UNIVERSIDAD INTERNACIONAL SEK

FACULTAD DE ARQUITECTURA E INGENIERÍAS

Trabajo de fin de carrera titulado:

"PREDICCIÓN DE ATAQUES DE CYBER BULLYING MEDIANTE TÉCNICAS DE APRENDIZAJE PROFUNDO APOYANDOSE EN UN CORPUS DE ENTRENAMIENTO PARA LA CLASIFICACIÓN DE TEXTO EN ESPAÑOL"

Realizado por:

Paul David Cumba Armijos

Director del proyecto:

Ing. Diego Fernando Riofrío Luzcano, PhD.

Como requisito para la obtención del título de MASTER EN TECNOLOGÍAS DE LA INFORMACIÓN

Quito, noviembre 2018

DECLARACION JURAMENTADA

Yo, **PAUL DAVID CUMBA ARMIOS**, con cédula de identidad **060395361-3**, declaro bajo juramento que el trabajo aquí desarrollado es de mi autoría, que no ha sido previamente presentado para ningún grado a calificación profesional; y, que ha consultado las referencias bibliográficas que se incluyen en este documento.

A través de la presente declaración, cedo mis derechos de propiedad intelectual correspondientes a este trabajo, a la UNIVERSIDAD INTERNACIONAL SEK, según lo establecido por la Ley de Propiedad Intelectual, por su reglamento y por la normativa institucional vigente.

Paul David Cumba Armijos

C.C: 0603953613

DECLARATORIA

El presente trabajo de investigación titulado:

“PREDICCIÓN DE ATAQUES DE CYBER BULLYING MEDIANTE TÉCNICAS DE APRENDIZAJE PROFUNDO APOYANDOSE EN UN CORPUS DE ENTRENAMIENTO PARA LA CLASIFICACIÓN DE TEXTO EN ESPAÑOL”

Realizado por:

PAUL DAVID CUMBA ARMIJOS

Como requisito para la Obtención del Título de:

MAGÍSTER EN TECNOLOGÍAS DE LA INFORMACIÓN

Ha sido dirigido por el profesor

ING. DIEGO FERNANDO RIOFRIO LUZCANO, PhD.

Quien considera que constituye un trabajo original de su autor

Ing. Diego Fernando Riofrío Luzcano, PhD.

DIRECTOR

DEDICATORIA

Dedico este trabajo a mi esposa María Teresa Tenesaca, a quien, gracias a su amor y comprensión, ha sabido apoyarme y comprenderme para la consecución de nuestros objetivos, a mi hijo Matías Alexander, quien se constituye en el motor que impulsa día a día mi vida, a mi madre Susana quien gracias a su apoyo y constancia me hace poder ser una persona de bien para la sociedad, a mis hermanos Andrea, Rafael y Anthony, y a mis sobrinos Sebastián y Fabiana por estar ahí, siempre junto a mi…

AGRADECIMIENTO

El tener a mi lado personas valiosas que, con su aliento y apoyo hacen posible que las metas y objetivos que nos trazamos en la vida, sean más fáciles de alcanzar, es por eso que agradezco a mi familia, a mi director Diego Riofrío, y demás maestros, compañeros y amigos que, de una u otra manera, me supieron brindar el apoyo y el incentivo para culminar con éxito esta etapa de mi vida.

ÍNDICE GENERAL DE CONTENIDOS

DEDICATORIA IV
AGRADECIMIENTO V
ÍNDICE GENERAL DE CONTENIDOS VI
RESUMEN XIII
ABSTRACT XIV
CAPITULO I 15
INTRODUCCIÓN 15
1.1. El problema de la investigación 15
1.1.1. Planteamiento del problema 15
1.1.2. Formulación del problema 21
1.2. Objetivo general 21
1.2.1. Objetivos específicos 21
1.3. Justificación 22
1.4. Marco Teórico 23
1.4.1. Analítica de datos 23
1.4.2. Análisis de datos en redes sociales 28
1.4.3. Aprendizaje Automático (*Machine Learning*) 29
1.4.4. Aprendizaje profundo (*Deep Learning*) 31
1.4.5. Inteligencia artificial 33
1.4.6. Redes Neuronales Artificiales 34
1.4.7. Topología de las redes Neuronales 35
1.4.8. Redes Neuronales Convolucionales 36
1.5. Marco Conceptual 41
1.5.1. Cyber Bullying 41
1.5.2. Neurona Artificial 42
1.5.3. Python 42
1.5.4. Spyder 42
1.5.5. Keras. 43
1.5.6. Twitter 43

CAPITULO II 44
ESTADO DEL ARTE 44
2.1. Análisis de comportamiento en redes sociales 44
2.2. Análisis de sentimientos en las redes sociales 46
2.3. Uso de aprendizaje automático (*machine learning*) para predicción de sentimientos 47
2.3.1. Análisis de sentimientos en textos con aprendizaje profundo (*deep learning*) 49
2.3.2. Uso de aprendizaje automático (*machine learning*) para análisis de sentimientos en redes sociales 50
2.3.3. Predicción de cyber bullying mediante aprendizaje automático (*machine learning*) 54
2.4. Discusión. 63
CAPITULO III 67
SOLUCION ADOPTADA 67
3.1. Introducción 67
3.2. Descripción de la arquitectura 67
3.2.1. Análisis de solución 67
3.2.2. Diseño de solución 68
3.3. Modelo 70
3.3.1. Modelo de aprendizaje profundo para clasificación de texto 70
3.3.2. Recopilación de datos 71
3.3.3. Preparación de datos 74
3.3.4. Conformación del corpus 76
3.3.7. Ejemplo de predicción. 96
CAPITULO IV 97
4.1. Introducción 97
4.2. Resultados 98
4.2.1. División de datos para Entrenamiento y Pruebas 98
4.2.2. Entrenamiento de la red neuronal convolucional para la validación del modelo 99
4.2.3. Validación del modelo 104

4.2.4. Predicción de los conjuntos de pruebas 105
4.3. Discusión de resultados 106
CAPITULO V 108
CONCLUSIONES Y TRABAJOS FUTUROS 108
5.1. Conclusiones 108
5.2. Trabajos Futuros 110
Referencias Bibliográficas 111

Índice General de Tablas

Tabla 1: Clasificación de los métodos según de datos de análisis. Fuente: (Gandomi y Haider, 2015) 25
Tabla 2: Técnicas y métodos para la Analítica de datos. Fuente: (Rodríguez, Palomino, y Mondaca, 2017) 27
Tabla 3: Resultados en Semeval-2015 para frase y nivel de tweet sub tareas Rango muestra la posición absoluta del sistema de Severyn, Inc, y Moschitti (2015). Fuente: (Severyn et al., 2015) 54
Tabla 4: Resultados de la predicción de bullying que demuestran acuerdo y desacuerdo con la codificación humana para la asignación de categorías utilizando métodos de aprendizaje automático. Fuente: (Bellmore et al., 2015) 56
Tabla 5: Ejemplos de procesamiento de oraciones con los analizadores morfológicos MeCab2 y CaboCha3. Fuente: (Ptaszynski et al., 2017) 60
Tabla 6: Resultados de todos los clasificadores aplicados. Fuente: (Ptaszynski et al., 2017) 61
Tabla 7: Resultados obtenidos en la aplicación de los modelos de machine learning con la métrica F1 score. Fuente: (Agrawal y Awekar, 2018) 62
Tabla 8: Comparación del rendimiento de los modelos de aprendizaje profundo utilizados. Fuente: (Agrawal y Awekar, 2018) 63
Tabla 9: Necesidades Funcionales de la solución adoptada. Fuente: Desarrollo del investigador 68
Tabla 10: Palabras clave para la búsqueda de bullying . Fuente: desarrollo del investigador 74
Tabla 11: Palabras clave para la búsqueda de no bullying . Fuente: desarrollo del investigador 74
Tabla 12: Estructura del archivo corpus-bullying.csv. Fuente: Desarrollo del Autor 77
Tabla 13: Vocabulario ejemplo generado a partir de la distribución de frecuencias. Fuente: Desarrollo del investigador 90
Tabla 14: Representación numérica del tweet ejemplo 1. Fuente: Desarrollo del investigador. 93
Tabla 15: Representación numérica del tweet ejemplo 2. Fuente: Desarrollo del investigador. 93
Tabla 16: Representación numérica del tweet ejemplo 3. Fuente: Desarrollo del investigador. 94
Tabla 17: Textos para ejecutar los ejemplos de predicción. Fuente: Desarrollo del investigador. 96
Tabla 18: Resultado de la predicción de los ejemplos. Fuente: Desarrollo del investigador. 96
Tabla 19: División de datos para entrenamiento y pruebas. Fuente: Desarrollo del investigador 98

Tabla 20: Modelo seleccionado para la validación por cada iteración de entrenamiento. Fuente: Desarrollo del investigador..... 104
Tabla 21: Resultados de acierto y fallo en la predicción de Tweets por cada iteración. Fuente: Desarrollo del investigador..... 105
Tabla 22: Resultado de la probabilidad de confianza que se tiene en el modelo por cada iteración de validación. Fuente: Desarrollo del investigador. 106

Índice General de Figuras

Figura 1: Uso de aplicaciones y actividades en Internet a nivel mundial a julio de 2015. Fuente: (CEPAL, 2018) 16
Figura 2: Datos estadísticos de cyber bullying en niños españoles. Fuente: (Organización Mundial de la Salud, Fundación Mutua Madrileña, y Fundación ANAR, 2016) 17
Figura 3: Uso de las Redes sociales en Ecuador. Fuente: (Guerrero, 2015) 18
Figura 4: Ciclo de vida de la Analítica de Datos. Fuente: (Schmarzo, 2013) 24
Figura 5: Tipos de aprendizaje automático. Fuente: (Raschka, 2015) 30
Figura 6: Arquitectura de aprendizaje profundo en capas, aplicada a una CNN. Fuente: (LeCun et al., 2015) 32
Figura 7: Red Neuronal Artificial. Fuente:(Meneses-Bautista y Alvarado, 2017) 35
Figura 8: Modelo de arquitectura convolucional para el aprendizaje de una oración. Fuente: (Kim, 2014) 37
Figura 9: Estructura en capas de la Red Neuronal Convolucional para el entrenamiento. Fuente: Desarrollo del investigador 37
Figura 10: Diferencia entre las capas convolucionales 1D y 2D. Fuente: (Ackermann, 2018) 39
Figura 11 Arquitectura de aprendizaje automático hibrida para análisis de sentimientos. Fuente: (Gupta et al., 2017) 52
Figura 12: Validación de expresiones de “con bullying” y “sin bullying” del modelo bayesiano. Fuente: (Nixon et al., 2018) 58
Figura 13: Arquitectura de la solución adoptada. Fuente: Desarrollo del investigador 68
Figura 14: Usuarios de Redes Sociales en Ecuador, Fuente: (Del Alcázar, 2018) 69
Figura 15: Funcionamiento de la CNN para clasificación de tweets. Fuente: desarrollo del investigador 71
Figura 16: Panel de Aplicaciones de Anaconda Navigator. Fuente: (Anaconda, 2018) 72
Figura 17: Participación de Bullying en el corpus. Fuente: Desarrollo del investigador 79
Figura 18: Nube de palabras de los tweets etiquetados como bullying. Fuente: Desarrollo del investigador. 80
Figura 19: Nube de palabras de los tweets etiquetados como no bullying. Fuente: Desarrollo del investigador. 80
Figura 20: Top 100 de la distribución de frecuencias de las palabras de los tweets. Fuente: Desarrollo del investigador 82
Figura 21: Aproximación de la distribución de las palabras del corpus vs una distribución Zipf. Fuente: Desarrollo del investigador 83
Figura 22: Top 50 de la distribución de frecuencias de las palabras de los tweets bullying. Fuente: Desarrollo del investigador 84
Figura 23: Top 50 de la distribución de frecuencias de las palabras de los tweets no bullying. Fuente: Desarrollo del investigador 84

Figura 24: Representación de vectores word2vec, de la relación de similitud y palabras relacionadas en base a su cercanía. Fuente: https://samyzaf.com/ML/nlp/nlp.html 86
Figura 25: Esquema de entrenamiento del modelo. Fuente: Desarrollo del investigador... 87
Figura 26: Distribución de frecuencias acumuladas de las palabras de corpus. Fuente: Desarrollo del investigador..... 89
Figura 27: Diagrama de Flujo para la obtención de la matriz de incrustación. Fuente: Desarrollo del investigador..... 92
Figura 28: Proceso de entrenamiento de la CNN en Keras. Fuente: Desarrollo del investigador. 95
Figura 29: Resumen del modelo CNN entrenado. Fuente: Desarrollo del investigador. 95
Figura 30: Conjuntos de datos de pruebas y entrenamiento por cada iteración de división. Fuente: Desarrollo del investigador..... 99
Figura 31: Precisión y Perdida por cada Epoch, para la iteración 1 de entrenamiento. Fuente: Desarrollo del investigador..... 101
Figura 32: Precisión y Perdida por cada Epoch para la iteración 2 de entrenamiento. Fuente: Desarrollo del investigador..... 102
Figura 33: Precisión y Perdida por cada Epoch para la iteración 3 de entrenamiento. Fuente: Desarrollo del investigador..... 102
Figura 34: Precisión y Perdida por cada Epoch para la iteración 4 de entrenamiento. Fuente: Desarrollo del investigador..... 103
Figura 35: Tiempo de procesamiento por cada iteración de entrenamiento. Fuente: Desarrollo Investigador 103

RESUMEN

En el mundo actual, se destaca la importancia de poder avizorar acontecimientos futuros a fin, de tomar decisiones acertadas. El llegar a predecir y detectar factores que puedan tener afectación en aspectos de negocios, sociales o personales, puede definir el curso de las acciones que se puedan realizar en beneficio de la humanidad. Es por esto, que este trabajo investigativo tiene el objetivo de consolidar y preparar un cuerpo con expresiones de texto extraídas de Twitter, con el fin de utilizarlo como insumo esencial para entrenar una red neuronal convolucional (CNN), mediante técnicas de aprendizaje profundo. Como resultado de este entrenamiento, se genera un modelo de predicción de textos que puedan o no presentar signos de cyber acoso (*cyber bullying*), en los cuales se pueda manifestar agresión verbal grave, como insultos, ataques racistas, ataques homofóbicos, etc. Este modelo fue validado mediante técnicas de validación cruzada, obteniendo resultados satisfactorios que permiten concluir que el modelo generado es óptimo para realizar predicciones de textos para la identificación del cyber acoso.

ABSTRACT

Nowadays, it is extremely importance to be able to foresee future advances in order to make correct decisions. The ability to predict and detect factors that can impact the business or personal aspects, can define the course of actions that can be carried out for the benefit of humanity. For this reason, this research has the main goal to consolidate and prepare a corpus with text extracted from Twitter, with the propose of train a convolutional neuronal network (CNN), through deep learning techniques. As result of this training a text predict model was obtained, which could find situations of interpersonal violence, where verbal aggression is manifested, such as insults, racist attacks, homophobic attacks, etc. This model was validated through cross-validation techniques, obtaining satisfactory results that allow us to conclude that the generated model is optimal for predicting texts for the identification of cyberbullying.

CAPÍTULO I

INTRODUCCIÓN

1.1. El problema de la investigación

1.1.1. Planteamiento del problema

1.1.1.1. Diagnóstico

Los grandes avances de las tecnologías de la información y comunicación, han dado lugar a que aplicaciones como las redes sociales se introduzcan fácilmente como herramientas de uso cotidiano para trabajo, estudios o entretenimiento. Es así, que las redes sociales han dado un giro importante en la manera de comunicarse y compartir información, esto ha permitido desarrollar técnicas para analizar millones de datos que se generan día a día. El procesamiento de estos datos, se ha convertido en una pieza fundamental en la definición de estrategias, políticas, económicas o de marketing. (Suárez y Guerrero, 2013)

Estos sistemas computacionales permiten a los usuarios compartir y publicar información, lo que se convierte en un riesgo debido a que se exponen datos muy confidenciales, los cuales, pueden ser utilizados en nuestra contra, (Gómez, 2014). Se debe considerar que, además de los beneficios de las redes sociales, se puede utilizar estos medios de forma incorrecta y mal intencionada por *cyber* atacantes que se encuentran del otro lado de la pantalla.

En un informe, presentado por la Comisión Económica para América Latina y el Caribe CEPAL (2018), se indica que, la difusión de las tecnologías digitales, han permitido un incremento exponencial de los usuarios en internet. En el 2015 se estimó que en el mundo 3.174 millones de personas usaron Internet, lo que equivale al 43,4% de la población, y se prevé que la actividad en internet se incremente en un 60%.

El despliegue de redes de banda ancha ha posibilitado el consumo de aplicaciones multimedia a través de tabletas y teléfonos inteligentes, lo que ha conllevado a que un usuario en promedio, cuente con más de 60 aplicaciones, realice más de 44.000 búsquedas en *Google*, y más de 1.700 llamadas por *Skype*, envíe más de 2 millones de correos electrónicos, más de 300.000 mensajes a través de *WhatsApp* y más de 8.500 *tweets*, efectúe más de 1.800 publicaciones en *Tumblr* y 50.000 en *Facebook*, suba más de 1.900 fotos, más de 98.000 videos en *YouTube* y visualice más 655 horas de video por *Netflix*, tal como se puede apreciar en la Figura 1.

Figura 1: Uso de aplicaciones y actividades en Internet a nivel mundial a julio de 2015. Fuente: (CEPAL, 2018)

De acuerdo con el informe "*Cyber bullying* según el afectado", realizado por la Organización Mundial de la Salud, Fundación Mutua Madrileña, y Fundación ANAR (2016), uno de cada cuatro acosos ocurridos en las escuelas españolas, se producen utilizando medios tecnológicos. El 97% de los afectados posee nacionalidad española, sin

embargo, esto no quiere decir que sus antecesores hayan nacido en España, sino que, puede tratarse de casos de niños nacidos en España o que hayan adquirido la nacionalidad, pero sus padres sean inmigrantes. Es así que, el número victimas con familias de procedencia extranjera corresponde al 10,5% en *cyber bullying* y el 5,7% son otros tipos de acoso escolar (ver Figura 2).

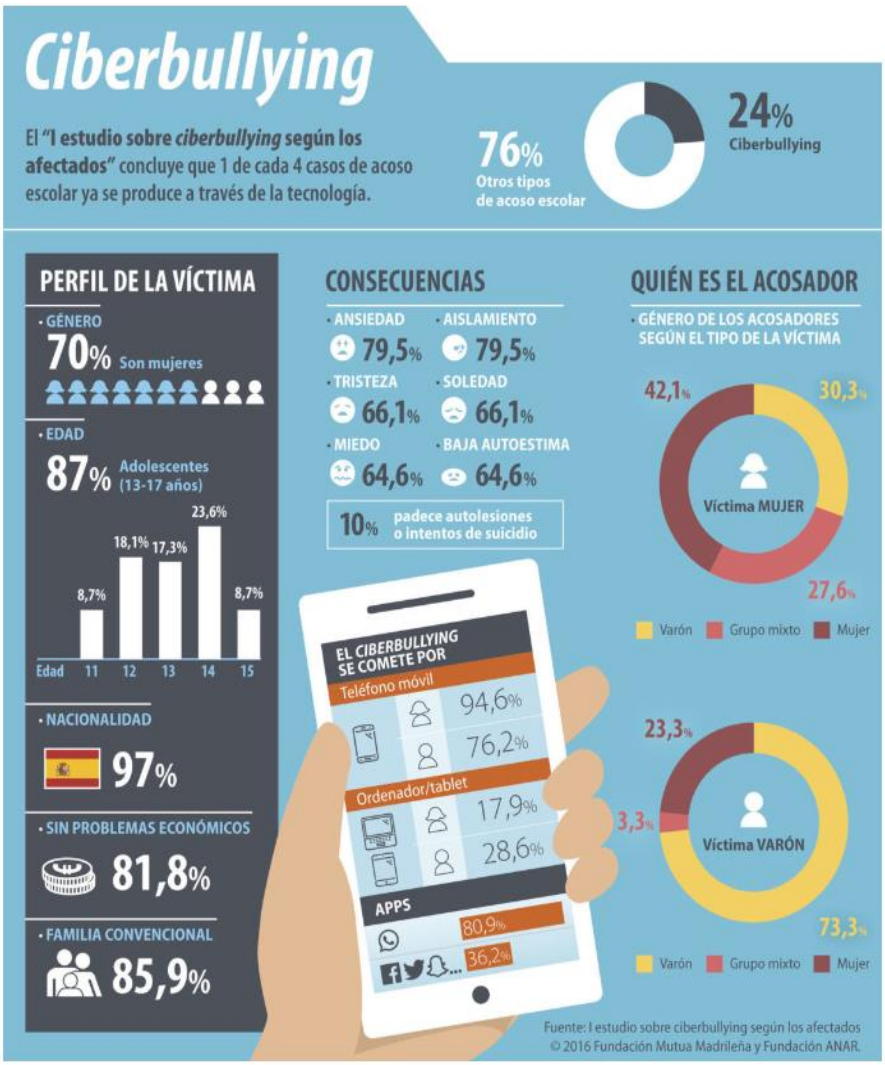

Figura 2: Datos estadísticos de cyber bullying en niños españoles. Fuente: (Organización Mundial de la Salud, Fundación Mutua Madrileña, y Fundación ANAR, 2016)

En América Latina, según un informe presentado por empresa especialista de seguridad ESET Latinoamérica (2014) se determinó en base a una encuesta realizada en esta región, el 30.7% de adolecentes sufrió de acoso a través de internet.

En el Ecuador, de acuerdo con la encuesta realizada por Guerrero (2015), se establece que, de una población de 4.995.474 de usuarios de internet, el 41,5 % cuenta con redes sociales, de las cuales el 68,9% se encuentran menores entre 12 y 17 años de edad, considerándose a este segmento el más vulnerable ante situaciones de acoso virtual o cualquier tipo de ataque cibernético, tal como se aprecia en la Figura 3.

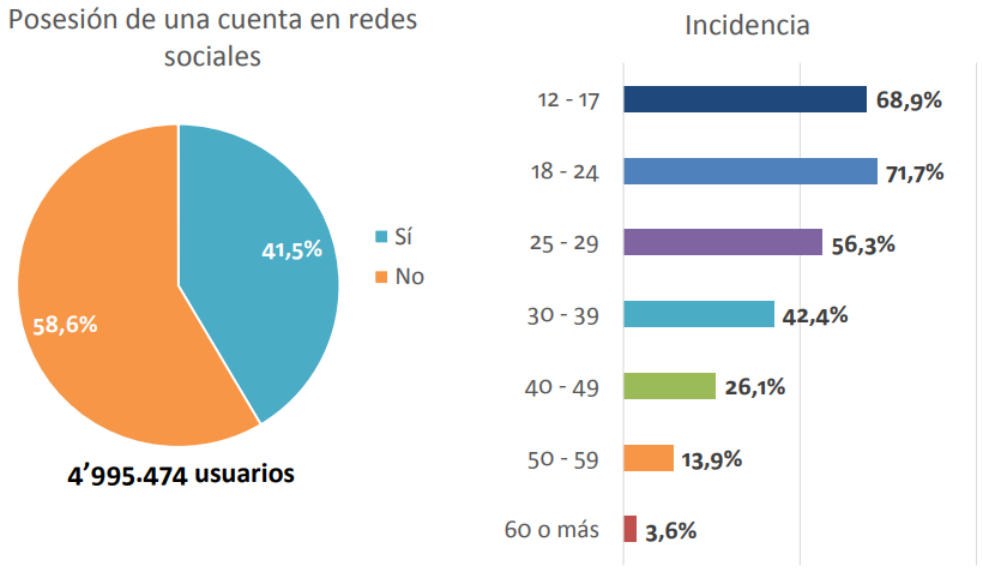

Figura 3: Uso de las Redes sociales en Ecuador. Fuente: (Guerrero, 2015)

Según Ortega y Ordoñez (2013), se ha detectado un alto nivel de *grooming*[1] y *cyber bullying*[2] en ciudades del Ecuador como: Quito, Manta y Guayaquil, determinando que el

[1] El sometimiento de los menores a las solicitudes y aproximaciones sexuales efectuadas por un adulto, la exposición no deseada a material sexual y, finalmente, el acoso. (Villacampa y Gómez, 2016)
[2] Es el daño constante y mal intencionado, realizado a una persona, la cual no es capaz de defenderse por sus propios medios y es realizado o ejecutado utilizando medios electronicos como el internet, telefonos moviles o computadoras. (Hinduja y Patchin, 2010)

27% de adolescentes sufre marginación, el 46% de hostigamiento, el 17% de agresión y el 10% sufre de extorción en las redes sociales. La incidencia de casos de cyber acoso en la población infanto juvenil, ha ido incrementándose de manera significativa.

El uso de las redes sociales para establecer ataques de *cyber bullying* era una práctica poco conocida en nuestro entorno, lo cual inquieta y preocupa a las autoridades de control en todas las partes del mundo. Esto se debe, a que existe una deficiencia de especialistas en investigación forense en el *cyber* espacio, que puedan entender la problemática de los ataques y agresiones que ocurren en las redes sociales, el cómo funciona la red, cómo funciona el sistema y cuáles son las vulnerabilidades. Estos, corresponden a los factores clave donde se debe dirigir los esfuerzos para combatir este tipo de acosos que cada vez llegan a ser más complejos e inclusive transfronterizos (Ramos y AFP, 2014).

Además, Girard (2017), manifiesta que, el peor de los castigos que puede sufrir un adolecente, es ser aislado de su grupo de pares, por ejemplo sus compañeros escolares. Es por esto, que existe una estrecha relación entre el *bullying* y el suicidio, e inclusive esta relación ha incrementado en los últimos años debido a la aparición del *cyber bullying*, que, a diferencia de las situaciones de violencia presencial, estas no finalizan cuando él o la adolecente regresa a su hogar y corresponden a situaciones estresantes en las que se cuestiona la orientación sexual, identidad de género, religión, raza, etnia, condiciona social, etc. Son factores de alto riesgo que tiene como resultado comportamientos suicidas.

Para prevenir esta problemática, es importante poder establecer mecanismos que, permitan avizorar y prevenir los acontecimientos inusuales e identificación temprana de acoso antes que las afecciones producidas sean estas físicas o emocionales puedan repercutir

gravemente en los niños y jóvenes, inclusive orillándolos a tomar decisiones tan radicales como el suicidio.

Si no se realizan trabajos de investigación que permitan tener la capacidad de identificar proactivamente los diversos tipos de agresiones que se establecen en las redes sociales, los índices de *cyber bullying* podrían incrementarse considerablemente, y con esto, podrían verse afectados millones de personas en sus entornos familiares, como por ejemplo que, los niños y adolescentes no quieran regresar a sus escuelas o colegios, baja autoestima, que se incremente el índice de suicidios, confusión en su orientación sexual, problemas de aprendizaje y concentración, homicidios fundamentados en el odio o venganza, etc.

Para que no ocurra este tipo de agresiones, la detección del *cyber bullying* y proveer medidas de control preventivo son los principales focos de acción para combatir al cyber acoso (Nandhini y Sheeba, 2015).

El uso de la minería de datos para el análisis y procesamiento de información expresada en lenguaje natural, permite maximizar la identificación de patrones que, a simple vista de los usuarios, padres o tutores escolares pasarían desapercibidos. (Aggarwal, 2015)

La analítica de datos mediante el uso de algoritmos de clasificación, permite establecer ciertas características de las expresiones con algún tipo de hostigamiento y con carácter agresivo hacia los usuarios de las redes sociales. (Mendoza, 2012).

De acuerdo a lo antes expuesto, se evidencia la importancia de conformar métodos, técnicas, algoritmos o modelos que posibiliten controlar los eventos de ataques en las redes sociales, es por esto que, la propuesta de este trabajo, es conformar un cuerpo que contenga

tweets, los cuales serán clasificados de acuerdo a sus expresiones como *bullying* o no *bullying*. Este cuerpo será utilizado para entrenar una red neuronal para la predicción de nuevos tweets e identificar si es que, un texto, tiene o no signos de *cyber* acoso.

