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.

Quito, noviembre 2018

UNIVERSIDAD INTERNACIONAL SEK

Introducción

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)

Problema de la Investigación

Uso de Internet

En 2015 se estimó que en el mundo 3.174 millones de personas usaron Internet, y se prevé que para el 2020 la actividad se incremente en un 60%.

Cyber bullying

Se ha detectado un alto nivel de grooming y cyber bullying en ciudades del Ecuador, como Quito, Manta y Guayaquil.

Acosos en escuelas

Uno de cada cuatro acosos ocurridos en las escuelas españolas, se producen utilizando medios tecnológicos.

Hostigamiento a adolecentes

Se determinando que el 27% de adolescentes ecuatorianos sufre marginación, el 46% de hostigamiento, el 17% de agresión y el 10% sufre de extorción en las redes sociales.

Acoso en internet

En América Latina, se determinó por una encuesta realizada en esta región, que el 30.7% de adolecentes sufrió de acoso a través de internet

Suicidio

Con aparición del cyber bullying se ha determinado que existe una estrecha relación entre el bullying y el suicidio la cual se ha incrementado en los últimos años.

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

Objetivos

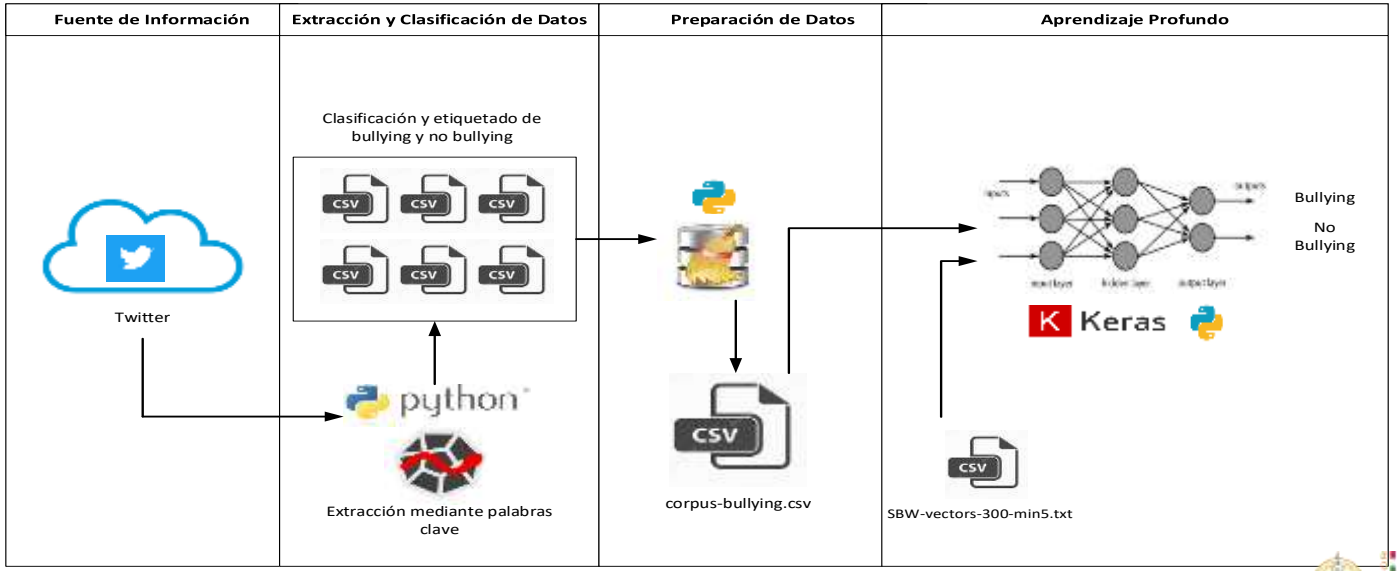
Crear un modelo de predicción de ataques de cyber bullying mediante técnicas de aprendizaje profundo apoyándose en un corpus de entrenamiento para la clasificación e identificación de texto en español con características de bullying

- 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*.

