

UNIVERSIDAD AUTÓNOMA DE ZACATECAS
"Francisco García Salinas"



**MINERÍA DE OPINIÓN: UN ANÁLISIS EN TIEMPO
REAL DE TWEETS PARA ZACATECAS**

Tesis para obtener el grado de:
Maestro en Ciencias del Procesamiento de la Información

Presenta

I.C.E. Luis Carlos Reveles Gómez

Director:

Dr. Huizilopoztli Luna García

Co-Directores:

Dr. José María Celaya Padilla

Dr. César Alberto Collazos Ordóñez

Asesores:

Dr. Pedro Daniel Alaniz Lumbreras

Dr. Julián González Trinidad

Zacatecas, Zac., 26 de abril de 2021.



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL



Zacatecas, Zac., 31 de mayo de 2021

C. Luis Carlos Reveles Gómez
Egresado de la MCPI
PRESENTE

At'n: Dr. Huizilopoztli Luna García
Responsable de la MCPI

Nos es grato comunicarle que después de haber sometido a revisión académica la propuesta de Tesis titulada “Minería de Opinión: Un Análisis en Tiempo Real de Tweets para Zacatecas”, presentada por el estudiante C. Luis Carlos Reveles Gómez y habiendo efectuado todas las correcciones indicadas por este Comité Tutorial, se **AUTORIZA** el documento de tesis para su impresión.

Sin más por el momento reciban un cordial saludo.

COMITÉ TUTORIAL

PROCESAMIENTO Y ANÁLISIS DE DATOS

Dr. Huizilopoztli Luna García

Dr. César Alberto Collazos
Ordóñez

Firmado digitalmente por Cesar A. Collazos
Fecha: 2021.05.31 21:29:53 -05'00'

Pedro Daniel
Alaniz
Lumbreras

Dr. Pedro Daniel Alaniz
Lumbreras

Firmado digitalmente por Pedro Daniel Alaniz Lumbreras
Fecha: 2021.06.01 08:53:43 -05'00'

Dr. José María Celaya Padilla

Dr. Julián González Trinidad

c.c.p. *Interesado*

c.c.p. *Responsable de la Maestría en Ciencias del Procesamiento de la Información*



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL

**COORDINACIÓN DE
INVESTIGACIÓN Y POSGRADO**

Carta de similitud núm. 051/IyP
Zacatecas, Zacatecas 27/mayo/2021

Dr. Huizilopoztli Luna García
Responsable de la MCPI – UAZ
Presente

Estimado Dr. Huizilopoztli,

Después de saludarlo, sirva el presente oficio para notificar que el documento

Minería de opinión: un análisis en tiempo real de tweets para zacatecas de Luis Carlos Reveles Gómez

Fue analizado con el software iThenticate de Turnitin, con la intención de detectar similitudes; el resultado en cuestión fue

23 % de similitud

De acuerdo a lo anterior, el porcentaje se considera **ACEPTABLE** de acuerdo a los estándares internacionales.

Atentamente

"Forjemos el Futuro con el Arte, la Ciencia y el Desarrollo Cultural"

Dr. Carlos Francisco Bautista Capetillo
Coordinador de Investigación y Posgrado
Universidad Autónoma de Zacatecas



SOMOS
ARTE, CIENCIA Y
DESARROLLO
CULTURAL



Zacatecas, Zacatecas a 31 de mayo del 2021
Carta Cesión de Derechos

A QUIEN CORRESPONDA

El que suscribe Luis Carlos Reveles Gómez alumno del Programa de Maestría en Ciencias del Procesamiento de la Información con número de matrícula **30113277**, adscrito a la Unidad Académica de Ingeniería Eléctrica de la Universidad Autónoma de Zacatecas, manifiesta que es autor intelectual del presente trabajo de Tesis bajo la dirección del Dr. Huizilopztli Luna García y Dr. José María Celaya Padilla y cede los derechos del trabajo titulado Minería de Opinión: Un Análisis en Tiempo Real de Tweets para Zacatecas a la Universidad Autónoma de Zacatecas para su difusión, con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o directores del trabajo. Este puede ser obtenido escribiendo al correo electrónico luisarlosreveles@uaz.edu.mx o estableciendo contacto con el responsable del programa de Maestría quien turnará la solicitud al autor y directores del trabajo de investigación. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo. Agradezco de antemano su atención a la presente, reciba un cordial saludo.

ATENTAMENTE

Reveles Gómez Luis Carlos

Luis Carlos Reveles Gómez

Agradecimientos

Especialmente a mis papás, José Refugio Reveles y Lorenza Gómez por su apoyo incondicional durante todo este tiempo, que a pesar de lo difícil nunca me dejaron de apoyar. A mi novia Selene Viramontes que siempre estuvo conmigo en cada momento apoyándome y animándome para seguir adelante. Gracias a todos mis amigos del CIAM de los cuales aprendí bastantes cosas con el tiempo y mi permanencia en el Centro de Investigación.

Gracias a todos los Drs. del CIAM, especialmente al Dr. José María Celaya Padilla, Dr. Huizilopoztli Luna García y al Dr. Jorge Isaac Galván Tejada, que siempre estuvieron ayudándome y apoyando mi proyecto de una u otra forma, por dedicarle el tiempo a mi investigación y la retroalimentación brindada tanto en lo académico como en lo personal, de lo cual me llevo una gran enseñanza.

Asimismo, quiero agradecer al Consejo Zacatecano de Ciencia, Tecnología e Innovación (COZCyT) por el apoyo económico que se me otorgó durante el último año de mis estudios de posgrado, a través del proyecto “Fortalecimiento de la Maestría en Ciencias del Procesamiento de la Información de la Universidad Autónoma de Zacatecas” con clave: ZAC-2018-05-01-125266.

Dedicatoria

Dedicada especialmente para mis padres, mis hermanas y mi novia.

Resumen

La red social Twitter se ha convertido en una excelente herramienta para conocer en tiempo real las opiniones que los usuarios expresan sobre una gran variedad de temas. El análisis formal de los textos en los tweets es objeto de numerosos estudios, derivado de ellos, se ha impulsado la aparición de tecnologías emergentes como la Minería de Opinión, donde está inerte el análisis de sentimientos; el cual se refiere al uso del procesamiento del lenguaje natural para identificar y extraer información subjetiva de los textos [1]. Por definición, el análisis de sentimientos busca generar herramientas automáticas capaces de extraer información subjetiva para crear conocimiento estructurado y procesable [2]. En otras palabras, se trata de una tarea de clasificación masiva de documentos de manera automática, en función de la connotación positiva o negativa del lenguaje utilizado en el documento.

Este trabajo se centra en realizar análisis de sentimientos de comentarios de Twitter georreferenciado a la ciudad de Zacatecas, como una clasificación de los tweets etiquetados con su polaridad, realizando una limpieza del texto de los tweets, así como la extracción de características propias del texto como polaridad positiva y negativa, utilizando el machine learning en especial los algoritmos de aprendizaje supervisado para realizar la clasificación. De los algoritmos utilizados se obtuvo que Random Forest tuvo un mejor accuracy al tener 0.977, después Árboles de Decisión con 0.9735 y SVM con 0.9551. Con los resultados obtenidos se puede concluir que la mejora del accuracy se logró gracias a las características que se fueron agregando, además se demuestra que los algoritmos de aprendizaje supervisado están clasificando los tweets de manera adecuada dado los resultados obtenidos.

Abstract

The Twitter social network has become an excellent tool to know in real time the opinions that users express on a great variety of topics. The formal analysis of the texts in tweets is the subject of numerous studies, derived from them, the emergence of emerging technologies such as Opinion Mining, where sentiment analysis is inert; which refers to the use of natural language processing to identify and extract subjective information from the texts [1]. By definition, sentiment analysis seeks to generate automatic tools capable of extracting subjective information to create structured and actionable knowledge [2]. In other words, this is a bulk document classification task automatically, depending on the positive or negative connotation of the language used in the document.

This work focuses on performing sentiment analysis of Twitter comments georeferenced to the city of Zacatecas, such as a ranking of tweets tagged with their polarity, cleaning up the text of tweets, as well as extracting characteristics typical of the text don positive and negative polarity, using machine learning especially supervised learning algorithms to perform the classification. From the algorithms used it was obtained that Random Forest had a better accuracy by having 0.977, then Decision Trees with 0.9735 and SVM with 0.9551. With the results obtained it can be concluded that the improvement of the accuracy was achieved thanks to the features that were added, in addition it is shown that the supervised learning algorithms are classifying the tweets appropriately given the results obtained.

Contenido General

Agradecimientos.....	i
Dedicatoria	ii
Resumen.....	iii
Abstract	iv
Índice de Figuras	v
Índice de Tablas.....	vi
1. Introducción.....	1
1.1 Antecedentes	1
1.2 Planteamiento del problema.....	4
1.3 Justificación	5
1.4 Preguntas de investigación	6
1.5 Objetivo General	6
1.5.1 Objetivos específicos	6
1.6 Hipótesis	7
1.7 Estructura de la tesis.....	7
2. Marco Teórico	8
2.1 Inteligencia Artificial	8
2.2 Aprendizaje automático (<i>Machine learning</i>)	9
2.2.1 Aprendizaje supervisado	9
2.2.2 Aprendizaje no supervisado	10
2.3 Proceso general de entrenamiento	10
2.3.1 Adquisición de datos	10
2.3.2 Preprocesamiento	10
2.3.3 Extracción de características.....	11
2.3.4 Entrenamiento	11
2.3.5 Validación	11
2.4 Algoritmos.....	11
2.4.1 K-ésimo Vecino más Cercano(K-NN).....	12
2.4.2 Naive Bayes	13
2.4.3 Árboles de Decisión	14

2.4.4	Bosques Aleatorios	15
2.4.5	Redes Neuronales	17
2.4.6	Máquinas de Soporte Vectorial.....	19
2.5	Evaluación de los algoritmos	20
2.6	Principales estudios relacionados.....	23
3.	Metodología y Propuesta de Investigación.....	29
3.1	Diseño experimental.....	29
3.2	Datos de entrada.....	30
3.3	Preprocesamiento	31
3.3.1	Extracción	31
3.3.2	Limpieza de datos y tokenización	32
3.4	Análisis de sentimientos u opiniones.....	32
3.4.1	Implementación	33
3.5	Caso de estudio: Zacatecas	35
3.5.1	Base de datos.....	35
3.5.2	Adquisición de datos	36
3.5.3	Preprocesamiento	38
3.6	Preparación de base de datos o procesamiento	39
3.7	Validación de polaridad.....	42
4.	Resultados	43
5.	Conclusiones	49
5.1	Productos de la investigación	50
5.2	Trabajos futuros.....	52
	Referencias.....	53
	Anexos	58

Índice de Figuras

Figura 2.1: Seudocódigo de algoritmo KNN.....	13
Figura 2.2: Diagrama general de un árbol de decisión.....	14
Figura 2.3: Seudocódigo de árboles de decisión.....	15
Figura 2.4: Comparación Árbol de decisión y Bosque Aleatorio	17
Figura 2.5: Modelo de neurona artificial estándar.....	17
Figura 2.6: División de dos clases por medio de SVM.....	19
Figura 2.7: Representación de posibles curvas ROC.....	22
Figura 3.1: Diagrama de flujo de la metodología para el análisis de sentimientos de Tweets.....	29
Figura 3.2: Ejemplo de un Tweet en formato JSON.....	30
Figura 3.3: Keys y Tokens de Twitter.....	37
Figura 3.4: Condiciones de búsqueda.....	37
Figura 3.5: Declaración de caracteres a eliminar.....	38
Figura 3.6: Obtención de características estadísticas.....	41
Figura 4.1: Rendimiento de la matriz de confusión en el conjunto de datos test del algoritmo Random Forest.....	45
Figura 4.2: Curvas ROC del desempeño del algoritmo SVM, donde el color rojo= curva ROC de las clases [-1,1], la naranja= curva ROC de las clases [-1,0] y azul= curva ROC de las clases [0,1]. ...	46
Figura 4.3: Curvas ROC del desempeño del algoritmo Arboles de Decisión, donde el color rojo= curva ROC de las clases [-1,1], la naranja= curva ROC de las clases [-1,0] y azul= curva ROC de las clases [0,1].	47
Figura 4.4: Curvas ROC del desempeño del algoritmo Random Forest, donde el color rojo = curva ROC de las clases [-1,1], la naranja= curva ROC de las clases [-1,0] y azul= curva ROC de las clases [0,1].....	48
Figura 5.1: Logo empresa ZTMAR.....	50
Figura 5.2: Poster presentado en jornadas de investigación.....	51
Figura 5.3: Artículo Publicado en V Jornadas Interacción Humano Computadora 2019.	51

Índice de Tablas

Tabla 1.1: Plataformas Sociales más usadas en México.	2
Tabla 2.1: Definiciones de Inteligencia Artificial.	8
Tabla 2.2: Funciones de activación de redes neuronales.....	18
Tabla 2.3: Kernels de SVM.....	20
Tabla 2.4: Matriz de confusión.	21
Tabla 2.5: Desempeño en base a la exactitud.....	22
Tabla 2.6: Máximo, mínimo y media de los valores de exactitud encontrados en cada enfoque.	24
Tabla 3.1: Comparativa entre Tweets sin preprocesamiento y con preprocesamiento.	31
Tabla 3.2: Base de Datos de los Tweets Recolectados.....	36
Tabla 3.3: Etiquetas de texto o Tags de cada palabra en Python.	40
Tabla 3.4: Correlación entre la polaridad arrojada por el algoritmo contra los resultados de las encuestas.	42
Tabla 4.1: Cantidad de Tweets Etiquetados.	43
Tabla 4.2: Desempeño de los Algoritmos con k-fold Cross Validation.....	43
Tabla 4.3: Desempeño de los Algoritmos con k-fold Cross Validation con las nuevas características agregadas.....	44
Tabla 4.4: Desempeño de Algoritmos Sin Validación Cruzada.....	45

1. Introducción

En este capítulo se describen en un contexto general, los antecedentes que motivaron el trabajo de investigación; se establece la problemática que se plantea resolver, asimismo, se describen los objetivos a resolver con la investigación, y se analiza la justificación e hipótesis del trabajo de investigación; finalmente, se presenta una síntesis del trabajo a realizar y se describe brevemente la estructura de la tesis.

1.1 Antecedentes

En el transcurso de los últimos años, el incremento del uso de las tecnologías ha ido en constante crecimiento, esto ha causado una adaptación en la forma de vivir de las personas, estas se vuelven más dependientes a un nuevo estilo de vida, donde las tecnología e Internet están siempre presentes. Según el INEGI, en la última encuesta realizada en 2019, el 73.5% de la población de seis años o más, utiliza el teléfono celular y, nueve de cada diez personas, disponen de celular inteligente, con lo cual tienen la posibilidad de conectarse a Internet. Entre las principales actividades de los usuarios de Internet en 2019, correspondieron a entretenimiento (91.5%), obtención de información (90.7%) y comunicación entre pares (90.6%), asimismo, 48.3 millones de usuarios Internet con celular inteligente (Smartphone) instalaron aplicaciones en sus teléfonos, de estos, el 86.4% instaló aplicaciones de mensajería instantánea, el 80.8% para acceder a redes sociales y el 69.6% instaló aplicaciones para acceder a contenidos de audio y video. Por otra parte, el 25.4% de los usuarios utilizaron su dispositivo para instalar alguna aplicación que les permitiera realizar transacciones financieras [3]. Desde hace un par de años la comunicación ha sufrido una nueva revolución, debido a la gran influencia del Internet y las redes sociales, dándose un cambio en la forma de comunicarnos a través de dispositivos inteligentes.

La comunicación es el medio que permite interactuar con los demás al enviar y recibir mensajes. Podemos decir que existen dos tipos de comunicación: la verbal y la no verbal. La comunicación verbal también llamada comunicación oral, tiene la capacidad de utilizar la voz para expresar lo que se siente o piensa a través de las palabras; los gestos y todos los recursos de expresividad de movimientos del hablante forman parte de aquello que inconscientemente acompaña a nuestras palabras, es decir, la comunicación no verbal [4]. Dentro de la comunicación verbal se encuentra la comunicación escrita, puede ser enviada vía correo, email, chat, entre otros; la efectividad de esta depende de los estilos de escritura, vocabulario usado, gramática, claridad y precisión del lenguaje [5]. Estamos asistiendo al crecimiento de las redes sociales como forma de comunicación tanto escrita como oral, dado que las redes sociales permiten la mensajería tanto de texto como de notas de voz, siendo muy flexibles para llevar a cabo una comunicación fluida.

Por otra parte, las redes sociales han tomado una gran importancia entre el uso del Internet, y la comunicación, hoy en día, se han considerado como materia de difusión masiva, debido a su alcance, características e impacto en la sociedad. Son utilizadas tanto por individuos, grupos de trabajo, organizaciones y empresas, dado que permiten lograr una comunicación interactiva y dinámica [6]. Las redes sociales son ya una necesidad para los usuarios, son estas, las que facilitan que el usuario este informado o bien para informar de sus propias actividades. Esto permite generar una gran cantidad de datos por cada individuo, lo cual se está dando en tiempo real. De acuerdo con [7] las redes sociales que más datos están generando son Facebook, YouTube, WhatsApp y Twitter.

En cuanto a los datos generados por las redes sociales, es posible mencionar que estos son un blanco para múltiples estudios, ya que se pueden estudiar las necesidades de los usuarios a través de metodologías y herramientas especializadas, encontrando soluciones para publicidad, comunicación, presupuesto, manejo de bases de datos y marketing [8]; a esto se le conoce como “Inteligencia de Negocios (BI, por sus siglas en inglés)”, una rama de la inteligencia de datos. Al concepto de BI es posible agregar el uso de tecnologías que permiten la integración y transformación de los datos en información, ya sea estructurada o desestructurada, esta información clasificada y analizada se convierte en conocimiento, que puede servir de base para la toma de decisiones en un negocio y actuar como un factor estratégico para la organización, que favorezca la ventaja competitiva de la empresa, además, de ofrecer la oportunidad de detectar problemas y buscar alternativas o soluciones [9]. En la Tabla 1.1 se muestran las redes sociales más usadas en México, con un porcentaje de usuarios que usan dichas aplicaciones.

Tabla 1.1: Plataformas sociales más usadas en México [10].

Plataforma	Porcentaje(%) de Usuarios
YouTube	96
Facebook	94
WhatsApp,	89
FB Messenger	78
Instagram	71
Twitter	61
Pinterest	46
LinkedIn	36
Snapchat	35

Una de las plataformas donde confluyen una cantidad importante de datos en forma de opiniones sobre diversos temas es *Twitter* (Tw). Tan solo en 2018, *Twitter* tenía un número total de usuarios mensuales activos de 330 millones y se generaban aproximadamente 500 millones de *Tweets* por día [11]. A diferencia de otras redes sociales como *Facebook*, en Tw todos los comentarios (*tweets*) de los usuarios permanecen de manera pública, lo cual facilita el uso y manejo de información. La red social *Twitter* se ha convertido en una excelente herramienta para conocer en tiempo real las opiniones que los usuarios expresan sobre una gran variedad de temas. *Twitter* permite identificar estos temas a través de los denominados

hashtag o etiquetas, que se caracterizan por comenzar con el carácter # y a continuación una cadena de una o varias palabras concatenadas [12]. El análisis formal de los textos en los *tweets* es objeto de numerosos estudios, derivado de ellos, se ha impulsado la aparición de tecnologías emergentes como la Minería de Opinión (OM, por sus siglas en inglés).