1.1.2. Formulación del problema

La falta de un cuerpo que consolide las expresiones utilizadas en los tweets para realizar ataques de *cyber bullying* en español, ha ocasionado que, las prácticas agresivas para maltratar, insultar, burlarse, calumniar, intimidar o amenazar a las víctimas en las redes sociales, sean difíciles de identificar debido a que no existen mecanismos que ayuden a analizar las características que posee una expresión de *bullying*.

1.2. Objetivo general

Crear un modelo de predicción de ataques de *cyber bullying* mediante técnicas de aprendizaje profundo apoyándose en un corpus de entrenamiento para la clasificación e identificación de texto en español con características de *bullying*.

1.2.1. Objetivos específicos

- Extraer los datos de la red social *Twitter* mediante el módulo de *Python tweepy* y la API de la misma red social para conformar un conjunto de datos.
- Clasificar los *tweets* en idioma español en *bullying* y *no bullying*, mediante la categorización del texto en base a palabras clave para la conformación de un corpus.
- Entrenar un modelo de aprendizaje profundo utilizando los datos del corpus para la identificación de ataques de *bullying* mediante el procesamiento de texto en lenguaje natural.

- Verificar la confiabilidad del modelo mediante la técnica de validación cruzada para comprobar la precisión de predicción de texto de *bullying* y *no bullying*.
- Analizar el corpus de *tweets* mediante técnicas de minería de texto para la identificación de las características y patrones utilizados en los *tweets* que han sido clasificados como *bullying* y *no bullying*.

1.3. Justificación

En la revisión de la literatura científica, explicada en el CAPÍTULO II, de los trabajos relacionados a los objetivos de este tema de investigación, no se ha encontrado investigaciones que permitan realizar predicciones de ataques de *cyber bullying* mediante el entrenamiento y aprendizaje con un corpus de palabras y expresiones explicitas y clasificadas con etiquetas de *bullying* y *no bullying* en español, mediante el cual, sea posible establecer un mecanismo de control e identificación temprana del *cyber* acoso, esto no solo en nuestro país, ya que el modelo podría motivar al desarrollo de aplicaciones de control en varios países de habla hispana.

Además, se ha realizado una revisión en nuestro medio y se ha identificado que, en la actualidad, dentro de la legislación del Ecuador, no existe marco legal que tipifique al *cyber* acoso o *cyber bullying* como un delito. Sin embargo, es importante que pueda llegar a determinarse y denunciarse, ya que existen faltas relacionadas como la extorsión, acoso, injurias, calumnia, discriminación, hostigamiento e intimidación, con el fin de causar daño a alguna persona. Para esto, es importante definir técnicas que permitan pronosticar cuando un texto tiene o no características de *cyber bullying*.

Además, en el país, no se ha encontrado estudios relacionados con la ciencia de datos que permita analizar la información que se genera en las redes sociales enfocándolo a identificar signos de *cyber* ataques en las redes sociales. El uso de estas técnicas para la analítica de la información, permite de manera oportuna la detección y determinación de patrones o comportamientos similares en un conjunto de datos, es así que, la aplicación de algoritmos que ayude a la abstracción del conocimiento basado en el aprendizaje profundo a partir de la conformación de un corpus de entrenamiento, posibilita identificar los *tweets* que reúna las características de textos con signos de *bullying*.

1.4. Marco Teórico

1.4.1. Analítica de datos

Para el procesamiento y analítica de datos, se ha desarrollado un campo denominado Ciencia de Datos, el cual conjuga una amplia gama de técnicas que provienen de diferentes disciplinas tales como: Ciencias de computación, Matemáticas, Estadística, Econometría, Investigación. (Provost y Fawcett, 2013).

El ciclo de vida de la Analítica de Datos, se encuentra detallado en la Figura 4, en la que se evidencia un ciclo de vida no lineal, compuesto de seis que son: Explorar, Preparar datos, Planificar modelo, Elaborar modelo, Comunicar, Utilizar.

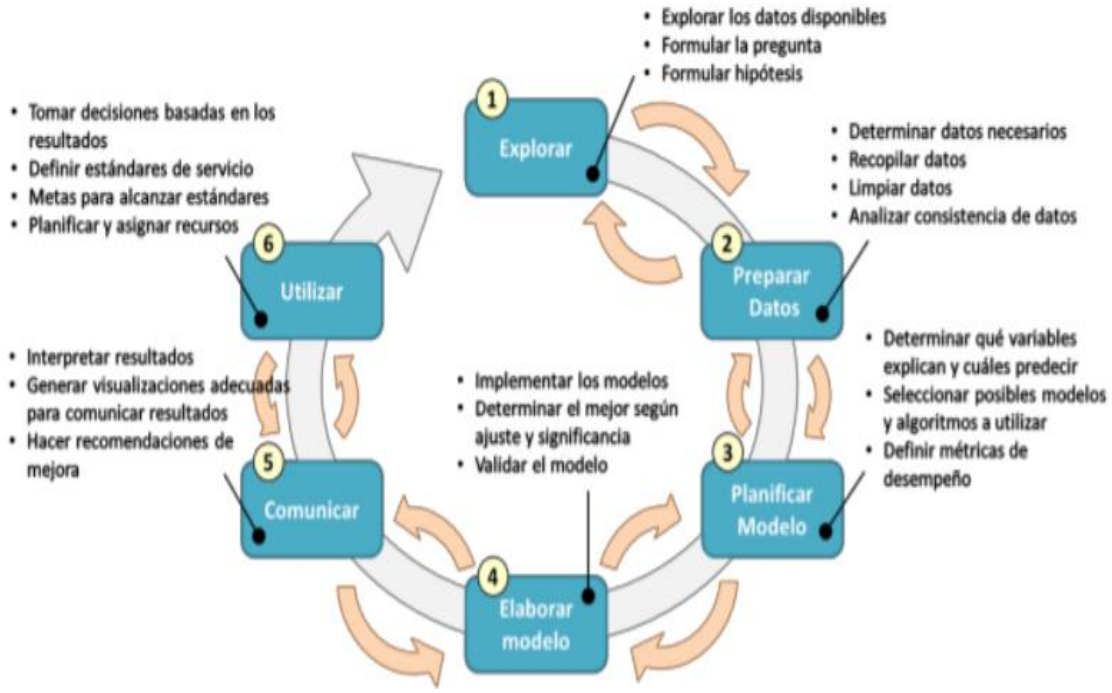

Figura 4: Ciclo de vida de la Analítica de Datos. Fuente: (Schmarzo, 2013)

1.4.1.1. Principales técnicas para la Analítica de datos

Para el estudio de la Analítica de datos, existe una amplia gama de métodos, los cuales, se componen de tres aspectos: adquisición y almacenamiento de los datos, limpieza y depuración de los datos y, la preparación para su análisis.

Además, estos métodos se clasifican de acuerdo al tipo de dato que se está analizando. Según lo expone Gandomi y Haider (2015), esta clasificación se detalla en la Tabla 1.

Tipo de datos	Ejemplos de técnicas de procesamiento según el tipo de datos
Texto	Extracción de la información: obtener datos estructurados de un texto, entidades y relaciones. Resumen de texto: resúmenes de documentos usando procesamiento de lenguaje natural. Respuesta a la pregunta: responde a preguntas formuladas en lenguaje natural usando procesamiento desarrollado para tales fines. Análisis de sentimiento: analiza un texto de opinión y genera una respuesta negativa o positiva.
Audio	Enfoque basado en transcripción: se genera una transcripción textual de contenido del audio a través de reconocimiento automático del habla con grandes diccionarios. El resultado se analiza con las técnicas analíticas de texto. Enfoque basado en la fonética: trabaja a partir de sonidos o fonemas que se traducen a una secuencia a partir del habla. La representación fonética de un término se busca en la secuencia.
Video	Arquitectura basada en el servidor: servidor dedicado al análisis de videos. Arquitectura basada en el borde: el video es analizado en forma local y sobre el video sin compresión de datos
Redes sociales	Analítica basada en contenido: se enfoca en los datos posteados por los usuarios, que son analizados posteriormente con técnicas descritas anteriormente, sea texto, audio o video. Analítica basada en la estructura: sintetiza los atributos estructurales de la red social y extrae inteligencia a partir de las relaciones de las entidades participantes. Las técnicas incluyen detección de comunidades, análisis de influencia social y predicción de enlaces.

Tabla 1: Clasificación de los métodos según de datos de análisis. Fuente: (Gandomi y Haider, 2015)

A continuación, se describen las diferentes técnicas que se utilizan para realizar la analítica de datos en función de la clasificación que fue descrita en la Tabla 1.

Técnicas / Métodos	**Descripción**	**Aplicaciones / Ejemplos**
Análisis espacial	Conjunto de técnicas que analizan las propiedades geométricas, topológicas y geográficas de un conjunto de datos	Regresiones espaciales (consumo vs. distancia a centros comerciales) Simulaciones (desempeño de una cadena de distribución con bodegas en distintos lugares) Identificación de líderes de opinión para focalizar campañas de marketing.
Análisis de redes	Conjunto de técnicas que caracterizan relaciones entre nodos discretos en un grafo o una red	Identificar cuellos de botella en flujos de información de una empresa Modelamiento de redes de transporte y predicción del tiempo de desplazamiento de un punto a otro Predicción de fenómenos como crimen, deserción escolar y universitaria, esperanza de vida post-operatoria, ventas. Sugerencias y recomendaciones de productos en función de historial pasado.
Aprendizaje automático (*Machine Learning*)	Subespecialidad de la Ciencia de la Computación (denominada históricamente "Inteligencia Artificial") que se ocupa del diseño y desarrollo de algoritmos que permiten inferir comportamientos basados en datos empíricos.	Procesamiento de lenguaje natural: reconocimiento de voz y lenguaje para interacción humano computador (ej.: Siri, Cortana, Alexa), y análisis de sentimientos en textos y redes sociales. Reconocimiento de patrones: texto manuscrito, procesamiento de imágenes y reconocimiento de caras para buscar sospechosos de crímenes. Detección de anomalías: detección de fraudes bancarios en base a actividad inusual en compras usando tarjeta de crédito.
Inteligencia Territorial	Metodologías de análisis espacial que, a través de tecnologías de información, combinan enfoques cualitativos, cuantitativos y espaciales, respetando además los enfoques de participación y aproximación	Indicadores espaciales de nivel de servicios públicos y privados. Análisis de brechas y crecimiento espacial de la oferta y la demanda de servicios. Accesibilidad urbana y rural, caracterización de los territorios en diferentes dimensiones geográficas,

Técnicas / Métodos	**Descripción**	**Aplicaciones / Ejemplos**
	global multidisciplinaria y multisectorial.	sociodemográficas, económicas y desarrollo humano.
Optimización	Técnicas numéricas de modelamiento para rediseñar y mejorar tanto procesos como sistemas complejos en múltiples dimensiones.	Localización óptima de recursos en hospitales, escuelas, centros productivos, bodegas. Producción: programación de maquinarias para la fabricación, manejo de inventarios.
Pruebas A/B	Una técnica en la que un grupo de control se compara con una variedad de grupos de prueba con el fin de determinar qué cambios mejorarán una variable objetivo dada. Por esto, esta técnica también se conoce como *Split testing* o *Buckettesting*.	Probar la efectividad de distintas campañas de marketing Probar la efectividad de un tratamiento médico, o un tipo de educación a través de experimentos naturales donde algunos sujetos participan de una intervención y otros no, por diferentes circunstancias. Se busca además que estos sujetos sean lo más parecido en sus características (pareo) para controlar la mayor cantidad de variables posible.
Simulación	Modelación del comportamiento de sistemas complejo para pronóstico, predicción y planificación de escenarios.	Pronóstico de los resultados financieros de una empresa dado circunstancias de incertidumbre Pronóstico del clima
Visualización analítica de datos	Forma de descubrir y entender patrones en grandes conjuntos de datos vía interpretación visual, para que así los usuarios pueden navegar y explorar los datos.	Análisis visual interactivo de componentes principales Infografías. Tableros de mando (*Dashboards*), para seguimiento y síntesis de ciertos fenómenos.

Tabla 2: Técnicas y métodos para la Analítica de datos. Fuente: (Rodríguez, Palomino, y Mondaca, 2017)

1.4.2. Análisis de datos en redes sociales

Según Sanz (2003), el análisis de las redes sociales (ARS), ha sido desarrollado, en base a la necesidad, de medir y analizar las estructuras sociales, que se establecen en función de relaciones o eventos que ocurren entre los actores que conforman una red. Trata de entender los fenómenos sociales que se establecen, analizando el comportamiento de las personas y la estructura relacional que surge, cuando los individuos u organizaciones, coinciden, interactúan, colaboran o se comunican, entre sí.

Entonces, Sanz (2003), define a la las redes sociales como un conjunto de relaciones interpersonales y sociales que conectan a individuos u organizaciones en grupos colaborativos, con el fin de establecer relación directa o indirectamente entre los actores, para, interactuar, comunicarse, intercambiar cosas, etc.

En cambio, Ramos (2015), define al ARS, como un consolidado conjunto de herramientas para la evaluación e implementación de programas que permiten la coordinación de grupos, con el fin de, entender las relaciones para conocer el desarrollo y evolución de coaliciones y los cambios que puede experimentar una red al modificar el número de actores en organizaciones comunitarias.

1.4.2.1. Principal uso del Análisis de Redes Sociales

Relaciones y Actores. – Cuando se habla de redes sociales, se establece claramente que existen vínculos entre entidades que se denominan actores. Estas interacciones pueden representar filiaciones (pertenencia a la misma asociación), relaciones interinstitucionales (convenios colaborativos), e interacciones (intercambio de información), (Wasserman,

1994). En cada relación, se establece un enfoque relacional distinto, independientemente de que la red este constituida por los mismos actores.

Se debe evaluar a las conexiones que relacionan a los actores en la red, estas, pueden ser dirigidas o no dirigidas. Además, es posible determinar la intensidad de esa relación por medio de la fortaleza que vincula a las entidades. (Ramos, 2015).

Obtención de métricas de cohesión. - Son medidas que permiten analizar propiedades estructurales de la red, y permiten comparar diferentes redes. Permiten evaluar el funcionamiento y el flujo de información entre los actores que se encuentran en la red (Ramos, 2015).

1.4.3. Aprendizaje Automático (*Machine Learning*)

En la era moderna, con el gran desarrollo de la tecnología, existe un recurso importante y de gran abundancia en el medio y es la gran cantidad de datos estructurados y no estructurados (Raschka, 2015).

Raschka (2015), también menciona que, a mediados del siglo XX, el aprendizaje automático, evoluciono como una rama de la inteligencia artificial, en la cual se involucró el desarrollo de varios algoritmos de aprendizaje para la obtención de conocimiento y la realización de predicciones. Esto, tuvo la visión de que, en lugar de requerir que los humanos establezcan reglas de aprendizaje y construyan los modelos manualmente, para analizar grandes cantidades de datos, el aprendizaje automático ofrece una eficiente alternativa para capturar el conocimiento de los datos y la mejora significativa en el rendimiento de los modelos predictivos para la toma de decisiones.

El aprendizaje automático, no solo tiene gran relevancia en la investigación informática, sino que también desempeña un papel muy importante en la vida cotidiana de las personas, aprovechando de beneficios como: filtros robustos en los correos electrónicos no deseados, software para el reconocimiento de voz y texto, motores de búsqueda web confiables, oponentes desafiantes de ajedrez, y quizá muy pronto, el desarrollo de vehículos de conducción automática y segura, (Raschka, 2015).

1.4.3.1. Tipos de aprendizaje automático

Según lo manifiesta Raschka (2015), existen tres tipos de aprendizaje automático, tal como se aprecia en la Figura 5.

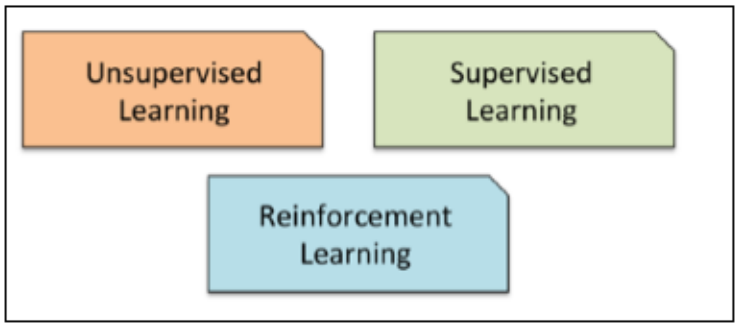

Figura 5: Tipos de aprendizaje automático. Fuente: (Raschka, 2015)

Aprendizaje no supervisado, se trabaja con datos sin etiqueta o con datos de estructura desconocida. Al utilizar este tipo de técnicas no supervisadas, se pueden explorar estructuras de los datos para extraer información importante sin la guía de una variable o resultado conocido o una función objetivo, (Raschka, 2015).

Aprendizaje supervisado, tiene como objetivo principal aprender y generar un modelo, a partir de datos de entrenamiento etiquetados que permitan realizar predicciones

sobre datos invisibles o futuros. El termino supervisado, se refiere a un conjunto de muestras donde ya se conocen las señales de salida deseadas (etiquetadas), (Raschka, 2015)

Aprendizaje reforzado, tiene como objetivo principal, el desarrollo de un agente que mejore su rendimiento en función de las interacciones con el entorno. Debido a que, cuando se realizan las predicciones, se utilizan funciones objetivo, se puede considerar que el aprendizaje reforzado es un campo del aprendizaje supervisado, con la diferencia de que el aprendizaje reforzado propone un concepto de retroalimentación, (Raschka, 2015).

1.4.4. Aprendizaje profundo (*Deep Learning*)

De acuerdo con, Schmidhuber (2015), el concepto de aprendizaje profundo o en ingles *deep learning*, apareció por primera vez en el año de 1986, y se deriva del aprendizaje automático. Posteriormente se introdujo en el uso de las redes neuronales artificiales, para revolucionar el reconocimiento de patrones y la predicción automática. Su principal enfoque es el de, formar relaciones en sistemas adaptativos, con largas cadenas que se vinculan fuertemente y se derivan de acciones y consecuencias. Los métodos de *deep learning* se componen de múltiples capas para aprender las características de los datos con múltiples niveles de abstracción (LeCun, Bengio, y Hinton, 2015).

Para Deng (2014), el aprendizaje profundo, también conocido como aprendizaje jerárquico, es una aplicación de las Redes Neuronales Artificiales (ANN), que corresponde a un sub campo del aprendizaje automático, utiliza muchos niveles de procesamiento y abstracción de información no lineal para el aprendizaje y la representación de características supervisadas o no supervisadas, clasificación y reconocimiento de patrones.

1.4.4.1. Arquitectura de aprendizaje profundo

Según LeCun et al. (2015), la arquitectura del *deep learning*, se constituye de un conjunto de múltiples capas de módulos, que están sujetas a aprendizaje mediante la computación de asignaciones y ejecución de funciones no lineales de entrada y salida. Cada módulo realiza una transformación. En la entrada de cada capa de abstracción se incrementa la selectividad antes de pasar a la siguiente capa.

En resumen, la arquitectura propuesta por LeCun et al. (2015), se constituye con la conformación de una red neuronal compuesta por múltiples capas no lineales que son extremadamente complejas en sus entradas, sensibles a detalles minúsculos, e insensibles a grandes variaciones irrelevantes como el fondo, ubicación, iluminación y objetos circundantes. Como se puede apreciar en la Figura 6, al utilizar esta arquitectura se tiene la capacidad de distinguir a los samoyedos de los lobos blancos.

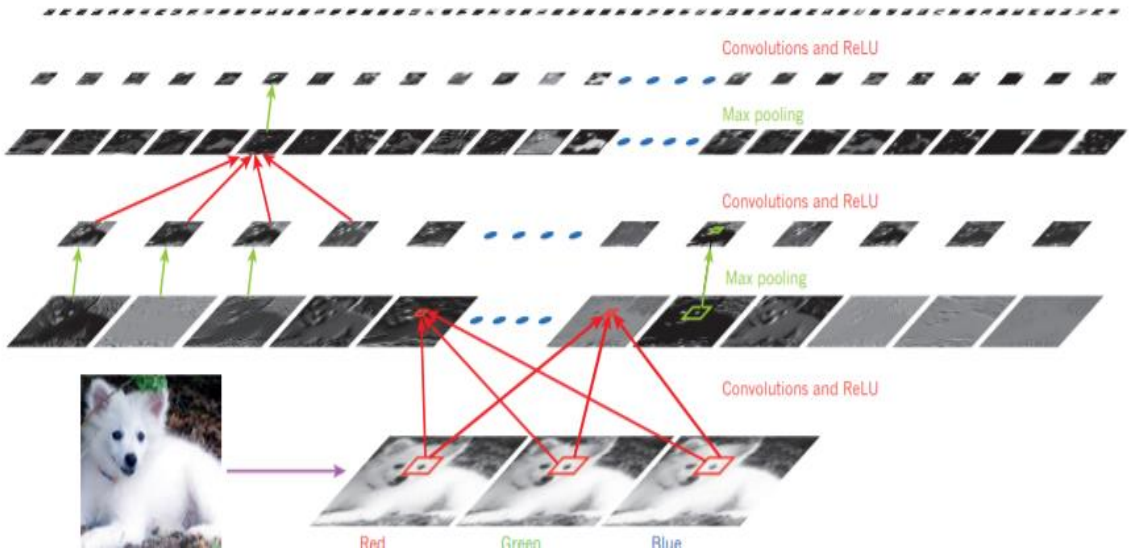

Figura 6: Arquitectura de aprendizaje profundo en capas, aplicada a una CNN. Fuente: (LeCun et al., 2015)

1.4.5. Inteligencia artificial

Según Romero, Dafonte, Gomez, y Penousal (2007), la Inteligencia Artificial (IA), se encarga del estudio de la inteligencia adaptada en elementos artificiales, y orientada desde el enfoque de ingeniería, a crear estos elementos para que se comporten de manera inteligente. En otras palabras, la IA, busca construir sistemas que sean utilizados por las maquinas, de forma que estas realicen y se comporten como una persona, y así se diría que son inteligentes. Tareas, tales como, el aprendizaje, la capacidad de adaptarse a los entornos cambiantes, la creatividad, etc., son actividades que generalmente se relacionan con la inteligencia. Adicionalmente, la IA, es considerada interdisciplinar, ya que, intervienen variadas disciplinas como la Psicología, las Tecnologías de la Información, la Neurociencia, la Física, las Matemáticas, la Ciencia del Conocimiento, etc.

La importante participación que tuvo Alan Turing en el desarrollo de uno de los primeros computadores, entre los años 40 y 50, fue muy considerable, ya que dio la pauta para que conceptos básicos como la computación, la arquitectura secuencial de computadores y la IA, sean considerados en ese entonces. El mismo Turing (1950), en su artículo *Computer Machine and Intelligence*, se cuestiona desde el inicio con la pregunta: ¿Pueden las maquinas pensar?. Con esto se aborda acciones que, hasta ese momento, eran propias de los seres humanos, para que, en un futuro, estas mismas acciones sean implementadas en las maquinas, dándoles la capacidad de aprender, crear y tomar decisiones.

1.4.6. Redes Neuronales Artificiales

Las Redes Neuronales Artificiales (RNA), son un conjunto de componentes para el procesamiento de la información que se basa en el funcionamiento del sistema nervioso biológico de los seres humanos. Estos, tratan de representar y emular el comportamiento del cerebro, para profundizar el conocimiento que se tiene respecto a él. Funcionan como, sistemas de procesamiento en paralelo y han demostrado, buenos resultado para la resolución de problemas complejos, (Crick, Barnett, Brenner, y Watts-Tobin, 1961).

Según, Palmer y Montaño, (1999), las RNA, consisten en sistemas que procesan información, cuya estructura y funcionamiento se basa y se inspira en las redes neuronales biológicas del cerebro. Estas, se componen de una gran cantidad de nodos o neuronas que se comunican entre sí y se encuentran estructuradas en capas, como se puede observar en la Figura 7. Todas las neuronas se conectan entre sí a través de enlaces de comunicación los cuales tiene un peso asociado. El aprendizaje que realizan las RNA y se constituye posteriormente en conocimiento, es determinado por los pesos.

En los últimos años, el estudio de las RNA, han llamado la atención de un gran número de investigadores. La razón principal de este interés radica en el hecho de que las RNA constituyen modelos flexibles de propósito general, que son capaces de aproximarse a prácticamente cualquier tipo de función lineal o no lineal (Cybenko, 1989). Debido a su flexibilidad en la aproximación de funciones, las RNA, son potentes métodos para la ejecución de tareas que requieren la estimación de variables continuas, la clasificación en base a patrones, y el pronóstico de estados futuros. (Kaastra y Boyd, 1996).

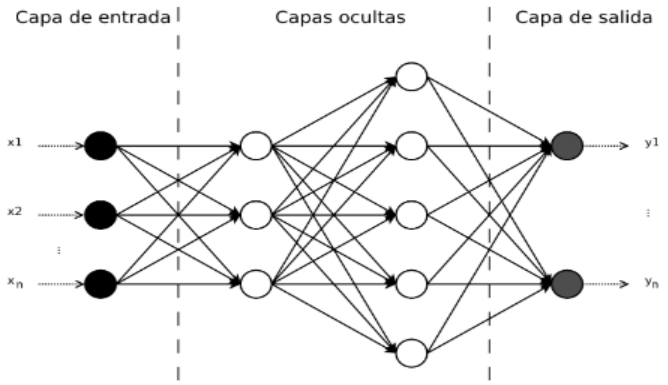

Figura 7: Red Neuronal Artificial. Fuente:(Meneses-Bautista y Alvarado, 2017)

1.4.7. Topología de las redes Neuronales

De acuerdo con, Matich, Ruiz, y Basualdo (2001), la topología de las redes neuronales, consiste en la arquitectura que presentan las mismas de acuerdo a la organización y disposición de sus neuronas. Esta organización, puede estar constituida por capas o agrupaciones de neuronas más o menos alejadas, y estas pueden ser de entrada o salida. En base a esto, tanto el número de capas, las neuronas que componen cada capa, el grado de complejidad y el tipo de conexiones entre neuronas, vienen a ser, los parámetros más fundamentales de una red neuronal.

De acuerdo a lo antes expuesto tenemos la topología de las redes neuronales puede ser:

Redes Monocapas, este tipo de redes esta formada, por neuronas que se encuentran conectadas en la unica capa que constituye la red. Este tipo de redes, utilizan tareas que se relacionan entre si, y a esto, se le conoce como autoasociación, es decir, regeneran información de entrada que se presenta dentro de la red de forma distoricionada o incompleta.

Redes Multicapas, son aquellas que se encuentran formadas por un conjunto de neuronas agrupadas en dos o más niveles o capas. Para distinguir la capa a la que pertenece una neurona, será necesario en fijarse en el origen de las señales que se recibe desde la entrada, hasta el destino de la señal de salida. Generalmente, todas las neuronas de una capa reciben señales de entrada desde una capa anterior, y envían señales de salida a una capa posterior, a esto se conoce como conexiones hacia adelante o *feedforward*, sin embargo, existen redes, cuyas capas posteriores pueden enviar señales a su capa anterior y a este proceso se le conoce como conexiones hacia atrás o *feedback*.

1.4.8. Redes Neuronales Convolucionales

Un claro ejemplo de una red neuronal con topología multicapa es la red neuronal convolucional, (CNN) o Convolutional Neural Network, por sus siglas en inglés, fueron creadas con el objetivo de utilizar información espacial entre los píxeles de una imagen, y fueron creadas para tareas de reconocimiento de video. Por lo tanto, se basan en la convolución discreta, (Khan, Rahmani, Shah, y Bennamoun, 2018). Sin embargo, trabajos recientes han demostrado que la CNN, han obtenido buenos resultados en el procesamiento de lenguaje natural. Los medios sociales y las redes se han convertido, en una fuente muy importante para los científicos de datos, ya que, en la actualidad, cada vez más personas comparten opiniones, mensaje, críticas y abordan diferentes temas en línea. Las CNN tiene la capacidad de extraer la semántica de textos de una manera muy esquematizada. (Kim, 2014).

