de los investigadores con respecto a la minería de datos. Para Aggarwal la minería de datos abarca todo el proceso de descubrimiento de conocimiento, mientras que para Bramer la minería de datos es una fase de este proceso.

Esta diferencia la hacen notar Han *et al*. (2012) quienes indican que la minería de datos es tratada por muchos como un proceso de descubrimiento de conocimiento en bases de datos (KDD por sus siglas en inglés Knowledge Discovery in Databases), mientras que para otros, la minería de datos es un simple paso esencial en este proceso de descubrimiento de conocimiento.

## **Datos estructurados y datos no estructurados**

Como se señaló en el apartado anterior, las técnicas de minería de datos trabajan con datos estructurados y para ello una vez que han sido recopilados, integrados y pre - procesados, estos deben de estructurarse mediante el proceso de “Preparación de datos” que se lo ve representado en la figura 2. O mejor aún como nos dicen Weiss *et al*. (2005) es posible recopilar datos ya estructurados en función de un diseño previo de minería de datos que se vaya a utilizar.

La mayoría de los datos que utilizan las empresas y organizaciones son de tipo texto y de tipo numérico. Weiss *et al*. (2005) mencionan que por regla general la minería de datos espera dos tipos de información: (a) ordenada numérica y (b) categórica. Los atributos numéricos corresponden por ejemplo, al peso o los ingresos, que claramente tienen valores mayores o menores y por ende están ordenados. Mientras que los atributos categóricos son códigos numéricos desordenados que tienen una definición para que puedan ser interpretados por las personas. Por ejemplo, los atributos de verdadero y falso, pueden ser representados por un uno o un cero; así como también, los géneros femenino y masculino que pueden ser representados por un código para su interpretación.

Weiss *et al*. (2005) concluyen que si los datos pueden describirse mediante una hoja de cálculo con su formato tabular, estos están en una fase altamente estructurada. (ver tabla 1).

**Tabla 1.**Ejemplo de una hoja de cálculo de datos médicos

|  |  |  |  |
| --- | --- | --- | --- |
| **Género** | **Altura** | **Peso** | **Código** |
| M | 175 | 65 | 3 |
| F | 141 | 72 | 1 |
| . . . | . . . | . . . | . . . |
| F | 160 | 59 | 2 |

*Nota.* Adaptado de *Text Mining (p.3)*, por Weiss et al., 2005, (https://doi.org/10.1007/978-1-84996-226-1). Springer Science+Business Media, Inc.

A diferencia de los datos estructurados, los datos no estructurados son textos. “El texto suele ser una colección de documentos no estructurados sin requisitos especiales para la composición de documentos”. (Weiss *et al*., 2005, p.12)

Todo tipo de revistas, artículos, documentos, e-books, blogs, sitios de noticias, post y conversaciones en redes sociales; son información no estructurada en lo concerniente a la minería de datos y sabemos bien que esta información es la que más abunda hoy en día.

Armstrong (2019) en el sitio web *statista* publica que hasta octubre de 2019 existen 1,72 mil millones de sitios web, la gran mayoría de estos sitios contienen información no estructurada.

Los reportes y estadísticas existentes sobre el uso de sitios web muestran que la mayor cantidad de tráfico se encuentra en las plataformas sociales. “Más de 4.500 millones de personas usan Internet a principios de 2020, mientras que los usuarios de las redes sociales han superado la marca de los 3.800 millones. Casi el 60 por ciento de la población mundial ya está en línea, y las últimas tendencias sugieren que más de la mitad de la población total del mundo usará las redes sociales a mediados de este año” (Kemp, 2020).

**Figura 3.***Plataformas sociales más usadas en el mundo*

Nota. Adaptado de DATAREPORTAL [Gráfico de barras], por Kemp, S., 2020, (https://bit.ly/3fMYMVr). Kepios.

Como se aprecia en la figura 3, el número de usuarios activos por mes es realmente impresionante. Kemp destaca que las redes sociales que lideran las estadísticas a enero de 2020 son: Facebook con de 2.449 millones de usuarios activos mensualmente, seguida de YouTube con 2.000 millones de usuarios activos, luego se tiene a Whatsapp con 1300 millones.

En estas redes sociales las personas comparten datos en distintos formatos: fotos, videos y lo que más nos interesa en esta investigación los textos. Textos que viene en forma de comentarios, estados, conversaciones vía chats, tweets, discusiones, post; etcétera. En (Internet-live-stats) se puede observar estadísticas en tiempo real de algunas redes sociales, el día que se está redactando este documento (8 de mayo de 2020) tenemos que: se han escrito en Twitter un aproximado de 617 millones de tweets, se han enviado 208 mil millones de emails, se han publicado 117 millones de posts en Tumblr y se han escrito más de 5 millones de blogs.

Todos estos textos que se publican en blogs, chats, emails, redes sociales, etcétera, es información que los humanos la podemos entender es decir tienen un conocimiento explícito, pero como ya se había mencionado anteriormente este tipo de datos no estructurados están en continuo crecimiento y se vuelve inhumano leer estas grandes cantidades de información. Es por eso que las empresas y organizaciones se ven en la necesidad de invertir en tecnologías que automaticen estos procesos con el fin de poder encontrar patrones o información potencialmente útil en los textos y es aquí en donde interviene otro campo de investigación que es la minería te texto que será abordada en el siguiente punto.

## **Minería de texto**

Feldman & Sanger (2007) citados por Taeho (2019) mencionan que “La minería de texto se define como el proceso de extracción del conocimiento implícito de los datos textuales” (p. 17).

De la misma forma Banchs (2013) afirma que la minería de texto es el proceso de descubrir conocimiento cuando los datos de origen son texto.

Estas definiciones tienen mucha similitud con el proceso KDD (descubrimiento de conocimiento en bases de datos) de minería de datos, tal como lo mencionan Justicia De La Torre *et al*. (2005). Ellos dicen que de manera similar a KDD, el descubrimiento de conocimiento en texto (KDT, por sus siglas en ingles Knowledge Discovery in Text) se puede definir como el proceso de extracción de información que previamente se desconocía y que puede ser potencialmente útil teniendo como fuente de datos un texto.

Para Weiss *et al.* (2005) los especialistas en minería de datos son conocidos como los “chicos de los números” porque sus datos están en forma numérica mientras que los especialistas en minería de texto esperan colecciones de documentos, donde los contenidos son legibles y tienen un significado obvio para las personas.

En la figura 4 se muestra un esquema general que se lleva a cabo en tareas de minería de texto.

**Minería de texto**

Preprocesamiento

descubrimiento

**Figura 4.**Esquema general de la minería de texto

*Nota.* Adaptada de *Generador de los grafos conceptuales a partir del texto en español* *(p. 40)* [Diagrama de flujo], por Hernández Cruz, 2007, (https://bit.ly/3fPTto4).

De la misma forma Hu & Liu (2012) en el artículo “Text analytics in social media” hablan de un framework tradicional para el análisis de texto:

*Nota.* Adaptada de *Mining Text Data (p.389)* [Diagrama de flujo]*,* por Hu & Liu, 2012, (https://doi.org/10.1007/978-1-4614-3223-4). Springer Science+Business Media, LLC.

Text Corpus

Stop Words Removal

Stemming

Tokenization

Twitter

Facebook

Blogger

Preprocesamiento

Representación

Descubrimiento de conocimiento

Vector Space Model

Bag of words

Clasificación

Agrupamiento

análisis de sentimientos

Detección de eventos

**Figura 5.**Framework tradicional para el análisis de texto

Como se puede observar en la figura 5 existe un consenso por parte de los investigadores en lo que ha análisis de textos se refiere. La primera fase, consta de la obtención del texto, el cual puede ser documentos de un tamaño considerable o micropost que los usuarios comparten en sus redes sociales. La segunda fase, y quisa la más importante en este proceso es el preprocesamiento del texto con el cual se va a obtener una representación o forma intermedia del mismo. A esa representación o forma intermedia ya es posible aplicarle los algoritmos de minería de texto como tal, con los cuales se obtendrá el tan ansiado conocimiento.

### **Pre – procesamiento**

Los datos textuales están en formato no estructurado o mejor dicho se encuentran en un formato que es muy complejo para que lo entiendan las máquinas; por esta razón, no es posible aplicar directamente técnicas de minería de texto sobre ellos. El preprocesamiento es una fase esencial para el proceso de descubrimiento de conocimiento y consiste en transformar el texto a una forma intermedia que permita la aplicación de técnicas para la extracción de patrones.

En la figura 5, dentro de la fase de preprocesamiento se encuentran algunos métodos como la eliminación de palabras de detención (Stop Words Removal) que consiste en eliminar las preposiciones, los artículos, los pronombres personales y demás palabras que se las considera sin sentido que comúnmente se denominan palabras vacías. Otro método de preprocesamiento es el Stemming el cual consiste en expresar las diferentes formas de las palabras a una sola conocida como la raíz por ejemplo: “leí”, “leeré” y “leyendo” puede ser reducida a su forma raíz que es “leer”.

Hu & Liu (2012) dicen que las palabras con formas variantes se las puede considerar palabras que tienen la misma característica.

Otro método de preprocesamiento de textos es la tokenización (tokenization) y consiste en dividir el flujo de texto en palabras o frases. A estas palabras o frases separadas se las conoce como *tokens.* (Weiss *et al*., 2005).

### **Representación o forma intermedia**

Una vez que el texto ha sido pre - procesado con los métodos que se mencionó anteriormente, ahora es posible representarlo a una forma intermedia. En la figura 6 se mencionan las principales formas intermedias.

Modelo de espacio vectorial (Vector Space Model) que consiste en la representación del texto por medio de vectores de términos. “En el modelo de espacio vectorial, cada palabra está representada por una variable que tiene un valor numérico que indica el peso (importancia) de la palabra en el documento” (Allahyari et al., 2017, p. 4). Un método para elaborar un VSM es TF-IDF o frecuencia de término – frecuencia inversa de documento (Term frequency – Inverse document frequency), consiste en una medida numérica para expresar cuantas veces aparecen los términos en los documentos y de esta forma determinar su relevancia.