En cuanto a la OM, donde está inerte el análisis de sentimientos; el cual se refiere al uso del procesamiento del lenguaje natural para identificar y extraer información subjetiva de los textos [1]. Por definición, el análisis de sentimientos busca generar herramientas automáticas capaces de extraer información subjetiva para crear conocimiento estructurado y procesable [2]. En otras palabras, se trata de una tarea de clasificación masiva de documentos de manera automática, en función de la connotación positiva o negativa del lenguaje utilizado en el documento, este análisis asigna un valor numérico al texto, siendo etiquetado como negativo o positivo con -1 y 1 respectivamente, o bien, 0's y 1's dependiendo mucho de las palabras que contenga el texto. Esto facilita mucho más realizar el análisis de datos aplicando técnicas de Inteligencia Artificial. En este contexto de la Inteligencia Artificial, el Aprendizaje constituye un sub-campo de la misma, que estudia los métodos de solución de problemas de aprendizaje por las computadoras. Son dos métodos de Aprendizaje: Supervisado y No Supervisado. Estos dos métodos derivados del machine learning, una de las disciplinas procedentes de la Inteligencia Artificial [13]. El supervisado constituye un algoritmo de aprendizaje basado en ejemplos donde el nuevo conocimiento es inducido a partir de una serie de ejemplos y contraejemplos y el no supervisado todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema [14]. Diversos investigadores han trabajado en estudios de la aplicación de la OM en la salud [15], negocios, documentos, turismo, todo desde datos proporcionados por *Twitter*.

En relación al aprendizaje supervisado, los algoritmos trabajan con datos “etiquetados”, intentado encontrar una función que, dadas las variables de entrada, les asigne la etiqueta de salida adecuada. El algoritmo se entrena con un “histórico” de datos y así “aprende” a asignar la etiqueta de salida adecuada a un nuevo valor, es decir, predice el valor de salida. La funcionalidad del aprendizaje supervisado se toma de base para desarrollar el presente trabajo de investigación, se usan datos de entrada para que los algoritmos aprendan y así obtener una salida etiquetada de forma adecuada. Si bien cada tipo de algoritmo tiene ventajas y desventajas, lo más usado en investigaciones relacionadas son los no supervisados. Esto es así debido a la dificultad que acarrea la obtención de bases de datos previamente clasificadas para la implementación de aprendizaje supervisado, sobre todo para nuevos temas de discusión [2].

Derivado de lo anterior, el propósito de esta investigación es desarrollar un algoritmo con el fin de obtener comentarios turísticos de la red social *Twitter* del estado de Zacatecas, crear una base de datos con estos comentarios y así aplicar técnicas de Inteligencia Artificial como lo es el Aprendizaje Supervisado antes ya mencionado, así como el análisis de sentimientos a estos comentarios.

1.2 Planteamiento del problema

Actualmente la actividad turística está caracterizada por un uso masivo e intensivo de información, dado el alto impacto que las Tecnologías de la Información y Comunicación (TIC) han tenido, tanto desde la perspectiva del consumidor como de la oferta turística en general [16]. Cada vez más, las experiencias positivas de los turistas en algún destino turístico buscan no sólo la repetición de visitas al mismo, sino además, la recomendación a los amigos, familiares y contactos de las redes sociales [17], aumentando la transferencia de datos georreferenciados a dicho lugar visitado, de igual manera si vivieron una mala experiencia. De manera inconsciente proporcionan información ya sea a sus perfiles de una red social o en forma de comentarios de las mismas, de esta manera es posible saber que tan bien la están pasando. Como consecuencia de la buena o mala satisfacción del turista, se generan mayores o menores ingresos a los lugares de visita [18]. Tener la información de que tan bien o mal la está pasando un turista puede ser aprovechado para ofrecer mejores servicios y con esto tener un balance entre la oferta y el consumidor [19].

Uno de los principales problemas que enfrenta la Minería de Opinión en las redes sociales, es que las personas se expresan de forma abierta y con múltiples expresiones coloquiales lo cual dificulta el análisis automático de los comentarios que se difunden. Además, muchos comentarios no ofrecen un sentimiento destacable por lo que se podrían considerar como neutros y no proporcionan información de utilidad en el análisis de la reputación [20]. Los medios sociales permiten que cada usuario se convierta en autor a través de las publicaciones en sus perfiles personales, de los comentarios y de los *tweets*. A diferencia de los textos tradicionales, los textos producidos en el ámbito virtual son mucho más mudables y dinámicos y la información es mucho más volátil, ya que implican la interacción constante de varios participantes que añaden, borran, o modifican su contenido. En general, el lenguaje mediado por el ordenador contiene características que no respetan la lengua estándar, que provienen de la intención del hablante de ahorrar la mayor cantidad de esfuerzo posible a la hora de escribir un mensaje, pero también de su deseo de expresarse de una manera original y creativa. La relajación lingüística y la actitud coloquial que caracteriza a la mayoría de las intervenciones que aportan los usuarios se asemejan al flujo de pensamientos del hablante, puesto que se suele evitar el uso de los signos de puntuación, de la mayúscula y de la organización rigurosa del contenido [21].

Tomando de base lo anterior, para que sea posible un análisis de texto en redes sociales varios autores [12][15][22][23] sugieren una serie de pasos a seguir con fines de que cada opinión, comentario o publicación sea pre-procesada, con esto se refieren a limpiar cada opinión de emoticones, signos de interrogación, o nombres de usuarios. Sin embargo, un aspecto importante es saber si los emoticones son necesarios en la opinión, y de ser así dejarlos. Por lo que el procesamiento automático de opiniones no es una tarea sencilla. Algunos de los problemas presentes en el tratamiento de las opiniones son: el uso de lenguaje informal, las abreviaturas, los errores ortográficos y tipográficos, el lenguaje irónico y sarcástico, el nivel de conocimiento del lenguaje, el nivel cultural, entre otros [24].

En relación a lo anterior si se explota de manera adecuada, georreferenciando para un lugar de interés, en este caso Zacatecas, se obtendrían grandes resultados, entre los cuales, saber la aceptación de las personas que visitan o viven en la ciudad, ubicar los lugares con mayor aceptación. Asimismo, de una herramienta que lea los datos obtenidos, los procese y proporcione una retroalimentación, ya sea a los turistas, al gobierno o para uso académico, según sea el caso.

Según [25], “la principal fuente de información para conocer los puntos fuertes y débiles de una organización es a través de las opiniones que generan los propios consumidores”. Por ello, hoy en día en el sector turístico y particularmente en la actividad turística de Zacatecas existe la necesidad de aprovechar la información contenida en las redes sociales por el gran contenido de opiniones que existen, realizar el análisis automático de dicha información, y de esta manera analizar grandes cantidades de datos simultáneamente. En este entorno una de las técnicas para este aprovechamiento del gran volumen de información existente es el análisis de sentimientos, perteneciente a la rama del Procesamiento del Lenguaje Natural (PLN), disciplina que combina el proceso de Inteligencia Artificial y la Lingüística Computacional para la recuperación de información, extracción de texto y detectar masivamente el significado residente en el lenguaje natural o humano. Cuyo objetivo principal es determinar si la parte del contenido es positiva, negativa o neutra [26].

1.3 Justificación

En la actualidad, el turismo se ha convertido en uno de los sectores con mayor impacto para los países en que su economía genera ingresos mediante este sector, tal es el caso del estado de Zacatecas, de acuerdo con [27] el estado cerró el año 2017, con una ocupación hotelera cercana al 60% con la llegada de 603 mil 84 turistas; lo que dio origen a una derrama económica de 1 mil 278 millones de pesos. La actividad turística en el estado de Zacatecas es un pilar importante para la economía de la entidad; por ello, continuamente detallan acciones encaminadas a fortalecer la actividad turística para consolidarla como uno de los sectores estratégicos más importantes de la economía en la generación empleos e ingreso con el fin de posicionar al Estado como uno de los principales destinos en turismo cultural y alternativo a nivel nacional [28]. Derivado de lo anterior y mediante el desarrollo de herramientas tecnológicas impulsar al Estado como un principal destino turístico.

En los últimos años la gran cantidad de comentarios y publicaciones que se han suscitado en las redes sociales, ha logrado que la Minería de Opinión (OM, pos sus siglas en inglés) tenga un gran crecimiento en su uso para investigaciones. Las empresas y organizaciones son las principales interesadas en aplicar esta herramienta, dado que les permite conocer la opinión de los usuarios [29]. En relación a lo anterior, utilizar la OM para realizar aplicaciones, herramientas tecnológicas, con el fin de conocer la opinión de las personas de una manera fluida. Existen varios usos para del análisis de opinión sin embargo el monitoreo de redes sociales esta entre las más populares para conocer los intereses de los usuarios [30]. No obstante, esta investigación tomará como base el monitoreo de redes sociales.

Asimismo, utilizando las aplicaciones en análisis de datos, para descubrir patrones y correlaciones en datos no estructurados, encontrar relaciones, dependencias y anomalías, se pretende usar la técnica de análisis de sentimientos, de esta manera se recolectarían las características y observaciones del proyecto, como lo son datos específicos de comentarios (positivos y negativos) de los turistas en el Estado de Zacatecas. Datos más específicos como la ubicación exacta, la hora de publicación y coordenadas. Con el propósito de generar una base de datos, y posteriormente aplicar las técnicas de *machine learning* [31], y obtener comentarios georreferenciados con una etiqueta correspondiente a negativo, neutro y positivo.

Es importante destacar que, para realizar OM de una red social, encontraremos con algunas restricciones, esto se refiere a que, aunque la mayoría de la información sea de manera pública los usuarios mantienen en privado varios aspectos, entre los cuales está la ubicación, si bien es importante, pero no del todo necesario, dado que al estar georreferenciando a una ubicación todos los datos nos servirán para llevar a cabo la investigación. Asimismo, el análisis de datos en conjunto con el *machine learning* tendrán un papel importante, dado que posiblemente no se tengan muchos datos con la ubicación presente, se puede hacer un estimador con los datos que se logren recolectar y contengan la ubicación.

1.4 Preguntas de investigación

¿Qué algoritmos de *machine learning* de aprendizaje supervisado permitirán obtener un mejor resultado en la clasificación de *tweets*?

¿Cuáles parámetros de entrada deben ser considerados para que los algoritmos arrojen los mejores resultados?

¿Cuál será el comportamiento de los algoritmos al omitir el texto del *tweet* como parámetro de entrada?

¿Qué tan precisos llegarán a hacer los algoritmos y que tan viable será replicarlos en otros lugares del estado o país?

1.5 Objetivo General

El objetivo principal de esta investigación es generar un modelo mediante la aplicación de las técnicas *machine learning* y análisis de sentimientos para clasificar *tweets* como positivos, negativos y neutros de manera automática y validar cuales de los algoritmos a utilizar nos brinda el mejor resultado de clasificación.

1.5.1 Objetivos específicos

- Desarrollar un algoritmo que permita analizar el *streaming* en tiempo real, mediante un lenguaje de programación de alto nivel.

- Extraer tweets específicamente con comentarios referentes al estado de Zacatecas, dichos tweets estarán referenciados por ubicación y palabra de búsqueda: Zacatecas.
- Crear una base de datos con los tweets obtenidos y extraer características para un análisis más profundo.
- Clasificar los tweets, mediante técnicas de inteligencia artificial y análisis de sentimientos.
- Crear una nueva base de datos con los tweets clasificados.
- Aplicar algoritmos de machine learning de aprendizaje supervisado, para clasificar de manera automática los tweets.
- Medir el desempeño de los algoritmos de machine learning para la clasificación de tweets.

1.6 Hipótesis

Mediante el uso de técnicas de inteligencia artificial y análisis de sentimientos es posible clasificar mensajes en redes sociales, para obtener la opinión de las personas en una zona georreferenciada, con el fin de detectar zonas de mayor influencia.

1.7 Estructura de la tesis

El resto del documento se organiza de la siguiente manera:

- **El Capítulo 2.** describe el marco teórico de la presente investigación, tratando dos los trabajos relacionados con los temas de análisis de sentimientos y recolección de datos de redes sociales, así como la aplicación del *machine learning*.
- **El Capítulo 3.** describe la metodología empleada en la investigación, así como el proceso de implementación del análisis de sentimiento y los algoritmos usados para la clasificación de los *tweets*.
- **El Capítulo 4.** describe los resultados obtenidos en la implementación de los algoritmos.
- **El Capítulo 5.** cierra el trabajo de investigación con conclusiones, consideraciones y trabajos futuros.

2. Marco Teórico

2.1 Inteligencia Artificial

Actualmente, la popularidad de la Inteligencia Artificial (IA) ha crecido considerablemente, sin embargo, pocos saben cómo se desarrolla o en qué consiste. Es a grandes rasgos, una ciencia que permite crear sistemas computacionales complejos capaces de realizar tareas específicas [32].

Cabe mencionar que no existe una definición exacta. Cada definición aporta conceptos o teorías distintas acerca de la interpretación del significado de “inteligencia”. Rouhiainen [33] define la IA como; “la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano”. Una de las ventajas de usar IA en los dispositivos, es que a diferencia de las personas estos no se cansarán. Además, pueden analizar grandes cantidades de información a la vez. Asimismo, los errores que generan las máquinas que son entrenadas con IA son significativamente menores a comparación de los humanos. En la Tabla 2.1 se presenta algunas definiciones que se han publicado para la IA.

Tabla 2.1: Definiciones de Inteligencia Artificial.

Definición	Autor
“La interesante tarea de lograr que los computadores piensen ... máquinas con mente, en su amplio sentido literal.”	Haugeland [34]-1985
“La automatización de actividades que vinculamos con procesos de pensamiento humano, actividades tales como toma de decisiones, resolución de problemas, aprendizaje.”	Bellman [34]-1978
“El arte de crear máquinas con capacidad de realizar funciones que realizadas por personas requieren de inteligencia.”	Kurzweil [34]-1990
“El estudio de cómo lograr que las computadoras realicen tareas que, por el momento, los humanos hacen mejor.”	Rich y Knight [34]-1991
“El estudio de las facultades mentales mediante el uso de modelos computacionales.”	Charniak y McDermott [34]-1985
“El estudio de los cálculos que permiten percibir, razonar y actuar.”	Winston [34] -1992
“Un campo de estudio que se enfoca en la explicación y emulación de la conducta inteligente en función de procesos computacionales.”	Schalkoff [34]-1990
“La rama de la ciencia de la computación que se ocupa de la automatización de la conducta inteligente.”	Luger y Stubblefield [34]-1993

En la Tabla 2.1, se aprecia como la IA tiene un concepto principal en su definición, el cual define a un objeto capaz pensar como un humano. Es así bajo este principio filosófico es como la ingeniería ha creado maquinas capaces de resolver problemas y aprender de ellos, se han creado maquinas que piensan como humanos.

La IA tiene sus orígenes desde siglo IV a.C. con el planteamiento de Aristóteles de cómo es que funciona la mente humana. Si bien en ese tiempo aún no se le denominaba IA, ese planteamiento se considera el primero para el futuro desarrollo de máquinas inteligentes. Con el avance del tiempo fueron apareciendo mitos en diversas civilizaciones antiguas, un ejemplo de ella fue la egipcia; logrando así, el desarrollo de aparatos y figuras con movimiento y sonidos propios (autómatas). Es hasta el siglo XX donde la aparición de la primera computadora da el gran salto para lo que hoy en día conocemos como IA, dado que en este siglo el desarrollo tecnológico, así como la gran cantidad de investigaciones en matemáticas y lógica llevó a la idea de crear máquinas que piensan [34].

El desarrollo de esta tecnología tiene como principio desglosar la inteligencia humana en todos sus procesos, desde los más simples a más complejos. Tiene el propósito de representarlos en lenguaje lógico digital. Estos pasarán a forma de algoritmos y programas cuya función es replicar el comportamiento de la inteligencia humana en una máquina [32].

La naturaleza de la IA es tan flexible que se puede dividir en diversas áreas de estudio. En cada una de ellas ha surgido de la necesidad de innovar en las investigaciones científicas e informáticas, desarrollando así, nuevas técnicas de programación además, de diseño y desarrollo de algoritmos para satisfacer las necesidades del área de aplicación [35]. Por lo tanto, la IA tiene diversas ramas de estudio, en una de ellas aparece el Aprendizaje Automático (*Machine Learning* en inglés). En esta parte de la IA se estudia los procesos de aprendizaje, así como su modelado, para llevar a cabo el desarrollo de algoritmos y facilitando trasladarlos a computadoras y crear sistemas inteligentes [36].

2.2 Aprendizaje automático (*Machine learning*)

Como ya se mencionó anteriormente, el Machine Learning (ML) surge de una de las tantas ramas de la IA, específicamente del paradigma del aprendizaje. Son algoritmos que aprenden conforme a los datos que se le están mostrando. Es decir, los algoritmos identifican patrones en los datos de entrada, por lo que, si tenemos una gran diversidad de datos recibidos en el algoritmo, éste estará aprendiendo a reconocer patrones dentro de todos los datos de entrada. Una vez que el algoritmo adquirió conocimiento, al recibir datos nuevos de entrada, este será capaz de evaluar la entrada con los patrones aprendidos para finalmente, obtener una relación entrada-salida [37][38]. En el machine learning existen dos tipos de aprendizaje en donde podemos encontrar el supervisado y no supervisado [33].

2.2.1 Aprendizaje supervisado

Los algoritmos de aprendizaje supervisado se dividen en dos conjuntos de datos. Uno es utilizado en el entrenamiento y el otro es para realizar pruebas. Estos están basados en modelos predictivos. Además, “en estos algoritmos podemos encontrar métricas de evaluación para determinar si el algoritmo está realizando bien la tarea” [38]. El aprendizaje

supervisado tiene una relación entrada-salida, por lo que etiqueta datos a la salida conforme a los que aprendió a la entrada. Al hacer esto al modelo le ayudará a brindar solución al problema a tratar así como a problemas similares [39].

2.2.2 Aprendizaje no supervisado

A diferencia del supervisado, los algoritmos no supervisados, solamente toman en cuenta el modelo predictivo y los datos no necesariamente deben de estar etiquetados, dado que, en estos modelos no es necesario para llevar a cabo el entrenamiento [38]. “Los algoritmos de agrupamiento permiten agrupar objetos de acuerdo a su distancia llamado Clustering, siendo este un proceso de agrupar datos en clases o clusters (grupos) de tal forma que los objetos de un clúster tengan una similitud alta entre ellos, y baja con objetos de otros clusters” [40]. Cabe resaltar, que el aprendizaje no supervisado se emplea cuando los datos no están etiquetados y la única manera de clasificarlos es agrupándolos según sus similitudes [13].

2.3 Proceso general de entrenamiento

Los algoritmos de ML para ser aplicados llevan un proceso el cual se le conoce como “entrenamiento”. A esto se refiere a que el algoritmo aprenda sobre los datos de entrada y con base en la información que contiene los datos y poder tomar una decisión a la salida. Sin embargo, el proceso general de entrenamiento de los modelos consiste en: adquisición de datos, preprocesamiento, extracción de características, entrenamiento y validación. Cada uno de ellos es elemental para el funcionamiento de los modelos de aprendizaje automático, los cuales se describen a continuación.

2.3.1 Adquisición de datos

La adquisición de datos se realiza con base a cada problemática, la cual consta de reunir los datos de interés necesarios para su posterior análisis. Los datos serán almacenados, creando bases de datos [41]. Estas son tablas formadas por columnas y renglones donde las columnas representan las características de los datos y los renglones son el número de observaciones. Estas son de estudio para las siguientes etapas.

2.3.2 Preprocesamiento

Su objetivo es la transformación del conjunto original de datos. En esta etapa se realiza un análisis previo, el cual consiste en la limpieza, normalización, en donde se eliminan manualmente las características que presentan un alto porcentaje de datos faltantes (NA),

variables atípicas (outliers) e imputación de datos. Por último, se dividen en dos conjuntos: entrenamiento y prueba

2.3.3 Extracción de características

Se refiere al proceso de reducción de la cantidad de datos que no representan información significativa para la creación de los modelos basados en IA mejorando así su eficiencia. “También, en esta etapa es posible reducir la cantidad de datos y mejorar la eficiencia del modelo” [42].