Estado del Arte

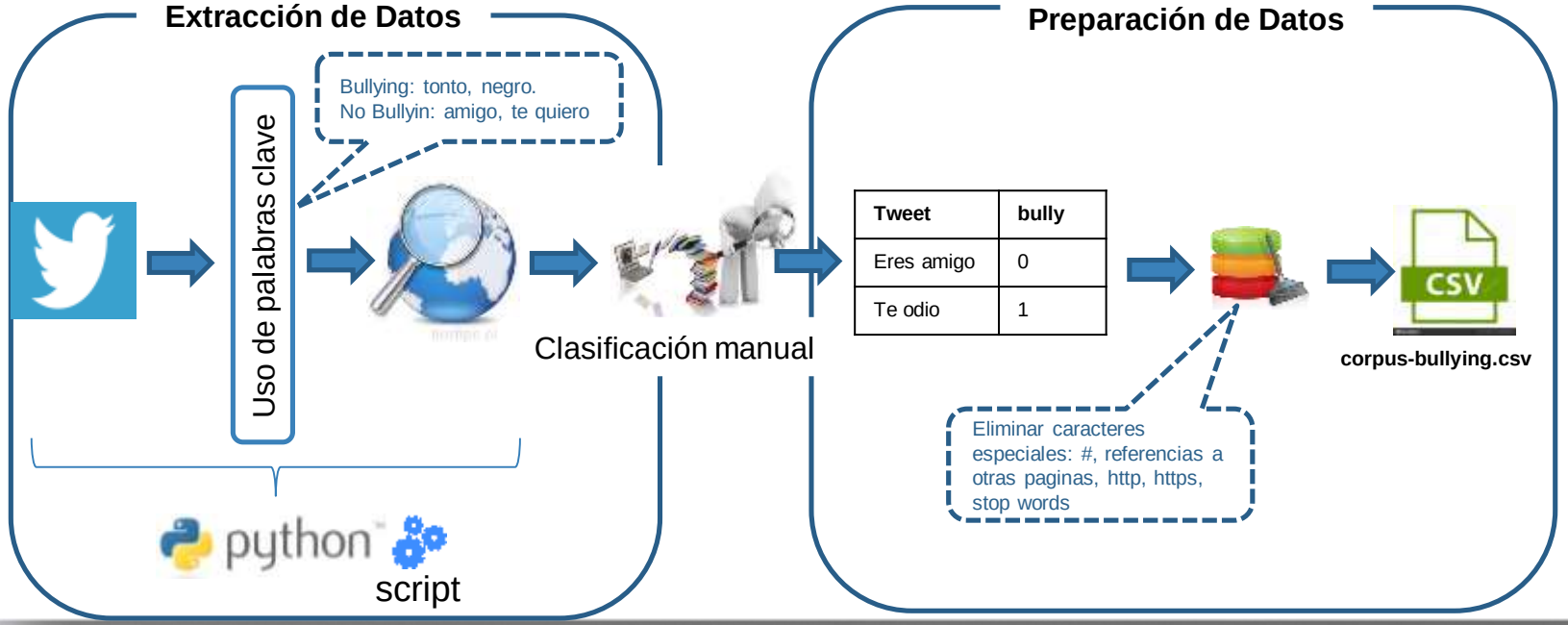
	Autores	Enfoque
Análisis de comportamiento en redes sociales	Galarsi, Medina, Ledezma, y Zanin (2011) Kaya y Alhajj (2018) Huiqi Zhang, Dantu, y Cangussu (2011) Squicciarini, Rajtmajer, y Griffin (2017) Pennebaker (2002) Hu y Liu (2015) Schiaffino y Amandi (2009)	Comportamiento sociológico y Humano, análisis de rumores en Twitter, aspectos que afectan al envejecimiento humano, análisis de personalidad, situación social y relaciones interpersonales, análisis de perfiles de consumidores para proveedores de servicios
Análisis de sentimientos en las redes sociales	Neri, Aliprandi, Capeci, Cuadros, y By (2012) Abrantes, Silva, y Fonseca (2012)	Análisis de sentimientos en redes sociales mediente la representacipon del conocimiento y la web usando minería de datos.
Uso de aprendizaje automático (*machine learning*) para predicción de sentimientos	Birjali, Beni-Hssane, y Erritali (2017) Romero (2017) Pang, Lee, y Vaithyanathan (2002)	Analizar sentimientos positivos, negativos o neutrales, se usa técnicas de aprendizaje automático como: Naive Bayes, Random Forest, Support Vector Machines, Recurrent Neural Networks y Maximum Entropy Classification
Análisis de sentimientos en textos con aprendizaje profundo (*deep learning*)	Amajd, Kaimuldenov, y Voronkov (2017) Chachra, Mehndiratta, y Gupta (2017) Ganesh B, Kale, Mankame, y Kulkarni (2018)	Análisis de sentimientos extrayendo las características lingüísticas de las expresiones utilizandas en los textos, utilizando varias técnicas de aprendizaje profundo como son: Convolutional Neural Network (CNN), y LSTM
Uso de aprendizaje automático (*machine learning*) para análisis de sentimientos en redes sociales	Chinthana, Madhushani, Marcus, Aberathne, y Premaratne (2013) Beigi, Hu, Maciejewski, y Liu (2010) Gupta, Pruthi, y Sahu (2017)	Analizar sentimientos positivos y negativos específicamente en redes sociales, mediante el uso de técnicas de aprendizaje automático, como: Naive Bayes, Maximum Entropy, KNearest Neighbors (KNN) y Support Vector Machines (SVM).
Uso de aprendizaje profundo (*deep learning*) para análisis de sentimientos en redes sociales	Lu, Sakamoto, Shibuki, y Mori (2017) Severyn, Inc, y Moschitti (2015)	Análisis de sentimientos expresados en redes sociales como Twitter. Las técnicas utilizadas son: CNN y RNN
Predicción de cyber bullying mediante aprendizaje automático (*machine learning*)	Bellmore, Calvin, Xu, y Zhu (2015) Nixon, Mercado, Faustino, Chuctaya, y Castro Gutierrez (2018) Caizaluisa y Riofrio (2018)	Clasificación de los tweets respecto a episodios de bullying en las diferentes categorías. Análisis de sentimientos para la detección de cyber bullying en las redes sociales en lenguaje español. Alertar por whasapp en base a la clasificacion de textos en tres categorias: SEXO, DROGAS y BULLYING
Predicción de bullying usando aprendizaje profundo (*deep learning*)	Ptaszynski, Kalevi, Eronen, y Masui (2017) Agrawal y Awekar (2018)	Predicción de cyber bullying, utilizando las redes neuronales convolucionales (CNN), en el idioma japonés. Detección del cyber bullying, comparar técnicas de Machine Learning y Deep Leraning