Otra forma de representación intermedia es la bolsa de palabras (Bag of Words). “Una representación basada en bolsa de palabras es aquella en la que se toman todas las palabras del documento, se eliminan sus relaciones tanto sintácticas como semánticas y se almacenan de tal modo que quede accesible para futuros tratamientos.” (Justicia de la Torre, 2017, p. 72). Hu & Liu (2012) dicen que en el modelo de bolsa de palabras, a las palabras se las representa como variables separadas que tienen un peso numérico de importancia variable.

Cifuentes (2016) menciona los N-Gramas para la representación de textos. En ellos los textos son separados en palabras únicas llamadas unigramas. En dos palabras llamadas Bi-gramas y en tres palabras llamadas Tri - gramas.

Otro método más avanzado de representar el texto en una forma intermedia, son las ontologías. Mizoguchi *et al.* (2006) citados en Justicia de la Torre (2017) dicen que una ontología es la forma de representar un marco conceptual, en el cual se incluye la descripción de los conceptos y las relaciones entre ellos.

Existen otras representaciones intermedias como: Jerarquía de conceptos o taxonomía de términos, grafos semánticos, red IS-A, grafos conceptuales, grafos conceptuales difusos, dependencia conceptual, ap-set, entre otras. Lo que hay que tener en cuenta para la elección de una forma intermedia es que técnica de minería de texto se quiere aplicar en la fase siguiente; por ejemplo: “Si se desea utilizar técnicas de minería basadas en lógica, será necesario una representación intermedia del texto sobre la que se pueda realizar dicha inferencia. Si por el contrario, lo que se pretende es jugar con la mayor cantidad de información contenida en el documento, habrá que decidir primero qué representación intermedia nos conviene.” (Justicia de la Torre, 2017, p. 74).

Como se observa la minería de texto presenta grandes desafíos y a diferencia de la minería de datos, esta requiere de la aplicación de técnicas de recuperación de información, inteligencia artificial y Procesamiento de Lenguaje Natural (NLP por sus siglas en ingles Natural Language Processing) lo que hace que este campo de investigación sea un verdadero desafío. En el siguiente punto que corresponde a la fase previa de obtención de patrones o conocimiento, se revisarán las principales técnicas de minería de texto.

## **Técnicas de minería de textos**

Técnicas de aprendizaje automático (Machine Learning) son ampliamente usadas en la minería de texto. Los modelos de aprendizaje automático se usan para reconocer patrones dentro de los datos textuales. Existen muchos algoritmos de aprendizaje automático, que de acuerdo con su utilidad o a la salida que producen se los clasifican en dos categorías principales que son; aprendizaje supervisado y aprendizaje no supervisado. Otra categoría es el aprendizaje por refuerzo.

Russell & Norvig dicen que; “el problema de aprendizaje supervisado consiste en aprender una función a partir de ejemplos de sus entradas y sus salidas.” Mientras que, “el problema de aprendizaje no supervisado consiste en aprender a partir de patrones de entradas para los que no se especifican los valores de sus salidas.” (2004, p. 740)

En otras palabras, se puede decir que los algoritmos de aprendizaje supervisado necesitan de conjuntos de datos etiquetados para ser entrenados, por ejemplo; si se quiere determinar si un comentario de cierto producto es positivo o negativo. El curso intensivo de aprendizaje automático impartido por Google menciona que primeramente se tiene que entrenar el modelo con datos que previamente hayan sido clasificados, en nuestro caso de ejemplo comentarios clasificados en positivos o negativos, estos se puede definir como:

en donde es el comentario, es la categoría a la que pertenece.

En este ejemplo identifica si el comentario es positivo o negativo. Lo que hace el modelo es aprender la relación que existe entre e , de esta manera se pueda deducir la categoría de ejemplos sin etiqueta, que en nuestro caso serían los nuevos comentaros que hacen los clientes. Un ejemplo no etiquetado se lo define como:

en donde es el comentario, y es la etiqueta a inferir.

En el mismo curso intensivo de aprendiza automático de Google dicen que un modelo de aprendizaje automático supervisado se encarga de establecer la relación que existe entre los atributos y las etiquetas y consta de dos fases; entrenamiento e inferencia.

El *entrenamiento* es la fase en la que se le indica al modelo los ejemplos etiquetados para que este aprenda la relación que existen entre el atributo y la etiqueta.

La *inferencia* es la fase en la que se le muestra al modelo ejemplos sin etiqueta que nunca han sido vistos con el objetivo de que realice la inferencia de la etiqueta.

Como se observa en el ejemplo mencionado anteriormente es un ejemplo de clasificación, es decir, dado un comentario identificar si este es positivo o negativo. Gran parte de las aplicaciones de minería de texto son problemas de clasificación, para resolver estos problemas generalmente se utilizan modelos de aprendizaje supervisado. Por otro lado, se tienen los problemas de agrupamiento (clustering), perfilado o profiling, agrupamientos de coocurrencias, entre otros, para los cuales se utilizan modelos de aprendizaje no supervisado.

En la figura 6 se muestra un proceso general del aprendizaje automático supervisado aplicado a la clasificación.

**Figura 6.**Marco general para el aprendizaje automático supervisado aplicado a la clasificación

*Nota.* Adaptado de *Natural Language Processing with Python* [Diagrama de flujo], por Bird et al., 2009, (https://www.nltk.org/book/). CC BY 3.0.

Entrenamiento

Predicción

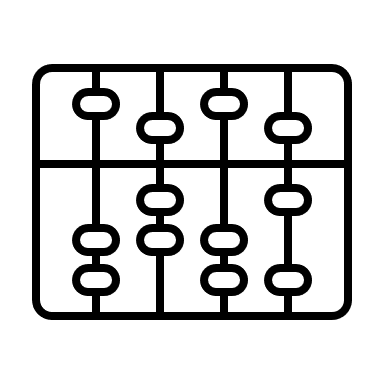
Entrada

características

características



Extractor de características

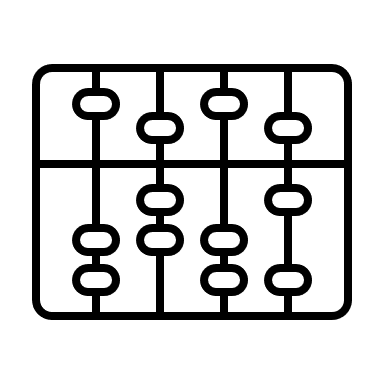


Algoritmo de Machine learning

Etiqueta



Extractor de características



Modelo de clasificación

Entrada

Etiqueta

El presente trabajo consta de un problema de clasificación, por este motivo en el siguiente punto se analizarán las principales técnicas de aprendizaje automático supervisado que se utilizan para la clasificación.

### **Clasificador Naïve Bayes (Naïve Bayes Classifier)**

Los modelos Naïve Bayes son utilizados en aplicaciones de clasificación de documentos con el propósito de asignar automáticamente documentos electrónicos individuales a una o más categorías en función de sus contenidos.

Pitigala et al. (2011) dicen que los modelos de clasificación Naïve Bayes se basan en el modelo de probabilidad posterior, además dicen que es un método rápido y robusto para la clasificación de texto. Dado un documento (d), ¿su probabilidad de pertenecer a una clase (c) es?. El objetivo de la clasificación de Naïve Bayes, es encontrar la clase óptima para un documento dado, es decir, la clase que ofrece la máxima probabilidad posterior.

Allahyari et al. (2017) dicen que el modelo Naïve Bayes es quizás el clasificador más simple y más utilizado. Modela la distribución de documentos en cada clase usando un modelo probabilístico, asumiendo que la distribución de diferentes términos son independientes entre sí. Dentro de la clasificación de texto, existen dos modelos de clasificadores Naïve Bayes. El primero es el Modelo Bernoulli Multivariado (Multi - Variate Bernoulli Model) en el cual un documento está representado por un vector de características binarias que indican la presencia o ausencia de palabras en el documento. El otro, es el Modelo Multinomial (Multinomial Model) en el que previamente un documento tiene que ser representado como una bolsa de palabras (Bag of Words) y el modelo se encarga de capturar la frecuencia de las palabras o términos. El objetivo de ambos modelos es encontrar la probabilidad posterior de una clase basada en la distribución de las palabras en el documento. La diferencia entre los dos modelos es que en el Modelo Bernoulli Multivariado no se toma en cuenta la frecuencia de las palabras; mientras que, en el Modelo Multinomial sí.

Un ejemplo se encuentra en el artículo de Reyes *et al*. (2014) en el cual se utiliza un modelo de clasificación bayesiana para determinar posibles plagios en textos científicos. En la figura 7 se presenta el flujo de su propuesta de solución.

**Figura 7.**Propuesta de solución

Documento para analizar

Análisis basado en

n - gramas

Modelo de espacio vectorial

Máxima subsecuencia común

Plagiado

No plagiado

*Nota.* Adaptado de *Técnica de clasificación bayesiana para identificar posible plagio en información textual* [Diagrama de flujo], por Reyes et al., 2014, (https://bit.ly/2ViGoKw).

En la tabla 2 se muestran los resultados obtenidos del clasificador Naïve Bayes.

**Tabla 2.**Resultados de la ejecución del algoritmo Naïve Bayes

|  |  |
| --- | --- |
| **Parámetro** | **Valor** |
| Porciento de instancias correctamente clasificadas | 97.67% |
| Porciento de instancias incorrectamente clasificadas | 2.34% |
| Promedio de verdaderos positivos (TP) | 0.98 |
| Promedio de falsos positivos (FP) | 0.01 |

*Nota.* Adaptado de *Técnica de clasificación bayesiana para identificar posible plagio en información textual*, por Reyes et al., 2014, (https://bit.ly/2ViGoKw).