2.3.4 Entrenamiento

Esta etapa consiste en dividir el conjunto de datos a trabajar en dos partes, llamadas “conjunto de entrenamiento” y “conjunto de prueba”. El primero es usado para, como su nombre indica, entrenar el algoritmo, y el segundo, para probar el rendimiento del mismo. Al modelo entrenado se le proporcionan los datos de prueba y predice el resultado conforme a su entrenamiento. El conjunto de datos puede ser dividido en un 70% para entrenamiento y un 30% para su validación, pero esto depende de la cantidad de datos que se dispongan, incluso se puede considerar un 80/20, 75/25, etc., el porcentaje de entrenamiento y prueba lo definirá el tamaño del conjunto de datos. Por ejemplo, si el conjunto de datos no tiene suficientes entradas, el 30% puede no ser suficiente información para clasificar las clases del conjunto de prueba [43].

2.3.5 Validación

La etapa de validación de los modelos, será con base a métricas estadísticas, las cuales se encargan de juzgar el desempeño de un algoritmo. Por lo tanto, la selección de una métrica es clave para discriminar y obtener un modelo predictivo. Unas de las métricas más importantes para evaluar el desempeño de un modelo son: sensibilidad, especificidad y área bajo la curva ROC.

Dentro de los algoritmos de inteligencia artificial, especializados en problemas de clasificación como el que se está abordando en este trabajo de investigación se pueden mencionar; knn, máquinas de soporte vectorial, naive bayes, árboles de decisión, random forest, redes neuronales, los cuales son explicados a continuación.

2.4 Algoritmos

Debido a las características intrínsecas al problema que se está tratando de resolver, en el cual los datos están etiquetados previamente, se usarán los algoritmos de aprendizaje

supervisado tales como; K-ésimo vecino más cercano, Naive Bayes, Árboles de decisión, Bosques aleatorios y Máquinas de soporte vectorial, los cuales tienen como característica la clasificación de clases. Los algoritmos tendrán como salida la clasificación automática de los *tweets*. Recordemos que todo algoritmo tiene su etapa de entrenamiento, por lo cual el 75% de la base de datos se utilizará para entrenar los algoritmos y el 25% de la base de datos se aplicará para medir el rendimiento de los algoritmos. A continuación, se describe el funcionamiento de los algoritmos de aprendizaje supervisado.

2.4.1 K-ésimo Vecino más Cercano(K-NN)

Los clasificadores K-NN se utilizan en tareas de clasificación de gran cantidad de datos en donde aparecen bastantes clases, aun si estas clases son complicadas o difíciles de entender, cuando el algoritmo encuentra datos con clases similares este funcionara de una forma eficaz. Sin embargo, cuando los datos son confusos y no se aprecia una similitud entre los datos el algoritmo tendrá problemas para poder clasificar [44]. Localizar a los vecinos más cercanos requiere una función de distancia, o una fórmula que mide la similitud entre dos clases [40].

Hay muchas maneras diferentes de calcular la distancia. Sin embargo, el algoritmo KNN utiliza la distancia euclidiana, que es la distancia que se mide entre dos puntos [44]. La distancia euclidiana se especifica mediante la siguiente fórmula:

$$d(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (2.1)$$

Donde P_1 es el punto uno, P_2 es el punto dos y (x, y) son las coordenadas de los puntos.

El equilibrio entre sobreajuste y subajuste de los datos de entrenamiento es un problema conocido como compensación de sesgo-varianza. La elección de una k grande reduce el impacto o la varianza causado por los datos ruidosos, pero puede sesgar que corra el riesgo de ignorar patrones pequeños pero importantes.

En la Figura 2.1, “se presenta un pseudocódigo para el clasificador K-NN básico. Tal y como puede observarse en el mismo, se calculan las distancias de todos los casos ya clasificados al nuevo caso, x , que se pretende clasificar. Una vez seleccionados los K casos ya clasificados, D_x^k más cercanos al nuevo caso, x , a éste se le asignara la clase (valor de la variable C) más frecuente de entre los K objetos, D_x^k ”[45].

COMIENZO

Entrada: $D = \{(x_1, c_1), \dots, (x_N, c_N)\}$

$x = (x_1, \dots, x_1)$ nuevo caso a clasificar

PARA todo objeto ya clasificado (x_i, c_i)

Calcular $d_i = d(x_i, x)$

Ordenar $d_i (i = 1, \dots, N)$ en orden ascendente

Quedarnos con los K casos D_x^k ya clasificados más cercanos a x

Asignar a x la clase más frecuente en D_x^k

FIN

Figura 2.1: Seudocódigo de algoritmo KNN [45].

2.4.2 Naive Bayes

Se trata de una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados. Está basada en el Teorema de Bayes, también conocido como teorema de la probabilidad condicionada [46].

Entre las características que poseen los métodos bayesianos en tareas de aprendizaje se pueden resaltar las siguientes:

- Cada ejemplo observado va a modificar la probabilidad de que la hipótesis formulada sea correcta (aumentándola o disminuyéndola). Es decir, una hipótesis que no concuerda con un conjunto de ejemplos más o menos grande no es desechada por completo, sino que lo que harán será disminuir esa probabilidad estimada para la hipótesis.
- Los métodos bayesianos permiten tener en cuenta en la predicción de la hipótesis el conocimiento a priori o conocimiento del dominio en forma de probabilidades.

Como anteriormente el algoritmo de Naive Bayes es clasificador probabilístico, este será utilizado para clasificar una nueva instancia de un documento D dentro de un conjunto finito C de clases predeterminadas. Esto significa que, dada una clase C y un conjunto de palabras W del nuevo documento a clasificar, se calcula la probabilidad de que dicho documento se clasifique dentro de la categoría C , así se tiene:

$$P(c|w) = \frac{P(w|c) P(c)}{P(w)} \quad (2.2)$$

Donde $P(C)$ es la probabilidad a priori de la clase y $P(W|C)$ es la probabilidad condicional la palabra W dada la clase C . En base a los datos observados en cada experimento, se conoce la probabilidad de una palabra dada una clase y la probabilidad de la clase.

2.4.3 Árboles de Decisión

Los Árboles de Decisión son diagramas lógicos, las cuales sirven para representar una serie de condiciones que ocurren de una forma consecutiva. Los Árboles de Decisión están compuestos por nodos interiores, nodos terminales y ramas que emanan de los nodos interiores como se puede apreciar en la Figura 2.2. Cada nodo contiene un atributo, y las ramas representan un valor distinto al atributo, si se realiza un seguimiento desde los nodos por cada rama al final convergen en un solo punto, donde se crea la separación de datos [31].

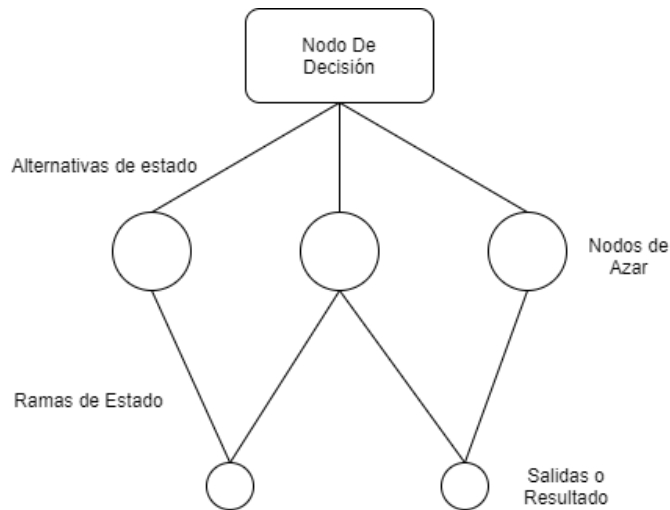


Figura 2.2: Diagrama general de un árbol de decisión.

“El árbol tiene una sola rama o nodo para todo el conjunto de datos de análisis, así como se muestra en la Figura 2.2. Conforme el grupo de datos S_b va peticionándose en subconjuntos S_s , se agrega un nodo con su respectiva rama para cada S_s . La partición se realiza en base a un criterio de valores o sobre un atributo de una característica determinada. Este criterio de partición es calculado automáticamente durante la etapa de crecimiento conforme a la ganancia de información basada en la entropía” [47].

Matemáticamente, “el algoritmo de Árboles de decisión se basa en la Entropía la cual se calcula con la ecuación (2.3) y la ganancia de la información con la ecuación (2.4). En teoría de la información, la entropía es la medida de la incertidumbre contenida en una variable aleatoria; mientras que la ganancia de información, es la reducción esperada en la entropía ocasionada por la partición del conjunto de muestras bajo el valor de cierta característica” [47].

$$\text{Entropía}(S_s) = \sum_{i=1}^c (-P_i \log_2 P_i) \quad (2.3)$$

$$G(S_v, A) = \text{Entropía}(S_s) - \sum_{v=1}^L \frac{|S_v|}{S_s} \text{Entropía}(S_v) \quad (2.4)$$

Donde $G(S_v, A)$ es la ganancia de información de una característica A relativa a un conjunto de muestras S_v , L es el número de salidas para el atributo A , S_v es cada uno de los subconjuntos correspondientes a la salida. En la entropía mide la homogeneidad del conjunto de datos, donde P_i es la proporción de S_s .

A continuación, se muestra el seudocódigo de la idea fundamental del denominado algoritmo TDIDT (Inducción de arriba hacia abajo de árboles de decisión)(ver Figura 2.3), el cual puede ser contemplado como base de la mayoría de los algoritmos de inducción de árboles de clasificación a partir de un conjunto de datos conteniendo patrones etiquetados [48].

```

Input:       $D$  conjunto de  $N$  patrones etiquetados, cada uno de los cuales está
               caracterizado por  $n$  variables predictoras  $X_1, \dots, X_n$  y la variable  $C$ 
Output:   Árbol de clasificación
Begin      TDIDT
               if todos los patrones de  $D$  pertenecen a la misma clase  $c$ 
                 then
                   resultado de la inducción es un nodo simple( nodo hoja)
                   etiquetado como  $c$ 
                 else
                   begin
                     1. Seleccionar la variable más informativa  $X_r$  con valores
                         $x_r^1, \dots, x_r^{nr}$ 
                     2. Particionar  $D$  de acorde con los  $n_r$  valores de  $X_r$  en
                         $D_1, \dots, D_{nr}$ 
                     3. Construir  $n_r$  subárboles  $T_1, \dots, T_{nr}$  para  $D_1, \dots, D_{nr}$ 
                     4. Unir  $X_r$  y  $n_r$  subárboles  $T_1, \dots, T_{nr}$  con valores
                         $x_r^1, \dots, x_r^{nr}$ 
                   end
                 endif
End        TDIDT

```

Figura 2.3: Seudocódigo de árboles de decisión [48].

2.4.4 Bosques Aleatorios

Los bosques aleatorios (*Random Forest* en inglés) son una combinación de árboles de decisión (ver Figura 2.4) de manera que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles del bosque. El error de generalización de un bosque de clasificadores de árboles

depende de la fuerza de los árboles individuales en el bosque y la correlación entre ellos [49]. “Los bosques aleatorios se pueden utilizar para una variable de respuesta categórica, denominada "clasificación", o una respuesta continua, denominada "regresión"” [50]. Desde un punto de vista computacional, los bosques aleatorios son atractivos porque

- manejar naturalmente tanto la regresión como la clasificación (multiclases);
- son relativamente rápidos de entrenar y predecir;
- dependen sólo de uno o dos parámetros de ajuste;
- tener una estimación incorporada del error de generalización;
- se puede utilizar directamente para problemas de gran dimensión;
- se puede implementar fácilmente en paralelo.

Estadísticamente, los bosques aleatorios son atractivos debido a las características adicionales que brindan, como

- medidas de importancia variable;
- ponderación de clase diferencial;
- imputación de valor perdido;
- visualización;
- detección de valores atípicos;
- aprendizaje sin supervisión.

Como se ha señalado anteriormente, el método Random Forest se basa en un conjunto de árboles de decisión, es decir, una muestra entra al árbol y es sometida a una serie de pruebas binarias en cada nodo, llamados split, hasta llegar a una hoja en la que se encuentra la respuesta. Esta técnica puede ser utilizada para dividir un problema complejo en un conjunto de problemas simples. En la etapa de entrenamiento, el algoritmo intenta optimizar los parámetros de las funciones de split a partir de las muestras de entrenamiento [51].

Para ello se utiliza la siguiente función de ganancia de información:

$$I_j = H(j) - \sum_{i \in \{1,2\}} \frac{|S_j^i|}{|S_j|} H(S_j^i) \quad (2.5)$$

Donde S representa el conjunto de muestras que hay en el nodo por dividir, y S^i son los dos conjuntos que se crean de la escisión. La función mide la entropía del conjunto, y depende del tipo de problema que se aborda.

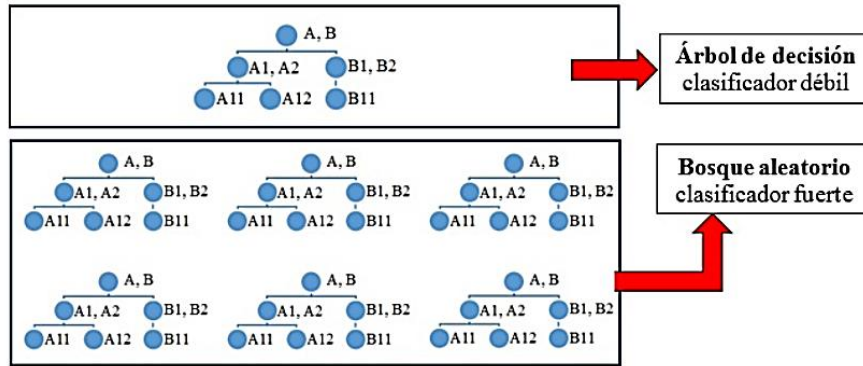


Figura 2.4: Comparación Árbol de decisión y Bosque Aleatorio [52].

2.4.5 Redes Neuronales

Una Red Neuronal Artificial (ANN, por sus siglas en inglés) es una relación entre un conjunto de señales de entrada y salida utilizando un modelo que simula el comportamiento de un cerebro biológico, respondiendo a los estímulos de las entradas sensoriales. El cerebro utiliza una red de células interconectadas llamadas neuronas con lo que se crea un mega procesador biológico, las redes de neuronas artificiales son utilizadas para resolver problemas de aprendizaje, simulando así al cerebro humano el cual es capaz de aprender conforme a las señales de entrada [46].

Una ANN típica con dendritas de entrada n (Figura 2.5), puede ser representada por la fórmula (2.6). Los pesos w permiten que cada una de las entradas n (denotadas por x_i) contribuyan en una cantidad mayor o menor a la suma de señales de entrada. El total neto es utilizado por la función de activación $f(x)$, y la señal resultante, $y(x)$, es el axón de salida:

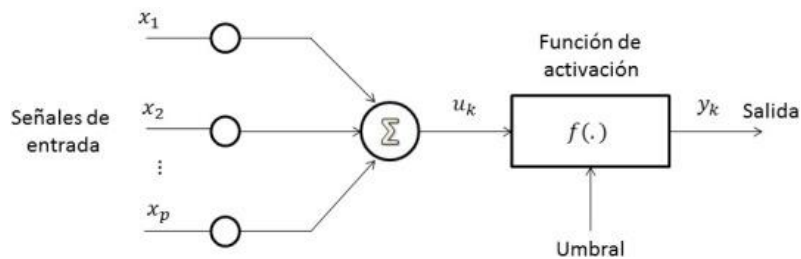


Figura 2.5: Modelo de neurona artificial estándar.

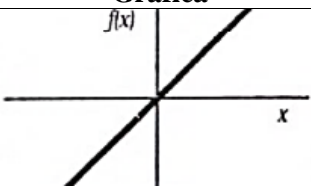
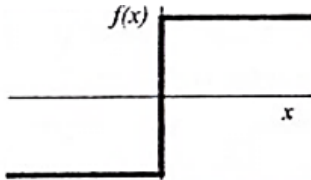
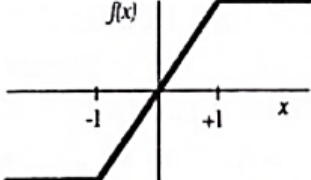
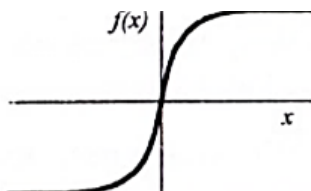
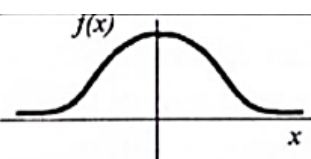
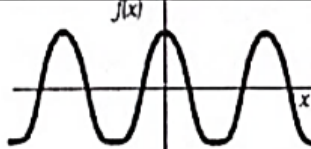
$$y(x) = f\left(\sum_{i=1}^n w_i x_i\right) \quad (2.6)$$

De acuerdo con Tobergte et al. [46], las ANN emplean neuronas como bloques de construcción para formar modelos complejos. A pesar de que existen variantes, cada una se puede definir en términos de las siguientes características:

- “Una función de activación, que transforma las señales de entrada combinadas de una neurona en una sola señal de salida que se transmitirá más adelante en la red.”
- “Una topología de red (o arquitectura), que describe el número de neuronas en el modelo, así como el número de capas y la forma en que están conectadas.”
- “El algoritmo de entrenamiento que especifica cómo se establecen los pesos de conexión para inhibir o excitar las neuronas en proporción a la señal de entrada”.

Existen varios tipos de funciones de activación para las redes neuronales (ver Tabla 2.2)

Tabla 2.2: Funciones de activación de redes neuronales [53].

	Función	Rango	Gráfica
Identidad	$y=x$	$[-\infty, \infty]$	
Escalón	$y = \begin{cases} 1, & \text{si } x \geq 0 \\ 0, & \text{si } x < 0 \end{cases}$	$[0,1]$	
	$y = \begin{cases} 1, & \text{si } x \geq 0 \\ -1, & \text{si } x < 0 \end{cases}$	$[-1,1]$	
Lineal a tramos	$y = \begin{cases} 1, & \text{si } x > 1 \\ x, & \text{si } -1 \leq x \leq 1 \\ -1, & \text{si } x < -1 \end{cases}$	$[-1,1]$	
Sigmoidea	$y = \frac{1}{1+e^{-x}}$	$[0,1]$	
	$y = \tanh(x)$	$[-1,1]$	
Gaussiana	$y = A e^{-Bx^2}$	$[0,1]$	
Sinusoidal	$y = A \sin(wx + \phi)$	$[-1,1]$	

Cabe mencionar que cada función de activación es elegida a cuestión de las especificaciones de lo que se desea obtener, sin embargo, la función Sigmoidea esta entre las más populares al momento de aplicar redes neuronales.

2.4.6 Máquinas de Soporte Vectorial

Las Maquinas de Vectores Soporte (SVM, por sus siglas en inglés), es una de las técnicas más eficaz del aprendizaje automático, dado que a pesar de su sencillez para desarrollarlo ha demostrado ser también un algoritmo robusto además de dar solución a problemas de la vida real [54]. Las máquinas de vectores de soporte pertenecen a una clase de algoritmos de *Machine learning* denominados métodos *kernel* y también se conocen como máquinas *kernel* [55]. En SVM es encontrar un plano que separe los grupos dentro de los datos de la mejor forma posible. Aquí, la separación significa que la elección del plano maximiza el margen entre los puntos más cercanos en el plano; éstos puntos se denominan vectores de soporte [31].

Para realizar un SVM se consideran dos casos:

- I. Caso linealmente separable
- II. Caso linealmente no se parable

Tomando como referencia el caso I, considerando dos clases, por ejemplo -1 y 1, el algoritmo buscara un hiperplano que separe a los datos en estas dos clases, este hiperplano está basado en la ecuación 2.7.

$$f(x)=x \cdot w+b \quad (2.7)$$

Dónde w son los pesos, al igual como se utilizan en las redes neuronales, y b la intercepción en el eje x . Una máquina de vectores de soporte construye un hiperplano óptimo en forma de superficie de decisión como se aprecia en la representación gráfica de la Figura 2.6, de modo que el margen de separación entre las dos clases en los datos se amplía al máximo.