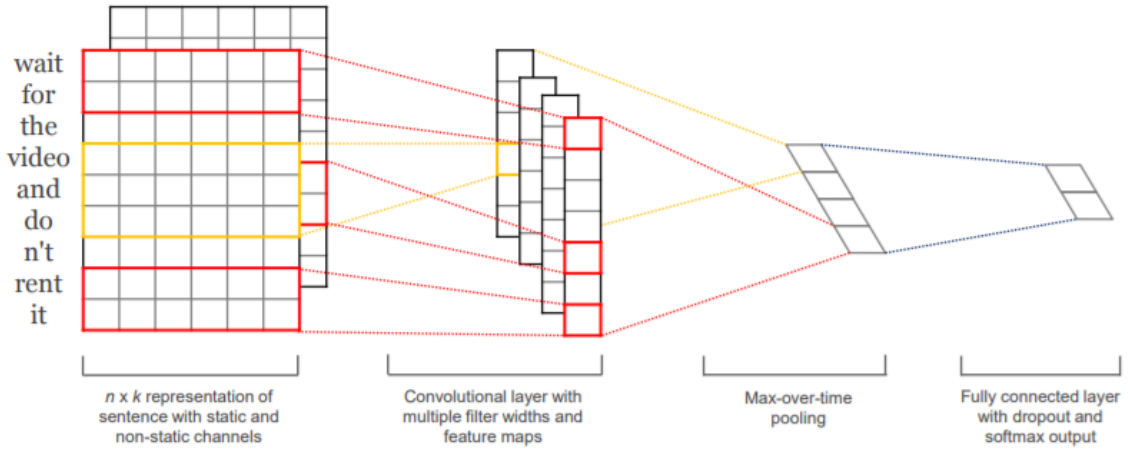

Figura 8: Modelo de arquitectura convolucional para el aprendizaje de una oración. Fuente: (Kim, 2014)

Como se puede apreciar en la Figura 8, Kim (2014), representa una arquitectura para el procesamiento de una oración a través de una red convolucional de una capa, en donde, la oración es representada por un vector de palabras de dimensión *k*, con longitud *n*. La capa de convolución, permite la extracción y mapeo de características ya que utiliza múltiples filtros para el paso de una capa a otra.

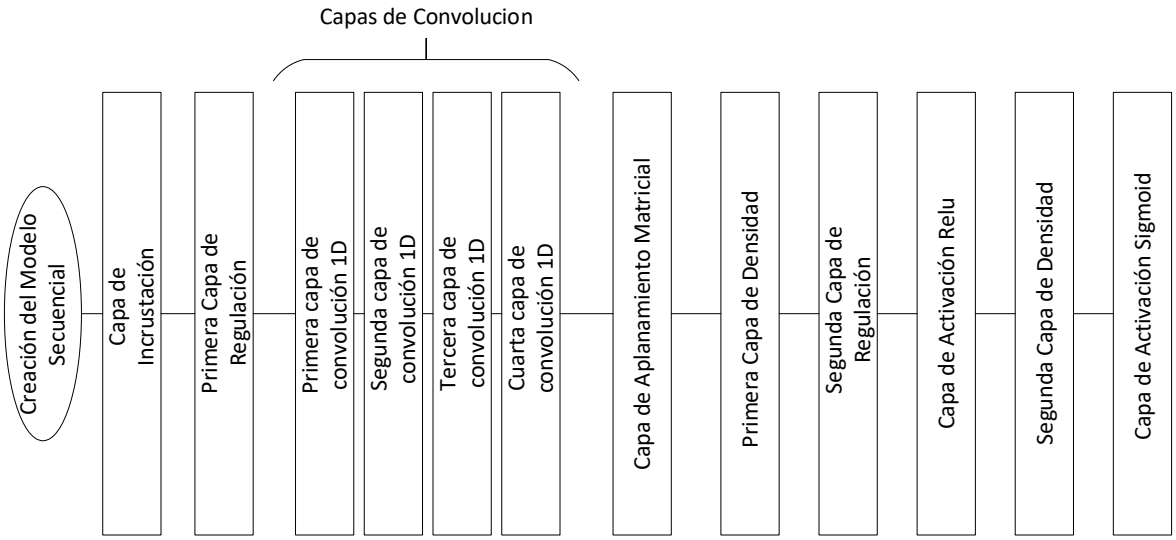

Figura 9: Estructura en capas de la Red Neuronal Convolucional para el entrenamiento. Fuente: Desarrollo del investigador

Según lo expuesto por Casas y Del Real (2017), las capas que constituyen una CNN, permiten resolver problemas de clasificación basándose en la reducción de las marices de entrada por el paso de las capas que conforman la red. Así, se puede observar en la Figura 9, la estructura en capas que se ha creado en este trabajo de investigación.

Modelo secuencial. – Permite la creación de una red neuronal con capas en secuencia, esto significa que las capas se encuentran, una después de la otra. (MkDocs, 2014)

Capa de incrustación. – Proporciona una representación de las palabras que son incrustadas mediante magnitudes vectoriales. Las CNN, permiten la incrustación de palabras previamente entrenadas. Esta capa se constituye, en la primera capa oculta de la red y sirve para aprender de las nuevas palabras que se están utilizando en los textos, para esto, proporcionan pesos que son aprendidos y utilizados por las capas posteriores para identificar las características de los textos. (MkDocs, 2014).

Primera capa de regulación. – Esta capa realiza una regulación por deserción, esto quiere decir que, ciertas neuronas que son seleccionadas al azar, son abandonas en el proceso en base a una probabilidad determinada, esto, se realiza para evitar que la red neuronal se sobrecaliente. (MkDocs, 2014).

Capas de Convolución. – El tipo de capa convolucional utilizado en la red neuronal es convolución 1D, constituida de una sola dimensión espacial, usada generalmente para el procesamiento de lenguaje natural. (MkDocs, 2014).

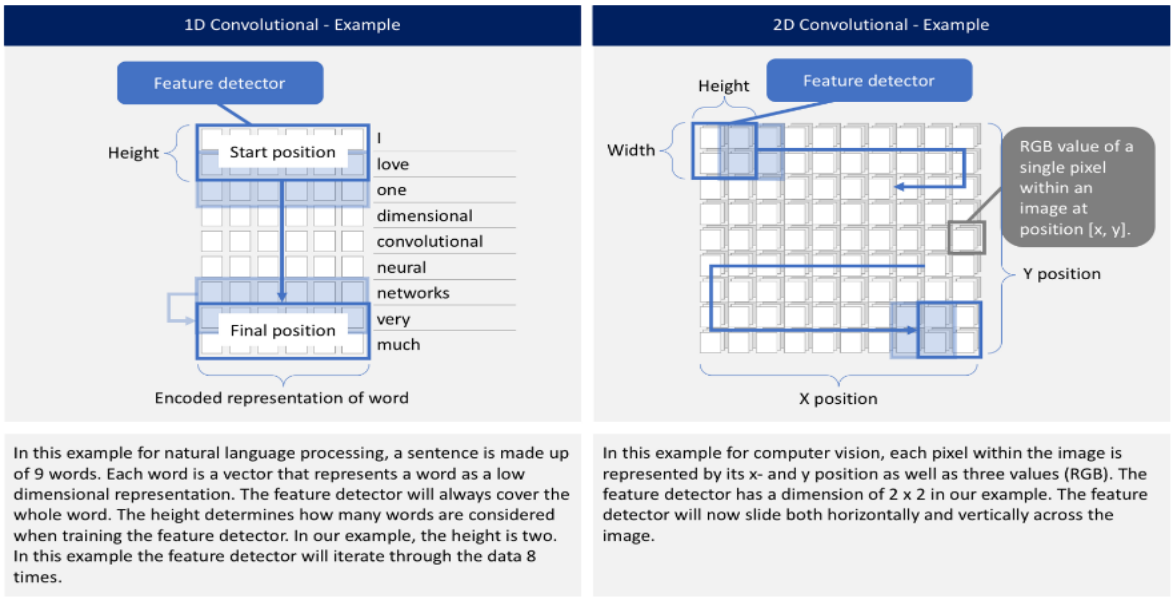

Figura 10: Diferencia entre las capas convolucionales 1D y 2D. Fuente: (Ackermann, 2018)

De acuerdo con la documentación de Keras MkDocs (2014), los parámetros utilizados, cuando se crea una capa de convolución son los siguientes:

Filtros (filters): Constituyen filtros o núcleos de convolución, se utilizan para el aprendizaje de las características de una capa a otra. En la primera capa se tiene una capa de neuronas representada por una matriz de 58 x 600.

Tamaño del núcleo (kernel_size): Constituye el tamaño del canal de convolución que se tiene para pasar de una capa hacia la otra capa. Para esta y todas las capas el tamaño es tres.

Relleno (padding): Se especifica el valor de 'valid', para que los datos sean rellenados con la misma longitud tanto en la entrada como a la salida de las capas.

Función de activación (activation): Especifica la función de activación que se usa en la capa de convolución, se utiliza "relu", para el aprendizaje de límites de decisión no lineales.

Saltos (strides): Son los saltos utilizados para el paso de una capa hacia la otra. Se especifica en esta y todas las capas el valor de uno.

Capa de aplanamiento matricial. – Esta capa, permite realizar un aplanado de la matriz dimensional, en otras palabras, transforma la matriz de entrada y se obtiene la salida un vector dimensional. Por ejemplo, si tenemos a la entrada de la capa una matriz de 52 x 75, se obtendrá a la salida a un vector unidimensional de 3900. (MkDocs, 2014).

Capas de densidad. – Capas densas, que se constituyen en las siguientes capas ocultas de la red, permite comunicar la capa de aplanamiento matricial con las funciones de activación. (MkDocs, 2014).

Funciones de activación. – De acuerdo con la documentación de Keras, MkDocs (2014), son funciones que se utilizan para asignar la transferencia de los valores de salida, de una capa a otra para realizar la predicción. Estas, pueden ser utilizadas por una capa de activación o por el argumento de activación por las capas permitidas, las activaciones disponibles son:

- **Softmax:** Esta función puede tomar valores en el rango de 0 a 1, y la suma de todas las probabilidades será igual a uno. Si esta función, se utiliza dentro de un modelo de clasificación múltiple, devuelve las probabilidades de la clase y objeto, lo cual corresponderá a una alta probabilidad.
- **Softplus:** Corresponde a una función muy similar a Relu, cuando esta tiende a cero. Allí, es donde softplus tiene mejor rendimiento, ya que, la derivada de la función es

más fácil y eficiente de calcular y además, corresponde a la función logística: $f'(x) = 11 + e{-}x$

- **Softsign:** Esta función corresponde a otra no lineal y puede considerarse como otra alternativa de Tanh, debido a que no se satura fácilmente.
- **Relu:** Esta función utiliza una unidad de rectificación Relu, la cual proporciona mejores resultados que las funciones Sigmoid y Tanh, ya que no se satura con valores altos para el cálculo de la probabilidad.
- **Sigmoid:** Esta función toma los valores reales dentro de un rango entre 0 y 1, dependiendo de la conversión que se realice se puede esperar en la salida valores entre -1 y 1.
- **Tanh:** Esta función también es conocida como tangente hiperbólica, se trata de la función de sigmoideo logístico, pero en diferente escala, por lo que, sus salidas van en el rango entra 0 y 1.
- **Hard_sigmoid:** Esta función tiene como objetivo, proporcionar un valor de probabilidad para uso en la binarización estocástica en los parámetros de la red neuronal. Corresponde a la función:

$$\sigma(x) = \text{clip}\,((x + 1) / 2, 0, 1) = \max(0, \min(1, (x + 1) / 2)).$$

- **Linear**: Se trata de una función de línea recta en donde la activación es proporcionada a la entrada y corresponde a la suma ponderada de los valores de la neurona.

1.5. Marco Conceptual

1.5.1. Cyber Bullying

Según lo definido por Hinduja y Patchin (2010), se lo conoce tambien como acoso cibernetico, y se entiende, como el daño constante y mal intencionado realizado a una persona considerada como victima, la cual no es capaz de defenderse por sus propios medios y es realizado o ejecutado utilizando medios electronicos como el internet, telefonos moviles o computadoras.

Las víctimas del cyber bullying, generalmente son seleccionadas por reunir ciertas características que las identifican como débiles tanto física y emocionalmente. Además, se sienten diferentes a las demás personas, por lo que se convierten en un blanco fácil para los agresores. Los ataques más frecuentes suelen ser, rumores, ofensas, insultos, amenazas, extorsiones e intimidaciones, esto mediante el uso de dispositivos o medios electrónicos, (Santillan, 2015).

1.5.2. Neurona Artificial

De acuerdo con Palmer y Montaño (1999)**,** la neurona artificial, se basa en una neurona bilógica, en donde se trata de imitar sus características más importantes. Esta, en general establece comunicación con demás neuronas vecinas, de quien recibe señales ponderadas por los pesos de las conexiones, la suma de estas señales ponderadas corresponde a una entrada total o neta para la obtención de una función de salida mediante la aplicación de funciones matemáticas.

1.5.3. Python

Python es un lenguaje de programación orientado a objetos el cual, presenta un intérprete de comandos completamente interactivo. Este lenguaje de programación ha presentado un importante potencial al ser usado en analítica y ciencia de datos, (Python Software Foundation, 2018).

1.5.4. Spyder

Es un entorno completo escrito y usado en Python, para científicos, ingenieros y analistas de datos. Tiene la capacidad de integrarse con facilidad a varios paquetes interactivos para la analítica, (Spyder Doc Contributors, 2018).

1.5.5. Keras.

Es una completa biblioteca de varios módulos de redes neuronales profundas, de código abierto escrita en Python. Fue desarrollada con la capacidad de permitir una interpretación inmediata de las características de los datos. Presenta un enfoque de uso fácil, modular y extensible. (MkDocs, 2014)

1.5.6. Twitter

Es un servicio de publicación de mensajes cortos en la web, que revoluciono la forma de comunicación en el aspecto social en internet, ha participado como plataforma fundamental en discusiones respecto a cambios políticos y sociales en todo el mundo, (Jorge et al., 2018). Tiene un enfoque, de microblogging, ya que permite a sus usuarios a través de la web o teléfono móvil, publicar mensajes cortos de 140 caracteres. A su vez, se constituye también en una red social, porque sus miembros poseen una página de perfil, con información personal y se pueden contactar con otros miembros. (Thelwall, Buckley y Paltoglou, 2011).

Esta red social permite y posibilita extraer información con facilidad, a diferencia de Facebook[3], que de acuerdo a sus políticas para desarrolladores, (Facebook, 2018), se requiere solicitar autorización de los usuarios de la red social para poder acceder a su información.

[3] https://www.facebook.com/

CAPÍTULO II

ESTADO DEL ARTE

Las tecnologías de la información, han permitido evolucionar la manera de establecer comunicación entre las personas, esto también, conlleva a un alto riesgo relacionado con los diferentes tipos de ataques no físicos que pueden sufrir las personas a través del uso de las redes sociales. Esto, ha incentivado a desarrollar nuevas técnicas en el ámbito de la detección de patrones y análisis de datos. Estos trabajos investigativos recopilan la importancia y el valor de explotar este tipo de información para generar conocimiento.

2.1. Análisis de comportamiento en redes sociales

De acuerdo con, Galarsi, Medina, Ledezma, y Zanin (2011), históricamente, el estudio del comportamiento humano ha sido abordado con gran interés por varias disciplinas. Este estudio, con el tiempo se ha direccionado a tratar de entender la influencia y el comportamiento de las personas enfocado en las aplicaciones de redes sociales.

Kaya y Alhajj (2018), realizan un análisis de las cuentas de Twitter de las empresas de telecomunicaciones, e identifican fuentes de rumores que son representados en grafos finitos con efectos de contorno mediante algoritmos de paso de mensajes.

Huiqi Zhang, Dantu, y Cangussu (2011) en cambio proponen un modelo probabilístico y estadístico para entender el comportamiento sociológico humano en las redes sociales, analizando registros de detalles de llamadas de teléfonos móviles. Para esto, se cuantifican los grupos sociales, relaciones y patrones de comunicación, con el fin de detectar cambios en el comportamiento humano. Los resultados experimentales demuestran que el modelo es efectivo.

Squicciarini, Rajtmajer, y Griffin (2017), proponen una clasificación de los principales aspectos que afectan al envejecimiento humano, basándose en técnicas de minería de datos en las redes sociales, además, desarrollan varias bases de datos longitudinales, con términos de los aspectos analizados y relacionados con el envejecimiento.

Pennebaker (2002), realiza un análisis del lenguaje natural que las personas utilizan a diario, para identificar claramente la personalidad, situación social y las relaciones interpersonales que se establecen entre los individuos que conforman una red social, inclusive, afirma que, este lenguaje se puede utilizar para evaluar la situación de la salud física y mental de estos individuos. Para esto utiliza la minería de datos como proceso esencial para la obtención de patrones y modelos aplicando métodos matemáticos, estadísticos y técnicas de aprendizaje automático *machine learning*.

Hu y Liu (2015), destaca la importancia de los servicios de redes sociales en los últimos años, ya que presentan nuevas oportunidades y desafíos tanto para los consumidores de información como para los proveedores de servicios. Con la creciente popularidad de las redes sociales, se tiene el potencial de extraer patrones accionables de una gran cantidad de datos para comprender el comportamiento del usuario y satisfacer su necesidade en el consumo de información. La creación de perfiles, se enfoca en describir mejor a los usuarios y sus relaciones en las redes sociales.

Schiaffino y Amandi (2009) presentan un enfoque para generar conocimiento mediante la extracción y recolección de información del usuario y sus publicaciones en redes sociales, a través de la minerias de datos. Esto es esencial para los sistemas de

recomendación parsonalizada, para la publicidad y el marketing efectivo. Por ejemplo, los usuarios de Twitter pueden etiquetar en sus cuentas con #hashtags de texto. Las etiquetas pueden proporcionar una descripción de los usuarios y utilizarse para facilitar la recuperación de información y otras aplicaciones. Yelp es otra aplicación, en la cual los perfiles precisos se definen en base a las preferencias en los alimentos, esto, mejorará significativamente en las recomendación de restaurantes para los usuarios.

2.2. Análisis de sentimientos en las redes sociales

Neri, Aliprandi, Capeci, Cuadros, y By (2012), estudian los sentimientos, a más de 1000 publicaciones hechas en Facebook, acerca de los noticieros de transmisión al público italiano. Comparan los sentimientos reflejados a la cadena Rai y la privada La7. Se utilizan datos proporcionados por Auditel[4], con respecto a audiencia y difusión, datos que son correlacionados con las métricas de dominio público de Facebook. Se aplican técnicas de minería de datos como son la clasificación y la clusterización, en base al procesamiento léxico y semántico de las publicaciones. Los resultados destacan la importancia de Facebook como plataforma para el marketing en línea, por lo que se concluye que, el monitoreo de las actividades de los medios sociales es una buena manera de medir la lealtad e intereses de los clientes, manteniendo un registro de su sentimiento hacia las marcas o productos.

Abrantes, Silva, y Fonseca (2012), en los resultados de su investigación presentan la generación de un modelo para la representacion del conocimiento acerca de los sentimientos expresados en la Web. En este trabajo se desarrolla un prototipo de software que analiza los sentimientos de varias marcas de automoviles, estos datos son almacenados para el posterior

[4] http://www.auditel.it/

analisis morfológico y la detección de caracterísitcas del lenguaje utilizado en base a ontologías. Tambien se desarrollan metodologías para la extracción de opinion, la composición y la clasificación en base a la creación de un diccionario de terminos. Una vez realizado el analisis y generado del modelo, se obtiene como resultado que, existen más opiniones positivas que negativas para la marca de automovil Palio, con el 63% de opiniones positivas contra el 37% que se consideraron negativas. Lo contrario sucede con las opiniones realizadas para las marcas Gol y Corsa, que tuvieron el 47% y el 38% de opiniones positivas contra el 53% y el 62% que fueron negativas respectivamente.

2.3. Uso de aprendizaje automático (*machine learning*) para predicción de sentimientos

Birjali, Beni-Hssane, y Erritali (2017), realizaron una investigación para buscar formas de analizar automáticamente los sentimientos y opiniones que los usuario expresan en las redes sociales. Como fuente de información se usan datos extraidos de Twitter mediante la librería de Java Twitter 4J[5], con el principal objetivo de analizar los sentimiento de aquellos usuarios que pudieran pensar en el suicidio. Para esto, proponen la elaboración de un vocabulario asociado al suicidio y lo complementa con el conjunto de datos disponible en WordNet[6] con el fin de extraer las caractersiticas semanticas de los sentimientos expresados en los tweets. Se utiliza la herramienta Weka[7] para realizar el proceso de minería de datos utilizando algoritmos de aprendizaje automático. Como resultado se establece que los algoritmos propuestos basados en aprendizaje automático pueden realizar predicciones de ideas suicidas utilizando los datos de Twitter.

[5] Twitter4J es una biblioteca no oficial de Java para la API de Twitter. http://twitter4j.org/en/index.html
[6] WordNet® es una gran base de datos léxica de inglés. https://wordnet.princeton.edu/
[7] Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. https://www.cs.waikato.ac.nz/ml/weka/

Romero (2017) muestra la capacidad de precedir sentimientos mediante la utilización de las técnicas de aprendizaje automático más utilizadas. Se usaron cuatro diferentes tecnicas de apredizaje automático: *Naive Bayes, Random Forest, Support Vector Machines y Recurrent Neural Networks*. Estas técnicas fueron entrenadas con dos diferentes conjuntos de datos disponibles en la web, *Movie reviews* y *Standford Twitter Sentiment.* Además, se generó una bolsa de palabras para la extracción de las características mediante métodos semánticos, como son los uni-gramas y bi-gramas. Como resultado, se tiene que los métodos propuestos de aprendizaje automático para el análisis de sentimientos, muestran óptimos resultados, destacando a *Recurrent Neural Networks* que tuvo mejores resultados que las demás técnicas.

Pang, Lee, y Vaithyanathan (2002), presentan un enfoque para realizar clasificación de documentos en base a un análisis de sentimientos, su principal objetivo es precedir si los sentimientos en los documentos son positivos o negativos. Para esto, se utiliza tres métodos de aprendizaje automático que son: *Naive Bayes*, *Maximum Entropy Classification*, y *Support Vector Machines*. Para el entrenamiento y validación de los modelos se usa la base de datos *Internet Movie Review Database (IMDb)* del grupo de noticias *rec.arts.movies.reviews*[8]. Como resultado, en términos de rendimiento, *Naive Bayes* tiende a ser menos optimo, con el 81% de precisión, comparado con los resultados de la validación de *SVM* que es del 81.6%.

[8] https://www.imdb.com/Reviews/

2.3.1. Análisis de sentimientos en textos con aprendizaje profundo (*deep learning*)

Amajd, Kaimuldenov, y Voronkov (2017), comparan diferentes redes neuronales basadas en algoritmos de aprendizaje profundo para la clasificación de textos. Al igual que Pang, Lee, y Vaithyanathan (2002), se usa la base de *Internet Movie Review Database (IMDb)* para realizar el entrenamiento y validación de los modelos generados las siguientes técnicas de *deep learning*: *Convolutional Neural Network (CNN), y LSTM for Sequence Classification*. Una vez realizado el proceso de entrenamiento y validación, se muestra que las CNN funcionan de mejor manera, para el aprendizaje de palabras y que no requieren de una estructura sintáctica o semántica del lenguaje, además se demuestra que el uso de representaciones de datos vectoriales pre-entrenados puede mejorar la precisión de la clasificación.

En el trabajo de Chachra, Mehndiratta, y Gupta (2017), se genera un modelo de análisis de oraciones, para detectar las relaciones semánticas con el fin de predecir sentimientos expresados en los textos de documentos web. Se utiliza para esto las redes neuronales de convolución profunda (DCNN) para modelar oraciones y realizar análisis de sentimientos. Este trabajo, propone un proceso compuesto por diferentes módulos para analizar la relación sintáctica e indentificar las relaciones de las palabras que conforman las frases que constituyen una oración. Como resultado se obtiene una precisión del 80.69% en la validación realizada al modelo.

En el trabajo de Ganesh B, Kale, Mankame, y Kulkarni (2018), se ha realizado un analisis de sentimientos en los datos recoplidaos de la Web mediante VADER[9], en conjunto con un método de filtración de palabras. Para esto han evaluado los metodos de aprendizaje automático estadístico, *Document Term Matrix* (representación) y *Support Vector Machine* (clasificación). Posteriormente se entrenó los algoritmos de aprendizaje profundo CNN, con la utilización del algoritmo para la generación de matriz vectorial *Word2Vec.* Los resultados obtenidos muestran que la CNN, obtuvo mejores resultados que los metodos anteriores con una precición de entrenamiento del 96%, y una precición de validación superior al 85%.

2.3.2. Uso de aprendizaje automático (*machine learning*) para análisis de sentimientos en redes sociales

En el trabajo realizado por Chinthana, Madhushani, Marcus, Aberathne, y Premaratne (2013), se realiza un análisis enfocado a extraer los sentimientos que albergan actitudes positivas y negativas con el uso de técnicas de aprendizaje automático, sobre personas, organizaciones, eventos, lugares e ideas. Se utiliza un rastreador (*crawler*), cuya función es recopilar los datos desde Twitter conectándose a través de su API, con la opción de capturar los datos públicos en *streaming.* Una vez se han capturados los *tweets*, estos son procesados y almacenados en una base de datos para el posterior entrenamiento de las redes *Naive Bayes*, y *Maximum Entropy*, los cuales se utilizan como clasificadores, para la extracción de las características lingüísticas se utilizaron los métodos uni-gramas y bi-gramas.

[9] Valence Aware Dictionary and Sentiment Reasoner, es una herramienta escrita en Python de análisis de sentimientos basada en reglas y léxicos que están específicamente en sintonía con los sentimientos expresados en las redes sociales. https://github.com/cjhutto/vaderSentiment

Una vez revisados los resultados del proceso de evaluación, se determinó que la técnica de aprendizaje automático más adecuada y que presenta el indice más alto en precisión y rendimiento, fue el clasificador *Naive Bayes* utilizando a los *uni-gramas* y *bi-gramas* informativos como el método de extracción de características.

Beigi, Hu, Maciejewski, y Liu (2010), realizan un análisis en redes sociales cuyo objetivo principal es presentar el enfoque que se utiliza para clasificar los sentimientos de acuerdo a la actitud de un escritor hacia diversos temas en categorías positivas, negativas o neutrales. En la realización de la clasificación de sentimientos y el análisis de afectividad, para la identificación de actitudes, emociones y opiniones de las personas sobre diferentes entidades, tales como productos, servicios, individuos, empresas, organizaciones, eventos, etc., se utilizan procesamiento de lenguaje natural y lingüística computacional, apoyados en algoritmos de aprendizaje automático. Finalmente se realiza una analítica visual y se centra en brindar una manera intuitiva para entender la gran cantidad de publicaciones disponibles en las redes sociales.

La investigación realizada por Gupta, Pruthi, y Sahu (2017), ha encontrado que al utilizar varios enfoques de aprendizaje automático de manera hibrida, dan mejores resultados que al utilizarlos de manera aislada. Se realizan varios pre procesamientos de los tweets utilizando varias funciones para la preparación de los datos previo al ingreso a los procesos de entrenamiento, esto con la finalidad de obtener datos útiles para la entrada a los clasificadores. Para la generación del modelo, básicamente se centra en la utilización de dos algoritmos de aprendizaje automático, *KNearest Neighbors (KNN)* y *Support Vector Machines (SVM)* de manera hibrida, tal como se puede apreciar en la Figura 11.

La observación analítica que se obtiene como resultado de la evaluación de los modelos, en términos de precisión y la métrica *F*, demuestra que al utilizar un enfoque hibrido, presenta una mejora en los indicadores que al utilizar los algoritmos de manera individual.