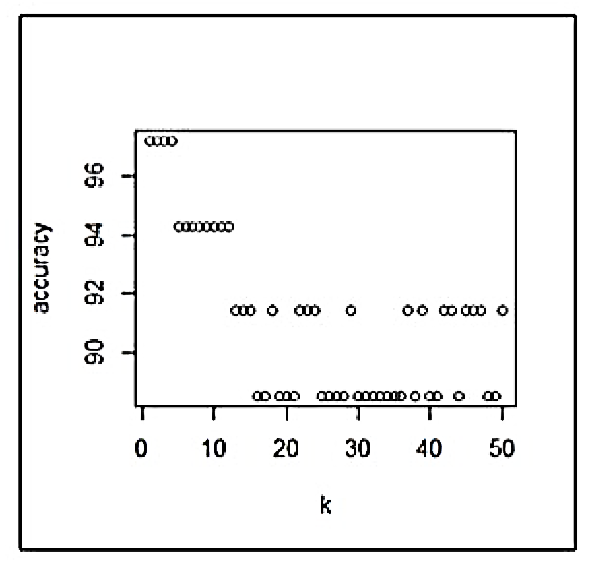
### **Clasificador Vecino Más Cercano (Nearest Neighbor Classifier or K-NN)**

El clasificador vecino más cercano es un clasificador basado en la proximidad, que utiliza medidas basadas en la distancia para realizar la clasificación. Allahyari *et al*. (2017). Así mismo, Manning *et al*. citados por Godoy, (2017) mencionan que en el clasificador vecino más cercano cuando una nueva instancia necesita ser clasificada, a esta se la compara con los ejemplos existentes, para ello usa una métrica de distancia y los ejemplos que estén más próximos son utilizados para asignar la clase a la que pertenece la nueva instancia.

Yu & Song (2010) dicen que para la aplicación del clasificador vecino más cercano, el texto tiene que estar representado en la forma intermedia de Modelo de espacio Vectorial (Vector Space Model) además mencionan que es un modelo muy efectivo para la categorización de texto. Al generar un vector de características para un texto desconocido, el modelo KNN buscará en todos los ejemplos de entrenamiento y comparará la similitud de los vectores de características con el objetivo de encontrar el ejemplo de entrenamiento K más cercano y finalmente asignar el texto desconocido al vecino K más cercano con mayor valor de clasificación.

Un ejemplo de K-NN para la clasificación automatizada de documentos textuales en categorías predefinidas se propone en el artículo de Moldagulova y Sulaiman (2017).

**Figura 8.**Tendencia de precisión de la clasificación en K de 1 a 50



Nota. Recuperado de Using KNN algorithm for classification of textual documents [Imagen], por Moldagulova & Sulaiman, 2017, (https://doi.org/10.1109/ICITECH.2017.8079924). IEEE.

En la figura 8 se observa los resultados de precisión obtenidos en valores de K de 1 a 50. El peor caso de precisión es del 88% aproximadamente. Con esto se puede deducir que, mientras más bajo es el valor de K, mejor es la precisión del clasificador.

### **Clasificadores de árbol de decisión (Decision Tree classifiers)**

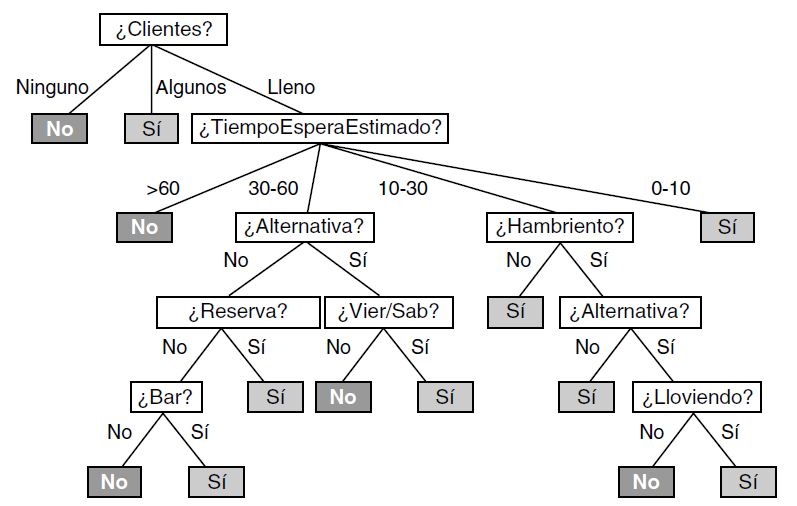
Los árboles de decisión son reglas especiales de decisión que se organizan en una estructura de árbol. Un árbol de decisión divide el espacio del documento en regiones no superpuestas en sus hojas, y se hacen predicciones en cada hoja. (Weiss *et al*., 2010)

Russell y Norvig (2004) dicen que en un árbol de decisión se toma como entrada un objeto o una situación descrita, la que es representada como un conjunto de atributos, para finalmente devolver una decisión. Una decisión, es el valor previsto de la salida dada la entrada.

Allahyari *et al*. (2017) dicen que en el caso de los datos textuales, las condiciones en los nodos del árbol de decisión se definen comúnmente en términos de términos en los documentos de texto. Por ejemplo; un nodo puede subdividirse en otros nodos (hijos) dependiendo de la presencia o ausencia de un término particular en el documento.

Yu & Song (2010) mencionan dos fases principales en los clasificadores de árboles de decisión; la primera, es la fase de crecimiento del árbol, en la que el algoritmo empieza con todo el conjunto de datos en el nodo raíz y mediante algún criterio de división el conjunto de datos se divide en subconjuntos. La segunda fase, es la fase de poda, en la cual el árbol adulto se corta para evitar un ajuste excesivo y mejorar su precisión.

Un ejemplo sencillo de cómo funcionan los árboles de decisión para algo cotidiano de la vida como esperar por una mesa en un restaurante se encuentra en el libro “Inteligencia artificial: un enfoque moderno” de Russell y Norvig (2004).



**Figura 9.**Árbol de decisión para decidir si esperar por una mesa

*Nota.* Recuperado de *Inteligencia Artificial: Un enfoque moderno (p. 745)* [Mapa conceptual] , por Russell y Norvig, 2004. Pearson Educación, S.A.

Pranckevičius y Marcinkevičius (2017) proponen un estudio comparativo de modelos de clasificación en documentos textuales demostrando que los árboles de decisión tiene una precisión baja en comparación con otros modelos como el Naïve Bayes y las Maquinas de Vectores de Soporte analizados en este documento.

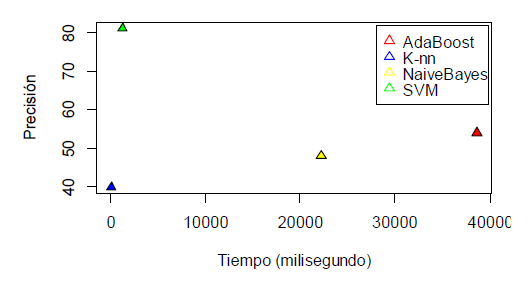
### **Máquinas de vectores de soporte (Support Vector Machines)**

Las máquinas de vectores de soporte (SVM) son algoritmos de aprendizaje supervisado que han sido utilizados ampliamente para problemas de clasificación de datos. Suthaharan dice que SVM utiliza un modelo matemático simple y lo manipula para permitir la división lineal del dominio, además menciona que la máquina de vectores de soporte se puede dividir en modelos lineales y no lineales. Se llama máquina de vector de soporte lineal si el dominio de datos se puede dividir linealmente (por ejemplo, línea recta o hiperplano) para separar las clases en el dominio original. Si el dominio de datos se puede transformar en un espacio llamado espacio de características, entonces se llama máquina de vectores de soporte no lineal. Matemáticamente, el modelado de una máquina de soporte lineal adopta la ecuación lineal:

y el modelo de una máquina de vector de soporte no lineal adopta la ecuación no lineal:

Haykin citado por Shafiabady et al. (2016) menciona que SVM posee una sobresaliente capacidad de generalización y que esta es aportada por la implementación del principio de minimización del riesgo estructural (SRM) que consiste en encontrar un hiperplano de separación óptimo que garantiza el error de clasificación más bajo.

En el artículo de Castellanos *et al.* (2017) desarrollan una biblioteca digital en la cual implementan SVM para la clasificación de documentos y demuestran su precisión en comparación con otros clasificadores como Naïve Bayes y K-NN.



**Figura 10.**Precisión de los clasificadores en términos de precisión y tiempo

*Nota.* Recuperado de *Biblioteca digital con técnicas de clasificación automática de documentos* [Imagen], por Castellanos et al., 2017, (https://bit.ly/3qeYwmQ)

En la figura 10 se observa que el mejor clasificador es SVM con un promedio del 80% de precisión en la clasificación de documentos textuales. También ocupa el segundo lugar en términos del tiempo que se demora el modelo en la clasificación.

### **Redes neuronales artificiales (Artificial Neural Network)**

Las redes neuronales artificiales son modelos de aprendizaje automático que tratan de simular el comportamiento de las neuronas biológicas. Para definir lo que es una red neural artificiales primero hay que definir lo que es una neurona. Dreyfus (2005) define a la neurona artificial como una función no lineal parametrizada y acotada. Mientras que una red de neuronas es la composición de las funciones no lineales de dos o más neuronas. La función de la neurona viene dada por:

donde son las variables y son los parámetros de la neurona.

Los parámetros se asignan a las entradas de las neuronas. La salida de la neurona es una combinación no lineal de las entradas , ponderada por los parámetros , que a menudo se denominan pesos. La función f se denomina función de activación. Es aconsejable que la función de activación sea una función sigmoidea.

Una neurona representada gráficamente se muestra en la figura 11.

**Figura 11.**Estructura de una neurona artificial.

*Nota.* Adaptadode *Neural Networks: Methodology and Applications (p. 2)* [Imagen], por Dreyfus, 2005, (https://doi.org/10.1007/3-540-28847-3). Springer-Verlag Berlin Heidelberg

*Precisión de los clasificadores en términos de precisión y tiempo*

Una red neuronal representada gráficamente se encuentra en la figura 12.

**Figura 12.**Estructura de una red neuronal con una capa oculta

*Nota.* Adaptado de *Inteligencia Artificial: Un enfoque moderno* *(p. 847)* [Imagen]*,* por Russell y Norvig, 2004*.* Pearson Educación, S.A.

Russell & Norvig, (2004) mencionan que la idea para el aprendizaje de las redes neuronales es ajustar los pesos de la red para así minimizar alguna medida del error que se produce con el conjunto de entrenamiento. La medida del error clásica es la suma de los errores cuadrados o también conocido como el error cuadrático medio. Uno de los problemas de las redes neuronales artificiales es que cuando existen demasiados parámetros en el modelo, las redes neuronales artificiales son sujeto de sobreajuste (overfitting). El enfoque para evitar el sobreajuste es intentar varias veces y quedarnos con el mejor modelo. Existen redes neuronales de una capa y multicapa. En redes multicapa totalmente conectadas como la de la figura 12, tenemos que preocuparnos por el número de capas ocultas y por su tamaño es decir la cantidad de neuronas de la capa.