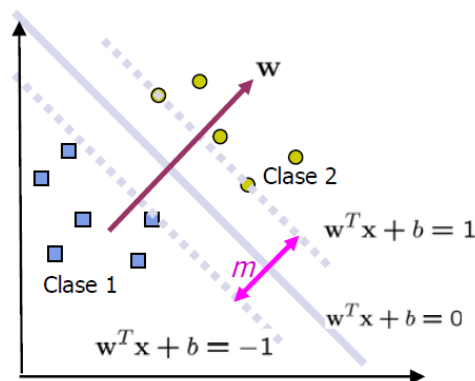


Figura 2.6: División de dos clases por medio de SVM.

Los vectores de soporte hacen referencia a un pequeño subconjunto de las observaciones de entrenamiento que se utilizan como soporte para la ubicación óptima de la superficie de decisión.

El entrenamiento de una máquina de vectores de soporte consta de dos fases:

- I. “Transformar los predictores (datos de entrada) en un espacio de características altamente dimensional. En esta fase es suficiente con especificar el *kernel*; los datos nunca se transforman explícitamente al espacio de características. Este proceso se conoce comúnmente como el truco *kernel*” [55].
- II. “Resolver un problema de optimización cuadrática que se ajuste a un hiperplano óptimo para clasificar las características transformadas en dos clases. El número de características transformadas está determinado por el número de vectores de soporte” [55].

Para construir la superficie de decisión solo se requieren los vectores de soporte seleccionados de los datos de entrenamiento. Una vez entrenados, el resto de los datos de entrenamiento son irrelevantes.

Entre los *kernels* populares (ver Tabla 2.3) que se emplean con las máquinas SVM se incluyen:

Tabla 2.3: Kernels de SVM [55].

Tipo de SVM	Kernel Mercer	Descripción
Función de base radial (RBF) o gaussiana	$K(x_1, x_2) = \exp\left(-\frac{\ x_1 - x_2\ ^2}{2\sigma^2}\right)$	Aprendizaje de una sola clase; σ es la anchura del <i>kernel</i>
Lineal	$K(x_1, x_2) = x_1^T x_2$	Aprendizaje de dos clases
Polinomial	$K(x_1, x_2) = (x_1^T x_2 + 1)^\rho$	ρ es el orden del polinomio
<i>Sigmoide</i>	$K(x_1, x_2) = \tanh(\beta_0 x_1^T x_2 + \beta_1)$	Es un <i>Kernel Mercer</i> solo para determinados valores β_0 y β_1

2.5 Evaluación de los algoritmos

Para cada algoritmo, se obtendrá las métricas de evaluación o bien llamadas métricas de evaluación de rendimiento para cuando son algoritmos de clasificación como lo es este caso. La matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real (ver Tabla 2.4). Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo las diferentes clases o resultados de la clasificación.

Tabla 2.4: Matriz de confusión.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

En la matriz de confusión, VP (verdaderos positivos) y VN (verdaderos negativos) representan respectivamente el número de ejemplos positivos y negativos clasificados correctamente, mientras que FP (falsos positivos) y FN (falsos negativos) representan respectivamente el número de ejemplos positivos y negativos clasificados incorrectamente [56].

La matriz de confusión nos proporciona los datos antes mencionados, con los cuales podemos obtener más métricas:

- I. Exactitud
- II. Precisión
- III. Sensibilidad
- IV. Especificidad

La Exactitud (ecuación 2.8) se refiere a lo cerca que está el resultado de una medición del valor verdadero. En términos estadísticos, la exactitud está relacionada con el sesgo de una estimación. También se conoce como Verdadero Positivo. Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos.

$$\text{Exactitud} = \frac{VP+VN}{VP+FP+FN+VN} \quad (2.8)$$

La Precisión (ecuación (2.9)) se refiere a la dispersión de los valores obtenidos a partir de mediciones repetidas de magnitud. Cuanto menor es la dispersión mayor la precisión, la precisión se representa de la siguiente manera:

$$\text{Precisión} = \frac{VP}{VP+FP} \quad (2.9)$$

La sensibilidad y la especificidad (ecuaciones (2.10) y (2.11)) son dos valores que nos indican la capacidad de nuestro estimador para discriminar los casos positivos, de los negativos. La sensibilidad es la fracción de verdaderos positivos, mientras que la especificidad, es la fracción de verdaderos negativos.

Por lo tanto, las fórmulas son muy similares:

$$\text{Sensibilidad} = \frac{VP}{VP+FN} \quad (2.10)$$

$$\text{Especificidad} = \frac{VN}{VN+FP} \quad (2.11)$$

Gráficamente las métricas de sensibilidad y especificidad, dan mucha información acerca del desempeño del algoritmo, estas graficas se les conoce como Curvas ROC (*Receiver Operating Characteristic*) (Figura 2.7).

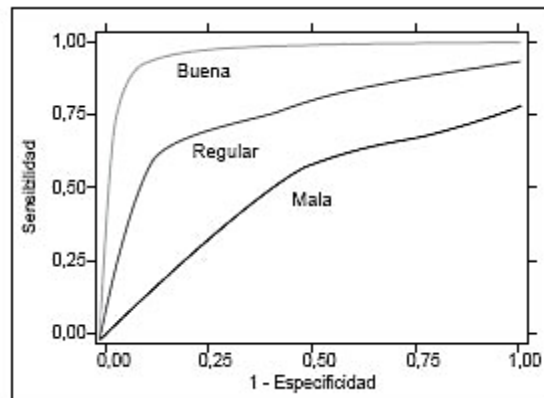


Figura 2.7: Representación de posibles curvas ROC.

El área bajo la curva nos indica una métrica antes calculada, la exactitud (*AUC*), si el área bajo la curva tiende a ser 1, entonces el algoritmo tiene un buen desempeño (ver Tabla 2.5).

Tabla 2.5: Desempeño en base a la exactitud.

Exactitud	Desempeño
[0.5-0.6)	Prueba mala
[0.6-0.75)	Prueba regular
[0.75-0.9)	Prueba buena
[0.9-0.97)	Prueba muy buena
[0.97-1)	Prueba excelente

2.6 Principales estudios relacionados

A continuación, se presentan los trabajos relacionados a la presente investigación, en las cuales se abarcan temas como lo es extracción de tweets, análisis de sentimientos y minería de opinión.

Sentiment Analysis Of Twitter Data Using NLTK In Python [57]

En el trabajo de investigación de Garg P. [57], su objetivo es clasificar los datos de Twitter en sentimientos (positivos o negativos) mediante el uso de diferentes clasificadores supervisados de aprendizaje automático en los datos recopilados de *twitter* para diferentes partidos políticos indios y mostrar qué partido político está teniendo un mejor desempeño para el público. Para clasificar los tweets en diferentes clases (positivas y negativas) el autor creo un clasificador que consta de varios clasificadores de aprendizaje automático. Para construir el clasificador uso una biblioteca de Python llamada Scikit learn. Realizando la clasificación por separado para 3 partidos políticos BJP, AAP y INC, obtuvo en sus resultados una precisión de alrededor del 70% para el clasificador MultinomialNB, lo que significa que el clasificador está funcionando correctamente. Otros resultados, que podemos encontrar son la confianza y la precisión. La clasificación que hicimos es básicamente como un voto. Por lo tanto, podemos aplicar una estrategia de votación para calcular el número máximo de votos para una característica mediante el uso del modo, y término como confianza. Se encontró que la clasificación de la confianza negativa es de hasta el 100%, mientras que para la clasificación positiva es del 85%, lo que es fiable. Se encontró que la precisión del modelo era del 70%, lo que significa que la mayoría de los resultados devueltos por el clasificador son relevantes.

Detección De La Polaridad De Las Opiniones Basada En Nuevos Recursos Léxicos [24]

El objetivo de la investigación de M. A. Amores Fernandez [24], consiste en desarrollar un sistema, a partir de las características de PosNeg Opinion, para la detección no supervisada de la polaridad de las opiniones a partir del empleo de nuevos recursos léxicos y que sea capaz de tratar la mayoría de los problemas presentes en las opiniones. Los resultados obtenidos son: la creación de los recursos SentiWordNet 4.0 y 4.1 para el idioma Inglés y SpanishSentiWordNet que es pionero en la puntuación de términos en español, los dos esquemas para la detección no supervisada de la polaridad de las opiniones, los recursos que permiten el manejo de jergas, emoticonos, palabras modificadoras y la negación, la herramienta PosNeg Opinion 3.0 que implementa el esquema finalmente propuesto auxiliándose de la biblioteca desarrollada PolarityDetection obteniendo satisfactorios valores de exactitud y F1 del 85%. Sin embargo, en esta investigación hace una comparativa interesante entre resultados del análisis de sentimientos con aprendizaje automático y en sistemas basados en recursos léxicos obteniendo los resultados de la Tabla 2.6.

Tabla 2.6: Máximo, mínimo y media de los valores de exactitud encontrados en cada enfoque.

Enfoque en la literatura	Máximo, Mínimo y Media de la Exactitud
Aprendizaje automático	91.5%, 64.1% y 75.3%
Basado en recursos léxicos	85%, 67% y 77%

Lo cual se concluye que el uso del machine learning da mejores resultados para la clasificación en texto.

A Case Study Of Spanish Text Transformations For Twitter Sentiment Analysis [58]

E. S. Tellez et al. [58], en el objetivo de su investigación identifican en un gran conjunto de combinaciones qué transformaciones de texto (lemmatización, derivación, eliminación de entidades, entre otros), tokenizadores (por ejemplo, palabras n-gramos) y esquemas de ponderación de tokens tienen el mayor impacto en la precisión de un clasificador (Support Vector Machine) entrenado en dos conjuntos de datos españoles. Teniendo como resultado que las mejores puntuaciones alcanzan 0.72 y las peores están por debajo de 0.43 de accuracy. La recaudación de los participantes está entre 0.59 y 0.61; se encuentra el mejor clasificador de sentimiento basado en n palabras (0.6051). La mejor configuración que utiliza q gramos, como un solo tokenizador, supera ese rango, es decir, 0.6330. Los clasificadores basados en la combinación de tokenizadores producen un rendimiento ligeramente mejor.

Minería De Opiniones Basado En La Adaptación Al Español De ANEW Sobre Opiniones Acerca De Hoteles [29]

C. Henríquez Miranda et al. [29], en su artículo muestran la construcción de un sistema de minería de opiniones en español sobre comentarios dados por clientes de diferentes hoteles. El sistema trabaja bajo el enfoque léxico utilizando la adaptación al español de las normas afectivas para las palabras en inglés (ANEW). Estas normas se basan en las evaluaciones que se realizaron en las dimensiones de valencia, excitación y el dominio. Para la construcción del sistema se tuvo en cuenta las fases de extracción, preprocesamiento de textos, identificación del sentimiento y la respectiva clasificación de la opinión utilizando ANEW. Los experimentos del sistema se hicieron sobre un corpus etiquetado proveniente de la versión en español de Tripadvisor. Como resultado final se obtuvo una precisión del 94% superando a sistemas similares.

Classifier Ensembles That Push The State-Of-The-Art In Sentiment Analysis Of Spanish Tweets [59]

Este artículo describe el sistema JACERONG propuesto para participar en la Tarea 1 de TASS 2017. Para tal evaluación, se implementaron dos métodos de combinación de clasificadores ampliamente utilizados debido a su demostrada capacidad de aumentar la exactitud de predicción, a saber: promediar y apilar. En primer lugar, clasificadores (relativamente) muy correctos utilizan algoritmos de aprendizaje supervisado para predecir una etiqueta de clase o estimaciones de probabilidad. Un clasificador de aprendizaje automático, o clasificador de primer nivel, recibe el vector de entidad y predice una etiqueta

de clase o estimaciones de probabilidad, es decir, la probabilidad de que el tweet sea de una clase determinada. Sea cual sea la predicción, se denomina predicción de nivel uno. Las máquinas vectoriales de regresión logística y soporte (SVM) con kernel 'lineal' son los algoritmos utilizados para desarrollar un enfoque de clasificación de aprendizaje supervisado; Scikit-learn es la biblioteca de aprendizaje automático utilizada. Luego, se combinan de manera óptima las predicciones de estos clasificadores con el fin de obtener una mejor predicción final. Por último, también se exploró cómo elegir cuales clasificadores constituyen un conjunto. Los resultados experimentales muestran que el sistema propuesto es el mejor clasificado en la evaluación del corpus InterTASS, de acuerdo con la métrica oficial de exactitud. Asimismo, los resultados indican que el desempeño predictivo sobre el conjunto de evaluación completo del Corpus General de TASS es superior al mejor resultado alcanzado en la evaluación de cuatro etiquetas de la edición anterior de TASS, en términos de la métrica oficial Macro-F1.

A Novel Framework For Aspect-Based Opinion Classification For Tourist Places [26]

En este artículo de M. Afzalet al. [26], realizan el análisis sobre métodos de tres tipos diferentes de técnicas de minería de opiniones, es decir, la minería de opiniones basada en tendencias, la minería de opiniones basada en aspectos y la minería de opiniones basada en frases, ha realizado que se extrajo información significativa de tweets, reseñas y blogs de viajes. Este documento identificó las limitaciones de que los métodos de clasificación de tendencias no clasifican las tendencias ambiguas, la integridad de los datos perturba la minería de opiniones y los métodos de extracción de aspectos no pueden identificar los aspectos de conferencia en las oraciones de opinión. Propusieron un marco de clasificación de opiniones basado en aspectos para el turismo que aborde estas limitaciones. El marco propuesto recopila datos de Twitter, extrae los aspectos tangibles y las tendencias luego clasifica las tendencias en tendencias positivas y negativas e identifica qué tweet tiene sentimiento positivo y qué tweet tiene un sentimiento negativo.

Minería De Opinión No Supervisada En Twitter [60]

En la investigación de J. A. Diaz-Garcia et al. [60], proponen un sistema que combina reglas de asociación, generalización de reglas y análisis de sentimientos para catalogar y descubrir tendencias de opinión en la red social Twitter. A diferencia de lo extendido, se usa el análisis de sentimientos para favorecer la generalización de las reglas de asociación. Para ello, primeramente, mediante minería de textos se resume un conjunto inicial de 1.7 millones de tuits captados de manera no dirigida en un conjunto de entrada para los algoritmos de reglas y análisis de sentimientos de 140718 tweets. Sobre este último conjunto se obtienen sets de reglas, estándar y generalizadas, fácilmente interpretables sobre personajes que el propio sistema revelará como interesantes.

Análisis De La Reputación De Un Destino Turístico En Las Redes Sociales [20]

Este Trabajo utiliza también Internet como recurso para obtener opiniones sobre los países que son destinos turísticos actualmente o que podrían ser objetivo de futuras campañas de marketing para una empresa. El hito principal de este proyecto será clasificar mensajes

obtenidos de la red social Twitter mediante algoritmos de aprendizaje automático de inteligencia artificial y de minería de textos. Esta clasificación se realizará en función de la reputación de un destino cuantificada mediante una puntuación positiva o negativa que exprese el sentimiento de cada comentario obtenido de la citada red social. Dentro de las técnicas de clasificación, este estudio va a analizar los algoritmos de aprendizaje automático (Machine Learning). Entre los algoritmos que más se utilizan en el análisis de sentimiento y que se han evaluado son los siguientes:

Clasificadores supervisados probabilísticos:

- Naïve Bayes (NB).
- Naïve Bayes Multinomial (NBM).

Clasificador supervisado lineal:

- Support Vector Machines (SVM)

Clasificadores no supervisados:

- SentiWordNet (SWN).

Obteniendo los siguientes resultados, en el algoritmo no supervisado SentiWordNet clasifica menos instancias correctamente que los algoritmos supervisados evaluados llegando a clasificar correctamente el 63% de las instancias de la base de datos de pruebas.

Los algoritmos supervisados, NB, NBM y SVM, tienen un comportamiento similar en la evaluación de esta base de datos, aunque Naive Bayes Multinomial es el que se comporta mejor con un 75% de instancias bien clasificadas, aunque le sigue muy de cerca con un 74% de Naive Bayes y SVM con un 72%.

Análisis De Sentimientos En Twitter [23]

El objetivo de este trabajo de investigación, es explicar los fundamentos teóricos sobre los que se asienta el análisis de sentimientos, su historia, aplicaciones y su relación con el procesamiento del lenguaje natural. Se ofrecerá una visión del estado del arte mediante un recorrido por los estudios publicados por decenas de autores y veremos los métodos más importantes que existen para desarrollar este tipo de soluciones.

Implementaremos un clasificador de sentimientos para los mensajes de Twitter basado en algoritmos de aprendizaje supervisado y se llevará a cabo un estudio comparativo con las técnicas más populares para el análisis de sentimientos a nivel de documento. Teniendo como resultado que el algoritmo de máquinas de vectores de soporte (SVM) es, sin ninguna duda y corroborando lo manifestado en multitud de trabajos de investigación, el mejor de los clasificadores para la realización de este tipo de tareas. En este caso, hemos probado su eficacia mediante el entrenamiento con unigramas, debido también a que los mejores resultados se suelen obtener con esta clase de características al tener 71% de accuracy. Las técnicas de ponderación normalizada contribuyen a aumentar la eficacia del modelo, mientras

que la reducción de características mediante streaming, aunque no de manera espectacular, también aporta una mejoría en el rendimiento global.

Análisis De Sentimientos Basado En Opiniones Turísticas [22]

F. N. Machado [22], en su trabajo de investigación realiza un sistema que permite realizar el análisis de las opiniones elaboradas por los usuarios de diversos sitios como restaurantes, hoteles, parques y playas de toda la isla de Tenerife. Las opiniones se obtienen a partir de una API que provee Google Maps que nos proporciona información valiosa del sitio como puede ser su localización, opiniones, idioma de éstas, etc. Todas estas opiniones serán almacenadas en una base de datos, de tal forma que más adelante sirvan como entrada al algoritmo de análisis. Los lugares elegidos fueron escogidos dentro de un radio que tiene como punto central un determinado lugar. Se utilizan técnicas avanzadas de procesamiento de lenguaje natural para resolver el problema. Por una parte, la obtención de la representación de las palabras mediante vectores numéricos usando el algoritmo Word2Vec. Por otra, técnicas de aprendizaje automático orientadas a la clasificación del sentimiento de las reseñas con objeto de determinar si una opinión tiene una connotación positiva o negativa. Sin embargo, el resultado no es tan exacto debido a que no se disponía de una cantidad suficiente de opiniones (a pesar de haber usado 924 opiniones en total). Para ir mejorando este resultado, se seguirá ejecutando todo el pipeline de scripts para recoger más opiniones. Es necesario tener en cuenta que algunas han perdido información debido a su descarte, bien porque estaban mal escritas o bien porque el Freeling es incapaz de analizarlas, y por tanto fueron descartadas; no es lo mismo una oración con diez palabras que con 5, algo de información se pierde y por tanto el entrenamiento pierde un poco de eficiencia. En general, el porcentaje de acierto es bajo (58%), debido sobre todo a la cantidad de palabras eliminadas a la hora de la conversión a formato Word2Vec. Sin embargo, es un porcentaje que está por encima de la mitad, por lo que, aunque sea mejorable, significa que para más de la mitad de las pruebas que se realizaron, se obtuvieron los resultados correctos.