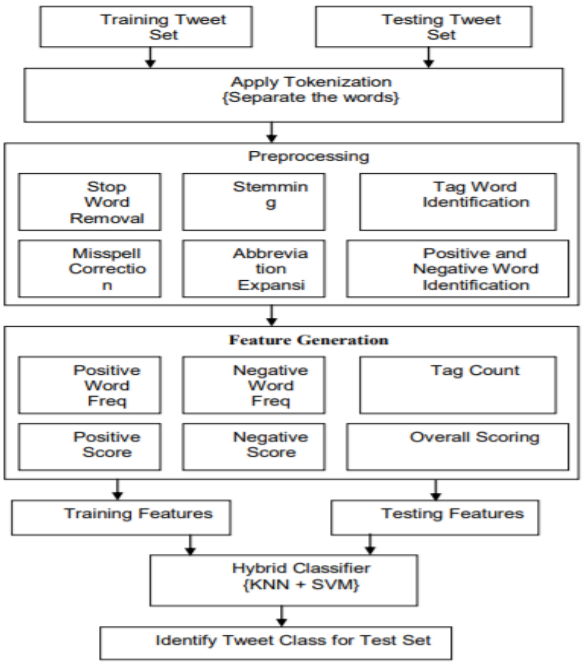

Figura 11 Arquitectura de aprendizaje automático hibrida para análisis de sentimientos. Fuente: (Gupta et al., 2017)

2.3.2.1. Uso de aprendizaje profundo (*deep learning*) para análisis de sentimientos en redes sociales

El trabajo realizado por Lu, Sakamoto, Shibuki, y Mori (2017), se enfoca en realizar un análisis de sentimientos expresados en Twitter, utilizando técnicas de *deep learning* (aprendizaje profundo). Estos tipos de análisis se lleva a cabo en un entorno experimental con conjuntos de datos de entrenamiento validación y prueba, en donde los resultados tienen dos tipos de polaridades globales, es decir: positivos y negativos. Generalmente, estos entornos experimentales, básicamente son poco prácticos para para sistemas del mundo real

ya que, en los entornos de producción, es inevitable tener tres tipos de polaridades: positivas, negativas y neutrales.

Para la generación de los modelos y la ejecución de las pruebas, se utilizan tres diferentes modelos: SVM[10], CNN[11] y RNN[12], con el objetivo de verificar que tan bien se desempeñan en el aprendizaje profundo para la predicción de sentimientos. El corpus utilizado para el entrenamiento y validación se constituye por datos extraídos de la red social Twitter, el mismo que ya se encuentra etiquetados con 503 *tweet* positivos, 774 *tweets* negativos y 526 *tweets* neutrales.

Para todos los modelos se utiliza la métrica de precisión y una validación cruzada de 10 iteraciones para la evaluación de cada modelo. Los modelos de CNN se ejecutan con 25 *epochs* y los modelos RNN ejecutan 100 *epochs*. Los resultados obtenidos en el proceso de validación reflejan buenos resultados en la precisión de los modelos, la precisión del modelo generado con la CNN alcanza el 67.6% y el RNN solo obtiene el 58.7%.

En el trabajo realizado por Severyn, Inc, y Moschitti (2015), se tiene como principal contribución la generación un modelo mediante técnicas de *deep learning* para el análisis sentimental en *Twitter*, se utiliza una red neuronal convolucional, se realizó un proceso de inicialización de pesos con matrices de vectores de palabras pre entrenados, que fueron utilizadas en la capa de incrustación, con el fin que la CNN, realice un entrenamiento preciso sin la necesidad de ingresar nuevas características. Los datos fueron entrenados y

[10] SVM stands for Support Vector Machine
[11] CNN stands for Convolutional Neural Network.
[12] RNN stands for Recurrent Neural Network.

comparados con el modelo disponible en la campaña oficial de evaluación del sistema de análisis de sentimientos de Twitter, organizado por Semeval (2015).

En la comparación de rendimiento, que se puede observar en la Tabla 3, entre el sistema oficial de Semeval (2015) y el sistema desarrollado por Severyn, Inc, y Moschitti (2015), se puede ver que, en la sub-tarea A, funciona bastante bien ubicándose en la primera posición, con el 84.79% de precisión, frente a otros modelos de conjuntos de pruebas, y para la sub-tarea B, se obtiene el 64.59% de precisión, ubicándose en el segundo lugar. En sí, el modelo presentado a nivel general, obtiene los mejores resultados con un AveRank de 4.3, que es la máxima puntuación de entre las 40 presentaciones realizadas.

Dataset	Score	Rank
Phrase-level subtask A		
LJournal'14	84.46	2
SMS'13	88.60	2
Twitter'13	90.10	1
Twitter'14	87.12	1
Sarcasm'14	73.65	5
Twitter'15	**84.79**	**1**
AveRank	2.0	1
Message-level subtask B		
LJournal'14	72.48	12
SMS'13	68.37	2
Twitter'13	72.79	3
Twitter'14	73.60	2
Sarcasm'14	55.44	5
Twitter'15	**64.59**	**2**
AveRank	4.3	1

Tabla 3: Resultados en Semeval-2015 para frase y nivel de tweet sub tareas Rango muestra la posición absoluta del sistema de Severyn, Inc, y Moschitti (2015). Fuente: (Severyn et al., 2015)

2.3.3. Predicción de cyber bullying mediante aprendizaje automático (*machine learning*)

El trabajo de Bellmore, Calvin, Xu, y Zhu (2015), se enfoca en tratar de responder a las cinco preguntas clave respecto a episodios de bullying. ¿Quién?, ¿Qué?, ¿Por qué?,

¿Donde? y ¿Cuando? Esto mediante la utilización de métodos de aprendizaje automático (machine learning), con la finalidad de comprender, este importante problema social y psicológico que se ha dado en los niños y jóvenes de los Estados Unidos.

Los métodos de machine learning fueron aplicados una vez se recopiló un total de 9,764,583 publicaciones de intimidación, realizadas en la red social Twitter entre el primero de septiembre del 2013 y el 31 de agosto del 2013. La gran mayoría de estas publicaciones fueron escritas por víctimas y reporteros los cuales hicieron referencia a algún episodio de ataque, generalmente, realizado en las escuelas.

Como se puede observar en la Tabla 4, se tiene el resultado de la clasificación de los tweets en las diferentes categorías, así como también, la precisión en la aplicación de las técnicas de machine learning, de acuerdo a las codificaciones:

- Traza de Bullying binaria: Si o No.
- Quien: El rol del autor.
- Que: La forma del bullying
- Por qué: La tipología del post

Human-coding	Total	Machine learning predicted as						Accuracy
Binary bullying trace								
		Yes	No					86%
Yes	2102	1555	547					
No	5219	503	4716					
Who: The author's role								
		Accuser	Bully	Defender	Reporter	Victim	Other	70%
Accuser	317	212	12	9	56	28	0	
Bully	303	24	165	4	45	65	0	
Defender	178	30	6	39	77	26	0	
Reporter	708	44	19	12	575	58	0	
Victim	589	19	16	3	63	488	0	
Other	7	0	0	0	6	1	0	
What: The form of bullying								
		General	Cyber-bullying	Verbal	Physical			91%
General	1857	1831	20	4	2			
Cyber-bullying	145	68	73	4	0			
Verbal	67	53	12	2	0			
Physical	33	32	0	0	1			
Why: The type of post								
		Accusation	Cyber-bullying	Denial	Report	Self-disclosure		72%
Accusation	316	196	0	2	54	64		
Cyber-bullying	16	4	0	0	7	5		
Denial	128	10	0	32	9	77		
Report	709	36	0	0	538	135		
Self-disclosure	933	48	0	3	132	750		

Tabla 4: Resultados de la predicción de bullying que demuestran acuerdo y desacuerdo con la codificación humana para la asignación de categorías utilizando métodos de aprendizaje automático. Fuente: (Bellmore et al., 2015)

Las publicaciones originadas en un estado en específico, se asoció directamente con la cantidad poblacional del estado, mientras que las temporalidades de las publicaciones revelan que estas se realizaron en mayor proporción los días entre semana, a diferencia de sábados y domingos, días en los cuales no se registró mayor actividad.

Con esto, se los autores procedieron al etiquetado de las trazas de bullying en los tweets, de acuerdo a las diferentes clasificaciones propuestas para dar respuesta a las preguntas clave. Es importante considerar la precisión que se tiene en el resultado de los modelos de machine learning, ya que esto está directamente relacionado con el tipo de codificación humana que se les dé a los modelos para el aprendizaje, y depende mucho de las características que se obtengan para cada clase.

El trabajo realizado por Nixon, Mercado, Faustino, Chuctaya, y Castro Gutierrez (2018), se traza como objetivo el utilizar técnicas de análisis de sentimientos para la

detección de cyber bullying en las redes sociales en lenguaje español. Para esto se utilizan técnicas de aprendizaje automático mediante el desarrollo de un clasificador Bayesiano, el cual fue entrenado con un conjunto de datos obtenido de la Facultad de Ciencias Exactas de la Universidad de San Agustín en Buenos Aires, (Ríos y Matias, 2015).

El conjunto de datos cuenta con un léxico de 2880 palabras las cuales han sido etiquetadas de manera manual basado en tres dimensiones de afectividad:

- Agradable (agradable, neutral, desagradable)
- Activación (activa, neutral, pasiva)
- Imaginabilidad (fácil de imaginar, neutral, difícil de imaginar)

Para la extracción de las características, se conformó una bolsa de palabras para la creación de un vocabulario y así asociar las palabras ofensivas y determinar las características de las expresiones para alimentar el modelo bayesiano.

Como resultado, el software proporciona el valor de sentimiento del texto o frase que está en el rango de 1 a 3, donde 1 es negativo, 3 es positivo y 2 es neutral. Además de la aceptación, que está en el rango de 0% a 100%, donde 0% es un texto o frase que tiene una alta probabilidad de contener intimidación y 100% una baja probabilidad de tenerlo.

Se realizó una validación individual para un caso de “sin bullying” introduciendo al sistema la expresión “pero que inteligente eres” clasificándola como el valor de sentimiento positivo con una aceptación del 95%. Posteriormente para un caso de “con bullying”, para lo cual se ingresa la expresión “eres la más estúpida e idiota”, asignando el modelo el valor de sentimiento negativo con una aceptación del 16%.

Figura 12: Validación de expresiones de “con bullying” y “sin bullying” del modelo bayesiano. Fuente: (Nixon et al., 2018)

El trabajo de investigación de Caizaluisa y Riofrio (2018), tiene como principal objetivo, desarrollar un aplicación para el control parental para dispositivos mobiles en android, mediante la generación alertas que son distribuidas por mensajes de emisor-receptor en Whatsapp, esto en base a la clasificación de textos utilizando tecnicas de machine learning con el entrenamiento del algoritmo Multinomial *Naíve Bayes.* Los datos para la conformación del corpus fuerón obtenidos de las paginas web asihablamos.com[13] y fepe55.com[14], considerando tres categorias para los datos de entrenamiento: SEXO, DROGAS, *BULLYING* y NEUTRAL. Se realizarón dos intentos de validación para que el modelo realice la prediccion de los textos en base a cada categoría, teniendo en el primer y segundo intento para BULLYING se obtuvo un 91.66% y 100% respectivamente, mientras que para DROGAS se obtuvo un 3.70% para los dos intentos. En cambio, en las categorías SEXO y NEUTRAL se obtuvo un 41.66% y un 100% en el primer intento, mientras que un 0%, en ambas categorías, en el segundo intento, lo cual evidencia una baja probabilidad de predicción para drogas y sexo.

[13] http://www.asihablamos.com/word/pais/EC/
[14] http://www.fepe55.com.ar/blog/2004/01/02/diccionario-los-codigos-de-la-droga/

2.3.3.1. Predicción de bullying usando aprendizaje profundo (*deep learning*)

Ptaszynski, Kalevi, Eronen, y Masui (2017), realizaron un trabajo de investigación, el cual consistió en detectar el *cyber bullying*, el mismo que ha sido reconocido como un problema social y ministerial en Japón (Mext, 2008), ya que, sus víctimas principalmente juveniles caen en depresión, se auto mutilan o incluso se suicidan.

Ptaszynski, Kalevi, Eronen, y Masui (2017), proponen generar un modelo de predicción de *cyber bullying*, utilizando las redes neuronales convolucionales (CNN) con técnicas de aprendizaje profundo y comparar con los modelos generados con otros métodos de predicción basados en fuerza bruta.

El conjunto de datos utilizado para el entrenamiento del modelo estaba en el idioma japonés, para lo cual tuvo que transcribir los textos a inglés y para esto se utilizó analizadores morfológicos de oraciones en japonés como son MeCab2[15] y CaboCha3[16], los cuales analizan las estructuras de los textos para procesar el conjunto de datos. Esto debido a que, cuanto más generalizada es una oración, menos únicos y frecuentes serán los patrones, por ejemplo, hay más patrones de Adjetivo y Nombre que "día agradable", tal como se puede ver en la Tabla 5.

[15] https://taku910.github.io/mecab/
[16] https://taku910.github.io/cabocha/

Sentence: 今日はなんて気持ちいい日なんだ！
Transcription in alphabet: *Kyōwanantekimochiiihinanda!*
Glosses: Today TOP what pleasant day COP EXCL
Translation: What a pleasant day it is today!

Preprocessing examples

–TOK: *Kyō* | *wa* | *nante* | *kimochiii* | *hi* | *nanda* | *!*
–POS: N | PP | ADV | ADJ | N | AUX | SYM
–TOK+POS: *Kyō*_N|*wa*_PP|*nante*_ADV|*kimochi_ii*_ADJ| *hi*_N| *nanda*_AUX|*!*_SYM
–CHNK: *Kyō_wa* | *nante* | *kimochi_ii* | *hi_nanda!*
–DEP: ⋆0_3D_*Kyō_wa* | ⋆1_2D_*nante* | ⋆2_3D_*kimochi_ii* | ⋆3_−1D_*hi_nanda!*

Tabla 5: Ejemplos de procesamiento de oraciones con los analizadores morfológicos MeCab2 y CaboCha3. Fuente: (Ptaszynski et al., 2017)

Para la obtención de las características se utilizan las palabras de los conjuntos de datos para la conformación de diferentes *bags of words* como son: Bolsa de Palabras / Fichas, Bolsa de *Lemmas*, Bolsa de *POS*, Bolsa de *Chunks*. Las oraciones que fueron procesadas del conjunto de datos son usadas en la capa de entrada para la clasificación.

Los resultados en la comparación de los modelos se encuentran resumidos en la Tabla 6, en donde se puede evidenciar que para los modelos generados por *kNN*, *Naive Bayes*, fueron bajos, posiblemente no pudieron captar el lenguaje ruidoso utilizado en el *cyber bullying*, sin embargo, se tuvo un mejor desempeño al compararse con los arboles de decisión *J48* o *JRip*, los cuales obtuvieron un mejor puntaje. *Random Forest* obtuvo buenos resultado al compararse con *SPEC*, desafortunadamente, es ineficiente en el tiempo de procesamiento, inclusive al compararse con *SVN*. Al verificar los resultados del modelo propuesto *CNN*, en la línea base utilizando una sola capa oculta, este no tuvo buenos resultados, aunque fue mejor comparándose con los demás modelos en la generación de su línea base. Sin embargo, en la generación del modelo final compuesto con dos capas ocultas para, el tamaño de la matriz en la capa de reducción de 5x5, y el uso de una función

estocástica con descenso de gradiente, supero a todos los clasificadores en casi todos sus ajustes, teniendo una puntuación final igual al 93.5%.

		LEM +POS	TOK +POS	LEM	TOK	CHNK +NER	POS	DEP +NER	DEP	CHNK	LEM +NER	TOK +NER
kNN (k=1)	Prec	.639	.630	.644	.630	.578	.544	.593	.628	.576	.668	.663
	Rec	.636	.627	.640	.628	.546	.543	.529	.528	.550	.663	.659
	F1	.633	.625	.637	.626	.505	.542	.446	.427	.494	.658	.656
	Acc	.636	.627	.640	.628	.546	.543	.529	.528	.550	.663	.659
Naïve Bayes	Prec	.678	.671	.686	.682	.666	.570	.652	.672	.685	.708	.703
	Rec	.674	.669	.682	.678	.627	.569	.578	.555	.598	.705	.701
	F1	.673	.668	.681	.677	.599	.568	.511	.453	.539	.705	.701
	Acc	.674	.669	.682	.678	.627	.569	.578	.555	.598	.705	.701
JRip	Prec	.606	.614	.604	.603	.628	.553	.643	.505	.685	.699	.700
	Rec	.606	.613	.603	.603	.555	.553	.533	.510	.598	.675	.672
	F1	.606	.613	.603	.603	.469	.553	.408	.345	.539	.663	.658
	Acc	.606	.613	.603	.603	.555	.553	.533	.510	.598	.675	.712
J48	Prec	.672	.671	.683	.675	.615	.566	.652	.260	.645	.711	.707
	Rec	.671	.666	.681	.672	.548	.566	.533	.510	.517	.708	.704
	F1	.670	.663	.680	.669	.458	.566	.408	.344	.365	.706	.707
	Acc	.671	.666	.681	.672	.548	.566	.533	.510	.517	.708	.663
Random Forest	Prec	.816	.803	.818	.809	.662	.547	.623	.619	.639	.818	.809
	Rec	.809	.795	.809	.801	.632	.546	.607	.582	.580	.809	.802
	F1	.808	.794	.808	.799	.610	.544	.590	.540	.522	.807	.800
	Acc	.809	.795	.809	.801	.632	.546	.607	.582	.580	.809	.802
SVM linear	Prec	.777	.768	**.827**	.777	.679	.563	.651	.639	.606	.820	.781
	Rec	.777	.766	**.825**	.776	.645	.563	.615	.577	.603	.818	.781
	F1	.776	.766	**.825**	.775	.623	.563	.586	.531	.597	.818	.780
	Acc	.777	.766	**.825**	.776	.645	.563	.615	.577	.603	.818	.781
SVM plynomial	Prec	.262	.499	.262	.263	.260	.553	.260	.260	.260	.260	.260
	Rec	.512	.499	.512	.513	.510	.545	.510	.510	.510	.510	.510
	F1	.346	.450	.347	.348	.344	.528	.344	.344	.344	.344	.344
	Acc	.512	.499	.512	.513	.510	.545	.510	.510	.510	.510	.510
SVM radial	Prec	.797	.753	.262	.793	.260	.565	.260	.260	.260	.752	.779
	Rec	.771	.747	.512	.756	.510	.565	.510	.510	.510	.516	.778
	F1	.765	.746	.347	.746	.344	.565	.344	.344	.344	.358	.778
	Acc	.771	.747	.512	.756	.510	.565	.510	.510	.510	.516	.778
SVM sigmoid	Prec	.757	.746	.262	.752	.260	.562	.260	.260	.260	.260	.771
	Rec	.549	.736	.512	.538	.510	.562	.510	.510	.510	.510	.606
	F1	.425	.733	.347	.403	.344	.561	.344	.344	.344	.344	.530
	Acc	.549	.736	.512	.538	.510	.562	.510	.510	.510	.510	.606
SPEC (highest BEP)	Prec	.802	.786	.784	.770	.655	.614	.548	.591	.633	N/A	N/A
	Rec	.802	.786	.784	.770	.655	.614	.548	.591	.633	N/A	N/A
	F1	.802	.786	.784	.770	.655	.614	.548	.591	.633	N/A	N/A
	Acc	.775	.780	.600	.765	.790	.635	.525	.640	.548	N/A	N/A
SPEC (highest F1)	Prec	.807	.756	.713	.724	.563	**.528**	.500	**.491**	.490	N/A	N/A
	Rec	.798	.839	.885	.842	.879	**.946**	.982	**1.000**	1.000	N/A	N/A
	F1	.803	.796	.790	.778	.686	**.677**	.663	**.658**	.658	N/A	N/A
	Acc	.808	.784	.770	.766	.603	**.550**	.510	**.491**	.490	N/A	N/A
CNN (1 hidden layer)	Prec	.700	.678	.724	.704	.506	.544	.616	.499	.496	.720	.739
	Rec	.700	.678	.724	.704	.510	.544	.525	.505	.509	.716	.738
	F1	.700	.677	.724	.704	.433	.544	.393	.461	.374	.714	.738
	Acc	.700	.678	.724	.704	.510	.544	.525	.505	.509	.716	.738
CNN (2 hidden layers)	Prec	**.867**	**.824**	.820	**.833**	**.936**	.500	**.929**	N/A	**.899**	**.847**	**.849**
	Rec	**.866**	**.821**	.819	**.831**	**.935**	.500	**.927**	N/A	**.893**	**.847**	**.849**
	F1	**.866**	**.821**	.819	**.830**	**.935**	.495	**.927**	N/A	**.892**	**.847**	**.849**
	Acc	**.866**	**.821**	.819	**.831**	**.935**	.500	**.927**	N/A	**.893**	**.847**	**.849**

Tabla 6: Resultados de todos los clasificadores aplicados. Fuente: (Ptaszynski et al., 2017)

En la investigación realizada por Agrawal y Awekar (2018), se aborda en plantear una solución a la problemática existente en la detección del *cyber bullying*, mediante la utilización de modelos basado en el aprendizaje profundo. Para esto, se han realizado varios experimentos utilizando tres diferentes conjuntos de datos, obtenidos de tres diferentes

fuentes del mundo real: *Formspring* con 12 mil *posts* aproximadamente, *Twitter* con 16 mil *posts* aproximadamente y *Wikipedia* con 100 mil *posts* aproximadamente.

Se ha experimentado diversos modelos tradicionales que aplican técnicas de machine learning como son: Regresión Logística, *SVM*, *Random Forest* y *Naive Bayes*, así como también se han utilizado modelos de redes neuronales aplicando técnicas de aprendizaje profundo como son: *CNN*, *LSTM*, *BLSTM*, *BLSTM*. Además, se utilizó varios métodos para la generación de características basado en la representación de palabras, como: bolsa de caracteres n -grama, bolsa de palabra unigram, incrustaciones GloVe, incrustaciones SSWE.

Al experimentar con los modelos tradicionales de machine learning, se utilizó representaciones de palabras basadas en uni-gramas y n-gramas, cuyas funciones han demostrado ser poderosas para la obtención de características lingüísticas y sintácticas, los lex-iconos creados a mano y las incrustaciones de palabra y párrafos. Al verificar los resultados obtenidos en la Tabla 7 y compararlos con los resultados de los modelos de aprendizaje profundo, se tiene un resultado significativamente menor, en lo que respecta a la métrica F1 score[17].

Dataset	label	Character n-grams				Word unigrams			
		LR	SVM	RF	NB	LR	SVM	RF	NB
Formspring	bully	0.448	0.422	0.298	0.359	0.489	0.463	0.264	0.025
Twitter	racism	0.723	0.676	0.752	0.686	0.738	0.772	0.739	0.617
	sexism	0.729	0.688	0.720	0.647	0.762	0.758	0.755	0.635
Wiki	Attack	0.694	0.677	0.674	0.655	0.711	0.686	0.730	0.659

Tabla 7: Resultados obtenidos en la aplicación de los modelos de machine learning con la métrica F1 score. Fuente: (Agrawal y Awekar, 2018)

[17] Calcula la medida f, la media armónica de precisión y recuperación. https://faroit.github.io/keras-docs/1.2.2/metrics/

Para la comparación de las diferentes técnicas de aprendizaje profundo se han utilizado las siguientes nomenclaturas:

- **Conjuntos de datos con sobre muestreo de publicaciones de acoso:** F + (*Formspring*), T + (*Twitter*), W + (*Wikipedia*)
- **Medidas de rendimiento:** P (Precisión), R (Recuperación), F1 (*F1 Score*)
- **Modelos Aprendizaje Profundo:** *M1* (*CNN*), *M2* (*LSTM*), *M3* (*BLSTM*), M4 (*BLSTM* con atención)

Al verificar en la Tabla 8 los resultados obtenidos de rendimiento de los diferentes modelos de aprendizaje profundo, se tiene que el modelo *CNN* presenta mejores resultados. Si comparamos la métrica de *F1 score*, que vendría a ser la media entre la precisión y la recuperación, tenemos el 91% para *bully*, el 93% para racismo, el 92% para sexismo y el 87% para ataques.

Dataset	Label	P				R				F1			
		M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
F+	bully	0.93	0.91	0.91	0.90	0.90	0.85	0.81	0.91	0.91	0.88	0.86	0.91
T+	racism	0.93	0.91	0.92	0.90	0.94	0.80	0.95	0.96	0.93	0.85	0.93	0.93
	sexism	0.92	0.84	0.88	0.88	0.92	0.93	0.94	0.92	0.92	0.88	0.92	0.90
W+	Attack	0.92	0.70	0.90	0.87	0.83	0.54	0.81	0.86	0.88	0.61	0.85	0.87

Tabla 8: Comparación del rendimiento de los modelos de aprendizaje profundo utilizados. Fuente: (Agrawal y Awekar, 2018)

2.4.Discusión.

Con la revisión de la literatura de los trabajos realizados por varios autores, y que se encuentran relacionados con los objetivos planteados en esta tesis, se puede concordar con la importancia en la utilización de técnicas para el análisis de los datos. Se ha podido constatar que, al utilizar estas técnicas, es posible descubrir casos como: analizar perfiles de

usuarios en redes sociales para realizar marketing efectivo, analizar sentimientos y hasta identificar cuando existe un ataque de *cyber bullying*. Si se establece una analogía entre los tres casos expuestos, se puede observar que, en los diferentes trabajos de investigación en muchos de los casos se utilizan las mismas técnicas, pero con fines distintos.

Por ejemplo, si se considera el trabajo de Pennebaker (2002), en el cual, se establece una colaboración interdisciplinaria entre la Psicología y Ciencia de los Datos, para entender el comportamiento de las personas, a través de los medios sociales. Este autor concuerda con Hu y Liu (2015), quien a su criterio ve la oportunidad de utilizar estos datos expuestos, para establecer un perfilamiento en base al comportamiento de los usuarios de las redes sociales, esto con el fin de conocer sus gustos, relaciones, datos que según Schiaffino y Amandi, (2009), pueden ser usados para publicidad y marketing efectivo.

Al establecer una relación entre el comportamiento social de la gente y la manera de expresarse, es que, aparece la necesidad de analizar los sentimientos de las personas, para esto, en los trabajos de Lu, Sakamoto, Shibuki, y Mori (2017) y Severyn, Inc, y Moschitti (2015), se realiza un análisis de sentimientos, con datos expresados en las redes sociales como *Twitter* y *Weibo* respectivamente, ambos utilizan diferentes técnicas tanto de *machine learning* como de *deep learning*, y coinciden en el uso de la red neuronal convolucional *CNN*, lo que se diferencia son los datos de entrenamiento y la arquitectura utilizada en la red, pero concuerdan en que, los modelos generados por esta red neuronal genera buenos resultados para el procesamiento del lenguaje natural.

De acuerdo con Torregrosa y Crespo (1984), especifica que las emociones y sentimientos humanos son una construcción social y el resultado de las relaciones sociales,

y según Ortega Mora y Ordoñez Ordoñez (2013), el *bullying* es un fenómeno multicausal, generado a partir de factores individuales, familiares, socioculturales y escolares, entonces, es posible ubicar a las expresiones de *bullying* como sentimientos negativos utilizados por un agresor para intimidar, acosar y atacar a una víctima. De acuerdo a esto es que los autores Bellmore, Calvin, Xu, y Zhu (2015), Ptaszynski, Kalevi, Eronen, y Masui (2017) y Agrawal y Awekar (2018), realizan trabajos de investigación enfocados en establecer métodos para la detección de bullying canalizado a través de diferentes redes sociales. De igual manera, se utilizan varias técnicas de *machine learning* y *deep learning* para la generación de modelos de predicción, en todos los trabajos coinciden en que, el utilizar una red neuronal convolucional, se puede obtener buenos resultados en la precisión de entrenamiento y validación del modelo. En todos estos trabajos, se considera como base de datos para el entrenamiento a conjunto de palabras en el idioma inglés, a diferencia del trabajo Ptaszynski, Kalevi, Eronen, y Masui (2017), quien debe realizar una transcripción del Japonés a Ingles para lo cual utiliza procesos analizadores morfológicos.

Únicamente el trabajo de (Nixon et al., 2018) realiza un entrenamiento de un modelo de machine learning utilizando un conjunto de datos en español para la generación de un modelo con una red neuronal bayesiana, este trabajo se enfoca en utilizar una bolsa de palabras ya etiquetadas en tres dimensiones de afectividad: agradable, activación e imaginabilidad, que se basan generalmente en un análisis de sentimiento, al considerar la aceptación para la predicción de *bullying*, se puede evidenciar que es del 16% lo cual representa un valor bajo. A diferencia del trabajo realizado en esta investigación, se ha clasificado de manera manual textos específicamente para el aprendizaje y predicción de *cyber bullying*, además se considera las comparaciones realizadas en los trabajos de

Bellmore, Calvin, Xu, y Zhu (2015), Ptaszynski, Kalevi, Eronen, y Masui (2017) y Agrawal y Awekar (2018), quienes concuerdan en que al usar un modelo generado por una CNN, se puede obtener mejores resultados.