Feldman & Sanger, (2009) citados por Godoy, (2017) dicen que en los clasificadores de texto que están basados en redes neuronales artificiales la capa de entrada está formada por unidades que representan los términos existentes en los documentos, mientras que en las capas ocultas y de salida están representadas las categorías de interés. En los problemas de clasificación, el número de neuronas de la capa de salida está dado por la cantidad de categorías en las que se quieren clasificar los datos de entrada.

Un ejemplo lo encontramos en el artículo de García *et al*. (2018) en el cual utilizan una red neuronal artificial y el modelo Naïve Bayes para clasificar tweets (ver tabla 3).

**Tabla 3.**Exactitud y error promedio por ejecución

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Clasificador Bayesiano** | | **Red Neuronal** | |
| **Ejecución** | **Exactitud** | **Error** | **Exactitud** | **Error** |
| 1 | 0.842 | 0.157 | 0.917 | 0.082 |
| 2 | 0.842 | 0.157 | 0.915 | 0.084 |
| 3 | 0.841 | 0.158 | 0.922 | 0.077 |
| 4 | 0.841 | 0.158 | 0.919 | 0.080 |
| 5 | 0.848 | 0.151 | 0.920 | 0.079 |
| 6 | 0.842 | 0.157 | 0.919 | 0.080 |
| 7 | 0.843 | 0.156 | 0.914 | 0.085 |
| 8 | 0.841 | 0.158 | 0.917 | 0.082 |
| 9 | 0.842 | 0.157 | 0.915 | 0.084 |
| 10 | 0.836 | 0.163 | 0.917 | 0.082 |

Nota. Adaptado de Aplicación de una red neuronal artificial para la

clasificación automática de tuits en español, por García et al. 2018. (https://bit.ly/2VlC6C9)

Los resultados indican que las redes neuronales tiene mejor precisión y una menor tasa de error en todas las ejecuciones de la clasificación de los tweets en comparación con el modelo de Naïve Bayes.

La figura 13 representa un diagrama resumen del proceso a seguir para resolver problemas de clasificación de corpus textuales obtenidos de redes sociales.

**Figura 13.**Procesamiento de textos

*Nota.* Principales métodos para resolver problemas de clasificación de textos de redes sociales. [Diagrama de llaves].

Corpus textual

Twitter

Facebook

Instagram

Blogs

Pre - procesamiento

Stop Word Removal

Stemming

Tokenization

Representación

Modelo de espacio vectorial

Bag of words

N-gramas

Ontologías

Modelos de aprendizaje automático supervisado para la clasificación

Naive Bayes Classifier

Nearest Neighbor Classifier (K-NN)

Decision Tree Classifier

Support Vector Machine

Artificial Neuronal Network

Minería de textos

Por otro lado, se tiene la opción de desarrollar una aplicación propia para el análisis de textos. Para ello existen librerías de programación que permiten realizar el preprocesamiento, la representación y el descubrimiento de conocimiento.

Entre otras, las principales herramientas de programación para el análisis de textos son:

* *Natural Language Toolkit (NLTK):* Es un conjunto de herramientas que permiten la escritura de programas en Python. En su sitio web (https://www.nltk.org/) se menciona que NLTK está conformado por un gran conjunto de bibliotecas las que permiten realizar tokenización, derivación, etiquetado, procesamiento de texto para la clasificación, análisis y razonamiento semántico.
* *TextBlob Simplified Text Processing:* Es una biblioteca de Python para el procesamiento de datos textuales. En su sitio web (https://textblob.readthedocs.io/) indican que TextBlob “proporciona una API simple para sumergirse en tareas comunes de Procesamiento del Lenguaje Natural (PNL) como el etiquetado de parte del discurso, extracción de frases nominales, análisis de sentimientos, clasificación, traducción y más.”
* *Scikit-learn:* es una biblioteca de Python para el aprendizaje automático. En su sitio web (https://scikit-learn.org/) describen que Scikit-learn cuenta con algoritmos para el aprendizaje supervisado como; modeles lineales, máquinas de vectores de soporte, k-vecinos más cercanos, naïve bayes, árboles de decisión, modelos de redes neuronales, entre otros.
* *TensorFlow*: En (https://www.tensorflow.org/) destacan que tensorFlow es la principal biblioteca de código abierto para la construcción de modelos de aprendizaje automático. Cuenta con librerías para el procesamiento de textos y está disponible para los lenguajes de programación Python y JavaScript; así como, para dispositivos móviles y de IOT.

Existen muchas más librerías de programación para el procesamiento y análisis de datos textuales como; Gensim, CoreNLP, spaCy, Pattern, VADER, Scikit-learn, PyNLPI.

## **Aplicaciones de la minería de texto**

Las aplicaciones de la minería de texto son muy diversas: resúmenes de documentos, síntesis de información, filtrado de emails, personalización de perfiles web y publicaciones digitales, análisis de redes sociales, detección de comunidades web, extracción de ideas principales, obtención de principales tópicos de discusión de artículos científicos, detección de plagios, organización bibliográfica y de autores, análisis de patentes, sistemas inteligentes de mercado vía web, sistemas de detección automática de comportamiento antisocial, realización de pronósticos, entre otras. Una explicación detallada de algunas de estas aplicaciones se encuentra en Justicia de la Torre, (2017).

En la tabla 5 se presenta un resumen de los diferentes usos de la minería de texto clasificados por área de aplicación.

**Tabla 4.**Aplicaciones de la minería de texto.

|  |  |
| --- | --- |
| **Área de aplicación** | **Uso** |
| Web Semántica | Servicios web de descubrimiento  Aprendizaje de ontologías |
| Redes Sociales | Filtrado de Email  Personalización de perfiles web  Detección de Comunidades Web  Teorías Sociales |
| Opinion Mining, Sentimental Analysis | Clasificación de Opiniones  Hotspots en foros  Predicciones  Detección de Comportamiento Antisocial  Encuestas de Opinión |
| Síntesis, Organización | Summarization  Revisiones sistemáticas  Obtención de Titulares  Extracción de Ideas útiles y nuevas  Discusión de temas principales  Organización de documentos detección de plagios  Búsqueda de contradicciones |
| Minería de Fuentes Abiertas, Tesauros | Identificación de temas  Exploración de tesauros |
| e-Commerce | Toma de decisiones |
| **Área de aplicación** | **Uso** |
| Marketing | Localización de "trozos" de información  Análisis de mercado |
| e-Learning | Herramientas colaborativas |
| Help Desk | Generación de casos modelo  Detección de fallos |

*Nota.* Adaptado de *Nuevas Técnicas de Minería de Textos: Aplicaciones,* por Justicia de la Torre, 2017. (https://digibug.ugr.es/handle/10481/46975). CC BY 3.0

### **Análisis de sentimientos**

El análisis de sentimientos o minería de opinión resulta ser una tarea desafiante pero útil, por ejemplo; las empresas y organizaciones siempre quieren conocer las opiniones que los consumidores tienen acerca de sus productos o servicios, de la misma forma, los clientes también quieren conocer las opiniones de otros usuarios antes de comprar un producto o usar un servicio.

Las empresas podrían destinar personal para que se encargue de leer y analizar todas las opiniones y resulta ser una tarea sencilla ya que los seres humanos poseemos la capacidad de interpretar la sintaxis y semántica de las oraciones. Pero esto se vuelve inviable cuando la cantidad de información es exageradamente grande. Por esta razón surge la necesidad de automatizar este proceso mediante hardware y software que es capaz de realizar cálculos y procesar datos en cuestión de segundos.

No es una tarea sencilla para las computadoras ya que estas no tienen la capacidad de interpretar el lenguaje oral o escrito de los humanos. Pero esto ha ido cambiando con el paso del tiempo y gracias a los avances tecnológicos hoy en día se puede dotar a las computadoras de cierta inteligencia que permite la interpretación del lenguaje.

El análisis de sentimientos o la minería de opinión para Liu & Zhang (2012) es el estudio computacional de las opiniones, valoraciones, actitudes y emociones de las personas hacia entidades, individuos, problemas, eventos, temas y sus atributos.

Una definición muy parecida la dan Medhat *et al*. (2014) que dicen que el análisis de sentimientos o la minería de opinión es el estudio computacional de las opiniones, actitudes y emociones de las personas hacia una entidad. La entidad puede representar individuos, eventos o temas.

Bravo *et al*. (2014) mencionan que el análisis de sentimientos es un problema de clasificación cuyos enfoques actuales se centran principalmente en una dimensión de opinión popular que es muy difícil de clasificar de forma independiente y para ello proponen una taxonomía basada en tres métodos: Polaridad (Polarity), Fuerza (Strength) y Emoción (Emotion).

* *Polaridad (Polarity). –* Mencionan que los métodos orientados a la polaridad tienen como objetivo extraer información de polaridad de un pasaje de texto y normalmente devuelven una variable categórica cuyos valores posibles son: positivos, negativos y neutros.
* *Fuerza (Strength). –* Mencionan que los métodos orientados a la fuerza devuelven puntajes numéricos que indican la intensidad de los sentimientos positivos y negativos expresados en un pasaje de texto.
* *Emoción (Emotion). –* Mencionan que estos métodos se centran en extraer emociones o estados de ánimo de un pasaje de texto. Las categorías en las que se puede clasificar un mensaje utilizando un método orientado a la emoción son: tristeza, alegría, sorpresa, entre otras.

Según Baviera (2017) el análisis de sentimientos aplicado a grandes volúmenes de datos tiene una gran dificultad ya que evaluar el tipo de emoción o la polaridad no siempre tiene la misma interpretación incluso si el proceso es realizado manualmente, los resultados varían según la persona que los analiza. Además menciona que el desarrollo de técnicas de análisis de sentimientos aplicadas al español aún se encuentra en fase de maduración.

## **Herramientas para la minería de texto**

En la actualidad existe una gran variedad de herramientas software que han sido desarrolladas con el fin de analizar grandes cantidades de texto y así obtener información valiosa de los mismos.