Técnicas de Análisis de Sentimientos Aplicadas a la Extracción de Opiniones en el Lenguaje Español [61]

R. Germán et al.[61], en su proyecto proponen analizar distintas técnicas de Análisis de Sentimiento aplicadas a opiniones expresadas en el lenguaje español, evaluar sus resultados para distintos casos reales, y realizar mejoras a las mismas. Evaluación de técnicas de análisis de sentimientos aplicadas a opiniones escritas en español/castellano. Las técnicas principales a ser evaluadas son: Naive Bayes, SVM, Entropía Máxima y combinaciones de las anteriores con el uso de diccionarios/léxicos. Además, se agregará al menos una más, a seleccionar entre Random Forest, Redes Neuronales Profundas y K-NN, teniendo como criterio su grado de complementación con las anteriores (en el lenguaje español), para definir enfoques híbridos. Posteriormente entrenaron un modelo usando Naive-Bayes y aún están realizando ajustes y validaciones para asegurar la calidad de los resultados.

Después de analizar los trabajos mencionados, se pudo observar una clara tendencia de uso de herramientas de inteligencia artificial las cuales solo toman en cuenta como características el texto, por lo que en este trabajo se desea explorar el uso de otras características como lo

son: cantidad de palabras positivas, negativas, tipos de palabras, así como la estadística descriptiva de la cantidad de palabras y cantidad de caracteres de cada comentario, además explorar el uso del algoritmo propio de nombre SvCr como método de selección.

3. Metodología y Propuesta de Investigación

3.1 Diseño experimental

Para obtener de manera automática la polaridad de los *tweets* de los usuarios en coordenadas georreferenciadas y, asimismo, aplicar técnicas de inteligencia artificial y análisis de sentimientos. Se muestra la descripción de la metodología para el análisis de sentimientos en la Figura 3.1. Donde se extraerán *tweets* desde *Twitter*, por medio del lenguaje de programación de alto nivel; *Python*, y posteriormente realizar una limpieza de *tweets*. Debe señalarse que una vez obtenidos los *tweets* preprocesados, con el fin de aplicar el análisis de sentimientos, de manera que se genere una base de datos con *tweets* analizados. Por último, se aplicarán algoritmos de *machine learning* presentados en el capítulo anterior a la base de datos, a dichos algoritmos se le medirá su desempeño, con el propósito de obtener un sistema automático de análisis de sentimientos. Cada etapa es explicada a detalle a continuación.

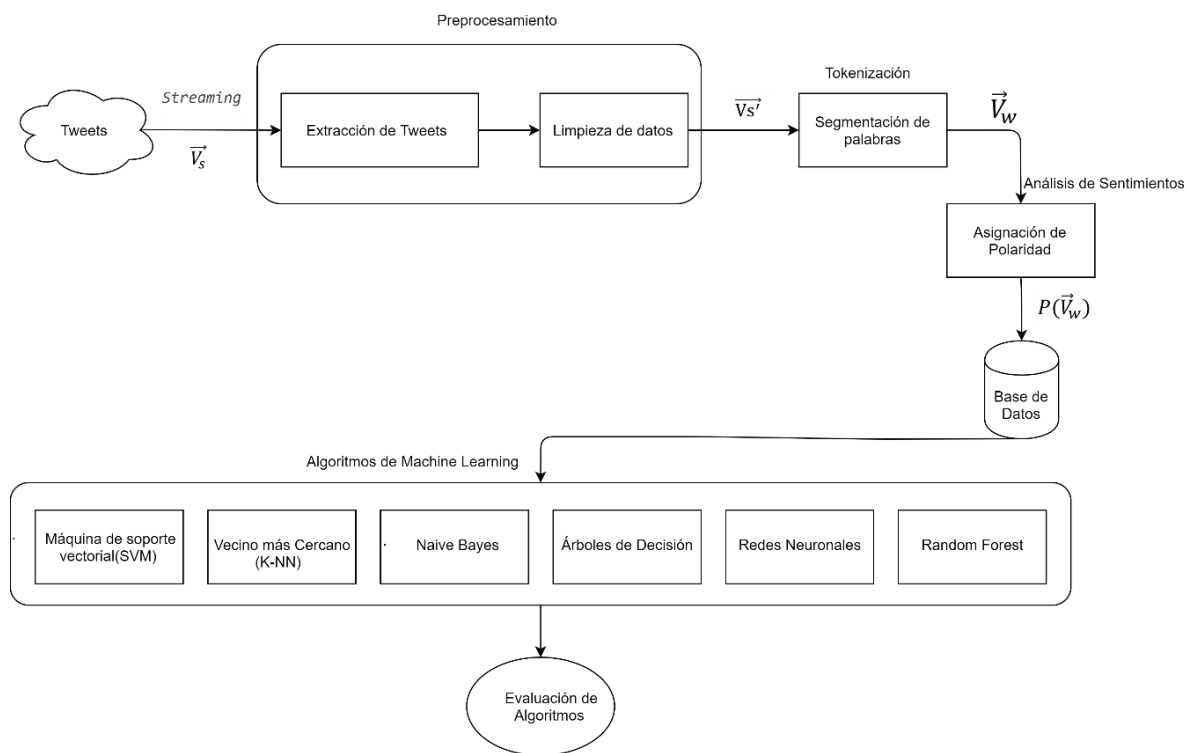


Figura 3.1: Diagrama de flujo de la metodología para el análisis de sentimientos de Tweets.

3.2 Datos de entrada

Los datos de entrada serán *Tweets*. Estos datos vienen codificados en una estructuras de datos ([char array]), mejor conocido como Notación de Objetos en JavaScript (JSON) [62].

La Figura 3.2 ilustra un elemento de datos de ejemplo, que es una cadena JSON estructurada que contiene información sobre un *tweet* y el usuario que lo publicó. Además, si el *tweet* es un *retweet*, el contenido original del *tweet* también se incluye en un campo "*retweeted_status*". Para hashtags, menciones de usuario y *URL* contenidas en el texto del *tweet*, se incluye un campo de "entidades" para proporcionar información detallada, como el ID del usuario mencionado y las *URL* expandidas [63].

```
{
  "text": "RT @sengineland: My Single Best... ",
  "created_at": "Fri Apr 15 23:37:26 +0000 2011",
  "retweet_count": 0,
  "id_str": "59037647649259521",
  "entities": {
    "user_mentions": [
      {
        "screen_name": "sengineland",
        "id_str": "1059801",
        "name": "Search Engine Land",
      }
    ],
    "hashtags": [],
    "urls": [
      {
        "url": "http://selnd.com/e2QPS1",
        "expanded_url": null
      }
    ]
  },
  "user": {
    "created_at": "Sat Jan 22 18:39:46 +0000 2011",
    "friends_count": 63,
    "id_str": "241622902",
    ...
  },
  "retweeted_status": {
    "text": "My Single Best... ",
    "created_at": "Fri Apr 15 21:40:10 +0000 2011",
    "id_str": "59008136320786432",
    ...
  },
  ...
}
```

Figura 3.2: Ejemplo de un Tweet en formato JSON [63].

De los *tweets* se extraerán las características más relevantes considerados para la investigación, como pueden ser; "*text*", "*created_at*", "*coordenadas*", "*lugar*". Para acceder a obtener *tweets*, es necesario crear una cuenta como desarrollador en *twitter*, y de esta manera obtener la *API* de *twitter*. La cuenta permite crear aplicaciones, por medio de las

cuales se proporciona la *API Key* y el *API Token*, de no ser así, ingresar a los datos de *Twitter* se torna muy complejo.

3.3 Preprocesamiento

En cuanto al preprocesamiento, se conforma de dos partes; la extracción y la limpieza de los datos. Por ende, es considerada una de las fases más importantes, dado que, es aquí donde se obtendrán los datos para ser procesados y clasificados. Es dado que, al fallar en esta fase, los resultados serán erróneos. Si se generan errores en esta fase, abre la posibilidad de arrojar un mal procesamiento y una alteración en el clasificador. La importancia del preprocesamiento se puede observar en la Tabla 3.1. Dado que el texto de un *tweet* sin preprocesamiento contiene más caracteres irrelevantes, como el usuario, y enlaces web.

Tabla 3.1: Comparativa entre Tweets sin preprocesamiento y con preprocesamiento.

Sin Preprocesamiento	Preprocesados
@MireyaQuintos señala que Zacatecas tiene mucho potencial, hay mucha experiencia, acción que no se ve en otras Entidades Federativas #ArchivosIZAI y #NoHayTiempoQuePerder https://t.co/0gdI5cx3Cd	señala que zacatecas tiene mucho potencial, hay mucha experiencia, acción que no se ve en otras Entidades Federativas archivosal y nohaytiempoqueperder
Pensaran que la foto es de #Italia, #Roma, #Venecia, pero no... Es mi hermoso #Zacatecas ♥ https://t.co/Auoa2HLMLn	pensaran que la foto es de italia, roma, venecia, pero no es mi hermoso zacatecas ♥
https://t.co/RQIKLQOCuF	<i>Tweet</i> Descartado
@chcrfd0 ??????????????????	<i>Tweet</i> Descartado

El procedimiento para llegar a la columna del procesamiento de *tweets* de la Tabla 3.1 se describe en las siguientes etapas.

3.3.1 Extracción

Con el objetivo de extraer *tweets* en tiempo real, se codifico un programa, en el cual se le agrego las condiciones necesarias para tener acceso a *Twitter*, cabe resaltar, que para que sea posible la conexión es necesario contar con la *API* de la plataforma antes mencionada.

Nuestro código tiene la capacidad de recolectar solo los *tweets* de interés, buscándolos por palabras clave y coordenadas.

Como ya se mencionó anteriormente, los *Tweets*, vienen codificados en JSON [62], cada *tweet* pasara por un filtro, dando como resultado, tener solo un vector con un objeto de todos los componentes de JSON ($\overrightarrow{V_{s1}}$). Cabe resaltar lo más importante de un *tweet*, es el contenido en el objeto de “*text*” ya que es parte fundamental para realizar el análisis de sentimientos. Ahora bien, será necesario verificar que el *tweet* no tenga habilitado la opción interna de texto extendido, esto lo hace *Twitter* por automático ya que si un usuario ingresa más de 140 caracteres se considera ya texto extendido, por ende, aparece una nueva estructura (“*extended_tweet*”) en JSON donde está el texto completo del *tweet*, por lo que se tendría que extraer esa propiedad del *tweet* diferente para que nos dé como resultado el texto completo.

3.3.2 Limpieza de datos y tokenización

La limpieza de los datos, en este caso, de nuestros *tweets* de interés es de suma importancia, referirnos a limpieza de *tweets*, es hablar de quitar caracteres que se encuentran solos, por ejemplo, si en un *tweet* aparece solo un comentario con un punto “.”, un *emoji*, o cualquier letra o palabra, ese tipo de *tweets* no dicen nada como tal por lo que es necesario quitarlos. *Python* cuenta con librerías para analizar texto (NLTK), por lo que para limpiar los *tweets* es necesario pasar a realizar la tokenización, un token corresponde a una palabra o signo de puntuación, por lo tanto, tokenizar se refiere a dividir una frase en cada palabra, con esto se identifica cantidad de palabras, tipo de palabra (adjetivos, verbos, etc), y por medio de *Python* poder eliminar lo que no tiene relevancia.

3.4 Análisis de sentimientos u opiniones

El análisis de sentimientos (AI de emoción), es el proceso automatizado de analizar datos de texto y clasificar las opiniones de los usuarios como positivas, neutrales o negativas. El análisis de sentimientos permite detectar el afecto en las conversaciones en línea, ayudando a comprender cómo se sienten los usuarios sobre productos, servicios, o estadías.

Por lo general, además de identificar la opinión, estos sistemas extraen tres atributos de la expresión, por ejemplo:

- I. Polaridad: si el hablante expresa una opinión positiva o negativa
- II. Asunto: de lo que se habla
- III. Titular de la opinión: la persona o entidad que expresa la opinión.

En este análisis se puede aplicar a tres diferentes niveles de alcance:

- I. El análisis a nivel de documento, obtiene el sentimiento de un documento o párrafo completo.
- II. El análisis a nivel de oración, obtiene el sentimiento de una sola oración.
- III. El análisis a nivel de sub-oración, obtiene el sentimiento de las sub-expresiones dentro de una oración.

3.4.1 Implementación

Dentro de las plataformas que nos permiten realizar el análisis de sentimientos, el lenguaje de programación *Python* cuenta con librerías de texto, de este modo permiten realizar análisis de sentimientos, tales como *Textblob* [64] y *Vader Sentiment* [65]. La primera se deriva de NLTK; que es una plataforma para crear algoritmos en *Python* para poder trabajar con datos de lenguaje humano [66], siendo NLTK la base para varias librerías. *Textblob* tiene características muy importantes para el procesamiento de textos, cuenta con su propio traductor de textos, dado que el lenguaje nativo del procesamiento de textos de esta librería es el inglés. *Textblob* cuenta con un rango de polaridad de sentimientos de -1 a 1, mientras que *Vader Sentiment* arroja a la vez que tan positivo, negativo y neutro es un texto. Además, *Vader Sentiment* proporciona un extra, el *compound*, que es una métrica que calcula la suma de todas las clasificaciones de léxico que se han normalizado entre -1 (más extremo negativo) y +1 (más extremo positivo).

Tal como estas dos librerías, existen más herramientas de software no libre para realizar el análisis de sentimientos. Casas [67], describe algunas de ellas, estas son:

- “*QuickSearch*; te da un resumen instantáneo de tu marca en línea. Es un buscador de redes sociales que ofrece cobertura extensa en *Facebook*, noticieros, *blogs* y foros”
- “*Hootsuite Insights*; Automáticamente analiza todas las plataformas de redes sociales noticieros, foros y blogs para revelar percepciones de influencers, historias, tendencias y sentimientos”.
- “*Textalytics*; es un motor de análisis de texto en lenguaje natural con múltiples funcionalidades, entre las que se encuentra el análisis de sentimientos en varios idiomas: español, inglés y francés. Además de analizar la polaridad asociada al texto, oración, concepto y entidad, es capaz de identificar otros aspectos relevantes, como la objetividad/subjetividad del mismo, así como el uso de la Ironía”.
- “*NCSU Tweet Visualizer*; Esta herramienta gratuita para análisis de sentimiento en *Twitter* es interesante. Escribe una palabra clave y *Tweet Visualizer* te enseña los *tweets* relacionados más recientes que han sucedido en la última semana”.
- “*RapidMiner*; Una plataforma de software de datos que permite minar texto para ayudar a las marcas a realizar análisis de sentimientos. Las reseñas y *posts* en redes sociales se pueden analizar al igual que publicaciones oficiales y documentos”.
- “*MeaningCloud*; El API de análisis de sentimiento implementa un análisis detallado y multilingüe de contenido de diferentes fuentes. Determinando si el sentimiento refleja de manera positiva, negativa o neutral – o si no es posible detectar. Las frases son identificadas y evaluadas según la relación entre ellas”.

- “*Google Prediction API*; *Google* ofrece una serie de documentos online para desarrolladores sobre cómo realizar análisis de sentimiento a través de su *API de Machine learning*”.

No obstante, el funcionamiento base de los softwares antes mencionados, se basa en librerías de análisis de sentimientos tales como de *Textblob*, y *Vader sentiment*. Las librerías como tal, nos ayudaran a crear nuestro propio código abierto, para una manipulación de datos especifica. De esta manera no dependeremos de un servicio de ningún software que no sea *Textblob* para realizar el análisis de sentimientos.

Por otra parte, centrándonos en la librería *Textblob*, y entender el funcionamiento es de suma importancia. La librería contiene una biblioteca de patrones [68], la cual contiene una gran cantidad de palabras, esta biblioteca está basada en los trabajos de Tom De Smedt y Walter Daelemans [69], como se muestra a continuación.

- “Los adjetivos tienen una polaridad (negativa / positiva, -1.0 a +1.0) y una subjetividad (objetiva / subjetiva, +0.0 a +1.0)”.
- “La fiabilidad especifica si un adjetivo fue etiquetado a mano (1.0) o inferido (0.7).”
- “Las palabras están etiquetadas por sentido, por ejemplo, ridículo (lamentable) = negativo, ridículo (humorístico) = positivo”.
- “La identificación de Cornetto (identificación de unidad léxica) y la identificación del conjunto de datos de Cornetto se refieren a la base de datos léxica de Cornetto para holandés”.
- “La identificación de WordNet se refiere a la base de datos léxica de WordNet3 para inglés”.
- “Las etiquetas de part-of-speech (pos), son un conjunto de etiquetas: NN = sustantivo, JJ = adjetivo, VB= verbo, etc.”.

Al ingresar el texto a *Textblob*, por automático este compara con los adjetivos almacenados en la biblioteca, y en base a lo anterior asigna una polaridad (positiva o negativa), por ejemplo, como se muestra en las siguientes sentencias en *Python* de la palabra “*great*”:

```
<word form="great" cornetto_synset_id="n_a-525317" wordnet_id="a-01123879"
pos="JJ" sense="very good" polarity="1.0" subjectivity="1.0" intensity="1.0"
confidence="0.9" />
```

```
<word form="great" wordnet_id="a-01278818" pos="JJ" sense="of major
significance or importance" polarity="1.0" subjectivity="1.0" intensity="1.0"
confidence="0.9" />
```

```
<word form="great" wordnet_id="a-01386883" pos="JJ" sense="relatively large in
size or number or extent" polarity="0.4" subjectivity="0.2" intensity="1.0"
confidence="0.9" />
```

```
<word form="great" wordnet_id="a-01677433" pos="JJ" sense="remarkable or out  
of the ordinary in degree or magnitude or effect" polarity="0.8" subjectivity="0.8"  
intensity="1.0" confidence="0.9" />
```

Por último, la biblioteca tiene varias sentencias de la misma palabra, ya que está diseñada para todos los casos posibles en donde se pueda utilizar la palabra, tanto para decir algo positivo o bien negativo, esa parte va especificada en “*sense*”. Al calcular el sentimiento de una sola palabra, *Textblob* utiliza el promedio, de *polarity*, *subjectivity* e *intensity*, dando como resultado lo siguiente:

```
Textblob(“great”).sentiment  
## Sentiment(polarity=0.8, subjectivity=0.75)
```

3.5 Caso de estudio: Zacatecas

Con el fin de validar la metodología propuesta se decidió realizar varios experimentos. En estos experimentos se restringieron a la ciudad de Zacatecas dado que es nuestro punto de interés ya que, al ser originario de este lugar, como surge la necesidad de conocer la opinión de las personas que visitan Zacatecas para investigaciones a futuras poder mapear la aprobación de las personas en lugares específicos (bares, parques, museos etc.), en este trabajo solo se clasificará de manera automática los comentarios positivos o negativos. En las siguientes sub secciones se describen las etapas de este caso de estudio.

3.5.1 Base de datos

Con el preprocesamiento y el análisis de sentimientos de cada *tweet*, se generó una base de datos, como se ejemplifica en la Tabla 3.2. Esta base de datos cual contiene 13,928mil *Tweets*, los cuales se utilizarán para entrenar algoritmos de aprendizaje profundo (*machine learning*).

Tabla 3.2: Base de Datos de los Tweets Recolectados.