En base a lo antes expuesto, es importante considerar que la selección de la técnica de minería para realizar el análisis de los datos, dependerá mucho de la necesidad o el enfoque se le quiera dar. Esto permitirá establecer conclusiones claras que permitan tomar decisiones prudentes y proponer soluciones a los diferentes problemas planteados. Por lo tanto, este tema de investigación proporciona como aporte un corpus clasificado y etiquetado con expresiones en el lenguaje español, para entrenar y generar un modelo que permita predecir e identificar nuevos textos con signos de *cyber bullying*. Además, los resultados obtenidos en este trabajo pueden ser discutidos y analizados en futuras investigaciones para obtener un beneficio potencial para implementar aplicaciones que permitan fusionar métodos de ciencias sociales y ciencias de la computación para alertar el acoso escolar en el Ecuador.

CAPÍTULO III

SOLUCIÓN ADOPTADA

3.1. Introducción

La utilización de la tecnología para el procesamiento y entendimiento del lenguaje de las personas, se ha convertido en los últimos años, en un factor muy importante en la analítica de datos. Las técnicas de aprendizaje de máquina, así como también el aprendizaje profundo, son muy utilizadas para abstraer situaciones que suceden en nuestro entorno natural y generar procesos que permitan automatizar funciones que anteriormente eran de exclusividad de los seres humanos (Broncano, 1995). El uso y la evolución de las tecnologías de información y comunicación, ha provocado que se produzcan millones de datos que pueden ser de gran utilidad para análisis y la generación de conocimiento. Entre estos datos, se encuentran los comentarios y vivencias expresadas por la gente en las redes sociales o blogs, entre los cuales, se puede materializar ataques de acoso y violencia.

De acuerdo a esto, es importante comprender, el lenguaje natural y llegar a predecir cuándo un mensaje, post o chat, es utilizado para acosar o agredir a otra persona, por medio de un medio digital. Para esto, en este trabajo de investigación, se ha entrenado una red neuronal, para el aprendizaje de expresiones naturales que han sido extraídas de *Twitter* y que tenga la capacidad de predecir cuándo un texto tiene o no, signos de ataques de *bullying*.

3.2.Descripción de la arquitectura

3.2.1. Análisis de solución

La solución propuesta, requiere cumplir las siguientes necesidades funcionales:

Necesidad Funcional	Descripción de la necesidad funcional
NF. 1	Extraer y consolidar un corpus con textos expresados en lenguaje natural, los textos deben ser clasificados y etiquetados de la siguiente manera: Cuando corresponden a un ataque de bullying con uno (1) Cuando no corresponden a un ataque de bullying con cero (0)
NF. 2	Crear un modelo de predicción, mediante el entrenamiento de una red neuronal, para el aprendizaje profundo de textos que son bullying y no bullying.

Tabla 9: Necesidades Funcionales de la solución adoptada. Fuente: Desarrollo del investigador

3.2.2. Diseño de solución

A continuación, se describe el diseño de la solución adoptada, la cual, tiene como objetivo definir la arquitectura y la estrategia técnica requerida, para la extracción de los datos de texto (*tweets*) y el desarrollo de la Red Neuronal Convolucional usando un módulo de aprendizaje profundo.

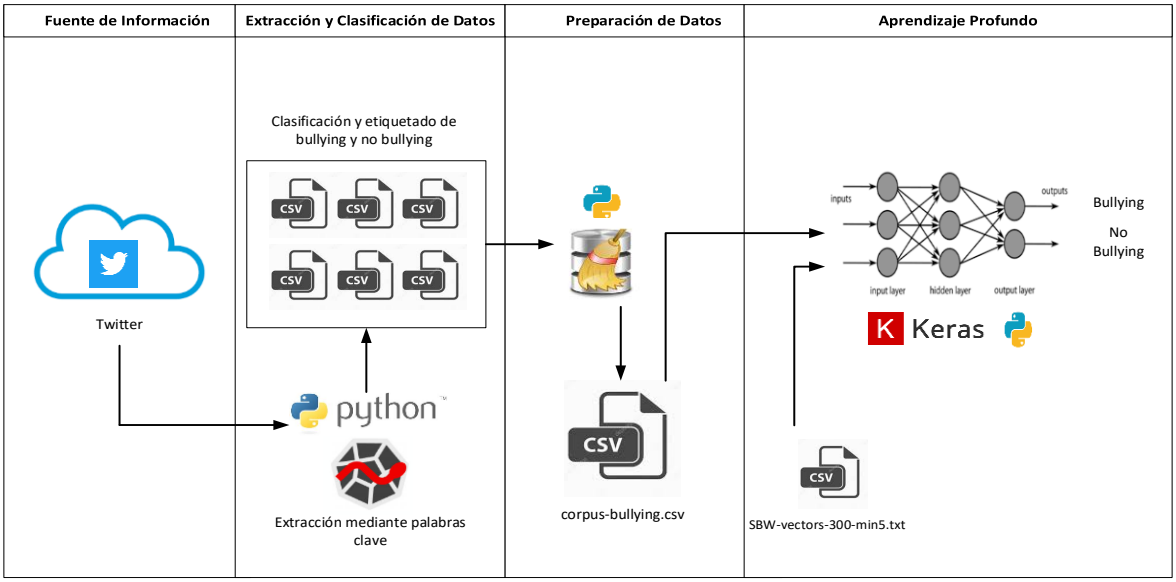

Figura 13: Arquitectura de la solución adoptada. Fuente: Desarrollo del investigador

En la Figura 13, se puede observar una arquitectura estructurada en capas, lo cual permite establecer un flujo, controlado e independiente de la información. Cada capa, determina un proceso o procesos utilizados para la preparación de la red neuronal, capas que son detalladas a continuación.

Fuente de datos: En esta capa, se definen las fuentes, de las cuales se ha obtenido la información, para entrenar el modelo. Para la selección de la fuente de información, se ha considerado el volumen de usuarios de redes sociales en el país, según el reporte de Del Alcázar (2018), y tal como se puede observar en la Figura 14, a julio del 2018, las redes sociales con más usuarios en el son Facebook, *Twitter*, *LinkedIn* e *Instagram*.

En base a esto, se ha utilizado como fuente de datos a *Twitter*, ya que, cuenta con 800 mil usuarios activos en el Ecuador.

Figura 14: Usuarios de Redes Sociales en Ecuador, Fuente: (Del Alcázar, 2018)

Extracción y clasificación de datos: Esta capa, define el esquema de extracción de los datos; para esto se ha utilizado un script desarrollado en *Python*, el cual interactúa con la red social *Twitter* a través de su API para la creación de aplicaciones. Posteriormente, se realiza

un proceso de revisión y clasificación manual de los datos, los cuales son utilizados en el proceso de enseñanza de la red neuronal.

Preparación de los datos: En esta capa, se definen los procesos preparatorios que se usaron para mejorar la calidad de los datos. Esto, con el fin de que no ingresen palabras o símbolos que puedan ocasionar ruido en el entrenamiento del modelo. Para esto, también se utilizó un script desarrollado en *Python*.

Aprendizaje profundo: En esta capa, se definen los procesos para realizar el entrenamiento de la red neuronal para la obtención del modelo de predicción. Se utiliza varios programas desarrollados en *Python*, con un módulo de inteligencia artificial denominado *Keras*[18], el cual tiene la capacidad de permitir una interpretación inmediata de las características de los datos. Presenta un enfoque de uso fácil, modular y extensible. (MkDocs, 2014).

Para el desarrollo e implementación de la arquitectura definida, se ha utilizado en todas sus capas, la plataforma *Anaconda*[19], con *Spyder*[20] como entorno de desarrollo y el lenguaje programación *Python*, debido a que, es flexible, presenta un intérprete de comando interactivo y tiene un alto potencial en el análisis y ciencia de datos.

3.3. Modelo

3.3.1. Modelo de aprendizaje profundo para clasificación de texto

Según lo manifestado por Kim (2014), las redes neuronales convolucionales (*CNN*), han tenido un importante crecimiento y mejora en la clasificación de oraciones, esto debido

[18] https://keras.io/getting-started/faq/
[19] https://www.anaconda.com/
[20] https://www.spyder-ide.org/

a que este tipo de redes pueden capitalizar la distribución de las palabras en una oración, convirtiéndolas en una matriz de vectores. Aunque originalmente las *CNN* fueron diseñadas para la clasificación de imágenes, estas han tendido un importante desenvolvimiento en otras tareas, inclusive en el procesamiento de lenguaje natural, (Collobert y Weston, 2011). En base a esto, se ha utilizado el modelo *CNN*, para realizar la clasificación de los *tweets*, usando las ventajas que prestan las librerías de aprendizaje profundo de *Python*.

Este tipo de red neuronal, tiene la capacidad de analizar los datos expresados secuencialmente en los textos de los *tweets*, debido a su estructura en capas para realizar convolución temporal y la serie de núcleos y filtros utilizados para la extracción de las características del corpus (Kim, 2014)

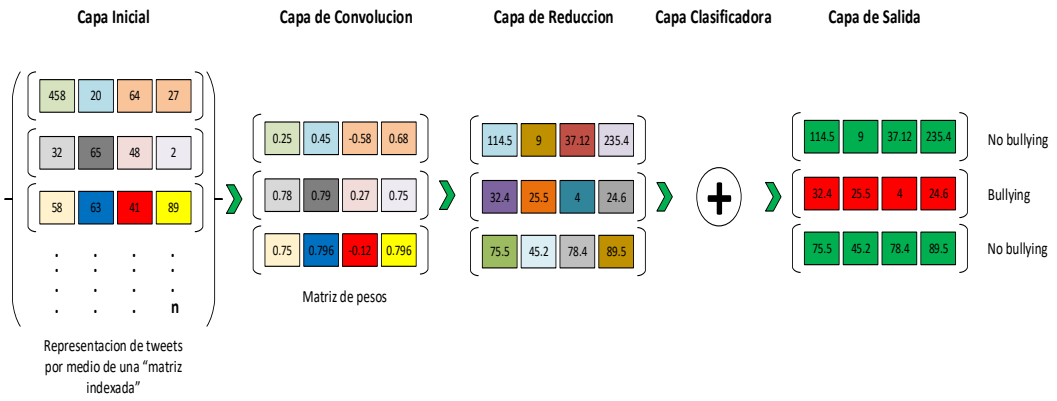

Figura 15: Funcionamiento de la CNN para clasificación de tweets. Fuente: desarrollo del investigador

3.3.2. Recopilación de datos

De acuerdo con lo establecido en el diseño de la solución descrito en la Figura 13, se tiene las capas de fuentes extracción y clasificación de datos. Para esto se realizó la recopilación de los datos, mediante un *script* en el lenguaje de programación *Python*, el cual, interactúa con la API de *Twitter* para la obtención de la información. Este *script* fue

desarrollado en la aplicación *Spider* 3.2.6[21], disponible en la plataforma *Anaconda Navigator*[22], como se puede observar en la Figura 16.

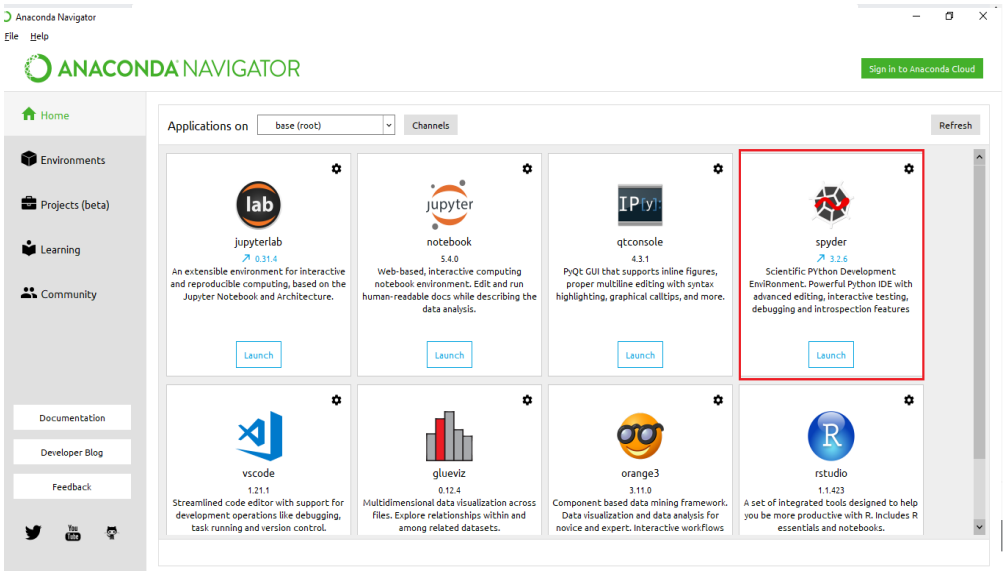

Figura 16: Panel de Aplicaciones de Anaconda Navigator. Fuente: (Anaconda, 2018)

La librería que se usó para interactuar entre *Python* y la API de *Twitter* se denomina *tweepy*[23], esta, permite la obtención de atributos de los objetos almacenados en *Twitter*. El acceso a Twitter se estableció mediante la creación de una *APP* en el entorno de desarrollo de la red social. La *APP* genera credenciales de comunicación mediante claves de consumo y *tocken* de acceso, estas credenciales son utilizadas por la librería *tweepy* para establecer la comunicación para la extracción de datos, como se muestra en el siguiente código fuente:

```
# Access Token for Twitter App
access_token = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
access_token_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'

# Consumer Keys
consumer_key = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
```

[21] Ambiente de desarrollo analítico e IDE para Python con características de edición avanzada, pruebas y depuración interactiva.
[22] Interfaz gráfica de usuario cuya distribución permite iniciar aplicaciones y administrar fácilmente paquetes, entornos y canales CONDA sin usar comandos de línea de comandos.
[23] http://www.tweepy.org/

```
consumer_secret = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth, wait_on_rate_limit=True,
wait_on_rate_limit_notify=True, compression=True)
```

3.3.2.1.Palabras clave utilizadas para la extracción de datos

Para la búsqueda y extracción de los *tweets*, se consideró, el estudio realizado por Del Rey, Genevat, y Ortega (2002), en el cual, se analizaron la etiquetas que se utilizaron en los jóvenes estudiantes entre 8 y 14 años de edad para nominar el fenómeno de *bullying*. Como instrumento se usaron tarjetas que representaban situaciones concretas de violencia interpersonal. En el grupo de agresión verbal, se agruparon desde insultos, ataques racistas y violencia verbal grave.

Se utilizó un listado de palabras clave para la extracción de datos descritas en la Tabla 10 y Tabla 11 para bullying y no bullying, respectivamente. La query realizada establece los parámetros *lang="es"* y *location = "EC" y tweet_mode = 'extended'* para la obtención del tweet completo, como se muestra en el siguiente código fuente:

```
limit=10000 #numero del limite de tweets a obtener
query="(eres AND basura)" #palabras clave para la búsqueda del tweet
since="2013-01-01" #fecha desde
until="2018-12-31" #fecha hasta

try:
    tweets = tweepy.Cursor(api.search,
                           q=query,
                           since=since,
                           until=until,
                           lang="es",
                           location = "EC",
                           tweet_mode='extended').items(limit)

    with open(file_name, 'wt', encoding="utf8") as f:
        writer = csv.writer(f, delimiter = "|")
        writer.writerows([tweet.author.screen_name, tweet.created_at,
tweet.full_text] for tweet in tweets)
```

Palabras clave bullying			
care asno	care culo	care huevo	care indio
eres regalada	homosexual	gay	maldita
maricon	mamerto	mojigato	mojigata
nerd	ojala mueras	orejas elefante	orejon
te den culo	te odio	tetas chuecas	tetas chuecas
care simio	care verga	eres gorda	care pene
borrego	maldito	marica	muerete
narizon	negro	apestoso	nalga
parece cerda	parece marrana	parece marrano	parece morza
eres deprimente	eres discapacitado	eres patetico	eres tonto

Tabla 10: Palabras clave para la búsqueda de bullying . Fuente: desarrollo del investigador

Palabras clave no bullying		
gusta	me gustas	linda
gustoso	eres mejor	lindo
pago entradas	animo	enamorarse
hola	amiga	frio
amigo	cuidate	calor
te quiero	gustoso conocerte	tener autoestima
te amo	como estas	vive feliz

Tabla 11: Palabras clave para la búsqueda de no bullying . Fuente: desarrollo del investigador

3.3.3. Preparación de datos

Es importante que los datos que ingresen al entrenamiento del modelo sean de calidad y que las palabras o símbolos que no agregan valor al significado del texto sean eliminados, esto debido a que palabras de uso común, como “el”, “la”, “un”, en el procesamiento de lenguaje natural, son conocidas como palabras vacías o “*stop words*” y que, en el aprendizaje del modelo, pueden ocupar espacio en la base de datos y tiempo de procesamiento

innecesario. Para esto, se utilizó funciones de limpieza para los textos en los tweets, con el fin de eliminar texto no relevante para la clasificación. Las acciones realizadas a los tweets fueron las siguientes:

- Conversión de texto a minúsculas.
- Reemplazar dos o más puntos por un espacio.
- Eliminar todos los caracteres especiales utilizados en el tweet. Se dejan únicamente letras.
- Eliminar las referencias de vínculos externos “http”, “https”, “www”.
- Eliminar las menciones a otros usuarios “@usuario”.
- Eliminar las palabras vacías (*stop words*).

Se han excluido las palabras que se encuentran en la librería *NLTK* el diccionario stopwords en español, y se han añadido a dicho diccionario las siguientes palabras: ‘rt’, ‘d’, ‘l’, ‘s’, ‘c’, ‘g’, ‘qu’, ‘n’, ‘m’, ‘re’, ‘lt’, ‘gt’, ‘x’, ‘q’, ‘xq’, ‘cl’.

A continuación, se describe el conjunto de stopwords que se han eliminado del corpus:

```
{'a', 'al', 'algo', 'algunas', 'algunos', 'ante', 'antes', 'c', 'cl',
'como', 'con', 'contra', 'cual', 'cuando', 'd', 'de', 'del', 'desde',
'donde', 'durante', 'e', 'el', 'ella', 'ellas', 'ellos', 'en', 'entre',
'era', 'erais', 'eran', 'eras', 'eres', 'es', 'esa', 'esas', 'ese',
'eso', 'esos', 'esta', 'estaba', 'estabais', 'estaban', 'estabas',
'estad', 'estada', 'estadas', 'estado', 'estados', 'estamos', 'estando',
'estar', 'estaremos', 'estará', 'estarán', 'estarás', 'estaré',
'estaréis', 'estaría', 'estaríais', 'estaríamos', 'estarían', 'estarías',
'estas', 'este', 'estemos', 'esto', 'estos', 'estoy', 'estuve',
'estuviera', 'estuvierais', 'estuvieran', 'estuvieras', 'estuvieron',
'estuviese', 'estuvieseis', 'estuviesen', 'estuvieses', 'estuvimos',
'estuviste', 'estuvisteis', 'estuviéramos', 'estuviésemos', 'estuvo',
'está', 'estábamos', 'estáis', 'están', 'estás', 'esté', 'estéis',
'estén', 'estés', 'fue', 'fuera', 'fuerais', 'fueran', 'fueras',
```

```
'fueron', 'fuese', 'fueseis', 'fuesen', 'fueses', 'fui', 'fuimos',
'fuiste', 'fuisteis', 'fuéramos', 'fuésemos', 'g', 'gt', 'ha', 'habida',
'habidas', 'habido', 'habidos', 'habiendo', 'habremos', 'habrá',
'habrán', 'habrás', 'habré', 'habréis', 'habría', 'habríais',
'habríamos', 'habrían', 'habrías', 'habéis', 'había', 'habíais',
'habíamos', 'habían', 'habías', 'han', 'has', 'hasta', 'hay', 'haya',
'hayamos', 'hayan', 'hayas', 'hayáis', 'he', 'hemos', 'hube', 'hubiera',
'hubierais', 'hubieran', 'hubieras', 'hubieron', 'hubiese', 'hubieseis',
'hubiesen', 'hubieses', 'hubimos', 'hubiste', 'hubisteis', 'hubiéramos',
'hubiésemos', 'hubo', 'l', 'la', 'las', 'le', 'les', 'lo', 'los', 'lt',
'm', 'me', 'mi', 'mis', 'mucho', 'muchos', 'muy', 'más', 'mí', 'mía',
'mías', 'mío', 'míos', 'n', 'nada', 'ni', 'no', 'nos', 'nosotras',
'nosotros', 'nuestra', 'nuestras', 'nuestro', 'nuestros', 'o', 'os',
'otra', 'otras', 'otro', 'otros', 'para', 'pero', 'poco', 'por',
'porque', 'q', 'qu', 'que', 'quien', 'quienes', 'qué', 're', 'rt', 's',
'se', 'sea', 'seamos', 'sean', 'seas', 'sentid', 'sentida', 'sentidas',
'sentido', 'sentidos', 'seremos', 'será', 'serán', 'serás', 'seré',
'seréis', 'sería', 'seríais', 'seríamos', 'serían', 'serías', 'seáis',
'siente', 'sin', 'sintiendo', 'sobre', 'sois', 'somos', 'son', 'soy',
'su', 'sus', 'suya', 'suyas', 'suyo', 'suyos', 'sí', 'también', 'tanto',
'te', 'tendremos', 'tendrá', 'tendrán', 'tendrás', 'tendré', 'tendréis',
'tendría', 'tendríais', 'tendríamos', 'tendrían', 'tendrías', 'tened',
'tenemos', 'tenga', 'tengamos', 'tengan', 'tengas', 'tengo', 'tengáis',
'tenida', 'tenidas', 'tenido', 'tenidos', 'teniendo', 'tenéis', 'tenía',
'teníais', 'teníamos', 'tenían', 'tenías', 'ti', 'tiene', 'tienen',
'tienes', 'todo', 'todos', 'tu', 'tus', 'tuve', 'tuviera', 'tuvierais',
'tuvieran', 'tuvieras', 'tuvieron', 'tuviese', 'tuvieseis', 'tuviesen',
'tuvieses', 'tuvimos', 'tuviste', 'tuvisteis', 'tuviéramos',
'tuviésemos', 'tuvo', 'tuya', 'tuyas', 'tuyo', 'tuyos', 'tú', 'un',
'una', 'uno', 'unos', 'vosostras', 'vosostros', 'vuestra', 'vuestras',
'vuestro', 'vuestros', 'x', 'xq', 'y', 'ya', 'yo', 'él', 'éramos'}
```

3.3.4. Conformación del corpus

El corpus fue estructurado en un archivo denominado "*corpus-bullying.csv*", con separación de comas (csv), en éste se consolidaron los textos depurados mediante el proceso de limpieza descrito en el punto anterior.

Previo a la ejecución al proceso de limpieza, se realizó una revisión y constatación del texto que contiene cada tweet con el fin de realizar la clasificación manual del mismo y etiquetar con cero (0) y uno (1), a los tweets "*no bullying*" y "*bullying*" respectivamente.

Un tweet fue clasificado como bullying o no bullying, considerando lo expuesto por Del Rey, Genevat, y Ortega (2002) y las palabras clave definidas en la Tabla 10 y Tabla 11. Si, el contenido del texto expresa signos de insultos, ataques racistas y/o violencia verbal grave y además incluía alguna de las palabras clave de "bullying" el tweet fue etiquetado con uno (1). Si, el contenido del texto no expresaba algún signo de ataque e incluía alguna de las palabras clave de "no bullying", el tweet fue etiquetado con cero, como se muestra en la Tabla 12.

Id	Text Tweet	Bully
1	minis comprendi volver enamorarse necesario vaciar corazon recuerdos viejos	0
2	tonto	1
3	miedo enamorarse alguien mayor amor conoci amor vida solo persona mas	0
4	cuenta alguien dice quiere enamorarse tener relacion refiere quiere rompa	0
5	si ven van confiar kelly cara de perra	1
6	da chingo miedo alguien mas vea vi brillas manera tan hermosa imponente	0
7	dicen malas lenguas juego juegas gente real v	0
8	saw miedo enamorarse alguien mayor amor conoci amor vida solo persona mas	0
9	ven aqui hijo mas tonto afilar mazas ven	1
10	historiador instruido as perra gorda	1

Tabla 12: Estructura del archivo corpus-bullying.csv. Fuente: Desarrollo del Autor

3.3.5. Análisis del corpus

El corpus se estructura de un total de 83,400 tweets, los mismos que fueron recopilados utilizando las palabras clave, descritas en la Tabla 10 y Tabla 11. Estos tweets, describen textos que poseen signos de *bullying* y *no bullying*, los mismos que han sido revisados y etiquetados respectivamente, tal como se explicó anteriormente.

En el trabajo realizado por Kim (2018), se hizo un análisis de sentimientos en twitter. Este trabajo consta de varias partes, y específicamente en la parte tercera se analiza visualmente el corpus público "*sentiment140*[24]", en el cual se expresan *tokens* de tweets etiquetados con sentimiento positivo y negativo. Estos textos son analizados mediante la distribución de las palabras utilizadas en las expresiones, basándose en lo que se manifiesta en la ley de Zipf (1950).

Se ha analizado el "*corpus-bulling.csv*", para lo cual se leyó dicho corpus mediante la librería pandas[25], la cual es utilizada para el análisis de datos. El conjunto de tweets se almacena en un objeto de conjunto de datos, conocido como *data frame*. Al obtener el resumen del mismo se tiene lo siguiente:

```
RangeIndex: 83400 entries, 0 to 83399
Data columns (total 3 columns):
id              83400 non-null int64
tweet_text      83400 non-null object
bully           83400 non-null int64
dtypes: int64(2), object(1)
memory usage: 1.9+ MB
```

***Range Index*:** Especifica el rango de los valores de entrada en el conjunto de datos, en este caso 83400 de 0 a 83399.

***Data Columns*:** Corresponde a los nombres de las columnas con su respectivo tipo de dato y cuantos valores no ausentes se encuentran en cada una.

***Dtypes*:** Indica los tipos de datos de las columnas existentes en el conjunto de datos.

24 http://help.sentiment140.com/for-students/
25 https://pandas.pydata.org/

***Memory usage*:** Indica la cantidad de memoria que está usando el conjunto de datos.

Como se explicó anteriormente, el conjunto de datos del "*corpus-bulling.csv*" tiene un total de 83400 *tweets*, los cuales fueron clasificados manualmente. De estos, 16247 han sido etiquetados como "*bullying*" y 67153 como "*no bullying*", según se muestra en la Figura 17.

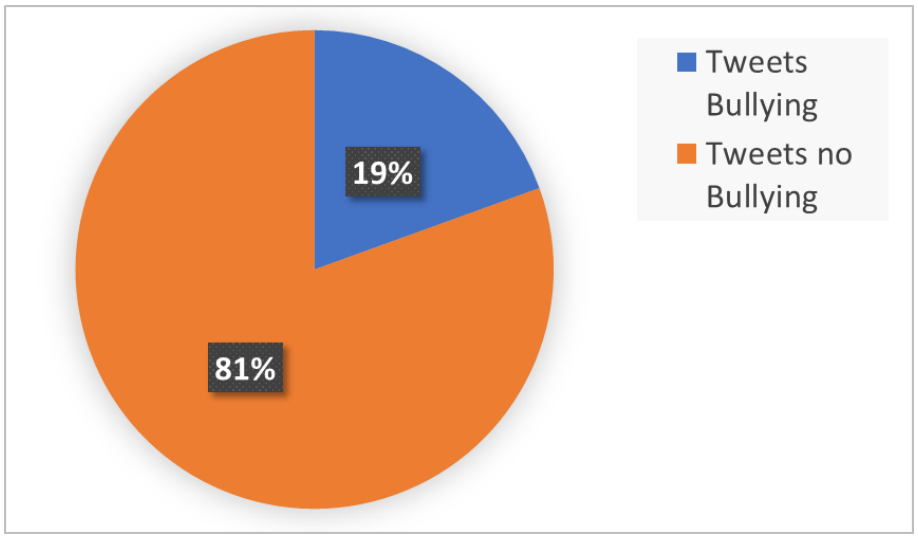

Figura 17: Participación de Bullying en el corpus. Fuente: Desarrollo del investigador

Este corpus contiene un sin número de palabras, como se puede apreciar en la Figura 18 y Figura 19, las cuales representan a un análisis visual de las palabras más importantes que constituyen el corpus. Estas figuras permiten identificar las palabras que más se usan dentro de las expresiones de los *tweets*, donde aquellas palabras que aparecen más veces, son representadas por la palabra con mayor tamaño en la nube de palabras, por ejemplo, en el caso de los tweets etiquetados con bullying, la palabra que más se usa es "tonto" y para

los tweets etiquetados como no bullying se tiene a “hola”, “amigo”. Para esta visualización, se usó la librería de *Python worldcloud*[26] (nube de palabras).