* *MonkeyLearn:* en la página oficial (<https://monkeylearn.com/>) dan a conocer su software mediante el eslogan; “Capacite modelos personalizados de aprendizaje automático para obtener temas, sentimientos, intenciones, palabras clave y más” Esta es una herramienta de pago orientada a pequeñas, medianas y grandes empresas. Ofertan planes de 300$ y 999$ por año.
* *IBM Watson:* IBM Watson tiene un gran número de herramientas de pago desarrolladas con inteligencia artificial y aprendizaje automático. Watson Natural Language Understanding (<https://www.ibm.com/cloud/watson-natural-language-understanding>) es una herramienta que ofrece el servicio de procesamiento de lenguaje natural para analizar textos. Watson Natural Language Classifier (<https://www.ibm.com/cloud/watson-natural-language-classifier>) es una herramienta que permite personalizar modelos de aprendizaje automático para analizar y etiquetar datos textuales.
* *Amazon Comprehend:* El gigante tecnológico Amazon también ofrece herramientas para el análisis y procesamiento de corpus textuales. En su sitio web oficial  
  (<https://aws.amazon.com/es/comprehend/>) mencionan que “Amazon Comprehend utiliza el aprendizaje automático para ayudar a descubrir la información y las relaciones en sus datos no estructurados. El servicio identifica el idioma del texto; extrae frases, nombres de lugares, personas, marcas o eventos clave; comprende el grado de positividad o negatividad del texto; lo analiza mediante tokenización y categorías gramaticales; y organiza automáticamente una colección de archivos de texto por tema.” Igual que las herramientas anteriores, esta también es de pago y sus precios varían de acuerdo a la cantidad de texto que se desea procesar.
* *Google Cloud NLP:* Otro gigante de la tecnología como lo es Google también ofrece herramientas que permiten el análisis de textos. En su sitio web oficial (<https://cloud.google.com/natural-language>) se destaca que “Natural Language usa el aprendizaje automático para revelar la estructura y el significado del texto. Puedes extraer información sobre personas, lugares y eventos, y comprender mejor las opiniones de las redes sociales y las conversaciones de los clientes. Natural Language te permite analizar texto y, también, integrarlo a tu almacenamiento de documentos en Cloud Storage.” Los precios varían dependiendo de las funcionalidades que se usen.

Existen otras empresas que han desarrollado herramientas como Aylien, Thematic, MeaningCloud, Lexalytics entre otras. La gran mayoría de las herramientas mencionadas en este apartado son de pago y están orientadas a las PYMES y las grandes empresas. Aunque vale destacar que también ofrecen versiones o planes gratuitos con el objetivo de poder probar sus productos.

Para finalizar esta sección, véase la tabla 4 en la cual se presenta un resumen comparativo de las herramientas de minería de texto ofertadas por las más grandes empresas tecnológicas.

**Tabla 5.**Herramientas para la minería de texto

|  |  |  |  |
| --- | --- | --- | --- |
| **Herramienta** | **Empresa** | **Que ofrece** | **Precio** |
| MonkeyLearn | MonkeyLearn | Agrupación temática.  Análisis de sentimientos.  Análisis de texto.  Clasificación taxonómica.  Consultas booleanas.  Detección del idioma.  Etiquetado.  Filtrado de documentos.  Modelado predictivo.  Presentación gráfica de datos.  Resumen. | Para equipos $299 al mes.  Para negocios $999 al mes. |
| IBM Watson | IBM | Análisis de texto.  Extracción de metadatos de contenido.  Categorización de contenido.  Identificación de conceptos de alto nivel.  Análisis de emociones.  Análisis de sentimientos. | El precio se calcula de acuerdo a las unidades de datos que queramos procesar. Una unidad de datos es un grupo de 10000 caracteres o menos. $0.003 es el costo por unidad de dato. Hay que recalcar que el costo por unidad varía de acuerdo al servicio que necesitemos. |
|  |  |  |  |
| **Herramienta** | **Empresa** | **Que ofrece** | **Precio** |
| Amazon Comprehend | Amazon | Extracción de frases claves.  Análisis de opiniones.  Análisis sintáctico.  Reconocimiento de entidad.  Detección de idioma  Clasificación personalizada.  Modelado de tema.  Compatibilidad con varios idiomas.  Comprensión de datos médicos. | El precio se calcula de acuerdo con las unidades que necesitamos procesar. Una unidad corresponde a 100 caracteres. $0,0001 es el costo por unidad. Hay que recalcar que el costo por unidad varía de acuerdo al servicio que necesitemos. |
| Google Cloud NLP | Google | Análisis sintáctico.  Análisis de entidades.  Extracción de entidades personalizadas.  Análisis de opiniones.  Clasificación de contenido.  Multilingüe.  Información sobre la estructura espacial.  Asistencia para conjuntos de datos grandes. | El precio varía de acuerdo a las unidades que necesitamos procesar y al servicio que necesitemos. Una unidad corresponde a 1000 caracteres o menos. Por ejemplo el precio para el análisis de opiniones en textos de 5001 a 1000000 caracteres cuesta $1.00. |

*Nota.* Comparativa de herramientas disponibles en el mercado para minería de texto.

## **Trabajos relacionados**

En este apartado se presentan algunos trabajos de investigación que utilizan técnicas de aprendizaje automático supervisado para realizar análisis de sentimientos en corpus textuales en español, específicamente corpus textuales de redes sociales.

Para la selección de los trabajos se realizó una búsqueda avanzada en la biblioteca virtual de la Universidad Técnica Particular de Loja. Las cadenas de búsqueda que se utilizaron fueron: análisis AND sentimientos AND redes AND sociales.

Los criterios de inclusión que se utilizaron fueron:

* Estudios realizados desde el año 2017 hasta el presente.
* Estudios cuyo idioma fuera el español.
* Estudios que pertenezcan a la categoría Academic Journals.
* Estudios más citados por otros investigadores.

### **TR01: Análisis de sentimientos de tweets en español utilizando SVM y CNN. (Rosa et al., 2017)**

En este trabajo se presenta un clasificador híbrido de sentimientos de polaridad para tweets, el cual fue puesto a prueba en la competición SEPLN (TASS). El TASS se celebra desde el 2012 y es una propuesta de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). Para el desarrollo de su trabajo contaron con un aproximado de 8000 tweets y lo particionaron en: 85% para el entrenamiento y 15% para el desarrollo, (véase tabla 6). Luego de un preprocesamiento de los tweets realizaron la clasificación utilizando tres enfoques: Primero utilizaron un clasificador basado en máquinas de vectores de soporte (SVM), luego un clasificador basado en redes neuronales convolucionales (CNN) y finalmente realizaron una clasificación utilizando una combinación de estos dos algoritmos. Los resultados de estos experimentos se presentan en la tabla 7.

**Tabla 6.**Tamaño de los corpus utilizados

|  |  |  |
| --- | --- | --- |
|  | **Entrenamiento** | **Desarrollo** |
| P | 2782 (39 %) | 576 (36 %) |
| N | 2295 (32 %) | 524 (33 %) |
| NEU | 721 (10 %) | 151 (10 %) |
| NONE | 1346 (19 %) | 338 (21 %) |
| Total | 7114 | 1589 |

*Nota.* Adaptado de RETUYT in TASS 2017: Sentiment

Analysis for Spanish Tweets using SVM and CNN, por

Rosa et al., 2017, (http://arxiv.org/abs/1710.06393)

**Tabla 7.**Resultados sobre el corpus de validación

|  |  |  |
| --- | --- | --- |
| **Clasificación** | **M-F1** | **Acierto** |
| svm | 49.9 | 61.7 |
| cnn4 | 50.2 | 64.1 |
| svm\_cnn | 50.9 | 64.7 |

*Nota.* Adaptado de RETUYT in TASS 2017: Sentiment

Analysis for Spanish Tweets using SVM and CNN, por

Rosa et al., 2017, (http://arxiv.org/abs/1710.06393)

Como se observa en los resultados de la tabla 7, el clasificador híbrido es el que presenta una mejor taza de acierto por encima del clasificador basado en máquinas de vectores de soporte (SVM) y el clasificador basado en redes neuronales convolucionales (CNN).

### **TR02: Un Análisis de Sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de #BlackFriday. (Saura et al., 2018)**

Ellos realizan un trabajo el cual consiste en medir los sentimientos de los consumidores que utilizan la red social Twitter en relación con el evento #BlackFriday que se celebra cada año. Para ello trabajaron con un corpus de aproximadamente 2200 tweets y utilizaron específicamente la librería MonkeyLearn que contiene algoritmos de para la clasificación de sentimientos. Después del análisis obtuvieron que el 60.2% de los tweets son categorizados como neutros, el 32.1% como tweets positivos y el 7.7% como negativos. En la figura 14 se muestra el proceso del desarrollo metodológico que implementaron.

**Figura 14.**Desarrollo del proceso metodológico

*Nota.* Adaptado de *Un Análisis de Sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de #BlackFriday*, por Saura et al., 2018 (https://bit.ly/37lu1Db). revistaESPACIOS.com

Tweets

n = 2204

0.521

0.563

0.654

Negative = 170

Neutral = 1327

Positive = 707

### **TR03: Maternidad en Perú a través del uso del Sentiment Analysis en Facebook. (Seperak et al., 2019)**

En el trabajo presentado por Seperak *et al*. (2019) se propone el uso del aprendizaje automático supervisado para conocer la percepción actual que tienen los peruanos sobre la maternidad y la mujer. Contaron con un corpus textual de aproximadamente 28.000 comentarios que hacían relación a la temática extraído de grupos abiertos al público general.

El conjunto de entrenamiento lo conformaron con alrededor de 500 comentarios positivos, 500 negativos y 500 neutros clasificados manualmente. Luego de la fase de preprocesamiento de los textos, utilizaron el algoritmo Conditional Random Fields para la clasificación. Los resultados se muestran en la tabla 8.