Usuario	Ubicación	Coordenadas	Fecha	Tweet	Polaridad(Out)
@deCultura_r	NA	NA	Sat Oct 12 00:32:42	#MáscarasMexicanas ???? Máscas	0
@livejoss	NA	NA	Fri Oct 11 01:36:04	Plateros de Fresnillo vs Mineros de	0
@Ecodiarioza	NA	NA	Thu Oct 10 19:30:00	#Entorno Recorrerán el área metr	0
@rg_luis	Facultad De Ir	[[[-102.56546	Thu Oct 10 19:30:41	Zacatecas de nuevo	0
@Migue1CG	NA	NA	Thu Oct 10 19:34:31	Los tránsitos pendejos deberían m	0.416666667
@ElCharlieBr	NA	NA	Thu Oct 10 19:37:56	@discapacidadcom ¿¿¿Cómo es p	0.083333333
@ElCharlieBr	NA	NA	Thu Oct 10 19:41:24	@discapacidadcom @bienestarm	0.48828125
@Ecodiarioza	NA	NA	Thu Oct 10 19:45:00	#Entorno Debido al conflicto con	-0.041666667
@portal_miné	NA	NA	Thu Oct 10 19:46:10	Pierden 500 mdp por bloqueo a M	-0.166666667
@danielals10	NA	NA	Thu Oct 10 19:46:33	Ahí va la lista, empecemos a dar d	0
@PorticoOnli	NA	NA	Thu Oct 10 19:46:38	#PorSiNoLoViste #Zacatecas Aum	0
@PorticoOnli	NA	NA	Thu Oct 10 19:47:09	#HoyEnPortico: Aumentará tarifa	0.5
@gabiloo	NA	NA	Thu Oct 10 19:47:34	@Artglez @David_Digital Gracias.	0
@Pande_zac	NA	NA	Thu Oct 10 19:47:54	Se cuidan su piel, su alimentación	-0.1
@somosmoti	NA	NA	Thu Oct 10 19:50:18	#Michoacán #Morelos #Nayarit #	0
@rolandom	Zacatecas, Mé	[[[-102.85015	Thu Oct 10 19:50:55	Panorámica Zacatecas #zacatecas	0
@PorticoOnli	NA	NA	Thu Oct 10 19:52:12	#OrgulloZacatecano El Instituto de	0.2375
@gerardorguz	NA	NA	Thu Oct 10 19:52:39	@Cinemex pues no quería venir a	0
@vickypulido	NA	NA	Thu Oct 10 19:56:14	Delegación de alumnos y maestro	0
@rg_luis	Ingeniería en	[[[-102.56457	Thu Oct 10 21:24:09	De los lugares que más me gusta e	0.5
.
.
.
.

Esta base de datos cuenta con 6 características y 13,928 observaciones. La primera característica es el usuario, siguiendo con la ubicación, la cual es muy importante para detectar de qué lugar está publicando o bien si se refiere al lugar como tal en su *tweet*, esta característica cuenta con muchos NA, debido a que no todos los usuarios de *Twitter* tienen habilitado para compartir su ubicación. Nuestra tercera característica, muy relacionada con la anterior son las coordenadas de la ubicación, además de agregar la fecha de publicación. La característica más importante de nuestra base de datos son los *tweets*, y al final tenemos la polaridad o salida de nuestra base de datos. Cabe destacar que esta base de datos fue creada desde cero, con *tweets* georreferenciados al municipio de Zacatecas.

3.5.2 Adquisición de datos

Como ya se mencionó anteriormente, los datos de entrada son los *tweets*. Para la adquisición, se recurrió a realizar un algoritmo, utilizando el lenguaje de programación *Python* como herramienta principal. Nuestro algoritmo de adquisición cuenta con 3 fases:

- a) Fase 1: Conexión con la API de *Twitter*.
- b) Fase 2: Condiciones de Búsqueda.
- c) Fase 3: Adquisición y almacenamiento de los datos.

Para la fase 1, fue necesario crear una cuenta como desarrollador en la plataforma de *Twitter*. Dado que es un requisito de suma importancia, ya que esta cuenta te permite obtener los

accesos de una aplicación para conectarse a dicha plataforma, tal y como se muestra ver Figura 3.3.

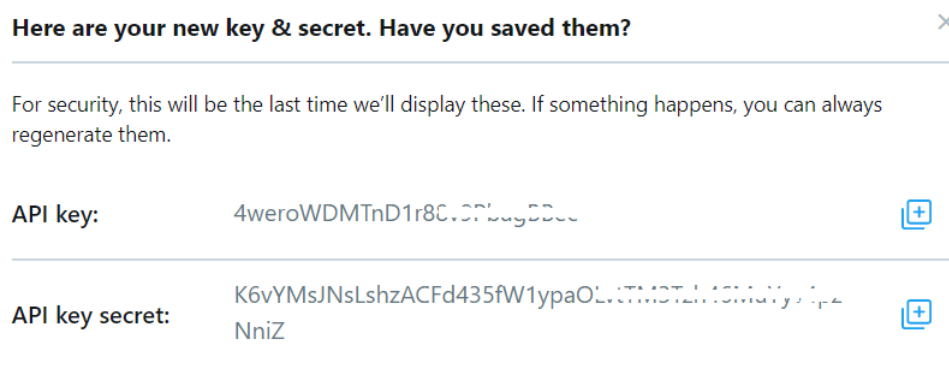


Figura 3.3: Keys y Tokens de Twitter.

Tweepy es una librería de *Python* utilizada específicamente cuando se trabaja con las API's de *Twitter*. Dicha librería contiene los requisitos para conectarse a la red social. Una vez aprobadas las credenciales de la API es posible interactuar con los datos(*tweets*).

La Fase 2, incluyó el establecimiento de las condiciones de búsqueda, para el desarrollo de este tema de investigación se restringió la búsqueda a solo tweets generados en la región de Zacatecas, así mismo se establecieron una lista de palabras clave (*Zacatecas*, *#zacatecas*, *zacatecas*, *#Zac*). Dichas condiciones fueron implementadas en el mismo algoritmo, código que se muestra en la Figura 3.4, la condición de búsqueda o filtro, contiene las palabras clave y coordenadas de los *tweets* que se desean extraer.

```
twitterStream = Stream(auth, listener(),tweet_mode='extended')
twitterStream.filter(locations=[-102.850156, 22.617916,-102.538543, 22.843297],
track = ['Zacatecas','#zacatecas','zacatecas','#Zac'])
```

Figura 3.4: Condiciones de búsqueda.

En la Figura 3.4, podemos observar las líneas de código implementadas para realizar la búsqueda de *tweets*. La primera línea de código, tiene como función mostrar el modo extendido de los *tweets*, este modo corresponde a mostrar si es que lo hay el texto mayor a 140 caracteres. En la segunda línea se colocó las coordenadas referentes al municipio de Zacatecas, y además las palabras clave mostradas en el *track*.

De saltar este paso, y recolectar los *tweets*, se extraerían todos los *tweets* que se estén publicando al instante, sin importar la procedencia de donde fueron escritos, por lo que no

solo nos interesan esos *tweets* procedentes de la zona, por lo que el filtro de coordenadas estará limitando a georreferenciar los datos y el filtro de palabras clave solo para obtener los referenciados a Zacatecas.

Para la última fase, es necesario guardar nuestros *tweets* recolectados, para esto se consideró el formato de texto .txt, para después analizar los datos recolectados. Antes de guardar los datos, se tomó a consideración guardar solo las características de importancias de los *tweets*, como lo es el texto extendido, las coordenadas, el lugar, fecha, y descartando todos aquellos *retweets*. Con estas características se guardaron los *tweets*, lo que le corresponde una línea a cada uno con sus respectivas características. Al tener el archivo .txt con todos los datos, es posible empezar a trabajar con ellos pasando así a la limpieza de los datos.

3.5.3 Preprocesamiento

En cuanto a la limpieza de los datos, la etapa más importante para realizar el análisis de datos. Es aquí donde se normalizan nuestros *tweets*, con esto nos referimos a quitar caracteres irrelevantes del texto. En la limpieza de los datos retomamos el archivo generado por el algoritmo de adquisición mencionado anteriormente, tomando la columna donde esta solo el texto y comenzar a analizar cada observación. Aquí se generó otro algoritmo, el cual contiene otras librerías como lo son: *sys*, *nltk*, *re*, *string*, *goslate*, *time*, *Textblob*, *os*. De estas librerías se usaron funciones específicas los cuales nos ayudaron a limpiar el texto tal como se muestran en el código de la Figura 3.5. Se comenzó con eliminar las urls, así como los hashtags (#) y las menciones de usuarios (@), de los *tweets*.

```
url = 'http[s]?://(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[*\(\)])|(?% [0-9a-fA-F][0-9a-fA-F]))+'
url2= '(www\.)?(?:[a-zA-Z]|[0-9]|[$-_@.&+]|[*\(\)])|(?% [0-9a-fA-F][0-9a-fA-F]))+'
emoticons= "[\&\-\.\√()=:;]+"
patron = re.compile('_|#|@|[\i?]|'+url+'|'+url2+'|'+emoticons)
entradaM=patron.sub('',entrada)
```

Figura 3.5: Declaración de caracteres a eliminar.

En el fragmento de código que mostramos en la Figura 3.5 podemos observar las variables declaradas como *url* y *url2*, las cuales contienen los caracteres posibles que hay después del *http*, así como de *www*, de esta manera eliminamos el contenido de un *url*. Una tercera variable la cual se declaró con el nombre *emoticons*, contiene los caracteres con los cuales se puede crear una carita feliz (☺) o una cara triste (☹) entre otros. De esta manera poder eliminar si es que existe algún *emoji*. Utilizando la función *re.compile* y *.sub* se eliminaron estos caracteres irrelevantes del texto que en el código corresponde a la variable *entrada*.

Para realizar el análisis de sentimientos, se estandarizo todo el texto a minúsculas, para realizarlo se recurrió a una función de la librería *textblob*, que convierte los caracteres del texto en minúsculas. Hecho lo anterior el texto pasa a revisión de cantidad de palabras, y es aquí donde se define una condición, si no contiene palabras se marca como texto descartado y salta a una nueva línea para analizar una nueva línea de texto. Si contiene más de 3 palabras entonces el texto pasa a ser analizado. Primero detectar el lenguaje, si es español, inglés u otro. Dependiendo del lenguaje que este escrito el texto; por ejemplo, si está en español, pasa a ser traducido en inglés y con esto pasa por la función de corrección del texto, así como después asignarle una polaridad de sentimiento. Si el texto viene en inglés, solo pasa a la corrección y asignación de polaridad sin ser traducido.

Con el texto en inglés, se sacan el tipo de palabras que contiene, como lo son verbos, adjetivos, pronombres, etc., se crea un contador el cual verifica el número total de palabras de cada tipo, para cada texto, y al final obtener características propias del texto, por ultimo cada texto entra a la asignación de polaridad, correspondiente a la librería *textblob*, especificando nos referimos a la función *sentiment.polarity*. Como último paso y con todas las características reunidas se genera una nueva base de datos la cual tiene características propias del texto.

3.6 Preparación de base de datos o procesamiento

En base al algoritmo de limpieza de datos, el cual nos arrojó una nueva base de datos, la cual contiene solo características propias del texto, esta base de datos, será usada para entrenar los algoritmos de aprendizaje supervisado. La base de datos contiene 13,939 observaciones, y 37 características. Entre las cuales se tienen el texto en español, en inglés, cantidad de palabras, cantidad de caracteres, y todos los *tags* del texto analizado. Entendemos como *tags*, a las etiquetas que tienen las palabras ver la Tabla 3.3.

Tabla 3.3: Etiquetas de texto o Tags de cada palabra en Python[70].

Tag	Significado	Tag	Significado
CC	Conjunción de coordinación	PRP\$	Pronombre posesivo
CD	Dígito cardinal	RB	Adverbio
DT	Determinador	RBR	Adverbio comparativo
EX	Existencial (allí)	RBS	Adverbio superlativo
FW	Palabra extranjera	RP	Particular
IN	Preposición/junta subordinación	TO	Particular de infinitivo
JJ	Adjetivo	UH	Interjección
JJR	Adjetivo comparativo	VB	Verbo
JJS	Adjetivo superlativo	VBG	Verbo tiempo pasado
LS	Marcador de lista	VBG	Verbo gerundio
MD	Modal (podría, podrá)	VBN	Verbo participio
NN	Sustantivo singular	VBP	Verbo presente
NNS	Sustantivo plural	VBZ	Verbo tercera persona
NNP	Sustantivo propio	WDT	wh-determiner
NNPS	Sustantivo propio plural	WP	Pronombre de pregunta (quién, qué)
PDT	Predeterminado	WP\$	Posesivo con pronombre (cuyo)
POS	Posesivo	WRB	Wh-abverb (donde, cuando)
PRP	Pronombre personal		

La base de datos fue importada a R-Studio, ya que este es un software de código abierto, en el cual se puede trabajar muy bien la parte estadística. Una vez importado, se agregaron más características a la base de datos, esto con el fin de tener características, mayor correlacionadas entre sí y con esto obtener un mejor resultado. Las nuevas características son estadísticas y corresponden a medidas de tendencia central entre la cantidad de los *tags* que le corresponden a cada observación. Teniendo como resultado una base de datos con 38, ya que de las 37 que se tenían se quitaron dos, dado que no contaban con ningún dato (*NA*). De las 38 características, una corresponde a nuestra salida, la cual consta de la polaridad de sentimiento del texto. La Figura 3.6 muestra fragmento del código realizado en R-Studio para obtener las características estadísticas antes mencionadas.

```

datos<-DataTWCompleto
datos<-datos[!is.na(datos$V3),]
Texto=datos$V1
Palabras=datos$V3
caracter=c()
media=c()
desviacion=c()
des_tipos=c()
varianza=c()
media_tipos=c()
for(i in seq_along(datos$V1)){
  caracter[i]=str_length(Texto[i])
  media[i]=mean(c(Palabras[i],caracter[i]))
  desviacion[i]=sd(c(Palabras[i],caracter[i]))
  varianza[i]=var(as.numeric(datos[i,c(4:36)]))
  media_tipos[i]=mean(as.numeric(datos[i,c(4:36)]))
  des_tipos[i]=sd(as.numeric(datos[i,c(4:36)]))}
caracter1=data.frame(caracter,media,desviacion,media_tipos,varianza,des_tipos)
datos=cbind(datos,caracter1)

```

Figura 3.6: Obtención de características estadísticas.

Después de realizar el código de la Figura 3.6, en el cual sacamos las medidas de tendencia central. Pasamos a preparar la base de datos, para así aplicar los algoritmos de aprendizaje supervisado. La base de datos esta desbalanceada, por lo que tenemos menos observaciones con una salida de -1 que corresponde a comentarios negativos, a las otras dos salidas, por lo que sacamos una muestra con todas las observaciones negativas (-1), teniendo un valor de 1919 observaciones negativas, 1919 positivas y 1919 neutras, teniendo un total 5757 observaciones.

El primer algoritmo que se realizo fue SVM, en conjunto con la técnica del *k-fold Cross Validation* que es una de las maneras más convenientes de seleccionar un modelo de aprendizaje automático frente a un problema o un conjunto de datos en particular. Ya que la Validación Cruzada aplica de manera iterativa cada modelo sobre todo el conjunto de entrenamiento tomando diversas (k) particiones, podemos afirmar que el resultado en promedio nos da una idea precisa de cuál es el desempeño del modelo de clasificación sobre el conjunto de datos entero, de modo que podemos caracterizar de manera adecuada cada algoritmo [71].

Se realizó la experimentación con tres valores de k, correspondiendo a 5,10 y 15, probando cada uno de los códigos con este valor y evaluando su desempeño como tal, además de tomar un 75% de los datos para entrenamiento, cabe destacar que la toma de este valor oscila entre el rango aplicado por varios autores en diversos trabajos al usar el análisis de datos, siendo este de 70% a un 80% para entrenamiento.

Al algoritmo de máquina de soporte vectorial se le adecuaron los parámetros que mejor resultados arrojaban, un *kernel* de tipo radial, y un tipo de clasificación tipo “*C-classification*”. Cabe destacar que el *kernel* y el tipo de clasificación se definen mucho según los datos a analizar, por lo que al realizar las pruebas correspondientes y dando como mejores resultados esos dos parámetros.

Para KNN se determinó un valor de $k=3$, dado que estamos clasificando en 3 clases como se mencionó anteriormente. Y de esta manera se probó el algoritmo, siendo este valor el único que se le modificó. Para el caso de Naive Bayes, Árboles de Decisión y Random Forest, se consideró los parámetros predefinidos de los algoritmos. Para realizar la predicción en todos los algoritmos utilizamos la función *predict*, con parámetro de clasificación de tipo clase.

3.7 Validación de polaridad

Como forma de validar la polaridad que se obtiene del análisis de sentimientos con la polaridad real del texto, se realizó una encuesta (Anexo II) a 10 personas con algunos tweets recolectados, de los cuales de forma aleatorio se tomaran 10 tweets negativos, 10 positivos y 10 neutros, se calculó la correlación entre la salida del algoritmo de análisis de sentimientos y la encuesta. Obteniendo como resultado una alta correlación entre el algoritmo y la encuesta, para esto podemos ver la Tabla 3.4 la cual muestra la correlación de la encuesta contestada por 5 personas diferentes.

Tabla 3.4: Correlación entre la polaridad arrojada por el algoritmo contra los resultados de las encuestas.

	Algoritmo	Encuesta1	Encuesta2	Encuesta3	Encuesta4	Encuesta5
Algoritmo	1.0000000	0.9891005	0.9965576	0.9609877	0.9649013	0.9690032

En el siguiente capítulo se muestran los resultados obtenidos de esta sección.

4. Resultados

Mediante el desarrollo del sistema para realizar el análisis en tiempo real de *tweets* descrito en el capítulo anterior, fue posible extraer específicamente comentarios referentes a Zacatecas. El algoritmo permaneció ejecutándose por 15 días el cual recolecto acerca de 13,928 *tweets*. De lo cual se pudo obtener un primer resultado, de los *tweets* recolectados se les aplico el análisis de sentimientos.

Tabla 4.1: Cantidad de Tweets etiquetados.

Negativos	Neutro	Positivos
-1	0	1
1931	6027	5970
13.86%	43.27%	42.86%

En la Tabla 4.1 podemos observar que de los *tweets* recolectados un 13.86% son negativos y un 42.86% son positivos, por lo que, en una primera instancia sabemos que en Zacatecas los comentarios de los usuarios tienden a ser de neutros a positivos.

En base a los *tweets* ya recolectados se elaboró una base de datos, de la cual en base al texto se obtuvo nuevas características las cuales nos permitiría aplicar un análisis más profundo. Las características mencionadas en el capítulo anterior, nos sirvieron para hacer las primeras pruebas. Aplicando algoritmos de *machine learning* para poder clasificar los *tweets* de manera automática. Aplicamos los algoritmos a la par de la técnica de validación cruzada. De lo cual se obtuvieron los resultados mostrados en la Tabla 4.4 la cual describe el algoritmo y su desempeño.

Tabla 4.2: Desempeño de los algoritmos con k-fold cross validation.

Algoritmo	Exactitud		
	k=5	k=10	K=15
SVM	0.5564061	0.5628974	0.5624258
KNN	0.5580284	0.5723903	0.5714641
Naive Bayes	0.4579648	0.4660582	0.4551603
Arboles de Decisión	0.5056781	0.5040544	0.4991934
Random Forest	0.5571022	0.5728383	0.5689170

Como sabemos, para que un algoritmo tenga un desempeño excelente, la exactitud debe de ser lo más cercano a uno. En la Tabla 4.2 podemos ver, que, para los 3 valores de k , la exactitud está por debajo del 0.6 en todos los algoritmos, por lo que los algoritmos no tienen el desempeño esperado.

Dado los resultados del desempeño de los algoritmos con la base de datos implementada, nos arroja que las características no describen bien nuestras salidas, por lo que, las características para un comentario negativo son muy parecidos a los datos de un comentario positivo, nuestras características no son muy descriptivas. Para solucionar esto se generaron 4 nuevas características de las cuales salieron del texto de los *tweets*. Las nuevas características son las siguientes; número de palabras negativas y positivas en un mismo comentario, así como la positividad y la negatividad, estas últimas dos características se calcularon sacando la polaridad de cada palabra positiva que aparece en el comentario, realizando una sumatoria de las polaridades, así mismo para las palabras negativas, se realizó el mismo proceso. Volviendo a replicar la experimentación para evaluación de los algoritmos, obteniendo los resultados de la Tabla 4.3.

Tabla 4.3: Desempeño de los algoritmos con k-fold cross validation con las nuevas características agregadas.