Figura 18: Nube de palabras de los tweets etiquetados como bullying. Fuente: Desarrollo del investigador.

Figura 19: Nube de palabras de los tweets etiquetados como no bullying. Fuente: Desarrollo del investigador.

26 https://amueller.github.io/word_cloud/generated/wordcloud.WordCloud.html

Análisis de distribución lingüística.

Según lo manifestado por la ley de Zipf (1950), en la comunicación lingüística de las personas, se utilizan un número reducido de palabras todo el tiempo, mientras que la gran mayoría se la usa en muy raras ocasiones. Este uso de palabras, sigue una distribución que puede ser representado por:

$$f(r) = \propto \frac{1}{r^{\propto}} \; ; \quad \propto \approx 1$$

En donde $f(r)$ representa la frecuencia que más es usada y $\propto$ es un número que se aproxima a uno. Lo que quiere decir que, la frecuencia de cada palabra, es inversamente proporcional a su rango de ubicación según su frecuencia, así que, la palabra más frecuente su utilizará dos veces más que la segunda palabra más frecuente y tres veces más que la palabra más utilizada, y así sucesivamente (Zipf, 1950).

Se puede observar en la Figura 20, la distribución según el rango de frecuencias de las palabras de los tweets de bullying y no bullying. La línea roja entre cortada, representa la distribución Zipfziana, y como se puede observar, la utilización de las palabras en el corpus, se asemejan a dicha distribución.

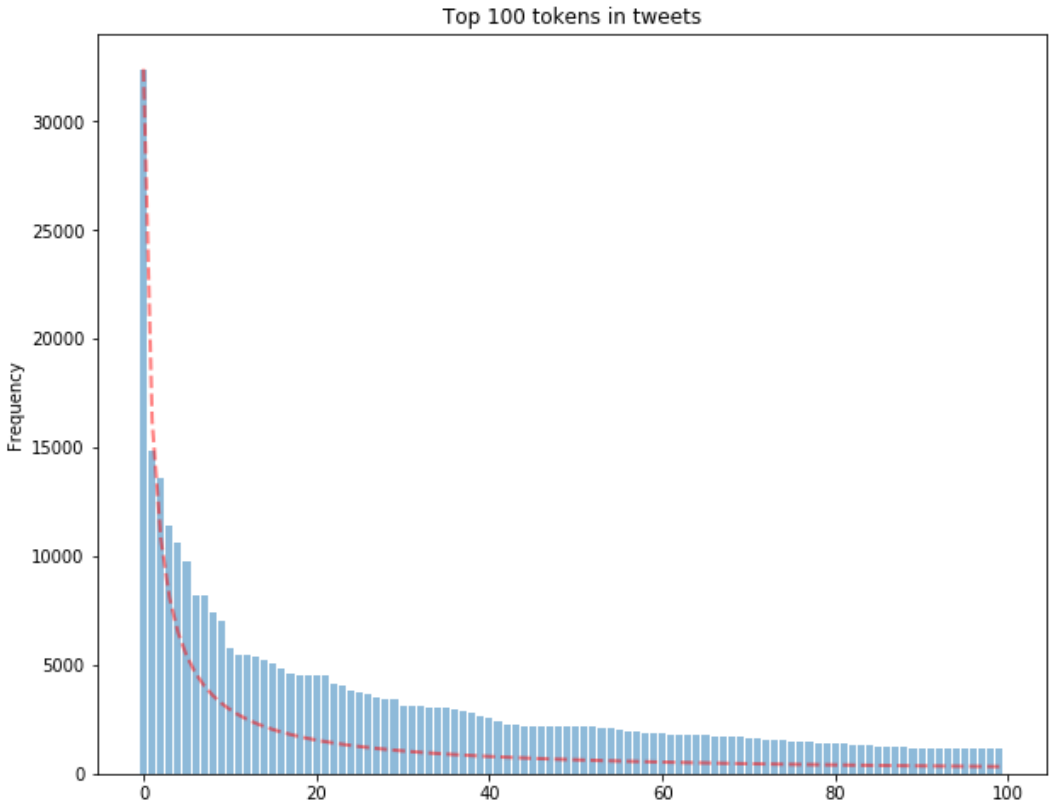

Figura 20: Top 100 de la distribución de frecuencias de las palabras de los tweets. Fuente: Desarrollo del investigador

Otra representación para verificar que la distribución del corpus, se cumple con lo manifestado por la ley de Zipf (1950) se muestra en la Figura 21. Como se puede apreciar, en el eje “X”, se tiene el rango de frecuencias de las palabras utilizadas en el corpus, y en el eje “Y”, su frecuencia absoluta. Al graficar los puntos en el plano observamos una curva casi lineal, aproximada a la línea roja, que representa a la distribución de Zipf (1950). Existe una desviación de la curva de la distribución del corpus, que sobrepasa la distribución esperada, lo que quiere decir que, en el corpus existen palabras con un rango más alto de frecuencia en su uso, esto puede significar que, en los tweets del corpus, se usan palabras frecuentes

más veces que expresiones normales, y esto, puede deberse a que se utilizaron palabras clave para la obtención del corpus.

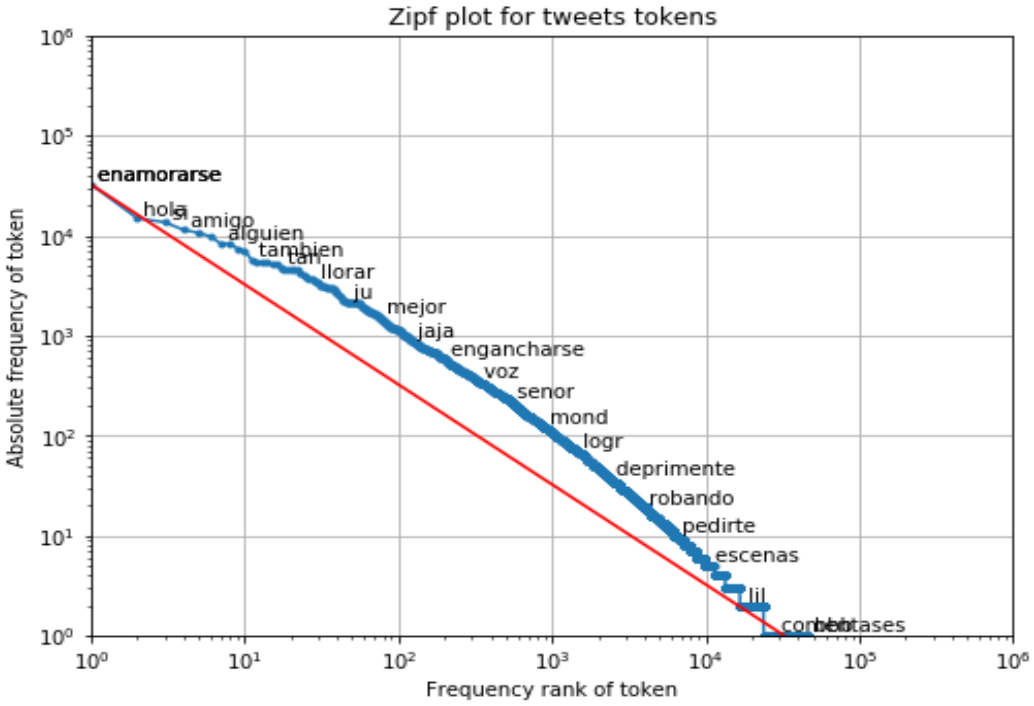

Figura 21: Aproximación de la distribución de las palabras del corpus vs una distribución Zipf. Fuente: Desarrollo del investigador

El la Figura 22 y Figura 23 se puede observar las distribuciones de acuerdo a los rangos de frecuencias usadas en el corpus para los tweets etiquetados como bullying y no bullying respectivamente. Se puede verificar, también que estas se asemejan mucho a una distribución zipfziana.

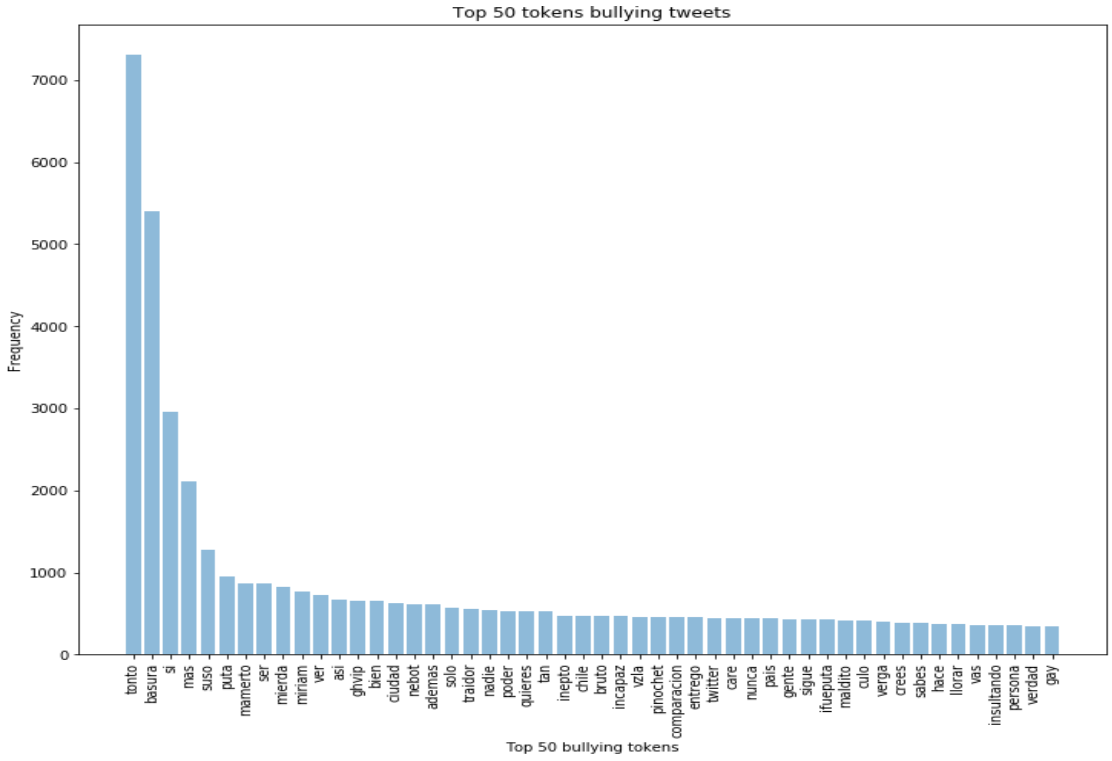

Figura 22: Top 50 de la distribución de frecuencias de las palabras de los tweets bullying. Fuente: Desarrollo del investigador

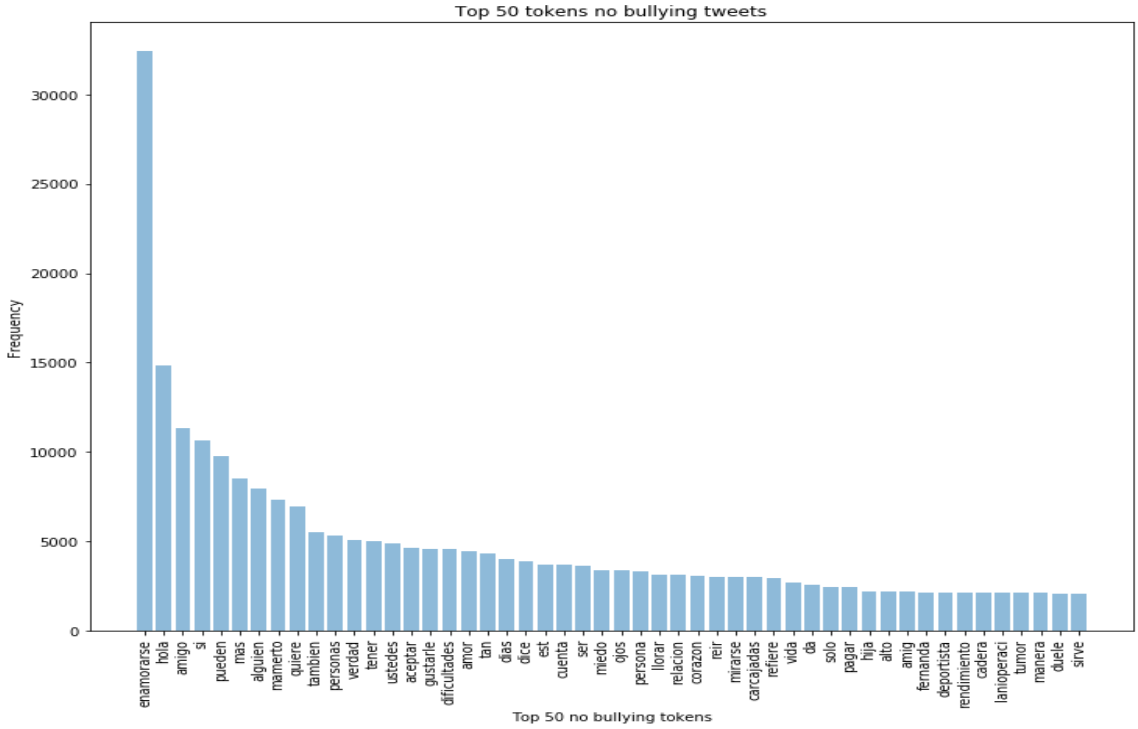

Figura 23: Top 50 de la distribución de frecuencias de las palabras de los tweets no bullying. Fuente: Desarrollo del investigador

Según el análisis realizado al corpus, se puede evidenciar que el mismo, cumple con las características de las expresiones lingüísticas que utilizan las personas para comunicarse, es decir, se utilizan en proporción las palabras necesarias para establecer comunicación. Esto de acuerdo a lo manifestado por Zipf (1950). Tal como se puede observar en la Figura *22* y Figura *23*, se observa que las palabras utilizadas en el *"corpus-bullying"*, se asemejan mucho a una distribución zipfziana. En base a esto, se puede concluir que el corpus estructurado, puede ser utilizado por la red neuronal convolucional para la extracción de características en base a la relación lingüística que se usa en las expresiones de *bullying* y *no bullying*.

3.3.6. Entrenamiento del modelo

Para el entrenamiento y aprendizaje profundo, se utilizó el modelo Red Neuronal Convolucional (*CNN*), ya que, como se mencionó anteriormente, este tipo de red neuronal permite el procesamiento secuencial de los tweets, por las diferentes capas que se estructuran dentro de las red neuronal, y en conjunto con técnicas de aprendizaje profundo, se ha efectuado con éxito el procesamiento del lenguaje natural y el análisis de los textos para la predicción de bullying y no bullying (Kim, 2014). Esta red neuronal fue implementada, utilizando el módulo de Python que permite la extracción y el aprendizaje profundo de Python denominado Keras[27].

Para el aprendizaje de las características, se utilizó la incrustación de palabras en español, representadas vectorialmente a través de un modelo lingüístico. Este cuerpo fue consolidado en el archivo "SBW-vectors-300-min5.txt", desarrollado por Cardellino (2016),

[27] Keras: Es una aplicación escrita en Python, que permite la implementación de redes neuronales de alto nivel para el aprendizaje profundo. https://keras.io/

a partir de diferentes recursos publicados en la web, y está conformado por 1.4 mil millones de palabras en español.

Las palabras recopiladas en el corpus de Cardellino (2016), fueron procesadas, estableciendo analogías entre las palabras y relaciones de proximidad, utilizando el algoritmo *word2vec*, desarrollado por un grupo de investigadores de *Google*, (Mikolov, Chen, Corrado y Dean, 2013). Esto con el fin de obtener un matriz con incrustación de 1000563 palabras con una dimensión de 300 vectores. Estas palabras son representadas vectorialmente en el espacio dimensional, de forma que las palabras que son similares o se encuentran relacionadas, son representadas por puntos que se juntan los unos con otros, (Mikolov et al., 2013). En la Figura 24*:* se puede observar la representación vectorial de las palabras que son similares o están relacionadas se encuentran más cerca, las unas de las otras.

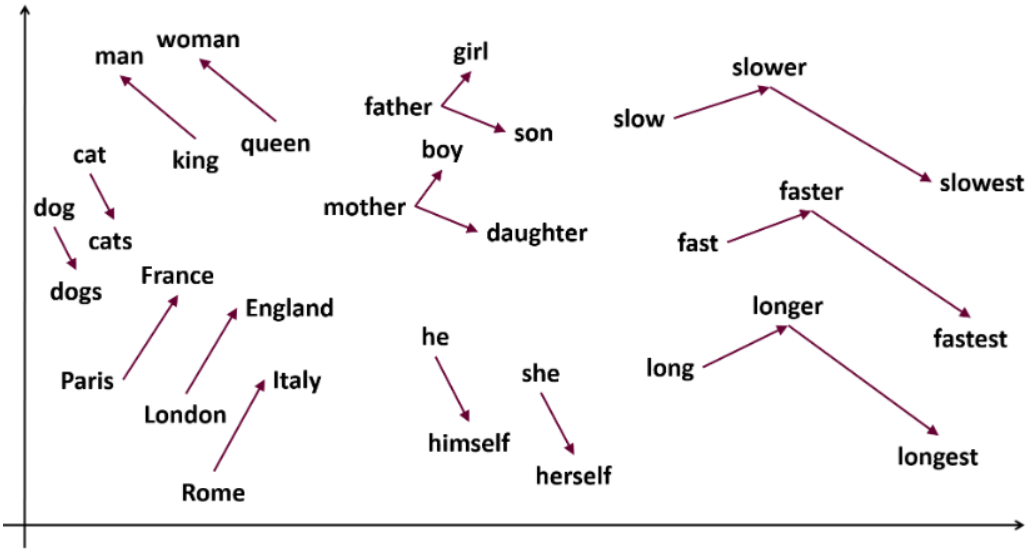

Figura 24: Representación de vectores word2vec, de la relación de similitud y palabras relacionadas en base a su cercanía. Fuente: https://samyzaf.com/ML/nlp/nlp.html

El corpus de Cardellino (2016), se utiliza en el proceso de entrenamiento de la red neuronal para realizar incrustaciones de palabra pre-entrenadas. La red neuronal utiliza una capa de incrustación para el procesamiento de lenguaje natural y se constituye en la primera capa oculta de la red, esto le permite a la *CNN*, extraer las características del texto de los tweets y aprender la relación semántica de las palabras que se encuentran en la matriz pre-entrenada.

En la Figura 25, se establecen los procesos que interactúan para realizar el entrenamiento de la *CNN*, como se puede observar, se tienen, como parte del proceso macro de entrenamiento de la red neuronal a cinco procesos que se detallan a continuación.

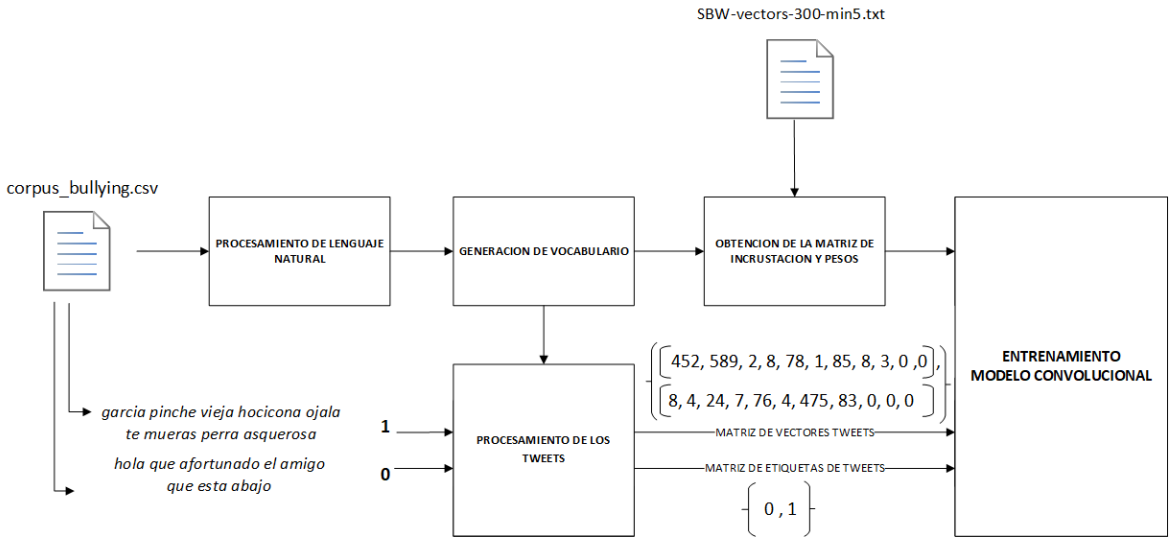

Figura 25: Esquema de entrenamiento del modelo. Fuente: Desarrollo del investigador

Procesamiento del lenguaje natural. - Este proceso permite procesar las palabras que constituyen el corpus, en el cual, se encuentran clasificados y etiquetados los diferentes tweets de *bullying* y *no bullying*. Para esto, primero se obtienen todas las palabras que se encuentran en el corpus-bullying.csv, esto con el objetivo de obtener la distribución de

frecuencias de ocurrencia en el corpus. Por ejemplo, si tomamos en consideración los textos de los tweets de la Tabla 12, al realizar el procesamiento de las palabras se tendría:

```
['minis', 'comprendi', 'volver', 'enamorarse', 'necesario', 'vaciar',
'corazon', 'recuerdos', 'viejos', 'tonto', 'miedo', 'enamorarse',
'alguien', 'mayor', 'amor', 'conoci', 'amor', 'vida', 'solo', 'persona',
'mas', 'cuenta', 'alguien', 'dice', 'quiere', 'enamorarse', 'tener',
'relacion', 'refiere', 'quiere', 'rompa', 'si', 'ven', 'van', 'confiar',
'kelly', 'cara', 'de', 'perra', 'da', 'chingo', 'miedo', 'alguien',
'mas', 'vea', 'vi', 'brillas', 'manera', 'tan', 'hermosa', 'imponente',
'dicen', 'malas', 'lenguas', 'juego', 'juegas', 'gente', 'real', 'v',
'saw', 'miedo', 'enamorarse', 'alguien', 'mayor', 'amor', 'conoci',
'amor', 'vida', 'solo', 'persona', 'mas', 'ven', 'aqui', 'hijo', 'mas',
'tonto', 'afilar', 'mazas', 'ven', 'historiador', 'instruido', 'as',
'perra', 'gorda']
```

Posteriormente, se realiza la obtención de la distribución de las frecuencias para lo cual se utiliza la librería NLTK[28] (*Natural Language Tool Kit*), mediante la función *FreqDist*, que permite la obtención del número de repeticiones de las palabras en un texto. Si consideramos a todas las palabras del ejemplo anterior, se tendría la siguiente distribución de frecuencias.

```
FreqDist({ 'enamorarse': 4, 'alguien': 4, 'amor': 4, 'mas': 4, 'miedo':
3, 'ven': 3, 'conoci': 2, 'mayor': 2, 'perra': 2, 'persona': 2, 'quiere':
2, 'solo': 2, 'tonto': 2, 'vida': 2, 'afilar': 1, 'aqui': 1, 'as': 1,
'brillas': 1, 'cara': 1, 'chingo': 1, 'comprendi': 1, 'confiar': 1,
'corazon': 1, 'cuenta': 1, 'da': 1, 'de': 1, 'dice': 1, 'dicen': 1,
'gente': 1, 'gorda': 1, 'hermosa': 1, 'hijo': 1, 'historiador': 1,
'imponente': 1, 'instruido': 1, 'juegas': 1, 'juego': 1, 'kelly': 1,
'lenguas': 1, 'malas': 1, 'manera': 1, 'mazas': 1, 'necesario': 1,
'real': 1, 'recuerdos': 1, 'refiere': 1, 'relacion': 1, 'rompa': 1,
'saw': 1, 'si': 1, 't': 1, 'tan': 1, 'tener': 1, 'v': 1, 'vaciar': 1,
'van': 1, 'vea': 1, 'vi': 1, 'viejos': 1, 'volver': 1,'minis': 1,})
```

De esta manera se realiza el procesamiento de todas las palabras del "*corpus-bullying.csv*". Se puede observar en la Figura 26, un gráfico de la distribución de frecuencias acumuladas de todo el corpus, en donde se inicia con la palabra más frecuente que es

[28] https://www.nltk.org/

enamorarse con 35000 veces aproximadamente y finaliza con la palabra menos frecuente que es aceptar, pero por cada palabra que se encuentra en el corpus ese va acumulando el valor de la frecuencia, estableciéndose una curva parecida a una función logarítmica, por lo que el corpus en total tiene 160 mil palabras aproximadamente.

El resultado de la distribución de frecuencias es un conjunto de palabras con su respectiva frecuencia de ocurrencia en los textos de los tweets, esta estructura es almacenada en un objeto *pickle*[29], el mismo que será utilizado para la generación del vocabulario.

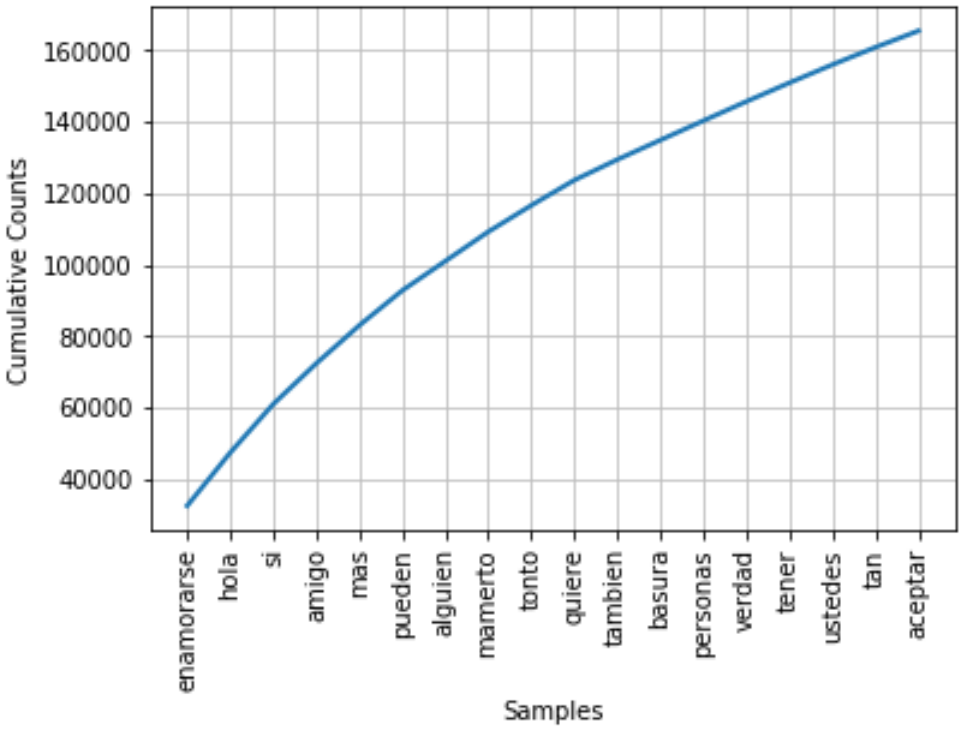

Figura 26: Distribución de frecuencias acumuladas de las palabras de corpus. Fuente: Desarrollo del investigador

29 Pickle: Socialización de objetos Python es el proceso mediante el cual una jerarquía de objetos de Python se convierte en un flujo de bytes. https://docs.python.org/3/library/pickle.html

Generación del vocabulario. - Este proceso genera un vocabulario, en base a la distribución de frecuencias de las palabras del corpus. Para esto, se toman las palabras más comunes en la distribución y se genera una matriz indexada la cual será utilizada para enviar los datos al modelo *CNN*. Esta matriz inicia con la palabra más frecuente asignando el indice uno (1), a la segunda palabra más frecuente, se le asigna el indice dos (2), y así sucesivamente, tal como se representa en la Tabla 13.