Solución Adoptada

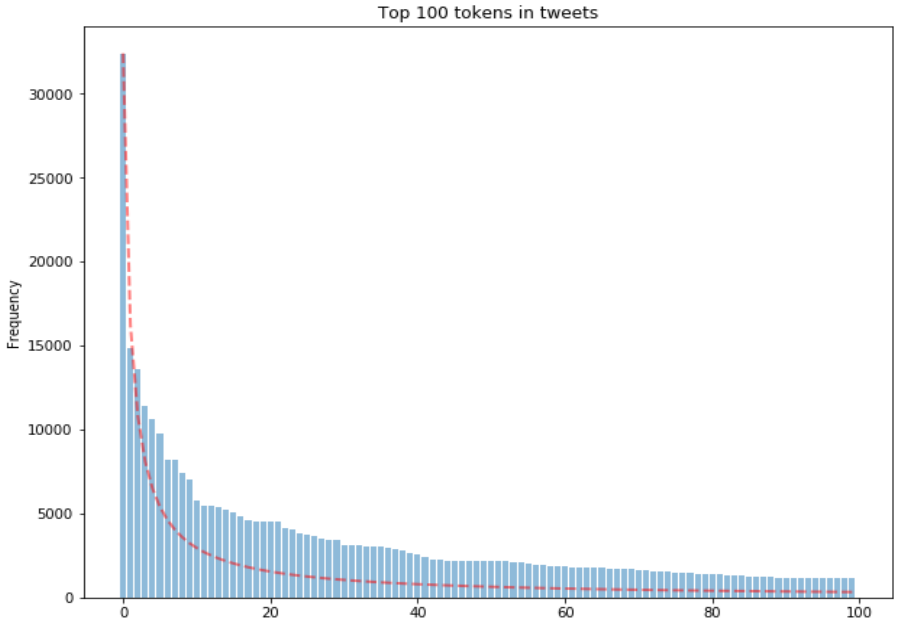


UNIVERSIDAD INTERNACIONAL SEK

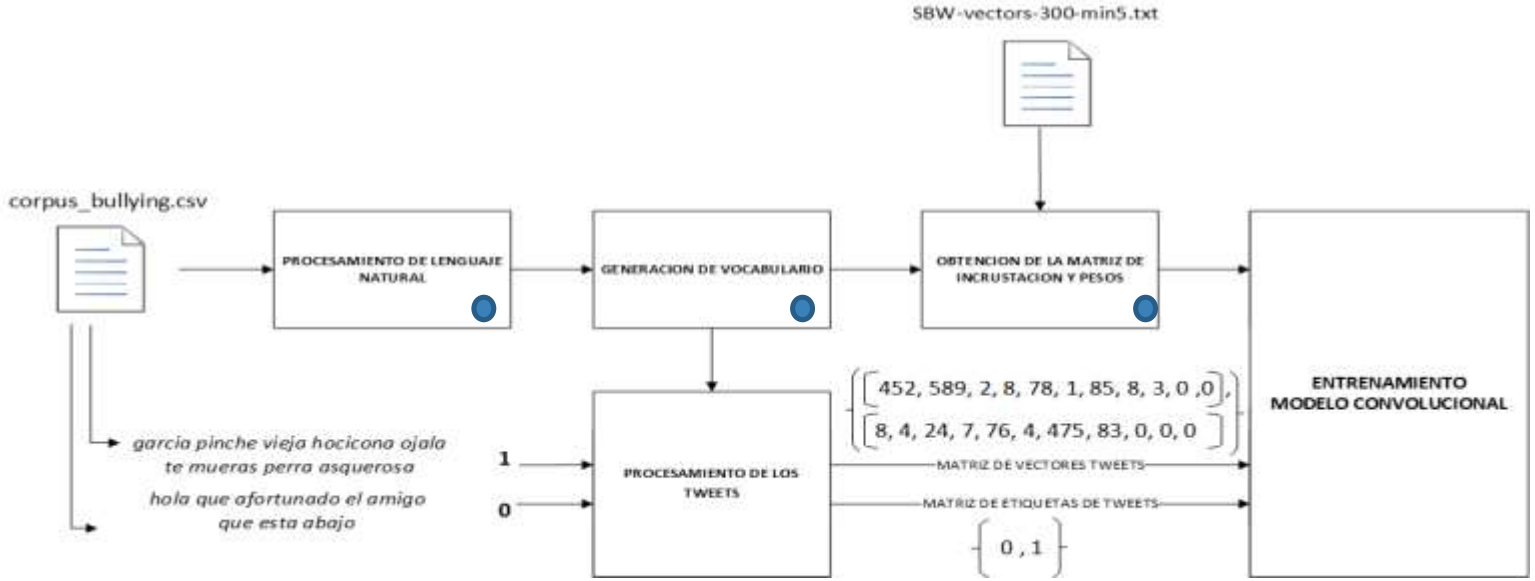
Conformación de Corpus