Algoritmo	Exactitud		
	k=5	k=10	K=15
SVM	0.956261	0.958101	0.9587948
KNN	0.6758719	0.6841715	0.6818743
Naive Bayes	0.892956	0.8872175	0.8945789
Arboles de Decisión	0.9691554	0.9705339	0.9693871
Random Forest	0.9751399	0.9739859	0.9749035

Al comparar los resultados de la Tabla 4.2 y la Tabla 4.3, se aprecia una gran mejora en el desempeño de los algoritmos, siendo *Random Forest* el algoritmo que mejor desempeño mostro, seguidos de Arboles de Decisión y Máquinas de Soporte Vectorial, sin embargo, KNN mostro no tener mucha mejora en comparación de la prueba pasada, esto se debe a que el algoritmo es dependiente del valor de k según las clases a clasificar, para nuestro caso $k=3$, es decir KNN no es tan automático como los otros algoritmos.

Probando los tres algoritmos con mayor desempeño, de manera individuales, pero esta vez sin usar la validación cruzadas se obtuvieron los resultados mostrados en la Tabla 4.4.

Tabla 4.4: Desempeño de algoritmos sin validación cruzada.

Algoritmo	Exactitud	P-Valor	Kappa
Random Forest	0.977	<2.2e-16	0.9655
Arboles de Decisión	0.9735	<2e-16	0.9603
SVM	0.9551	<2.2e-16	0.9327

La Tabla 4.4 nos muestra nuevamente que el algoritmo de *Random Forest* sigue dando mejor desempeño, además de observar que los tres algoritmos tiene un P_Valor (es una medida directa de lo verosímil que resulta obtener una muestra) igual y con un grado de concordancia entre las observaciones relativamente bueno (Kappa) ya que se acerca a un valor igual 1. Asimismo, en la Figura 4.1 podemos observar la matriz de confusión correspondiente al rendimiento del mejor algoritmo presentado en la Tabla 4.5.

		Actual		
		-1	0	1
Predicción	-1	33.37%	0%	0.68%
	0	0.46%	32.68%	0%
	1	0.86%	0%	31.65%

	Clase: -1	Clase: 0	Clase: 1
Sensibilidad	0.9619	1.0000	0.9700
Especificidad	0.9850	0.9932	0.9872
Precisión	0.9715	0.9861	0.9734

Figura 4.1: Rendimiento de la matriz de confusión en el conjunto de datos test del algoritmo Random Forest.

En la Figura 4.1 se presentan más de las métricas de evaluación del algoritmo, siendo estas obtenidas por la matriz de confusión del conjunto de datos de prueba o datos con los que el algoritmo no conoce, prediciendo con un alto porcentaje a cada clase correspondiente y teniendo un error de fallo acumulado de 2%, obteniendo excelentes resultados en el rendimiento del algoritmo de *Random Forest*.

A continuación, se presentan las curvas ROC de los algoritmos con mejor desempeño, mostrados anteriormente en la Tabla 4.4. Cabe mencionar que para calcular una curva ROC para multiclases, donde el número de clases es mayor a dos. Se calcula las curvas ROC

independiente para cada par de clases, teniendo al final una combinación de todas las clases contra todas las clases. Por ejemplo, de tener 3 clases tendremos al final una cantidad de curvas ROC= $3C2 = 3$, siendo C la fórmula de combinaciones entre ellas, véase dichas curvas ROC en la Figura 4.2.

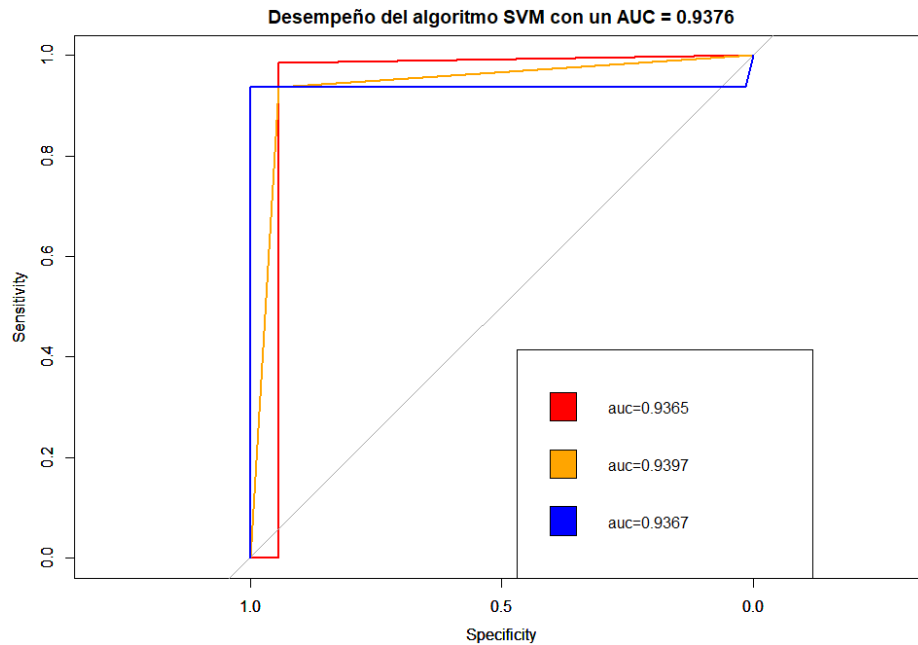


Figura 4.2: Curvas ROC del desempeño del algoritmo SVM, donde el color rojo= curva ROC de las clases [-1,1], la naranja= curva ROC de las clases [-1,0] y azul= curva ROC de las clases [0,1].

Como se mencionó anteriormente, otra de las métricas de gran ayuda para determinar el desempeño de un algoritmo son sus curvas ROC, en las cuales la métrica que sobresale es el Área Bajo la Curva (AUC), teniendo como representación de un excelente resultado cuando el AUC=1, las cuales como ya sabemos se calculan utilizando dos de las métricas que se obtienen de la matriz de confusión, en la Figura 4.2 se observa el comportamiento de estas curvas en el algoritmo SVM, las tres curvas presentan un AUC casi igual entre ellas, además de ser un valor arriba de 0.9 lo cual se puede considerar como un valor aceptable. Para obtener el desempeño del algoritmo como métrica de evaluación AUC, el promedio de los AUC de las curvas ROC procedentes de las clases que el algoritmo clasificó, nos dará el desempeño general del algoritmo, en este caso observamos tener un valor de 0.9376.

Para hacer una comparativa de los algoritmos con mejor desempeño, se calculó las curvas ROC de los algoritmos restantes, las cuales se presentan en las Figuras 4.3 y 4.4, en ambas Figuras podremos observar junto con la anterior como entre mejor desempeño tiene el algoritmo, estas curvas tienen una forma rectangular el cual tiene a tener el área igual a 1.

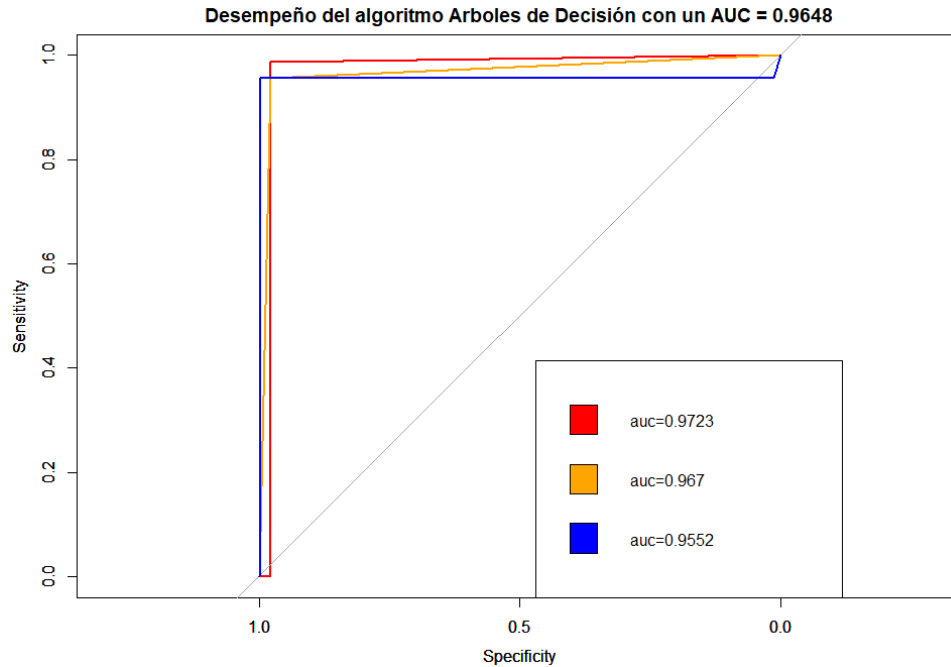


Figura 4.3: Curvas ROC del desempeño del algoritmo Arboles de Decisión, donde el color rojo= curva ROC de las clases [-1,1], la naranja= curva ROC de las clases [-1,0] y azul= curva ROC de las clases [0,1].

La Figura 4.3 muestra el desempeño del algoritmo Arboles de Decisión, que al ser comparado con la Figura 4.2, se puede apreciar una mejora en las curvas, asimismo, en la mejora del AUC del algoritmo, esto se debe a que en general el algoritmo de Arboles de Decisión presenta ser mejor que SVM. Finalmente, en la Figura 4.4 presentamos las curvas ROC correspondientes al algoritmo con mejor rendimiento de la investigación, obteniendo un valor AUC=0.9703, dicho valor al igual que en la exactitud del algoritmo tiende a acercarse más a tener un valor igual a 1.

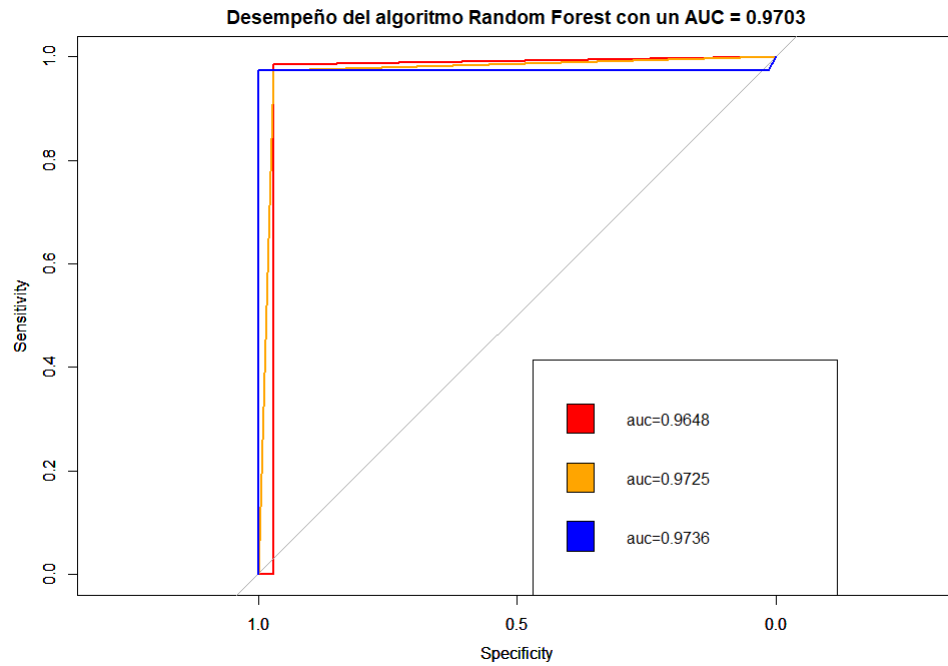


Figura 4.4: Curvas ROC del desempeño del algoritmo Random Forest, donde el color rojo = curva ROC de las clases [-1,1], la naranja= curva ROC de las clases [-1,0] y azul= curva ROC de las clases [0,1].

5. Conclusiones

De la presente investigación se puede concluir que, en un sistema de análisis de sentimientos en comentarios de *Twitter*, está compuesto de fases importantes, entre ellas está el preprocesamiento y la extracción de características, de estas influye mucho para que el algoritmo tenga el rendimiento adecuado. El omitir alguna de las fases afecta directamente en el rendimiento de los algoritmos de *machine learning*. Como ya se mencionó a lo largo del documento, esta investigación se enfocó en la extracción de *tweets*, el preprocesamiento y procesamiento de los *tweets*, utilizando los algoritmos de aprendizaje supervisado, los cuales algunos de ellos mostraron desempeños satisfactorios.

Los objetivos establecidos en esta investigación se cumplieron satisfactoriamente:

- Se desarrolló un sistema que permite analizar el *streaming* en tiempo real, mediante el lenguaje de programación Python.
- Se extrajo *tweets* específicamente con comentarios referentes al estado de Zacatecas. Dichos *tweets* están referenciados por ubicación (no importa lo que diga el *tweet*) y por palabra de búsqueda “Zacatecas”.
- Se creó una base de datos con los *tweets* obtenidos de los cuales se extrajo características para un análisis más profundo.
- Se clasificaron los *tweets*, mediante técnicas de inteligencia artificial y análisis de sentimientos. Obteniendo una base de datos con características del texto del *tweet*.
- Se obtuvo un accuracy de 0.97 a 0.95 en los algoritmos de aprendizaje supervisado, por lo que se puede clasificar adecuadamente y de forma automática los *Tweets*.
- Se obtuvo una correlación de 0.97 entre la encuesta aplicada a sujetos y los obtenidos automáticamente por el sistema.

La correlación entre la encuesta y los obtenidos por el sistema sugieren una alta concordancia entre el texto escrito y la opinión obtenida por el sistema, validando así el resultado obtenido automáticamente a través de la metodología propuesta

De los resultados obtenidos en esta investigación se puede concluir que es posible clasificar de manera automática los *tweets*, la limpieza de los datos y una extracción adecuada de características del propio texto son fundamentales para obtener buenos resultados. Dado que, si no se extraen las características adecuadas los algoritmos no presentan el rendimiento esperado. En el capítulo 4, se comprobó que el rendimiento de algunos algoritmos mejoró considerablemente de ser malo paso a ser muy bueno, esto debido a las características tomadas para realizar las pruebas.

Las características agregadas a las cuales se les nombro como positividad y negatividad en un mismo texto, resultaron ser clave para obtener el resultado que se esperaba, dado que un texto en este caso un comentario puede tener palabras negativas y también positivas se optó por sacar estas características y realizar las pruebas, lo cual se obtuvo una respuesta muy buena.

De los algoritmos, en la primera prueba se obtuvo un rendimiento muy bajo, sin embargo, en la segunda prueba con nuevas características se observó una mejor, entre los algoritmos con una mejora en su desempeño se encuentra *Random Forest*, SVM y Árboles de Decisión. El algoritmo KNN no presentó una mejora significativa, y esto como se comentó anterior mente se debe a que no es tan automático como los anteriores algoritmos, ya que depende del valor de “k”, que se le asigne al ejecutarlo.

Una contribución muy importante es que este sistema es automático en la clasificación del sentimiento, es decir, no se necesita realizar ningún análisis semántico, ya que las entradas del algoritmo son las palabras dentro del mensaje enviado por el usuario, estas palabras son usadas para generar las entradas necesarias para realizar la predicción. Lo anterior al no necesitar un análisis semántico propio de cada lengua o idioma, permite que la implementación en otros idiomas sea sencilla, transparente y robusta ante singularidades propias de cada lenguaje.

5.1 Productos de la investigación

Durante la realización de la presente investigación, se obtuvieron algunos productos, entre ellas la creación de la empresa ZTMAR S.A de C.V (ver Figura 5.1), la cual está enfocada a la creación de ciudades inteligentes, en donde el tema de investigación encaja perfectamente.



Figura 5.1: Logo empresa ZTMAR.

Así mismo, la presentación de un poster en el Concejo de Ciencia y Tecnología Zacatecas con el nombre de “Implementación de Análisis de Datos y Big Data para la Creación de Turismo Inteligente” (ver Figura 5.2), el cual se deriva de la investigación enfocada hacia la

creación de turismo inteligente, la cual tiene como base el análisis de sentimientos de comentarios turísticos en el estado.



Implementación de Análisis de Datos y Big Data para la Creación de Turismo Inteligente

Luis Carlos Reveles-Gómez*, Huizilopoztlil Luna-García*, José María Celaya-Padilla*, Joyce Selene Anaid Lozano-Aguilar**
luisCarlosreveles@gmail.com, hlugar@uaz.edu.mx, jose.celaya@uaz.edu.mx, joyce_lozag@uaz.edu.mx
*Maestría en Ciencias del Procesamiento de la Información, **Maestría en Ciencias de la Ingeniería.

Figura 5.2: Poster presentado en jornadas de investigación.

La publicación del artículo “Desarrollo de Prototipo de Aplicación Móvil para *Smart Tourism* basado en Diseño Centrado en el Usuario” (ver Figura 5.3), el cual se deriva de la investigación de la presentación de poster en las jornadas de investigación, dicho artículo tiene un enfoque hacia la creación de una aplicación móvil basada en turismo inteligente, dicha aplicación tiene como base el análisis de sentimientos de comentarios turísticos en el estado así como un mapeo de los sentimientos en tiempo real.

Desarrollo de Prototipo de Aplicación Móvil para Smart Tourism basado en Diseño Centrado en el Usuario

Luis C. Reveles-Gómez¹, Huizilopoztli Luna-García¹, José M. Celaya-Padilla², Hamurabi Gamboa-Rosales¹, Jorge I. Galván-Tejada¹, Carlos E. Galván-Tejada¹, José G. Arceo-Olague¹, Valeria Maeda-Gutiérrez¹, Joyce S. A. Lozano-Aguilar¹

¹ Centro de Investigación e Innovación Automotriz de México (CIAM), Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, 98000 Zacatecas, Zac, México.

² CONACYT – Universidad Autónoma de Zacatecas, Jardín Juárez 147, Centro, 98000 Zacatecas, Zac, México.

{luiscarloreveles, hlugar, jose.celaya, hamurabigr, gatejo, ericgalvan, arceojg, valeria.maeda, joyce_lozag}@uaz.edu.mx.

Resumen. En este artículo, se presenta la implementación de la norma ISO 9241-210:2010 (Human Centred Design for Interactive Systems) para el desarrollo de una aplicación móvil con el fin de fortalecer la experiencia del usuario al momento de utilizar la aplicación móvil in situ. Siguiendo las fases que la norma dicta para el desarrollo y evaluación de software y hardware con el propósito de obtener un prototipo funcional, y al término del proceso un producto. La implementación de la norma permitió generar un prototipo inicial validado por usuarios reales (turistas), por lo que, para un trabajo futuro se llevará a cabo el uso de técnicas de inteligencia artificial (AI) y análisis de datos, estas mismas, complementarán este trabajo, dando como resultado una aplicación para Smart Tourism completamente validada y funcional. Cabe destacar que el propósito es usar el Diseño Centrado en el Usuario (DCU), logrando así un prototipo de alta fidelidad.

Palabras Clave: Prototipos de Aplicación Móvil, Diseño Centrado en el Usuario, Destinos Turísticos Inteligentes, Smart Tourism.

Figura 5.3: Artículo Publicado en V Jornadas Interacción Humano Computadora 2019.

Asimismo, tiene la capacidad de mapear en tiempo real los comentarios buenos y malos de los lugares turísticos en Zacatecas, con la posibilidad de mandar notificaciones de los lugares con mejor aceptación por otros usuarios.

5.2 Trabajos futuros

Como trabajo futuro, se puede seguir trabajando con esta base de datos y mapear los comentarios positivos y negativos en tiempo, así como aplicar otros algoritmos como lo son las redes neuronales y redes neuronales convolucionales, además de extender el proceso no solo a Zacatecas, si no a cualquier parte del país, o bien aplicar el mismo proceso para algún producto o tema en particular, dado que con los resultados de esta investigación solo basta el texto para poder trabajar en ello y replicar el experimento.