Si consideramos como ejemplo la distribución de frecuencia obtenida a partir de los tweets de ejemplos descritos en el proceso anterior tendríamos el siguiente vocabulario:

Indice	Palabra	Indice	Palabra	Indice	Palabra	Indice	Palabra
1	'enamorarse'	16	'comprendi'	31	'confiar'	46	'lenguas'
2	'alguien'	17	'volver'	32	'kelly'	47	'juego'
3	'amor'	18	'necesario'	33	'cara'	48	'juegas'
4	'mas'	19	'vaciar'	34	'de'	49	'gente'
5	'miedo'	20	'corazon'	35	'da'	50	'real'
6	'ven'	21	'recuerdos'	36	'chingo'	51	'v'
7	'tonto'	22	'viejos'	37	'vea'	52	'saw'
8	'mayor'	23	'cuenta'	38	'vi'	53	'aqui'
9	'conoci'	24	'dice'	39	'brillas'	54	'hijo'
10	'vida'	25	'tener'	40	'manera'	55	'afilar'
11	'solo'	26	'relacion'	41	'tan'	56	'mazas'
12	'persona'	27	'refiere'	42	'hermosa'	57	'historiador'
13	'quiere'	28	'rompa'	43	'imponente'	58	'instruido'
14	'perra'	29	'si'	44	'dicen'	59	'as'
15	'minis'	30	'van'	45	'malas'	60	'gorda'

Tabla 13: Vocabulario ejemplo generado a partir de la distribución de frecuencias. Fuente: Desarrollo del investigador

Obtención de la matriz de incrustación. – Para esto, se utilizan valores aleatorios de una distribución normal, los cuales son almacenados en una matriz cuya dimensión está definida

en base a las mismas dimensiones de la matriz pre-entrenada con palabras en español "*SBW-vectors-300-min5.txt*" (Cardellino, 2016). Posteriormente, se va obteniendo los valores de los pesos de cada una de las palabras del vocabulario, desde la matriz pre-entrenada, en caso de que esta palabra no se encuentre, se mantendrá los valores obtenidos aleatoriamente. Finalmente, se obtendrá como resultado, una matriz de pesos, únicamente para las palabras que se encuentren en el vocabulario. En la Figura 27, se representa mediante un diagrama de flujo el proceso para la obtención de la matriz de incrustación utilizada en nuestra *CNN*.

Una matriz de incrustación, como se ha mencionado anteriormente, se constituye de las palabras incrustadas en espacios vectoriales, en donde las palabras más cercanas poseen una relación sintáctica entre sí, se representan por puntos numéricos que se encuentran juntos en el espacio. Esta matriz, constituye un elemento básico en el aprendizaje profundo de la *CNN*, ya que viene a ser la primera capa oculta para aprender la representación semántica de los vectores utilizada para extraer las características de las palabras que están relacionadas en los tweets, características que definen como se conforma las oraciones en el lenguaje natural. A continuación, se presenta un ejemplo de la matriz de incrustación utilizado en la red neuronal convolucional utilizadas en este trabajo de investigación.

```
embedding_matrix
Out[70]:
array([[-7.03187310e-03, -4.90282363e-03, -3.21814330e-03, ...,
         6.51163018e-05, -5.37901188e-03,  4.93670710e-03],
       [-2.28490000e-02, -5.46710000e-02,  2.73330000e-02, ...,
        -2.48050000e-02, -1.90700000e-03, -2.00910000e-02],
       [-9.19989863e-03,  1.58897332e-03,  1.11722335e-02, ...,
        -9.31219401e-03,  1.87904597e-03,  4.73601927e-03],
       [-1.86262950e-03,  1.05277042e-02,  1.46418454e-03, ...,
         2.53428914e-03,  7.34533696e-03,  9.38234619e-03],
       [-4.96159825e-03, -4.08589735e-03, -2.56768154e-03]])
```

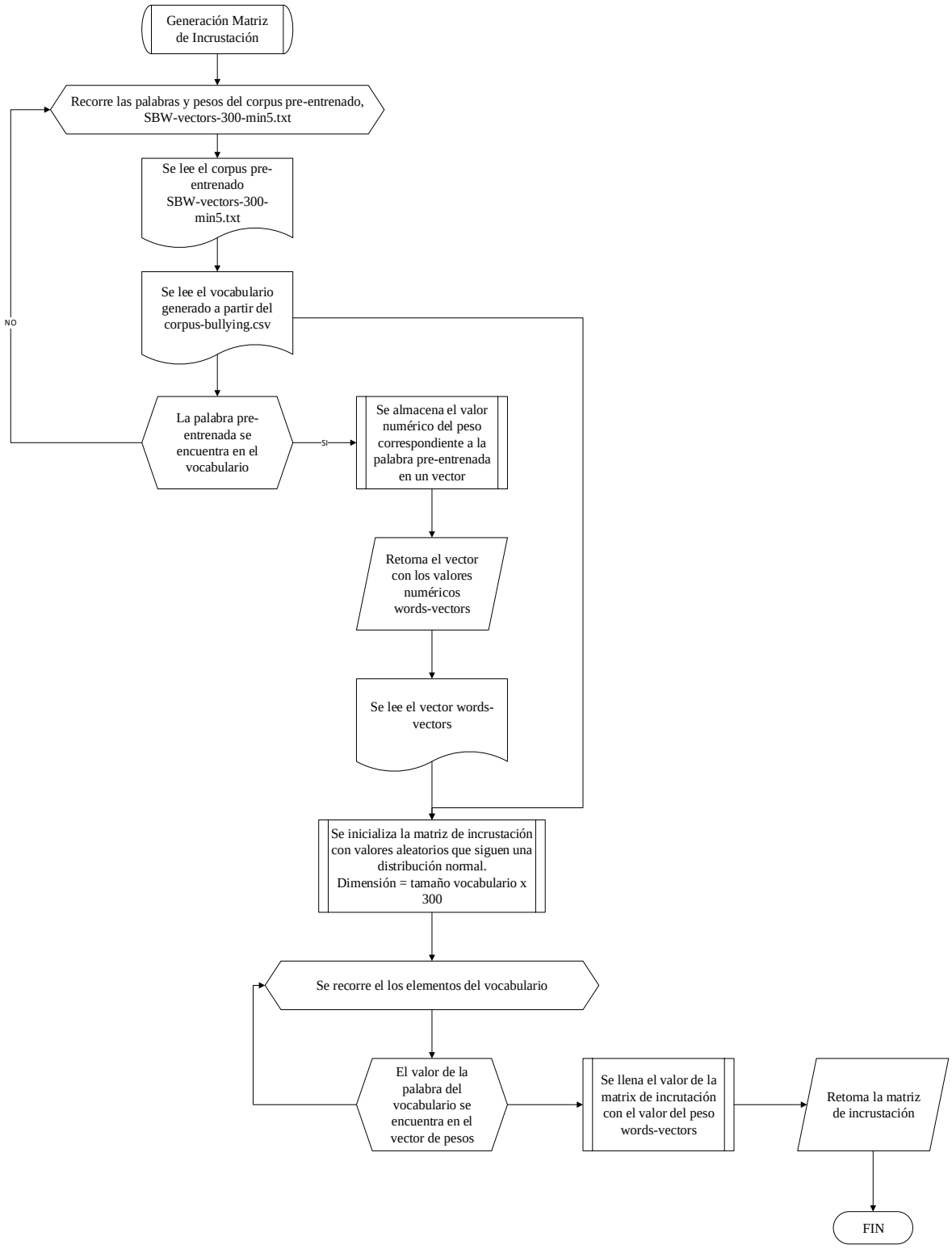

Figura 27: Diagrama de Flujo para la obtención de la matriz de incrustación. Fuente: Desarrollo del investigador.

Procesamiento de los *tweets*. – Este proceso descompone el corpus del conjunto de datos de entrenamiento en tres vectores.

- *tweets* [] – Se almacena los valores de los índices, que representan a las palabras del corpus en base a la frecuencia en el corpus.
- *labels* [] – Se almacenan los valores de las etiquetas, si es *bullying* uno (1) y si no es *bullying* cero (0).
- *Ids* [] – Se almacenarán los *ids* que identifican a cada *tweet*.

Los conjuntos de datos almacenados en las matrices de vectores *tweets* [] y *labels* [] son los que se ingresa al proceso de entrenamiento de la red *CNN* en *Keras*. La matriz que representa a los *tweets*, se la obtiene buscando las palabras que conforman el texto en el vocabulario y se ubica el índice numérico de la palabra que le corresponde. Por ejemplo, al procesar los *tweets* con Id. 1, 2, 3, descritos en la , y luego de buscar las palabras en el vocabulario de ejemplo descrito en la Tabla 13, se obtiene como resultado la siguiente representación de los *tweets*:

Representación numérica	15	16	17	1	18	19	20	21	22	0	0	0
Tweet 1	minis	comprendi	volver	enamorarse	necesario	vaciar	corazon	recuerdos	viejos			

Tabla 14: Representación numérica del tweet ejemplo 1. Fuente: Desarrollo del investigador.

Representación numérica	7	0	0	0	0	0	0	0	0	0	0	0
Tweet 2	tonto											

Tabla 15: Representación numérica del tweet ejemplo 2. Fuente: Desarrollo del investigador.

Representación numérica	5	1	2	8	3	9	3	10	11	12	4	0
Tweet 3	miedo	enamorarse	alguien	mayor	amor	conoci	amor	vida	solo	persona	mas	

Tabla 16: Representación numérica del tweet ejemplo 3. Fuente: Desarrollo del investigador.

Los números que se encuentran en la representación numérica, corresponden al índice de la ubicación de la palabra en el vocabulario, es decir que, si buscamos la palabra "minis" en el vocabulario de ejemplo en la Tabla 13, se tiene el índice 15, si buscamos la palabra "comprendi", se obtiene el índice 16, y así, para todas las palabras de cada uno de los tweets del corpus.

Los tweets se encuentran representados por índices numéricos, en los vectores correspondientes, estos deben tener las mismas dimensiones para todos los vectores que conforman la matriz que ingresará como entrada a la *CNN*, por lo que a todos los vectores se los rellena con valores de cero en sus elementos, hasta completar la longitud máxima de palabras que tiene el tweet más extenso en el corpus, y de acuerdo a lo validado, el *tweet* más extenso tiene un total de 60 palabras.

Entrenamiento del modelo. – Este proceso, definido anteriormente en el esquema de entrenamiento del modelo, permite entrenar la *CNN* utilizando las capas de la red neuronal implementada en Keras para el aprendizaje profundo. Una vez, se han procesado todos los tweets por las diferentes capas explicadas anteriormente, con sus respectivas etiquetas de *bullying* y *no bullying*, se procede con la creación del modelo entrenado. El modelo obtenido de este proceso posee una precisión del 98.5%, como se puede observar en la Figura *28*.

```
Epoch 00007: loss improved from 0.00664 to 0.00594, saving model to C:\Users\Paul David
\OneDrive\DESARROLLO FINAL TESIS\RED NEURONAL CONVOLUCIONAL/Models/
model-07-0.006-0.998-0.076-0.986.hdf5
Epoch 8/8
56871/56871 [==============================] - 5860s 103ms/step - loss: 0.0048 - acc: 0.9980 -
val_loss: 0.0758 - val_acc: 0.9850

Epoch 00008: loss improved from 0.00594 to 0.00478, saving model to C:\Users\Paul David
\OneDrive\DESARROLLO FINAL TESIS\RED NEURONAL CONVOLUCIONAL/Models/
model-08-0.005-0.998-0.076-0.985.hdf5
```

Figura 28: Proceso de entrenamiento de la CNN en Keras. Fuente: Desarrollo del investigador.

Al visualizar el resumen del modelo entrenado en la Figura 29, se puede apreciar los resultados de las matrices de salida que se tiene de pasar de una capa hacia otra. La red neuronal convolucional que se ha utilizado, está compuesta por doce capas. Se inicia con la capa de incrustación con una matriz de 60 x 300 y se finaliza con una matriz con una salida, con un total de parámetros entrenados de 30,591,376, que corresponden al total de elementos entrenados en cada capa de la red neuronal.

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, 60, 300)	27000300
dropout_3 (Dropout)	(None, 60, 300)	0
conv1d_5 (Conv1D)	(None, 58, 600)	540600
conv1d_6 (Conv1D)	(None, 56, 300)	540300
conv1d_7 (Conv1D)	(None, 54, 150)	135150
conv1d_8 (Conv1D)	(None, 52, 75)	33825
flatten_2 (Flatten)	(None, 3900)	0
dense_3 (Dense)	(None, 600)	2340600
dropout_4 (Dropout)	(None, 600)	0
activation_3 (Activation)	(None, 600)	0
dense_4 (Dense)	(None, 1)	601
activation_4 (Activation)	(None, 1)	0

Total params: 30,591,376
Trainable params: 30,591,376
Non-trainable params: 0

Figura 29: Resumen del modelo CNN entrenado. Fuente: Desarrollo del investigador.

3.3.7. Ejemplo de predicción.

Una vez se ha creado el modelo, este puede ser utilizado para realizar predicciones de nuevos textos que no formen parte del corpus, es decir, puede ser utilizado para predecir nuevos *tweets* para determinar si estos tienen signos de *bullying* o *no bullying*.

Para los ejemplos de predicción consideraremos los siguientes textos, los cuales serán almacenados bajo la siguiente estructura:

Id	Texto
1	Eres tan homosexual que ojala te mueras
2	Amigo es un gusto poder tener contacto contigo

Tabla 17: Textos para ejecutar los ejemplos de predicción. Fuente: Desarrollo del investigador.

El texto que con ID igual a 1, por lo que podemos observar, tiene signos de *bullying*, en cambio el otro texto no refleja signos de *bullying*.

Al cargar el modelo y ejecutar la predicción se tienen los siguientes resultados:

Id	Texto	Predicción	Probabilidad
1	Eres tan homosexual que ojala te mueras	1	99.999%
2	Amigo es un gusto poder tener contacto contigo	0	95.992%

Tabla 18: Resultado de la predicción de los ejemplos. Fuente: Desarrollo del investigador.

Como se puede apreciar en la Tabla 18, el modelo ha acertado de manera correcta la predicción de los textos, etiquetando al primer texto como *bullying* con uno (1) y al segundo como *no bullying* con cero (0), la probabilidad de certeza que tiene la predicción es del 99.999% y 95.992% respectivamente para cada uno de los *tweets* de ejemplo.

CAPÍTULO IV

VALIDACIÓN EMPÍRICA

4.1. Introducción

En el capítulo anterior, se explicó en detalle el modelo de solución adoptado para el entrenamiento de la Red Neuronal Convolucional (*CNN*), lo que ~~nos~~ permitió realizar predicciones de *bullying*, a partir de mensajes cortos de textos obtenidos de *Twitter*. Como en todo trabajo de investigación, el modelo propuesto, debe estar sujeto a casos experimentales, y pruebas de validación, ya que, es irrealista dar como aceptado un modelo, sin que antes este haya sido validado (Jesús San Isidro, 1998).

Es por esto que, este capítulo recopila el proceso de validación realizado para dar como aceptado el modelo generado por la CNN para la predicción de *bullying*. Además, se podrá verificar que el "*corpus-bullying"* conformado en este trabajo de investigación, es consistente y valido para el entrenamiento de la red neuronal.

Para la validación empírica del modelo, se ha seleccionado la técnica de validación cruzada (*cross validation*), la cual es muy popular para la selección de algoritmos de aprendizaje automático y aprendizaje profundo, según lo expone Arlot y Celisse (2010). Esta técnica, consiste en dividir el conjunto de datos del corpus-bullying, varias veces de manera aleatoria, con el objetivo de medir la confiabilidad de los modelos obtenidos con el conjunto de datos de entrenamiento. Posterior a esto, se realizarán pruebas de predicción con los datos de prueba, para verificar la estabilidad y confianza del modelo generado.

4.2. Resultados

4.2.1. División de datos para Entrenamiento y Pruebas

Para realizar la validación fue necesario dividir el conjunto total de datos del corpus-bullying, en dos subconjuntos, uno de entrenamiento y otro de pruebas. Este proceso se lo realizó en cuatro iteraciones, es decir se dividió el "*corpus-bullying*" cuatro veces de manera aleatoria. La proporción para realizar la división de datos es 90% de datos para entrenamiento (75059 *tweets*) y 10% de datos para pruebas (8341 *tweets*), del total de tweets del corpus-bullying, tal como se muestra en la Tabla 19.

	% División	Tweets
Train Data	90%	75059
Test Data	10%	8341
Total Corpus	**100%**	**83400**

Tabla 19: División de datos para entrenamiento y pruebas. Fuente: Desarrollo del investigador

Para realizar la división de los datos, se utilizó la librería de Python *sklearn.cross_validation*[30], según se muestra en el Algoritmo 1.

Algoritmo 1: *Proceso división de datos para entrenamiento y pruebas*

Asignar variables
tamaño_set_pruebas = 0.01
semilla_aleatoria = 100

function divide_data_train_test *(archivo_corpus,archivo_test, archivo_train, tamaño_set_pruebas, semilla_aleatoria)*
columnas = ['id','tweet_text','bully']
datos_corpus = lee archivo_corpus y transforma en dataframe pandas (columnas)
twets_bully = datos_corpus cuando bully es igual a 1
twets_no_bully = datos_corpus cuando bully es igual a 0
datos_entrenamieto_bully = dividir (tweets_bully con el 90%, semilla_aleatoria)
datos_pruebas_bully = dividir (tweets_bully con el 10%, semilla_aleatoria)
datos_entrenamieto_no_bully = dividir (tweets_no_bully con el 90%, semilla_aleatoria)
datos_pruebas_no_bully = dividir (tweets_no_bully con el 10%, semilla_aleatoria)

[30] http://scikit-learn.org/0.16/modules/generated/sklearn.cross_validation.train_test_split.html

```
    datos_entrenamiento = unir (datos_entrenamiento_bully,
datos_entrenamiento_no_bully)
    datos_pruebas = unir (datos_pruebas_bully, datos_pruebas_no_bully)

    convertir a csv (datos_entrenamiento, archivo_train)
    convertir a csv (datos_pruebas, archivo_test)
end function

i = 1
niteraciones = 4
for i  to niteraciones do
    divide_data_train_test (archivo_corpus,archivo_test, archivo_train,
tamaño_set_pruebas, semilla_aleatoria)
end for
```

Como resultado de la ejecución de las iteraciones de división de datos, se obtiene ocho conjuntos de datos, dos conjuntos, uno de entrenamiento y el otro de pruebas, esto por cada iteración de división, los cuales fueron guardados en archivos delimitados por comas (csv), tal como se aprecia en la Figura *30*

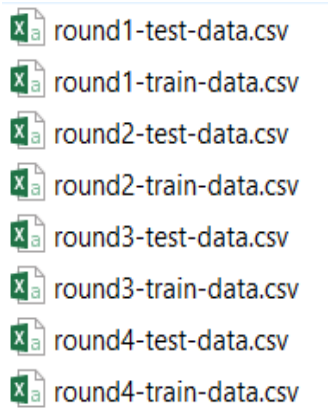

Figura 30: Conjuntos de datos de pruebas y entrenamiento por cada iteración de división. Fuente: Desarrollo del investigador.

4.2.2. Entrenamiento de la red neuronal convolucional para la validación del modelo

Una vez obtenidos los diferentes conjuntos de datos, se realizó el entrenamiento de la CNN, en base a la arquitectura definida en el punto 3.2.2, del Capítulo III.

La CNN fue entrenada cuatro veces, conforme al número de archivos creados para el "*training*" (ver Figura 30), y por cada entrenamiento se utilizó ocho *epochs*.

Un *epoch* de acuerdo a la documentación de Keras (2018), corresponde a la transmisión de datos completando un ciclo completo hacia adelante y hacia atrás por toda la red neuronal (*feedforward* y *feedback*). Esto se realiza para que la red neuronal, que en este caso tendría una topología multicapa y con un modo de entrenamiento secuencial, realice la actualización de los pesos sinápticos por todas las capas de esta (Yao, 1993).

Por cada iteración de entrenamiento y a la finalización de cada *epoch*, se estableció un punto de control o *check point* para que se almacene un modelo como resultado de ese *epoch*. Es decir que, por cada iteración de entrenamiento se ejecutaron ocho *epochs* y en cada punto de control se almacenaba un modelo por lo que se generó un total de 32 modelos con sus respectivas métricas de precisión y pérdida. Se debe considerar que, del proceso de división de datos se obtuvieron cuatro conjuntos diferentes de entrenamiento, con datos aleatorios, tal como se explica en el punto 4.2.1, de esta manera es que cada iteración de entrenamiento utiliza su archivo correspondiente de entrenamiento, por esta razón es que la precisión en cada iteración de entrenamiento varia.

Según Bertona (2005), el número de *epochs* a seleccionar para realizar el entrenamiento de una red neuronal, no está dado por una regla específica, sino que depende de la complejidad de los datos y de los valores de precisión y perdida que se va obteniendo en cada ciclo de entrenamiento.

Cada iteración de entrenamiento está representada por la Figura 31, Figura 32, Figura 33 y Figura 34. Como se observa, en el *eje x* se encuentra representado por los *epoch* que tuvo cada iteración de entrenamiento, es decir cada punto en este eje es un modelo generado

una vez finalizado el *check point*, en el *eje y*, se encuentran representadas las medidas de pérdida (barra de la figura) y precisión (línea de la figura) por cada *epoch*, y así se representan para todas las iteraciones de entrenamiento. Como se puede observar en las figuras mencionadas, en todas las iteraciones de entrenamiento, los valores de precisión y pérdida tienen una estabilidad a partir del *epoch* 5 aproximadamente, por lo que para los diferentes conjuntos de datos de entrenamiento los modelos alcanzan su mejor precisión y menor pérdida en el *epoch* 8.

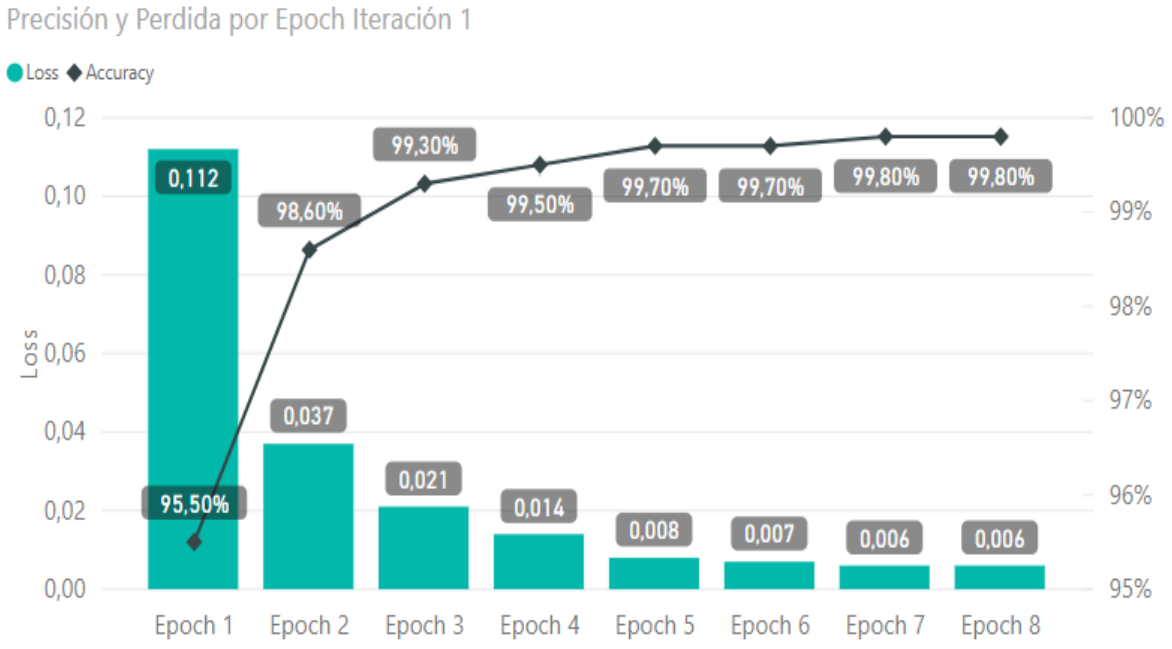

Figura 31: Precisión y Perdida por cada *Epoch*, para la iteración 1 de entrenamiento. Fuente: Desarrollo del investigador.

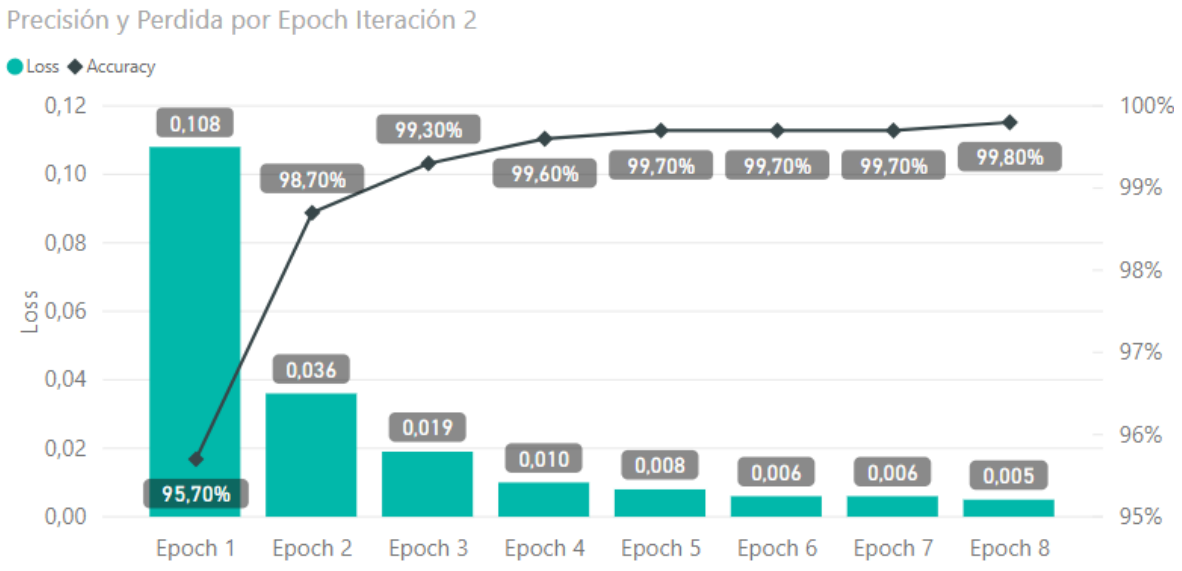

Figura 32: Precisión y Perdida por cada Epoch para la iteración 2 de entrenamiento. Fuente: Desarrollo del investigador.

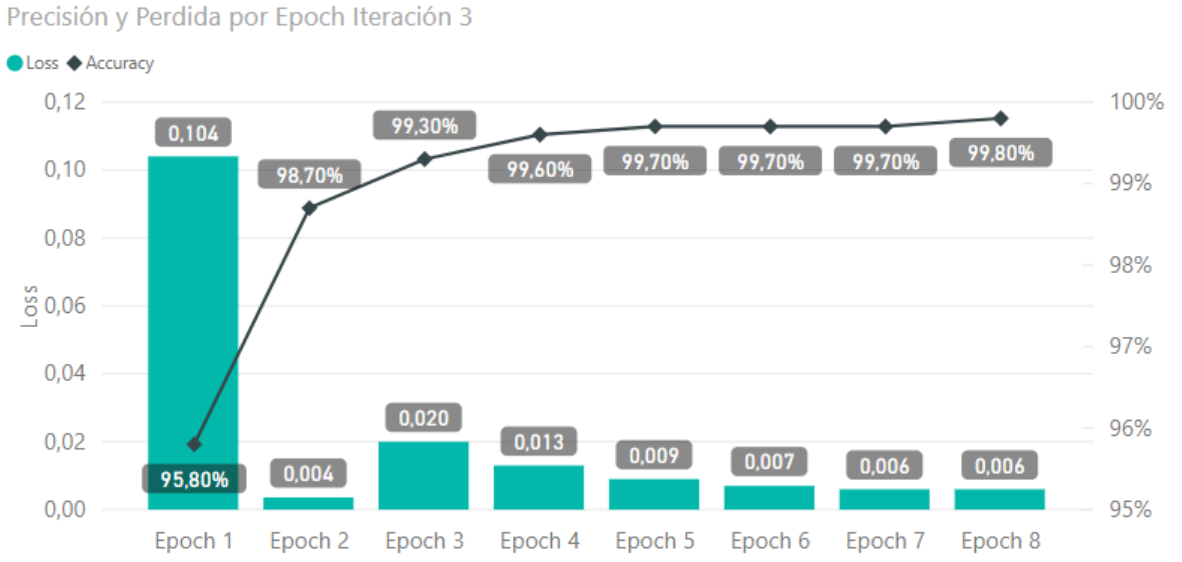

Figura 33: Precisión y Perdida por cada Epoch para la iteración 3 de entrenamiento. Fuente: Desarrollo del investigador.