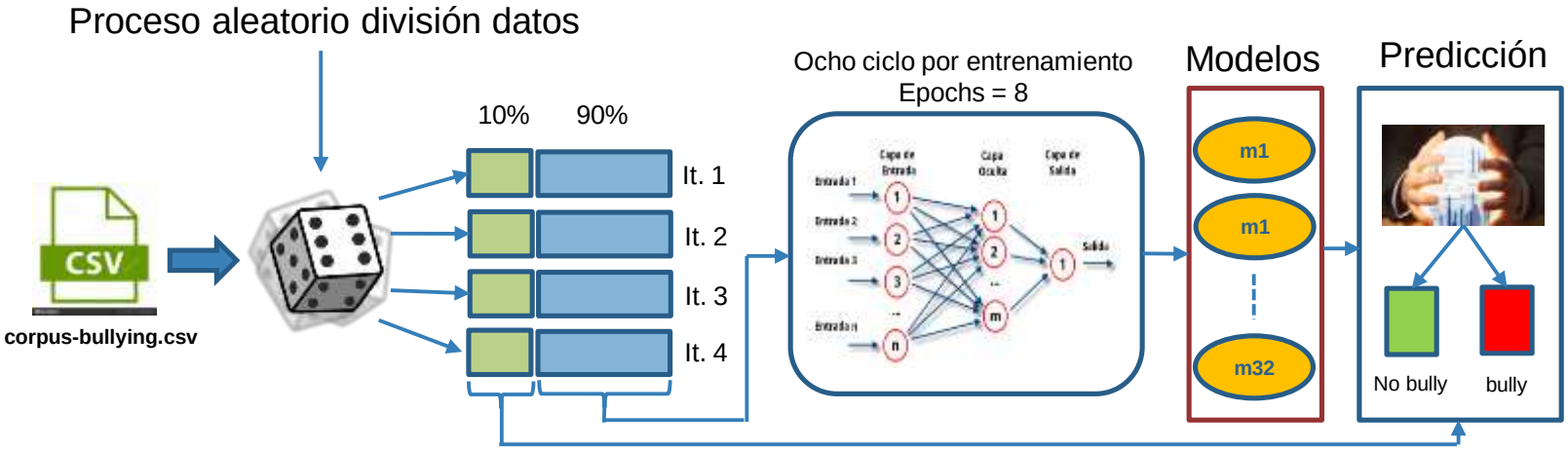
Análisis del corpus



Proceso de entrenamiento



Validación de Datos



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

Demostración.....!!