**Tabla 8**.  
Resultados de la clasificación

|  |  |  |
| --- | --- | --- |
|  | **Cantidad** | **Porcentaje** |
| Comentarios positivos | 20.858 | 72.16 % |
| Comentarios negativos | 8.049 | 27.84 % |
| Total | 28.907 | 100.00 % |

*Nota.* Adaptado de *Maternidad en Perú a través del uso del*

*Sentiment Analysis en Facebook*, porSeperak et al., 2019,

(<https://doi.org/10.4185/RLCS-2019-1370>).

En la tabla 8 podemos observar que la mayoría de los comentarios acerca de la mujer y la maternidad en el Perú son positivos con un 70.16%, mientras que los comentarios negativos representan un 27.84% del total.

### **TR04: Análisis de sentimiento en Instagram: polaridad y subjetividad de cuentas infantiles** **(Arantxa & Aguaded, 2020)**

Finalmente presentamos un trabajo muy actual de Arantxa y Aguaded en el cual se realiza un análisis de sentimientos en cuentas infantiles administradas por padres de la red social Instagram. Realizan una clasificación utilizando los métodos de polaridad y fuerza, es decir, aparte de determinar si un comentario es positivo o negativo a este se le asociado un peso, en donde un comentario muy negativo esta dado por (-1), un comentario muy positivo este dado por (+1) y (0) representa un comentario neutro. El corpus textual consta de 772 comentarios realizados en 100 fotos de 8 cuentas infantiles, 4 cuentas en el lenguaje inglés y 4 en el lenguaje español. Los resultados que aquí presentamos son de las cuentas en español. Después de la fase de preprocesamiento de los textos utilizan la librería TextBlob para realizar la clasificación y los resultados se muestran en la tabla 9.

**Tabla 9.**Polaridad en cuentas infantiles en español

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Cuenta** | **Número de**  **entradas**  **positivas** | **Número de**  **entradas**  **negativas** | **Número de**  **entradas**  **neuras** | **Media de**  **polaridad**  **positiva** | **Media de**  **polaridad**  **negativa** | **Valoración**  **media de**  **polaridad**  **de la cuenta** |
| 1 | 46 | 4 | 22 | 0.43 | -0.26 | 0.26 |
| 2 | 94 | 6 | 0 | 0.39 | -0.17 | 0.36 |
| 3 | 80 | 7 | 13 | 0.37 | -0.17 | 0.29 |
| 4 | 62 | 18 | 20 | 0.36 | -0.18 | 0.19 |
| Total | 282 | 35 | 55 | 0.39 | -0.19 | 0.27 |

Nota. Adaptado de Análisis de sentimiento en Instagram: polaridad y subjetividad de cuentas infantiles, por Arantxa y Aguaded, 2020, (<https://doi.org/10.1387/zer.21454>)

## **Comparación de trabajos relacionados**

Para finalizar este capítulo véase la tabla 10, en la cual se presenta un resumen comparativo de los cuatro estudios analizados.

**Tabla 10.**Trabajos relacionados

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Objetivos** | **Herramientas para la obtención y procesamiento del texto** | **Algoritmos para la clasificación** | **Resultados** | **Inconvenientes** |
| **TR01** | Construir un clasificador de sentimientos de tweets. | Twitter API.  Scikit-learn. | Support Vector Machines (SVM).  Convolutional Neural Networks (CNN).  Híbrido entre SVM y CNN. | El clasificador con mejor desempeño es el que utiliza una combinación de SVM y CNN. | Debido a que la clase (tweets neutros) tiene pocos ejemplos en el corpus y a la similitud que tienen con los tweets negativos y positivos, el sistema es poco efectivo en la detección de estos. |
|  | **Objetivos** | **Herramientas para la obtención y procesamiento del texto** | **Algoritmos para la clasificación** | **Resultados** | **Inconvenientes** |
| **TR02** | Identificar el sentimiento sobre las ofertas de #BlackFriday. | Twitter API.  MonkeyLearn. | No mencionan | Tweets positivos: 32,1%.  Tweets negativos: 7,7%.  Tweets neutros: 60,2%. | El tamaño de la muestra y el número limitado de empresas que utilizan el hashtag #BlackFriday. |
| **TR03** | Conocer cuál es la percepción y nuevos conceptos en la actualidad acerca de la maternidad en Perú. | Facebook API. | Conditional Random Fields. | Comentarios con sentimientos positivos hacia la maternidad y a la imagen de madre: 72,16%.  Comentarios negativos: 27,84%. | El proceso de extraer el corpus lo tuvieron que realizar dos veces ya que la primera vez se encontró que la mayoría de los resultados no correspondían al concepto “madre”. |
|  | **Objetivos** | **Herramientas para la obtención y procesamiento del texto** | **Algoritmos para la clasificación** | **Resultados** | **Inconvenientes** |
| **TR04** | Analizar la polaridad de comentarios publicados en fotos de cuentas infantiles gestionadas por padres en la red social Instagram. | TextBlob. | No mencionan | Entradas positivas: 75,81%.  Entradas negativas: 14,66%.  Entradas neutras: 9,4% | Corpus textual limitado. |

*Nota.* Tabla comparativa de los trabajos relacionados.

**Capítulo tres**

**Metodología de desarrollo**

En el presente capítulo se define la metodología con la cual se van a resolver los objetivos planteados en esta investigación. En la sección 3.1. se plantean las fases de la metodología......

## **Metodologías para la minería de texto**

No existe un estándar metodológico para resolver problemas de minería de textos. Sin embargo, investigadores como Verma, Ranjan & Mishra (2015) citados en Barrera, (2016), proponen una metodología general que contempla dos fases; la fase de refinación del texto en donde se transforman los documentos y se los representa en estructuras de datos; y otra fase de destilación del conocimiento, en donde a partir de las estructuras de datos generadas en la primera fase, se identifican los patrones que servirán de base de conocimiento.

Por otro lado, M.Sukanya (2012), citado en el mismo estudio de Barrera, (2016), propone una metodología más específica en la cual el proceso para la resolución de problemas de minería de texto consta de varias fases. En la primera fase se define el propósito de la minería de texto. En la segunda fase se realiza el proceso de recuperación de la información. En la tercera fase se realiza el preprocesamiento de los datos. En la cuarta fase se procede con la extracción de relaciones o patrones. Finalmente, como último paso se presentan los resultados para su posterior interpretación.

Para esta investigación se selecciona la metodología propuesta por M. Sukanya, es decir, la metodología más específica. Se hizo esta elección principalmente porque al dividir el problema en un conjunto de pasos más grande, se tiene un mejor control del proceso de desarrollo lo que facilita la obtención de resultados. En la figura 15 se muestra un diagrama de la metodología adaptada al contexto de esta investigación.

***Figura 15.***Metodología de desarrollo

*Nota.* Esta figura resume el proceso metodológico que se va a seguir para desarrollar la aplicación del presente trabajo de titulación [Diagrama de flujo].

Propósito de la minería de texto

Análisis de sentimientos

Recuperación de la información



Procesamiento del texto

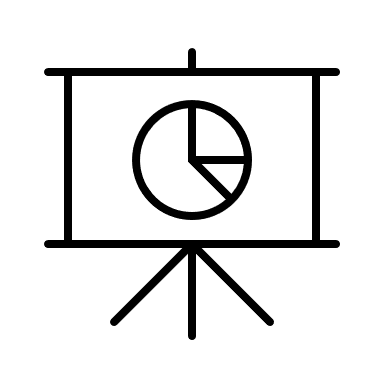
Stop Words Removal

Tokenization

Aplicación del método de minería de texto

Algoritmo de clasificación

Análisis de resultados e interpretación



## **Herramientas**

Para desarrollar la parte experimental de la presente investigación se trabaja con el lenguaje de programación Python. Se hizo esta elección porque Python es un lenguaje intuitivo, dinámico y fácil de interpretar con el cual se pueden resolver problemas más rápido que en otros lenguajes. Además cuenta con un sinnúmero de librerías y grandes comunidades que dan soporte. En lo que ha ciencia de datos y machine learning se refiere, Python es el lenguaje preferido por los científicos de datos.

Con el objetivo de organizar de mejor manera el código y que sea más entendible, se trabaja con la herramienta JupyterLab. JupyterLab es un entorno de desarrollo interactivo muy usado en flujos de trabajo en ciencias de datos y machine learning Kluyver et al., (2016).

Por otro lado, Visual Studio Code es el editor elegido para escribir el código fuente. Debido a la gran cantidad de IDEs y editores de código que existen en el mercado todos ellos con muy buenas características, la elección de VSC es netamente personal. Cabe recalcar que VSC cuenta con soporte para trabajar con Jupyter Notebooks de manera nativa.

Las librerías con las que se va a desarrollar el almacenamiento, pre – procesamiento del texto y la aplicación de los algoritmos son:

Almacenamiento:

* Pandas

Preprocesamiento:

* unidecode
* string
* re
* NLTK
* SPACY

Implementación de algoritmos:

* Scikit-Learn

Los algoritmos de clasificación que se van a testear son:

* Support Vector Machine
* Nearest Neighbors
* Decision Trees
* Neuronal Network
* Random Forest Classifier

Los algoritmos mencionados anteriormente fueron escogidos en base a los trabajos relacionados estudiados, en los cuales se encuentra un alto grado de usabilidad de estos algoritmos. En el apartado 2.8. se describen los trabajos relacionados.

## **Fase 1: Propósito de la minería de texto**

El propósito de la minería de texto es desarrollar un clasificador de sentimientos automatizado, el cual categorice comentarios de grupos de estudio de Facebook, como positivos, negativos o neutros (ver figura 16).

**Figura 16.**Clasificador de comentarios

*Nota.* Esta figura muestra a manera abstracta como va a funcionar el clasificador [Diagrama de flujo].



Corpus textual

Algoritmo de clasificación

Positivo

Neutro

Negativo

Origen de datos

Clasificación

## **Fase 2: Recuperación de la información**

Esta etapa empieza por seleccionar la fuente de información. En este caso, se está trabajando con la red social Facebook. En esta plataforma se crearon grupos de 4 a 6 estudiantes cada uno. Los estudiantes pertenecen a la materia de Inteligencia Artificial de la Modalidad Abierta y a Distancia de la Universidad Técnica Particular de Loja. El objetivo de estos grupos es que en ellos se debatan, concuerden y organicen el desarrollo de los trabajos autónomos. Se identificaron un total de 45 grupos de estudio. En los 45 grupos de estudio se identificaron un total de 265 post. Entiéndase por post a cualquier texto que es publicado dentro del grupo. Los estudiantes interactúan respondiendo al post, a estas interacciones se las denomina comentarios. En todos los grupos de estudio existe un total de 1980 comentarios cuya distribución se muestra en la tabla 11.