Referencias

- [1] C. Arcila-Calderón, F. Ortega-Mohedano, J. Jiménez-Amores, and S. Trullenque, "Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático/ Supervised sentiment analysis of political messages in Spanish: Real-time classification of tweets based on machine learnin," *El Prof. la Inf.*, vol. 26, no. 5, pp. 973–982, 2017, doi: 10.3145/epi.2017.sep.18.
- [2] P. D. Leonardo, "Análisis de sentimientos: aplicación sobre textos en redes sociales," Instituto Tecnológico De Buenos Aires, Trabajo Integrador, 2019.
- [3] Instituto Nacional de Estadística y Geografía, "Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en los Hogares (ENDUTIH)," *Inegi*, pp. 1–18, 2019.
- [4] A. Hernández Rios, "Unidad Temática 1: Comunicación Verbal y no Verbal," pp. 1–10, 2006.
- [5] D. V. Santos, *Comunicación Oral y Escrita*. 2012.
- [6] H. Hütt, "Las redes sociales: una nueva herramienta de difucion.," *Reflexiones*, vol. 91, no. 2, pp. 121–128, 2012, doi: 10.15517/rr.v91i2.1513.
- [7] R. Chari, "El Big Data y la publicidad personalizada en redes sociales," Universidad de Valladolid, Tesis Licenciatura, 2020.
- [8] I. H. Ting, H. J. Wu, and P. S. Chang, "Analyzing multi-source social data for extracting and mining social networks," *Proc. - 12th IEEE Int. Conf. Comput. Sci. Eng. CSE 2009*, vol. 4, pp. 815–820, 2009, doi: 10.1109/CSE.2009.418.
- [9] D. M. R. Ramírez, S. O. V. Núñez, E. M. Rojas, and H. B. R. Moreno, "Business intelligence and BigData," *Iber. Conf. Inf. Syst. Technol. Cist.*, vol. 2019-June, no. June, pp. 19–22, 2019, doi: 10.23919/CISTI.2019.8760628.
- [10] Y. Min Shum, "Situación digital, Internet y redes sociales México 2020," 14-Mar-2020. [Online]. Available: <https://yiminshum.com/social-media-mexico-2020/>. [Accessed: 09-Nov-2020].
- [11] J. Dorsey, "Twitter by the Numbers: Stats, Demographics & Fun Facts," 2020. [Online]. Available: <https://www.omnicoreagency.com/twitter-statistics/>. [Accessed: 09-Jan-2020].
- [12] A. Cardoso, L. Talame, M. Amor, and C. Neil, "Minería de Opiniones : Análisis de Sentimientos en una Red Social," pp. 1–5, 2019.
- [13] D. Hinestroza Ramírez, "El Machine Learning A Través De Los Tiempos, Y Los Aportes A La Humanidad," Universidad Libre Seccional Pereira, 2018.
- [14] H. Chaviano, "Técnicas de aprendizaje supervisado y no supervisado para el aprendizaje automatizado de computadoras," in *Memorias del primer Congreso Internacional de Ciencias Pedagógicas: Por una educación integral, participativa e incluyente*, Bolivia, 2015, pp. 549–564.
- [15] C. Colón Ruiz, I. Segura Bedmar, P. Martínez Fernández, and C. Colón Ruiz,

- “Sentiment analysis on health domain: analyzing patient comments on drugs,” *Proces. del Leng. Nat.*, no. 63, pp. 15–22, 2019, doi: 10.26342/2019-63-1.
- [16] Q. M. Morales-Ballagas, I. Cervantes-Alonso, and A. Rodríguez-Fernández, “LAS REDES SOCIALES EN LA HOTELERÍA,” *Doc. Espec.*, pp. 1063–1086, 2019.
- [17] C. M. Giraldo Cardona and S. M. Martínez María-Dolores, “Análisis de la actividad y presencia en facebook y otras redes sociales de los portales turísticos de las Comunidades Autónomas españolas,” *Cuad. Tur.*, no. 39, p. 239, 2017, doi: 10.6018/turismo.39.290521.
- [18] M. Guitart Roch, “Turismo y sostenibilidad en una ciudad inteligente,” Universidad Abierta de Cataluña, 2020.
- [19] S. F. Cornejo Cruz, “Proyecciones para el mejoramiento de la calidad de los servicios turísticos en la ciudad de Quevedo, año 2019.,” Universidad Técnica De Babahoyo, 2020.
- [20] J. F. Esteban, “Análisis de la reputación de un destino turístico en las redes sociales,” pp. 1–8, 2015.
- [21] C. Candale, “Las características de las redes sociales y las posibilidades de expresión abiertas por ellas. La comunicación de los jóvenes españoles en Facebook, Twitter e Instagram.,” *Colindancias Rev. la Red Hisp. Eur. Cent.*, no. 8, pp. 201–218, 2017.
- [22] F. N. Machado, “Análisis de sentimientos basado en opiniones turísticas,” Universidad de la Laguna, 2018.
- [23] J. C. Sobrino Sande, “Análisis de sentimientos en Twitter,” Universitat Oberta de Catalunya, 2018.
- [24] M. A. Amores Fernandez, “Detección de la polaridad de las opiniones basada en nuevos recursos léxicos,” p. 123, 2016.
- [25] E. Vallés, P. Rosso, A. Locoro, and V. Mascardi, “Análisis de Opiniones con Ontologías,” no. 41, pp. 29–37, 2010.
- [26] M. Afzaal and M. Usman, “A novel framework for aspect-based opinion classification for tourist places,” *10th Int. Conf. Digit. Inf. Manag. ICDIM 2015*, no. Icdim, pp. 1–9, 2016, doi: 10.1109/ICDIM.2015.7381850.
- [27] Zacatecas, “Zacatecas Deslumbrante cierra 2017 con cerca de 1 mil 280 mdp de derrama: Secretario Eduardo Yarto - Gobierno del Estado de Zacatecas,” 2017. [Online]. Available: <https://www.zacatecas.gob.mx/zacatecas-deslumbrante-cierra-2017-con-cerca-de-1-mil-280-mdp-de-derrama-secretario-eduardo-yarto/>. [Accessed: 09-Jan-2020].
- [28] Á. Román, “Estudio de Competitividad Turística del Destino Zacatecas,” Zacatecas, 2016.
- [29] C. H. Miranda, J. Guzmán, and D. Salcedo, “Minería de opiniones basado en la adaptación al español de ANEW sobre opiniones acerca de hotels,” *Proces. Leng. Nat.*, vol. 56, no. January, pp. 25–32, 2016.
- [30] C. Inc, “Crear una aplicación de análisis de opinión.” [Online]. Available:

<https://www.cloudera.com/tutorials/building-a-sentiment-analysis-application.html>.
[Accessed: 09-Nov-2019].

- [31] R. E. López, “Machine Learning - Libro online de IAAR.” [Online]. Available: <https://iaarbook.github.io/ML/>. [Accessed: 09-Nov-2019].
- [32] S. Marín García, “Ética E Inteligencia Artificial,” 2019.
- [33] L. Rouhiainen, *Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro*. Barcelona, 2018.
- [34] A. Serna A., E. Acevedo M., and E. Serna M., “Principios de la Inteligencia Artificial en las Ciencias Computacionales,” in *Desarrollo e innovación en ingeniería*, 2017, pp. 161–172.
- [35] M. Villén, “Big Data Analytics y la inteligencia Artificial,” 14-Apr-2019. [Online]. Available: <https://www.caminosmadrid.es/9938-2>. [Accessed: 09-Dec-2020].
- [36] C. Vegega, P. Pytel, and M. F. Pollo C., “Método basado en el emparrillado para evaluar los datos aplicables para entrenar algoritmos de aprendizaje automático,” in *Desarrollo e innovación en ingeniería*, 2017, pp. 106–138.
- [37] T. Baviera, “Técnicas para el análisis del sentimiento en Twitter,” *Rev. Dígitos* 1.3, pp. 33–50, 2016.
- [38] G. Ruiz Manosalva, “Modelo de análisis de datos utilizando técnicas de aprendizaje supervisado y no supervisado, para identificar patrones en la información generada por los pacientes, sometidos a juegos diseñados como un instrumento de apoyo terapéutico,” Universidad Jorge Tadeo Lozano, 2019.
- [39] X.-D. Zhang, “Machine Learning,” in *A Matrix Algebra Approach to Artificial Intelligence*, Singapore: Springer Singapore, 2020, pp. 223–440.
- [40] C. García and I. Gómez, “Algoritmos de aprendizaje: knn & kmeans,” *Univ. Carlos III Madrid*, 2006.
- [41] A. Gutiérrez, “Base de Datos clave,” *Cent. Cult. Itaca S.C.*, p. 36, 2010.
- [42] F. J. Martínez, “Se refiere al proceso de eliminar aquellas características que no representan información significativa para el proceso de la creación de modelos predictivos basados en técnicas de inteligencia artificial. También, en esta etapa es posible reducir la cant,” Universidad de La Rioja, 2003.
- [43] J. Zamorano Ruiz, “Comparativa y análisis de algoritmos de aprendizaje automático para la predicción del tipo predominante de cubierta arbórea,” Universidad Complutense De Madrid, Tesis Maestría, 2018.
- [44] M. Arriagada, “Comparación de métricas de distancia en el algoritmo K-Vecinos Más Cercanos para el problema de Reconocimiento Automático de Dígitos Manuscritos,” Pontificia Universidad Católica De Valparaíso, Tesis Licenciatura, 2015.
- [45] A. Moujahid, I. Inza, and P. Larrañaga, “Clasificadores K-NN.”
- [46] D. R. Tobergte and S. Curtis, *Machine learning with R*, Second Edi., vol. 53, no. 9. 2013.
- [47] K. Sánchez, “Monitoreo de salud estructural empleando análisis de componentes

principales con árboles de decisión y máquinas de soporte vectorial,” Centro De Investigación Y De Estudios Avanzados Del Instituto Politécnico Nacional, Tesis Maestría, 2015.

- [48] P. Larrañaga, I. Inza, and A. Moujahid, “Árboles de Clasificación.”
- [49] L. Breiman, “Random Forests,” *Machinelearning202.Pbworks.Com*, pp. 1–35, 1999.
- [50] A. Cutler, D. R. Cutler, and J. R. Stevens, “Ensemble Machine Learning,” *Ensemble Mach. Learn.*, 2012, doi: 10.1007/978-1-4419-9326-7.
- [51] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1201/9780429469275-8.
- [52] R. F. Medina-Merino and C. I. Ñique-Chacón, “Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python,” *Interfases*, vol. 0, no. 010, p. 165, 2017, doi: 10.26439/interfases2017.n10.1775.
- [53] J. Areli and T. Barrera, “Redes Neuronales.”
- [54] Y. Gala Garcia, “Algoritmos SVM para problemas sobre big data,” Universidad Autónoma de Madrid, 2013.
- [55] “Máquina de vectores de soporte (SVM) - MATLAB & Simulink.” [Online]. Available: <https://la.mathworks.com/discovery/support-vector-machine.html>. [Accessed: 03-Mar-2020].
- [56] L. J. M. Camaré, “Aprendizaje Automático a partir de Conjuntos de Datos No Balanceados y su Aplicación en el Diagnóstico y Pronóstico Médico,” INSTITUTO NACIONAL DE ASTROFÍSICA, OPTICA Y ELECTRÓNICA, 2008.
- [57] P. Garg, “Sentiment Analysis of Twitter Data using NLTK in Python,” Thapar University, 2016.
- [58] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, and E. A. Villaseñor, “A case study of Spanish text transformations for twitter sentiment analysis,” *Expert Syst. Appl.*, vol. 81, pp. 457–471, 2017, doi: 10.1016/j.eswa.2017.03.071.
- [59] J. A. Cerón-Guzmán and S. de Cali, “Classifier Ensembles That Push the State-of-the-Art in Sentiment Analysis of Spanish Tweets,” *CEUR Workshop Proc.*, vol. 1896, pp. 59–64, 2017.
- [60] J. A. Diaz-Garcia, M. D. Ruiz, and M. J. Martin-Bautista, “Minería de Opinión no Supervisada en Twitter,” *XVIII Conf. la Asoc. Española para la Intel. Artif. (CAEPIA 2018)*, pp. 1023–1028, 2018.
- [61] R. Germán, T. Sebastián, G. Pablo, L. M. Emilia, P. Andrés, and C. Damián, “Técnicas de Análisis de Sentimientos Aplicadas a la Extracción de Opiniones en el Lenguaje Español .”
- [62] “JSON.” [Online]. Available: <https://www.json.org/json-es.html>. [Accessed: 20-Feb-2020].
- [63] X. Gao *et al.*, “Supporting a Social Media Observatory with Customizable Index Structures: Architecture and Performance,” in *Cloud Computing for Data-Intensive Applications*, Springer New York, 2014, pp. 401–427.

- [64] S. Loria, "textblob Documentation," 2015.
- [65] C. J. Hutto and E. E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14).", *Proc. 8th Int. Conf. Weblogs Soc. Media, ICWSM 2014*, 2014.
- [66] "Kit de herramientas de lenguaje natural - documentación NLTK 3.5b1." [Online]. Available: <https://www.nltk.org/>. [Accessed: 09-Mar-2020].
- [67] A. Casas, "Sistema de análisis automático de sentimientos basado en procesamiento del lenguaje natural," Escuela Politécnica Superior de la Universidad Carlos III de Madrid, 2014.
- [68] S. Bump, "TextBlob: Simple, Pythonic, text processing--Sentiment analysis, part-of-speech tagging, noun phrase extraction, translation, and more." [Online]. Available: <https://github.com/sloria/TextBlob>. [Accessed: 20-Feb-2020].
- [69] T. De Smedt, "Modeling Creativity: Case Studies in Python," no. August, 2014.
- [70] "Análisis de Sentimientos con TextBlob y VADER en Python." [Online]. Available: <https://unipython.com/analisis-de-sentimientos-con-textblob-y-vader/>. [Accessed: 29-Sep-2020].
- [71] R. Delgado, "RPubs - Introducción a la Validación Cruzada en R," 18-Jul-2018. [Online]. Available: <https://rpubs.com/rdelgado/405322>. [Accessed: 04-Jun-2020].

Anexos

Anexo I

Códigos empleados en la investigación en el siguiente link.

https://drive.google.com/drive/folders/1CqR0emiqTP6skVMY8pYbcN_QZWBumENf?usp=sharing

Anexo II

Encuesta aplicada a 10 personas

UNIVERSIDAD AUTÓNOMA DE ZACATECAS
"Francisco García Salinas"

MAESTRÍA EN CIENCIAS DEL PROCESAMIENTO DE LA INFORMACIÓN

De los siguientes Tweets indicar con una Pos (Positivo), Neg (Negativo) o Neu (Neutro) como usted considere el mensaje:

Ejemplos de Tweets:

Positivo: Hoy es un día maravilloso! Tengo tanta felicidad, q nada ni nadie la pueden opacar! ¡IDIOTAS lo que piensan lo contrario! :) (5)

Negativo: Ya ni me llama la atención salir :((7)

Neutro: Maromas, en maromas se les va la vida (7)

Num.	Tweet	Polaridad
14231	Quien salva una vida, salva al mundo entero Sa ludo y Felicito a todos estos profesionistas que brindan día a día asistencia y apoyo para cuida r nuestra #Salud y salvar vidas	Pos
14888	Muy buen ejemplo d un marica, cuando no pue de o la ve perdida, empieza a echar culpas. As hhhh, a quien me recordoooo?????	Neg
2105	Ups, bueno ya me tocará conocer lo bueno de l a comida y lugares de Zacatecas	Neu
15762	te ves preciosa, espectacular y súper sexy en E xpreso te amo me encantas	Pos
4361	Zacatecas 24, Aguascalientes 29, León 29, Qu erétaro 30, San Luis Potosí 27, Guadalajara RA apto 28, Colima RA 31, Morelia 26, CDMX prob RA dispersas por la noche 25, Xalapa RA 24, V eracruz RA nocturna 30, Acapulco RA nocturna 31, Chilpancingo 31, Pachuca 21 grados	Neu
10681	Ns vems hoy en la feria d Villanueva Zacatecas pr armar el fieston #LPBSJ #MIENTRASNOME BUSQUES Los esperamos https://t.co/bQmklS GdbM	Pos
8412	próximo a lanzar su nuevo álbum!! Estén truch as banda!! #VIERNES #musica #mexico #rock #instrumental #Zacatecas https://t.co/NrAzcPrv f6	Pos
1818	Hoy y mañana disponible en #Reynosa #Tama	

	s armas o deveras sin son muy cabrones, denle s los protocolos de servicio a la ciudadanía para que den la atencion y no contraten chavos barrios o ñeros de aqui de mi delegacion iztapalapa contraten gente con criterio y sentido comun	Neu	/
11439	@variste_armon @KeanaBanana98 @SexIsNotReal @CBMCringe Ok we all know it's fictional but I mean you can't make black panther from central Africa without letting him be black. I'm this case he can be both white skinned or dark skinned because I've seen pic	Neu	/
<u>3854</u>	Y en Zacatecas la elección interna de @MorenaZacatecas @PartidoMorenaMx también terminó en violencia y acarreo: https://t.co/okR1ivASOJ	Neu	Neu X
10802	Recuerdos que vienen y otros más que van; sólo uno es el que me permite tener vida y ánimos de crecer.....	Neu	/
5175	No tienen que renunciar, con que no crean es suficiente, la iglesia no los obliga a nada wapos.	Neu	/
4364	México a través de la historia, Ciudad de México, México City https://t.co/14jlbRcx1n	Pos	X
9473	EL GAAASSS, Zacatecas, Mexico	Neu	/
16036	Cloon Cloon Cloon	Neu	/
3371	TIN TIN	Neu	/
16452	Hidalgo del Parral CAPTURAN EN PARRAL A HOMBRE BUSCADO EN ZACATECAS https://t.co/9N2h8FyXUb	Neu	/
4903	El esfuerzo y tomar riesgos si tienen recompensa.	Pos	X
12443	#Congreso #Historia #Masonería en #Zacatecas #México https://t.co/mNXpPJtqH9	Neu	/

	uilpas ?? Gente de #Matamoros voy el martes y estoy agendando citas. #CiudadVictoria y #Tampico reserva ya!!! #SanLuisPotosi y #Zacatecas por allá nos vemos pronto #Escort #Venus Princess a la orden mi amor	Neu	X
12448	Asi son las necesitadas de atención, pero ya no hay que hablar de ella, pa que darle importancia a una mocosa precoz	Neg	X
420	Una calle bien iluminada contribuye a tener un #ZacatecasSeguro para las familias. Ya hemos cambiado más de 3,600 lámparas en #Zacatecas y hoy ponemos en marcha el programa "Iluminando la Joya de la Corona", con el cambio de lámparas de vapor de sodio por focos tipo LED	Pos	/
10546	Me pasó lo que le pasa a la gente estúpida y ordinaria... olvide mis lentes en mi casa... de zacatecas... no vuelvo hasta dentro de 15 días	Neg	Neg /
13820	preparate un guion de apoyo cuando narres. Siempre estás desactualizado con información, dando datos que no son. No te duermas en tus l aureles ¡SALUDOS!	Neu	X
5696	que triste es verlo decir que en Baja California hay un "no problema" y que no toque ni con el pétalo de una crítica a su patrón #LopezObrador que todos sabemos que es él quien está sosteniendo esa aberrante violación a las leyes electorales. ¡muy lamentable!	Neg	/
2634	Esta estúpida Ana Navarro necesita unas cachetadas Zacatecas para que se le quite lo Pendejo!	Neg	/
10929	@jorgeramosnews @m_zamarripa @lopezobrador_ le invitamos formalmente a realizar una investigación en Sinaloa, vaya y parece ahí, escuché a la gente, entreviste al grupo delictivo, no se escude como siempre detrás de sus notas tendenciosas, pregúnteles que	Neu	X
4588	Muy lamentable que #Notimex sirva solo para legitimar ilegalidades...	Neg	/
15738	Afuera hace frío, no olviden una chaqueta antes de salir. Si pueden también abríguense.	Neu	X
16869	El demonio de la prepotencia y el maltrato a los meseros me está dominando.	Neg	/
11314	mi pregunta es asi son con los que no portamo		