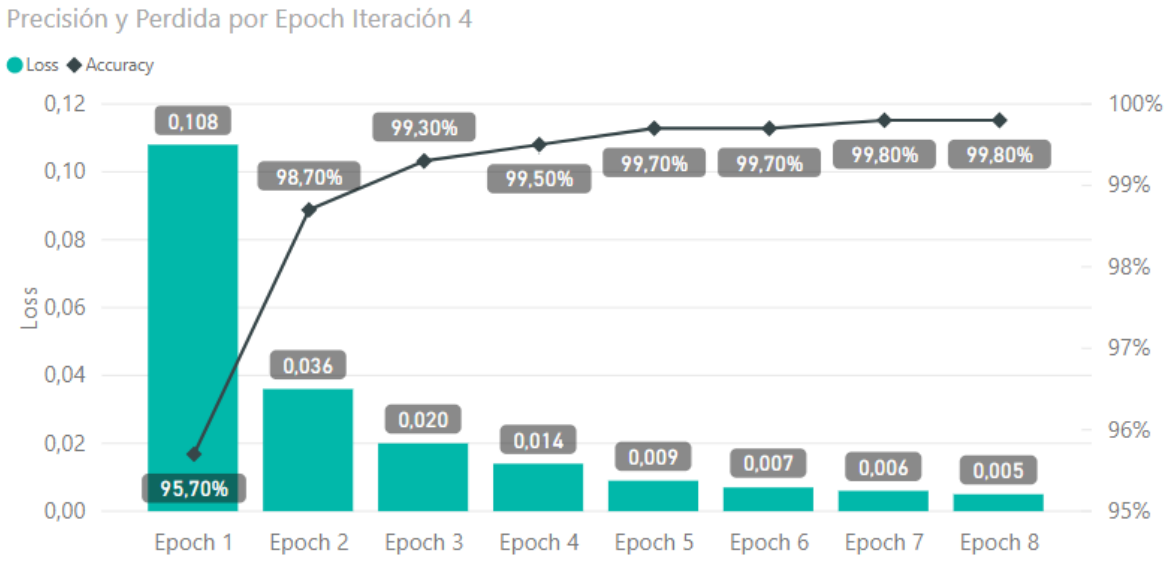

Figura 34: Precisión y Perdida por cada Epoch para la iteración 4 de entrenamiento. Fuente: Desarrollo del investigador.

En la Figura 35, se puede observar el tiempo que se tomó en procesar cada iteración de entrenamiento. Las iteraciones de entrenamiento dos, tres y 4 son bastantes estables y tomaron un tiempo aproximado de diez horas, a diferencia del tiempo de procesamiento en la iteración 1, el cual tuvo una variación del más del 100%.

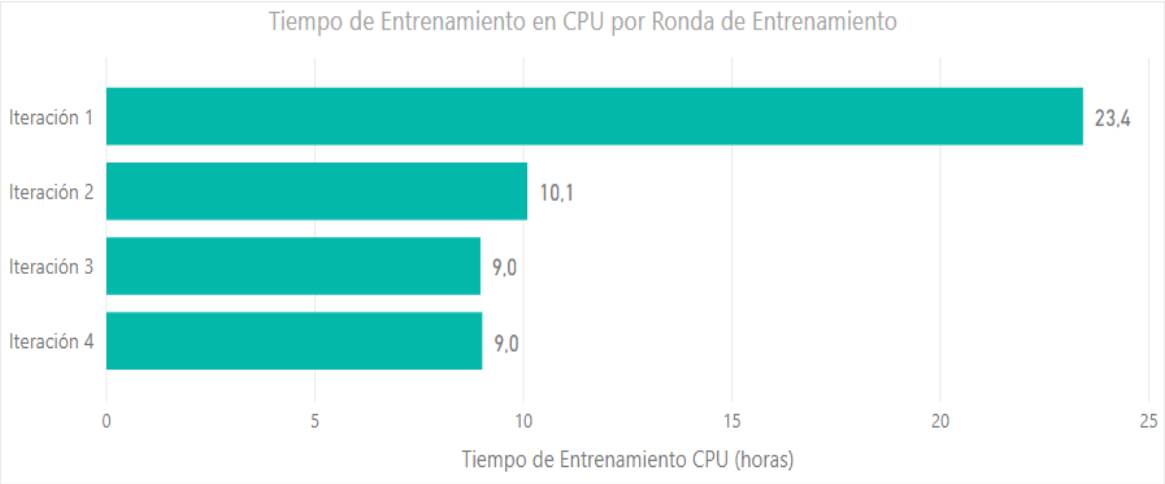

Figura 35: Tiempo de procesamiento por cada iteración de entrenamiento. Fuente: Desarrollo Investigador

4.2.3. Validación del modelo

Para realizar la validación de los modelos, se toma como referencia los valores de precisión y pérdida obtenidos en cada modelo generado por cada *epoch*, de cada iteración de entrenamiento, tal como se explicó anteriormente, cada *epoch* generó un modelo y se seleccionó para realizar la validación, aquel que posee la mayor precisión y la menor pérdida, tal como se presenta en la Tabla 20.

Iteración de Entrenamiento	Modelo Seleccionado	Precisión	Perdida
Iteración 1	Epoch 8	99.80%	0.006
Iteración 2	Epoch 8	99.80%	0.005
Iteración 3	Epoch 8	99.80%	0.006
Iteración 4	Epoch 8	99.80%	0.005

Tabla 20: Modelo seleccionado para la validación por cada iteración de entrenamiento. Fuente: Desarrollo del investigador

Con la selección de los mejores modelos (cuatro en total), especificados en la Tabla 20, se procede a realizar la validación de los datos de pruebas que fueron obtenidos en el punto 4.2.1, para lo cual se utiliza el siguiente algoritmo y se lo ejecuta por cada iteración.

Algoritmo 2: *Proceso de validación de los modelos con los datos de prueba*

Asignar variables
nombre_modelo = "Iteracion 1 - Modelo 8" #Cambiar el nombre del modelo por cada iteración
archivo_pruebas = "Iteracion 1 - pruebas bullying.csv" # Cambiar el nombre del archivo por cada iteración

modelo = llamar a la funcion de carga modelo en keras y enviar (nombre_modelo)

tweets = llamar a la funcion para procesar los tweets, enviar (archivo_pruebas)

prediccion = llamar al metodo predict del modelo, enviar (tweets)

convertir a csv (tweet, predicción, probabilidad)

4.2.4. Predicción de los conjuntos de pruebas

De los datos que fueron divididos para las respectivas pruebas, se realizó la predicción para realizar la validación cruzada del modelo, esto de acuerdo a las respectivas iteraciones de entrenamiento que generó los modelos descritos en la Tabla 20. La predicción fue realizada para cada conjunto de pruebas con el modelo respectivo que se generó en cada iteración de entrenamiento.

En la Tabla 21, se detallan los resultados obtenidos de acierto y fallo de cada predicción, utilizando los datos de pruebas. El porcentaje de fallo global que se tiene es de un promedio de 385 *tweets* equivalente al 1.15%, en cambio el de acierto es de 32,979 *tweets* que equivale a un 98.85%, de un total de 33,364 *tweets* de prueba.

	Fallo		**Acierto**	
Iteraciones	**% Fallo**	**Tweets Errados Predicción**	**% Acierto**	**Tweets Acertados Predicción**
Iteración 1	1.04%	87	98.96%	8,254
Iteración 2	1.16%	97	98.84%	8,244
Iteración 3	1.08%	90	98.92%	8,251
Iteración 4	1.33%	111	98.67%	8,230
	1.15%	**385**	**98.85%**	**32,979**

Tabla 21: Resultados de acierto y fallo en la predicción de Tweets por cada iteración. Fuente: Desarrollo del investigador.

La probabilidad de predicción, que se tiene como resultado por cada iteración de validación, se encuentra detallado en la Tabla 22, en la cual se tienen que en promedio la

probabilidad de predecir un tweet como “No Bullying” es del 99.66% y de que el tweet sea catalogado como “Bullying” es del 98,19%.

Iteraciones	**No Bullying**	**Bullying**
Iteración 1	99.61%	97.91%
Iteración 2	99.69%	99.54%
Iteración 3	99.68%	97.82%
Iteración 4	99.66%	97.46%
	99.66%	**98.19%**

Tabla 22: Resultado de la probabilidad de confianza que se tiene en el modelo por cada iteración de validación. Fuente: Desarrollo del investigador.

4.3. Discusión de resultados

Los resultados obtenidos con anterioridad, ~~nos~~ permiten evidenciar que los modelos obtenidos en las diferentes iteraciones de validación no tienen una alta varianza entre cada una de ellas, además de que en todas se presenta un alto porcentaje de aceptabilidad, teniendo los siguientes resultados: la probabilidad de que un texto sea bullying, es del 98.19% y de que el texto no sea bullying es del 99.66%, además el porcentaje de fallo en la predicción es del 1.15% y de acierto es del 98.85%. Por lo tanto, se concluye que el modelo generado es óptimo para realizar predicciones de textos para la identificación del *cyber bullying*.

Además se observa en la Tabla 22 que la probabilidad de predecir un *tweets* como *no bullying*, es mayor a la probabilidad de que este no lo sea. Esto puede deberse a que existió muchos más datos etiquetados como *no bullying* de la muestra total del *corpus*, lo que implica que el modelo aprendió, más casos lingüísticos expresados, sin signos de bullying

lo cual hace que incremente su probabilidad de predicción. Sin embargo, la diferencia del porcentaje de aciertos entre "*No Bullying*" y "*Bullying*" es muy pequeña.

CAPÍTULO V

CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Conclusiones

Este trabajo de investigación, presenta un aporte para el desarrollo de sistemas que permitan prevenir y alertar casos de *cyber bullying*, esto mediante el aprovechamiento del modelo generado a partir de un corpus con expresiones etiquetadas con *bullying* y *no bullying.*

Los datos extraídos desde la red social en *Twitter* utilizando el módulo *tweepy*, fueron esenciales para los procesos de etiquetado y entrenamiento de la Red Neuronal Convolucional. Se destaca la importancia de extraer datos focalizados utilizando palabras clave, las cuales permitieron obtener las expresiones que representaban situaciones concretas de violencia interpersonal. Además, este proceso permitió la obtención de gran cantidad de información, la misma que constituyó un pilar esencial para la preparación del modelo.

El proceso de clasificación manual de lo tweets, permitió discernir los mensajes realizados por personas comunes que expresan sus sentimientos en las redes sociales, proceso que duró aproximadamente tres meses en los cuales se leyó cada uno de los textos y se verificó si estos mostraban signos de acoso o no. Si realizamos una analogía, se podría decir que este es el proceso en el cual se le indica o se le enseña a la red neuronal convolucional, que expresión corresponde a *bullying* y cual no.

El proceso de entrenamiento y aprendizaje de la red neuronal convolucional, se realizó eficientemente utilizando técnicas de aprendizaje profundo, ya que, las capas de convolución de la red neuronal, con ayuda de la incrustación de una matriz pre-entrenada de

vectores con palabras en español, facilitó la extracción y la comprensión de las características de la relación lingüística de las palabras de las expresiones que poseían signos de *bullying*. En este proceso fue importante la preparación de datos que se realizó al corpus, puesto que, esto garantizó que el procesamiento del modelo sea eficiente en tiempo y calidad de entrenamiento.

Todo corpus que recopile textos debe poseer congruencia en las expresiones realizadas en base a la frecuencia en el uso de sus palabras, según lo manifiesta la Ley de Zipf (1950). Al realizar el análisis del *corpus-bullying.csv*, se puede concluir que las expresiones manifestadas en los tweets que fueron recopilados de diversos usuarios, se asemejan mucho a una distribución zipfziana, los cual garantiza que existen relaciones lingüísticas en las palabras usadas y que servirán para la extracción de características por la red neuronal convolucional, para determinar cuándo una expresión tiene signo de *bullying* o *no bullying.*

El uso de la técnica de validación cruzada permitió verificar que los modelos, generados a partir de un grupo de datos seleccionados al azar, poseen en promedio una precisión estable (98.19%) al momento de predecir los datos de prueba, esto ~~nos~~ permite concluir que el modelo es confiable en la certeza de predicción para nuevos textos a ser evaluados.

Además, el modelo de solución propuesto presenta una arquitectura diseñada por capas, esto permite establecer escalabilidad e independencia en los procesos que se ejecutan en cada una de ellas. También posibilita que la solución sea independiente y escalable, esto da cabida a que se puede usar diversas fuentes de datos, además de Twitter.

5.2.Trabajos Futuros

El estudio realizado en este trabajo de investigación, da pautas para que nazcan nuevas iniciativas en la contribución para el uso de técnicas o el desarrollo de aplicaciones que estén enfocadas a la generación de conocimiento. De esta forma varios trabajos pueden surgir como consecuencia del desarrollo de esta tesis, como, por ejemplo:

- Incluir en el corpus otras fuentes de datos, además de Twitter, esto para alimentar la red neuronal convolucional (CNN) con más características lingüísticas que puedan ser obtenidas de otras expresiones utilizadas en español, a fin de generar un modelo que pueda mejorar los valores de predicción.
- Generar un modelo recursivo que se generé en base a la retro alimentación del corpus, con los textos nuevos que ya hayan sido predichos con una alta probabilidad de acierto tanto para *bullying* y *no bullying*.
- Al ser una propuesta de una arquitectura en capas se podría automatizar el proceso de preparación de datos que se retro-alimente con los textos con fallo de acierto o baja probabilidad de predicción, utilizando otros servicios de calidad de datos, integrando tecnologías como: Informática Data Quality[31] o SAP Business Objects Data Services[32].
- Desarrollo de una aplicación que se integre con los servicios de correo electrónico en instituciones educativas que permita clasificar los mails, cuando estos puedan tener signos de agresión y los mismos sean automáticamente descartados de la bandeja de entrada.

[31] https://www.informatica.com
[32] https://www.sap.com

- Mejorar el trabajo de Caizaluisa y Riofrio (2018), en el módulo de predicción de los mensajes de *WhatsApp*, esto mediante la preparación de corpus específicos para las categorías de DROGAS y SEXO, e integrar al sistemas de alarmas y alertas, con un modelo generado por la red neuronal convolucional (CNN) utilizada en este trabajo de investigación. Esto permitirá mayor efectividad en señales que se generen para el control parental.

Referencias Bibliográficas

Abrantes, M., Silva, G., & Fonseca, L. (2012). *Sentiment analysis in social networks: a study on vehicles*.

Ackermann, N. (2018). Introducción a las redes neuronales convolucionales 1D en keras para secuencias de tiempo.

Aggarwal, C. C. (2015). *Data Mining*. (Springer, Ed.). New York.

Agrawal, S., & Awekar, A. (2018). Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms (pp. 141–153).

Amajd, M., Kaimuldenov, Z., & Voronkov, I. (2017). *Text Classification with Deep Neural Networks*.

Anaconda, I. (2018). Anaconda Navigator — Anaconda 2.0 documentation. Retrieved October 14, 2018, from http://docs.continuum.io/anaconda/navigator/

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection *. *Statistics Surveys*, *4*, 40–79.

Beigi, G., Hu, X., Maciejewski, R., & Liu, H. (2010). *An Overview of Sentiment Analysis in Social Media and its Applications in Disaster Relief*.

Bellmore, A., Calvin, A. J., Xu, J.-M., & Zhu, X. (2015). The five W's of "bullying" on Twitter: Who, What, Why, Where, and When. *Computers in Human Behavior*, *44*, 305–314.

Bertona, F. (2005). *ENTRENAMIENTO DE REDES NEURONALES BASADO EN ALGORITMOS EVOLUTIVOS*. UNIVERSIDAD DE BUENOS AIRES.

Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Machine Learning and Semantic Sentiment Analysis based Algorithms for Suicide Sentiment Prediction in Social Networks. *Procedia Computer Science*, *113*, 65–72.

Broncano, F. (1995). *La mente humana*. (Trotta, Ed.) (Primera Ed, Vol. 8). Madrid: Editorial CSIC-CSIC Press.

Caizaluisa, E., & Riofrio, D. (2018). *Desarrollo de una aplicación para el control parental de WhatsApp en dispositivos móviles Android*.

Cardellino, C. (2016). Spanish Billion Word Corpus and Embeddings. Retrieved October 6, 2018, from http://crscardellino.github.io/SBWCE/

Casas Martínez, Á., & Del Real, J. (2017). *Reconocimiento de imágenes con Redes Convolucionales en C*. Universidad de Sevilla.

CEPAL. (2018). *La nueva revolución digital La revolución digital. De la Internet del consumo a la Internet de la producción*. Ciudad de Mexico.

Chachra, A., Mehndiratta, P., & Gupta, M. (2017). Sentiment analysis of text using deep convolution neural networks. In *2017 Tenth International Conference on Contemporary Computing (IC3)* (pp. 1–6). IEEE.

Chinthana, S., Madhushani, T., Marcus, E. R., Aberathne, I. A. A. U., & Premaratne, S. (2013). Sentiment Analysis for Social Media. *Market Potential of Thumba Karawila*, *4*.

Collobert, R., & Weston, J. (2011). *A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning*.

Crick, F. H., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, *192*, 1227–1232.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314.

Del Alcázar, J. P. (2018). *ESTADO DIGITAL ECUADOR JULIO 2018*.

Del Rey, R., Genevat, R., & Ortega, R. (2002). Etiquetas verbales en el vocabulario de docentes, padres y madres para nominar el fenómeno bullying. *Revista Electrónica Interuniversitaria de Formación Del Profesorado, ISSN-e 1575-0965, Vol. 5, Nº. 4, 2002 (Ejemplar Dedicado a: Medios, Violencia y Escuela)*, *5*(4), 11.

Deng, L. (2014). Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197–387.

ESET Latinoamérica. (2014). Reporte Sustentabilidad 2014 - ESET.

Facebook. (2018). Política de la plataforma - Facebook for Developers.

Galarsi, M. F., Medina, A., Ledezma, C., & Zanin, L. (2011). *Comportamiento, historia y evolución, Behavior, history and evolution*. San Luis.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and

analytics, *35*(2), 137–144.

Ganesh B, B. H., Kale, M., Mankame, P., & Kulkarni, G. (2018). *Deep Learning for Digital Text Analytics : Sentiment Analysis*.

Girard, G. (2017). El suicidio en la adolescencia y en la juventud. *Revista de Formación Continuada de La Sociedad Española de Medicina de La Adolescencia*, *5*.

Gómez, L. (2014). LA VIDA PRIVADA EN LA SOCIEDAD DIGITAL. LA EXPOSICIÓN PÚBLICA DE LOS JÓVENES EN INTERNET. *Aposta. Revista de Ciencias Sociales*, *61*(E-ISSN: 1696-7348), 1–32.

Guerrero, J. (2015). *La Encuesta de Condiciones de Vida ECV 2014*.

Gupta, A., Pruthi, J., & Sahu, N. (2017). *International Journal of Computer Science and Mobile Computing Sentiment Analysis of Tweets using Machine Learning Approach*. *International Journal of Computer Science and Mobile Computing* (Vol. 6).

Hinduja, S., & Patchin, J. W. (2010). Bullying, Cyberbullying, and Suicide. *Archives of Suicide Research*, *14*(3), 206–221.

Hu, X., & Liu, H. (2015). Social Media, Mining and Profiling in. In *The International Encyclopedia of Digital Communication and Society* (pp. 1–6). Hoboken, NJ, USA: John Wiley & Sons, Inc.

Huiqi Zhang, Dantu, R., & Cangussu, J. W. (2011). Socioscope: Human Relationship and Behavior Analysis in Social Networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *41*(6), 1122–1143.

Jesús San Isidro, M. (1998). *Proceso de Validación de Modelos de Simulación*. (Editorial CIEMAT, Ed.) (844th ed.).

Jorge, M., Rebato, C., Arcos, E., Gonzalo, M., Pavan, B., Notario, E., … Acevedo, I. (2018). *Un recorrido por la herramienta que se convirtió en plataforma*. (Hipertextual, Ed.).

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, *10*(3), 215–236.

Kaya, M., & Alhajj, R. (Eds.). (2018). *INFLUENCE AND BEHAVIOR ANALYSIS IN SOCIAL NETWORKS AND SOCIAL MEDIA.* SPRINGER.

Keras. (2018). Keras Epoch. Retrieved from https://keras.io/getting-started/faq/#what-does-sample-batch-epoch-mean

Khan, S., Rahmani, H., Shah, S. A. A., & Bennamoun, M. (2018). A Guide to Convolutional Neural Networks for Computer Vision. *Synthesis Lectures on Computer Vision*, *8*(1), 1–207.

Kim, R. (2018). Twitter Sentiment Analysis. Retrieved October 15, 2018, from

https://github.com/tthustla/twitter_sentiment_analysis_part3/blob/master/Capstone_part3-Copy2.ipynb

Kim, Y. (2014). *Convolutional Neural Networks for Sentence Classification*. New York.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lu, Y., Sakamoto, K., Shibuki, H., & Mori, T. (2017). Are Deep Learning Methods Better for Twitter Sentiment Analysis?

Matich, D. J., Ruiz, C. A., & Basualdo, M. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Rosario.

Mendoza, E. (2012). *Acoso cibernético o cyberbullying: Acoso con la tecnología electrónica*. *Pediatría de México* (Vol. 14).

Meneses-Bautista, F. D., & Alvarado, M. (2017). *Pronóstico del tipo de cambio USD/MXN con redes neuronales de retropropagación Forecasting USD/MXN Exchange Rate with Backpropagation Neural Networks*.

Mext. (2008). *bullying on the net "manual for handling and collection of cases for schools and teachers."* Tokio.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

MkDocs. (2014). Keras Documentation. Retrieved October 14, 2018, from https://keras.io/

Nandhini, B. S., & Sheeba, J. I. (2015). Online Social Network Bullying Detection Using Intelligence Techniques. *Procedia - Procedia Computer Science*, *45*, 485–492.

Neri, F., Aliprandi, C., Capeci, F., Cuadros, M., & By, T. (2012). Sentiment Analysis on Social Media. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 919–926). IEEE.

Nixon, R., Mercado, M., Faustino, H., Chuctaya, C., & Castro Gutierrez, G. (2018). *Automatic Cyberbullying Detection in Spanish-language Social Networks using Sentiment Analysis Techniques*. *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 9).

Organización Mundial de la Salud, Fundación Mutua Madrileña, & Fundación ANAR. (2016). *Acabemos con el Bullying - Mutua Madrileña*.

Ortega, A., & Ordoñez, M. (2013). *MANIFESTACIONES DE LA AGRESIÓN VERBAL ENTRE ADOLESCENTES ESCOLARIZADOS*.

Palmer, P. A., & Montaño, M. J. . (1999). *¿Qué son las redes neuronales artificiales? Aplicaciones realizadas en el ámbito de las adicciones*. *ADICCIONES* (Vol. 11).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? Sentiment Classification using*

Machine Learning Techniques.

Pennebaker, J. W. (2002). What our words can say about us: Toward a broader language psychology. *Psychological Science Agenda*, *15*(1), 8–9.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, *1*(1), 51–59.

Ptaszynski, M., Kalevi, J., Eronen, K., & Masui, F. (2017). Learning Deep on Cyberbullying is Always Better Than Brute Force. *LaCATODA*, 35–53.

Python Software Foundation. (2018). Welcome to Python.org. Retrieved October 14, 2018, from https://www.python.org/

Ramos, I. (2015). Análisis de redes sociales: una herramienta efectiva para evaluar coaliciones comunitarias Social network analysis: an effective tool for evaluating community coalitions. *Rev. Salud Pública*, *17*(3), 323–336.

Ramos, P. (2014). El 82% del ciberacoso se da por redes sociales | El Comercio. Retrieved from http://www.elcomercio.com/actualidad/ciberacoso-redes-sociales-ciberbulliyng-fbi.html

Raschka, S. (2015). *Python Machine Learning Unlock deeper insights into machine learning with this vital guide to cutting-edge predictive analytics*. (R. Banerjee & S. Copestake, Eds.). BIRMINGHAM - MUMBAI: Packt Publishing Ltd.

Ríos, & Matias, G. (2015). A Spanish Dictionary of Affect in Language Spanish DAL : A Spanish Dictionary of Affect in Language.

Rodríguez, P., Palomino, N., & Mondaca, J. (2017). *El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe*.

Romero, J., Dafonte, C., Gomez, A., & Penousal, F. (2007). *INTELIGENCIA ARTIFICIAL Y COMPUTACIÓN AVANZADA*. (Fundación Alfredo Brañas, Ed.) (13/2007). Santiago de Compostela.

Romero, O. (2017). *Using Machine Learning Techniques for Sentiment Analysis*. Barcelona.

Santillan, M. L. (2015). Ciberbullying: perfil de víctimas y victimarios - Ciencia UNAM.

Sanz, L. (2003). Unidad de Políticas Comparadas (CSIC) Análisis de redes sociales: o cómo representar las estructuras sociales subyacentes ARTÍCULO Análisis de Redes Sociales: o como representar las estructuras sociales subyacentes.

Schiaffino, S., & Amandi, A. (2009). Intelligent User Profiling (pp. 193–216). Springer, Berlin, Heidelberg.

Schmarzo, B. (2013). Big Data: Understanding How Data Powers Big Business. *Wiley*.

Schmidhuber, J. (2015). Deep Learning. *Scholarpedia*, *10*(11), 32832.

Semeval. (2015). International Workshop on Semantic Evaluation 2015 < SemEval-2015 : Semantic Evaluation Exercises.

Severyn, A., Inc, G., & Moschitti, A. (2015). *Twitter Sentiment Analysis with Deep Convolutional Neural Networks*.

Spyder Doc Contributors. (2018). Spyder: el entorno de desarrollo científico de Python - Documentación - Documentación de Spyder 3. Retrieved October 14, 2018, from https://docs.spyder-ide.org/

Squicciarini, A., Rajtmajer, S., & Griffin, C. (2017). Positive and negative behavioral analysis in social networks. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(3), e1203.

Suárez, S., & Guerrero, F. (2013). La conversación sobre big data en Twitter. Una primera aproximación al análisis del discurso dominante. *Revista de Recerca i d'Anàlisi [Societat Catalana de Comunicació]*, *33*(2014-0444), 113–131.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406–418.

Torregrosa, J. R., & Crespo, E. (1984). *Estudios básicos de psicología social*. Hora.

Turing, A. M. (1950). *COMPUTING MACHINERY AND INTELLIGENCE*. *Computing Machinery and Intelligence. Mind* (Vol. 49).

Villacampa, C., & Gómez, M. (2016). NUEVAS TECNOLOGÍAS Y VICTIMIZACIÓN SEXUAL DE MENORES POR ONLINE GROOMING. *Revista Electrónica de Ciencia Penal y Criminología* .

Wasserman, F. (1994). Social network analysis: Methods and applications. In C. U. Press (Ed.). New York.

Yao, X. (1993). Evolutionary Artificial Neural Networks. *International Journal of Neural Systems*, *4*, 203–222.

Zipf, G. K. (1950). Human behavior and the principle of least effort. Cambridge, (Mass.): Addison-Wesley, 1949, pp. 573. *Journal of Clinical Psychology*, *6*(3), 306–306.