**Tabla 11.**Distribución de los comentarios

|  |  |  |
| --- | --- | --- |
| **Número de grupo** | **Número de post** | **Número de comentarios** |
| 1 | 4 | 5 |
| 2 | 17 | 51 |
| 3 | 4 | 6 |
| 4 | 29 | 63 |
| 5 | 9 | 28 |
| 6 | 12 | 17 |
| 7 | 27 | 276 |
| 8 | 26 | 105 |
| **Número de grupo** | **Número de post** | **Número de comentarios** |
| 9 | 5 | 22 |
| 10 | 26 | 94 |
| 11 | 7 | 10 |
| 12 | 3 | 9 |
| 13 | 13 | 29 |
| 14 | 16 | 97 |
| 15 | 5 | 10 |
| 16 | 13 | 43 |
| 17 | 9 | 30 |
| 18 | 4 | 3 |
| 19 | 6 | 14 |
| 20 | 6 | 16 |
| 21 | 8 | 201 |
| 22 | 10 | 48 |
| 23 | 14 | 35 |
| 24 | 23 | 24 |
| 25 | 4 | 3 |
| 26 | 4 | 18 |
| 27 | 4 | 15  5 |
| 28 | 15 | 33 |
| 29 | 17 | 38 |
| 30 | 3 | 4 |
| 31 | 17 | 60 |
| 32 | 18 | 34 |
| 33 | 6 | 9 |
| 34 | 2 | 4 |
| 35 | 3 | 9 |
| 36 | 3 | 1 |
| 37 | 12 | 23 |
| 38 | 11 | 15 |
| 39 | 3 | 2 |
| 40 | 9 | 15 |
| 41 | 2 | 20 |
| 42 | 3 | 7 |
| 43 | 9 | 25 |
| 44 | 8 | 7 |
| 45 | 10 | 17 |

*Nota.* Distribución de los comentarios en los diferentes grupos de estudio

en Facebook.

Para obtener la información se utiliza las herramientas para desarrollo que brinda la plataforma Facebook. Se crea una aplicación a la cual se la denomina *“GroupPost”.* Esto con el objetivo de acceder a la API Graph de Facebook Developers, ya que sin la creación de una aplicación esto no es posible. Una vez que se tiene acceso a la API Graph, se accede a la información de los grupos por medio de su ID. El ID de los grupos se lo encuentra en la URL de estos.

La API Graph de Facebook facilita en gran medida la descarga de la información ya que simplemente se necesita el ID del grupo y llenar unos cuantos parámetros para obtener los datos en distintos formatos. En este caso en particular, los parámetros utilizados dentro de la API Graph, específicamente en el método GET fueron, *feed* para obtener todos los posts y *comments* para obtener todos los comentarios que tenía cada post. El formato elegido para la descarga de los datos fue JSON. Un ejemplo de cómo se muestran los datos una vez elegidos los parámetros y el formato de salida se muestra en la figura 17.



**Figura 17.**

*Obtención de los Datos*

*Nota.* Comentarios en formato JSON.

## **Fase 3: Pre - procesamiento de texto**

Con los datos en formato JSON descargados, se procede con la etapa de pre - procesamiento. Primeramente, se construye un módulo para leer estos archivos y almacenar su contenido en un solo archivo de extensión CSV. Una vez que se tiene el corpus textual integrado se procede con la construcción de los módulos para el pre – procesamiento.

Estos módulos utilizan el kit de herramientas de lenguaje natural NLTK y funciones integradas de Python para procesar el texto. A continuación se detallan los pasos a seguir en el pre – procesamiento:

1. Se utiliza una función que viene integrada en el paquete de Python llamada *lower(),* la cual permite convertir el texto a minúsculas. El resultado del texto una vez que es procesado por esta función se muestra en la tabla 12.

**Tabla 12.**Texto en minúsculas

|  |  |
| --- | --- |
| **Texto** | **Texto en minúsculas** |
| Por favor pongamos en la portada algo que tenga relación con la asignatura. | por favor pongamos en la portada algo que tenga relación con la asignatura. |
| ¡Rodolfo, muy buenos días! Gracias por compartirnos la tarea... esperemos a ver que dicen sus colegas de equipo. | ¡rodolfo, muy buenos días! gracias por compartirnos la tarea... esperemos a ver que dicen sus colegas de equipo. |
| Bueno compañeros nos podremos comunicar en la noche el 27 digan que si ya que en fechas anteriores a mí se me dificulta no se a ustedes | bueno compañeros nos podremos comunicar en la noche el 27 digan que si ya que en fechas anteriores a mí se me dificulta no se a ustedes |

*Nota.* Texto que ha sido procesado con la función lower()

1. Se realiza la eliminación de acentos con otra función que viene integrada en el paquete de Python que se denomina *unidecode()*. Este método recibe como entrada el texto en minúsculas. El resultado del texto una vez que fue procesado por esta función se muestra en la tabla 13.

**Tabla 13.**Texto sin acentos

|  |  |
| --- | --- |
| **Texto** | **Texto sin acentos** |
| por favor pongamos en la portada algo que tenga relación con la asignatura. | por favor pongamos en la portada algo que tenga relación con la asignatura. |
| ¡rodolfo, muy buenos días! gracias por compartirnos la tarea... esperemos a ver que dicen sus colegas de equipo. | ¡rodolfo, muy buenos dias! gracias por compartirnos la tarea... esperemos a ver que dicen sus colegas de equipo. |
| bueno compañeros nos podremos comunicar en la noche el 27 digan que si ya que en fechas anteriores a mí se me dificulta no se a ustedes | bueno compañeros nos podremos comunicar en la noche el 27 digan que si ya que en fechas anteriores a mí se me dificulta no se a ustedes |

*Nota.* En esta tabla se observa el antes y el después del texto procesado con la función unidecode()

1. Se realiza la eliminación de signos de puntuación, se lo hace por medio de expresiones regulares. La biblioteca *re* que también está incluida en el paquete de Python permite la creación y aplicación de expresiones regulares. El resultado del texto una vez que fue aplicada la expresión regular se muestra en la tabla 14.

**Tabla 14.  
Texto sin signos de puntuación**

|  |  |
| --- | --- |
| **Texto** | **Texto sin signos** |
| por favor pongamos en la portada algo que tenga relación con la asignatura. | por favor pongamos en la portada algo que tenga relación con la asignatura |
| ¡rodolfo, muy buenos dias! gracias por compartirnos la tarea... esperemos a ver que dicen sus colegas de equipo. | rodolfo muy buenos dias gracias por compartirnos la tarea esperemos a ver que dicen sus colegas de equipo |
| **Texto** | **Texto sin signos** |
| bueno compañeros nos podremos comunicar en la noche el 27 digan que si ya que en fechas anteriores a mí se me dificulta no se a ustedes | bueno companeros nos podremos comunicar en la noche el 27 digan que si ya que en fechas anteriores a mi se me dificulta no se a ustedes |

*Nota.* En esta tabla se observa el antes y el después del texto procesado con las expresiones regulares.

1. Se realiza la lematización del texto. para ello se trabaja con la biblioteca spaCy que cuenta con métodos avanzados para el procesamiento de lenguaje natural. Específicamente se utiliza el método *spacy*.*load()* el cual requiere como parámetro el idioma del texto que se va a procesar en este caso *es\_core\_news\_sm.*

**Tabla 15.**Texto lematizado.

|  |  |
| --- | --- |
| **Texto** | **Texto Lematizado** |
| por favor pongamos en la portada algo que tenga relación con la asignatura | por favor poner en la portada algo que tenga relación con la asignatura |
| rodolfo muy buenos dias gracias por compartirnos la tarea esperemos a ver que dicen sus colegas de equipo | rodolfo muy buenos dias gracias por compartir la tarea esperar a ver que dicen sus colegas de equipo |
| bueno compañeros nos podremos comunicar en la noche el 27 digan que si ya que en fechas anteriores a mi se me dificulta no se a ustedes | bueno companeros nos poder comunicar en la noche el 27 digan que si ya que en fechas anteriores a mi se me dificulta no se a ustedes |

*Nota.* En esta tabla se muestra el texto antes y después de ser lematizado.

1. Se eliminan las stopwords. El kit de herramientas NLTK proporciona un corpus textual de palabras vacías a través del método stopwords.words(). Como parámetro de este método se especifica el idioma, en este caso, spanish. En la tabla se muestran los ejemplos anteriores son las palabras vacias.

**Tabla 16.**Texto sin stopwords

|  |  |
| --- | --- |
| **Texto** | **Texto sin stopwords** |
| por favor pongamos en la portada algo que tenga relación con la asignatura | favor poner portada relacion asignatura |
| rodolfo muy buenos días! gracias por compartirnos la tarea esperemos a ver que dicen sus colegas de equipo | rodolfo buenos dias gracias compartir tarea esperar ver dicen colegas equipo |
| bueno compañeros nos podremos comunicar en la noche el 27 digan que si ya que en fechas anteriores a mi se me dificulta no se a ustedes | bueno companeros poder comunicar noche 27 digan si fechas anteriores dificulta ustedes |

*Nota.* En esta tabla se muestra el texto antes y después de eliminar las stopwords.

1. Se segmenta el texto (tokenizing) se lo hace con la función *tokenize()* del kit de herramientas de lenguaje natural NLTK.

**Tabla 17.**Texto segmentado

|  |  |
| --- | --- |
| **Texto** | **Texto segmentado** |
| favor poner portada relacion asignatura | [favor] [poner] [portada] [relación] [asignatura] |
| rodolfo buenos dias gracias compartir tarea esperar ver dicen colegas equipo | [rodolfo] [buenos] [dias] [gracias] [compartir] [tarea] [esperar] [ver] [dicen] [colegas] [equipo] |
| bueno companeros poder comunicar noche 27 digan si fechas anteriores dificulta ustedes | [bueno] [companeros] [poder] [comunicar] [noche] [27] [digan] [si] [fechas] [anteriores] [dificulta] [ustedes] |

*Nota.* En esta tabla se muestra el texto antes y después de ser segmentado.

1. El siguiente paso una vez que ya se tiene el texto pre – procesado, es convertirlo a formato numérico. A esto se lo conoce como representación o forma intermedia tal como se lo explicó en el apartado 2.4.2. De todas las posibles formas intermedias se elige TF-IDF (Frecuencia de Término - Frecuencia de Documento Inverso). Esta elección se la hizo porque el enfoque de TF-IDF considera que las palabras que aparecen más en un documento en específico y menos en todos los documentos contribuyen de mejor manera al proceso de clasificación. Los documentos individuales en este caso son cada comentario que hay en los posts.

La biblioteca Scikit-Learn cuenta con la clase TfidfVectorizer que permite la conversión de texto a vectores de características TF-IDF.

En el Anexo 1 se encuentra el código utilizado para la esta fase de preprocesamiento.

## **Fase 4: Aplicación del método de minería de texto**

En esta fase se explica el proceso de experimentación con los distintos algoritmos de clasificación. Para esta etapa previamente se realiza la clasificación manual del conjunto de datos, es decir, a cada comentario se lo clasifica como positivo, negativo o neutro. Esta clasificación manual representa un verdadero reto dentro de los problemas de análisis de sentimientos con algoritmos supervisados ya que la clasificación en si va a depender de la persona que lo realice. Con el fin de que el algoritmo sea lo más preciso posible, esta clasificación la realizaron dos personas y luego mediante un consenso se resolvieron las ambigüedades que existían en algunos comentarios.

Para llevar a cabo la experimentación se tiene como entrada el conjunto de datos representado en su forma intermedia. El conjunto de datos almacenado en un DataFrame de la librería Pandas, contiene dos columnas. En la primera columna se encuentra el texto y en la segunda columna la etiqueta. Estos datos son los que se necesita para entrenar un algoritmo de aprendizaje supervisado tal como se lo explico en el apartado 2.5.

**Capítulo 4**

**Análisis de Resultados e Interpretación**

**REFERENCIAS**

Aggarwal, C. (2015). *Data Mining*. Springer, Cham. https://doi.org/10.1007/978-3-319-14142-8

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. http://arxiv.org/abs/1707.02919

Arantxa, V., & Aguaded, I. (2020). Análisis de sentimiento en Instagram: polaridad y subjetividad de cuentas infantiles. *ZER - Revista de Estudios de Comunicación*, *25*(48), 213–229. https://doi.org/10.1387/zer.21454

Armstrong, M. (2019). *How Many Websites Are There? | Statista*. https://www.statista.com/chart/19058/how-many-websites-are-there/

Banchs, R. E. (2013). *Text Mining with MATLAB®* (1.a ed.). Springer-Verlag New York. https://doi.org/10.1007/978-1-4614-4151-9

Barrera, M. C. (2016). Minería de texto en la clasificación de material bibliográfico. *Biblios*, *64*, 33–43. https://doi.org/doi.org/10.5195/biblios.2016.309

Baviera, T. (2017). Técnicas para el Análisis de Sentimiento en Twitter: Aprendizaje Automático Supervisado y SentiStrength. *Dígitos: Revista de Comunicación Digital*, *1*(3), 33–50. https://doi.org/10.7203/rd.v1i3.74

Bramer, M. (2016). *Principles of Data Mining* (3.a ed.). https://doi.org/10.1007/978-1-4471-7307-6

Bravo, F., Mendoza, M., & Poblete, B. (2014). Meta-level sentiment models for big social data analysis. *Knowledge-Based Systems*, *69*(1), 86–99. https://doi.org/doi.org/10.1016/j.knosys.2014.05.016

Castellanos Domínguez, M. I., Pérez Perdomo, E., Rivas Méndez, A., & Góngora Zaldívar, V. J. (2017). Biblioteca digital con técnicas de clasificación automática de documentos. *VIII Conferencia Científica Internacional de la Universidad de Holguín, Cuba*, *April*. https://www.researchgate.net/publication/316860869\_Biblioteca\_digital\_con\_tecnicas\_de\_clasificacion\_automatica\_de\_documentos

Cifuentes, F. (2016). *Clasificación automática de Tweets utilizando K-NN y K-Means como algoritmos de clasificación automática, aplicando TF-IDF y TF-RFL para las ponderaciones* [Pontificia Universidad Católica de Valparaíso]. http://opac.pucv.cl/pucv\_txt/Txt-8500/UCD8528\_01.pdf

Dreyfus, G. (2005). *Neural Networks: Methodology and Applications* (1.a ed.). Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/3-540-28847-3

García, A., Ríos, A., Tello, E., Barrón, J., & Díaz, A. (2018). *Aplicación de una red neuronal artificial para la clasificación automática de tweets en español*. *40*(130). http://www.itcelaya.edu.mx/ojs/index.php/pistas/article/view/1737

Godoy, A. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. *Investigacion Bibliotecologica*, 24. https://doi.org/10.22201/iibi.0187358xp.2017.71.57812

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3.a ed.). https://doi.org/10.1016/C2009-0-61819-5

Hu, X., & Liu, H. (2012). *Mining Text Data* (C. C. Aggarwal & C. Zhai (eds.); 1.a ed.). Springer-Verlag New York. https://doi.org/10.1007/978-1-4614-3223-4

Internet-live-stats. (s/f). *Internet Live Stats - Internet Usage & Social Media Statistics*. https://www.internetlivestats.com/

Justicia De La Torre, C., Martín-Bautista, M. J., Síanchez, D., & Vila, M. A. (2005). *Text mining: intermediate forms on knowledge representation*. 7. https://www.researchgate.net/publication/221399082\_Text\_mining\_intermediate\_forms\_on\_knowledge\_representation

Justicia de la Torre, M. del C. (2017). *Nuevas Tecnicas De Minería De Textos: Aplicaciones* [Universidad de Granada]. https://digibug.ugr.es/handle/10481/46975

Kluyver, T., Benjamin Ragan-Kelley, Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). *Project Jupyter | Home*. Jupyter Notebooks -- a publishing format for reproducible computational workflows. https://jupyter.org/

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Moldagulova, A., & Sulaiman, R. B. (2017). Using KNN algorithm for classification of textual documents. *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, 665–671. https://doi.org/10.1109/ICITECH.2017.8079924

Mourya, S. K., & Gupta, S. (2012). *Data Mining and Data Warehousing*. https://search.proquest.com

Pitigala, S., Li, C., & Seo, S. (2011). A comparative study of text classification approaches for personalized retrieval in PubMed. *2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2011*, 3. https://doi.org/10.1109/BIBMW.2011.6112503

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*, *5*(2), 221–232. https://doi.org/10.22364/bjmc.2017.5.2.05

Reyes, G. C., González, Y. G., & Farias, G. L. (2014). Técnica de clasificación bayesiana para identificar posible plagio en información textual. *Revista Cubana de Ciencias Informáticas*, *8*(4), 130–144. http://scielo.sld.cu/scielo.php?script=sci\_arttext&pid=S2227-18992014000400008&lng=es&tlng=es

Rosa, A., Chiruzzo, L., Etcheverry, M., & Castro, S. (2017). *RETUYT in TASS 2017: Sentiment Analysis for Spanish Tweets using SVM and CNN*. *October*, 77–83. http://arxiv.org/abs/1710.06393

Russell, S. J., & Norvig, P. (2004). *Inteligencia Artificial: Un Enfoque Moderno* (2.a ed.). PEARSON EDUCACIÓN, S.A.,.

Saura, J., Reyes, A., & Palos, P. (2018). Un Análisis de Sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de #BlackFriday. *Espacios*, *39*, 16. http://www.revistaespacios.com/a18v39n42/18394216.html

Seperak, R., Cerellino, L., Ochoa, J., Torres-Valer, A., & Dianderes, C. (2019). Maternidad en Perú a través del uso del Sentimient Analysis en Facebook. *Revista Latina de Comunicacion Social*, *74*, 1031–1055. https://doi.org/10.4185/RLCS-2019-1370

Shafiabady, N., Lee, L. H., Rajkumar, R., Kallimani, V. P., Akram, N. A., & Isa, D. (2016). Using unsupervised clustering approach to train the Support Vector Machine for text classification. *Neurocomputing*, *211*, 4–10. https://doi.org/10.1016/j.neucom.2015.10.137

Taeho, J. (2019). *Text Mining* (1.a ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-91815-0

Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of Predictive Text Mining* (1.a ed.). Springer-Verlag London. https://doi.org/10.1007/978-1-84996-226-1

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2005). *Text Mining* (1.a ed.). Springer-Verlag New York. https://doi.org/10.1007/978-0-387-34555-0

Yu, W., & Song, X. (2010). Research on text categorization based on machine learning. *ICAMS 2010 - Proceedings of 2010 IEEE International Conference on Advanced Management Science*, 3. https://doi.org/10.1109/ICAMS.2010.5552917

**Apéndice**

**Apéndice 1:** Código para el pre – procesamiento del texto.

1. **Convertir a minúsculas.**

datos['minusculas'] = datos['MENSAJES'].str.lower()

1. **Eliminar acentos.**

datos['acentos'] = datos['minusculas'].apply(unidecode)

1. **Eliminar signos de puntuación**

reg\_signos = re.compile('[%s]' % re.escape(string.punctuation))

datos['puntuacion'] = datos['acentos'].replace(reg\_signos, '')

1. **Lematizar texto**

nlp = spacy.load('es\_core\_news\_sm')

datos['lemas'] = datos['acentos'].apply(nlp)

1. **Eliminar stopwords**

stop\_words = set(stopwords.words('spanish'))

datos['stopwords'] = datos['tokens'].apply(lambda msg: [item for item in msg if item not in stop\_words])

1. **Segmentar el texto**

toktok = ToktokTokenizer()

tokenizer = nltk.RegexpTokenizer(r'\w+')

datos['tokens'] = datos['lemas'].apply(toktok.tokenize)

1. **Vectorizar el texto**

vectorizer = TfidfVectorizer(max\_features=2500, min\_df=7, max\_df=0.8)

processed\_features\_vector = vectorizer.fit\_transform(processed\_features).toarray()