DESDE 2023
https://revista.ccaitese.com

RICT Revista de Investigación Científica, Tecnológica e Innovación

Publicación Semestral RICT Vol. 2 No. 4 (2024) P. 64 – 71

ISSN: 2992-7315

Redes Neuronales Recurrentes para la detección de noticias falsas cuentas Bot en Twitter

Recurrent Neural Networks for fake news and Bot accounts detection on Twitter

Sandra Paulina Castillo Cárdenas [a,b, d], Francisco Jacob Ávila Camacho [a,c]

[a] *División de Ingeniería en Sistemas Computacionales, Tecnológico Nacional de México /TES Ecatepec, 55210, Ecatepec, Estado de México, México.* [b] *Programa Investigadoras e Investigadores COMECYT, Consejo Mexiquense de Ciencia y Tecnología, 50120, Toluca, Estado de México, México.* [c] *Centro de Cooperación Academia Industria, Tecnológico Nacional de México / TES Ecatepec, 555210, Ecatepec, Estado de México, México.* [d] *Doctorado en Sistemas Computacionales, Universidad Da Vinci, 11520, CDMX, México*

Resumen

Este artículo presenta el desarrollo e implementación de un modelo de Inteligencia Artificial (IA) para la detección de cuentas bot y noticias falsas en Twitter (ahora X). El modelo combina técnicas de Análisis de Sentimiento (AS), Procesamiento de Lenguaje Natural (PLN), Redes Neuronales Simples (RNS) y Redes Neuronales Recurrentes (RNR), diseñadas para identificar bots a nivel de tweet y diferenciar entre noticias verdaderas y falsas. El objetivo principal es proporcionar un sistema capaz de detectar cuentas operadas por bots de manera rápida y eficiente, al mismo tiempo que genera un repositorio y un catálogo de medios que diseminan información falsa. El proyecto fue desarrollado en la plataforma Google Colab, utilizando bibliotecas especializadas para el análisis de datos y el procesamiento de texto en Python, como NLTK, TextBlob y spaCy. El modelo también integra la herramienta Botometer, un algoritmo basado en IA que facilita la identificación de bots mediante el análisis de patrones de comportamiento y contenido en cuentas sospechosas. El sistema emplea algoritmos de aprendizaje automático para procesar grandes volúmenes de tweets, y su implementación permite la detección automatizada de bots en tiempo real. Los resultados obtenidos muestran una precisión del 75.96% en la detección de bots, validando la eficacia de las Redes Neuronales Recurrentes utilizadas en el modelo. Adicionalmente, se analizaron métricas como retweets y likes, lo que evidencia la funcionalidad y viabilidad del enfoque propuesto para combatir la desinformación en redes sociales.

Palabras clave: Análisis de Sentimientos, Bots, Redes Neuronales Recurrentes y Redes Sociales.

Abstract

This paper presents the development and implementation of an Artificial Intelligence (AI) model for the detection of bot accounts and fake news on Twitter (now X). The model combines Sentiment Analysis (SA), Natural Language Processing (NLP), Simple Neural Networks (SNN) and Recurrent Neural Networks (RNR) techniques, designed to identify bots at the tweet level and differentiate between true and fake news. The main objective is to provide a system capable of detecting bot-operated accounts quickly and efficiently, while generating a repository and catalog of media that disseminate false information. The project was developed on the Google Colab platform, using specialized libraries for data analysis and text processing in Python, such as NLTK, TextBlob, and spaCy. The model also integrates the Botometer tool, an AI-based algorithm that facilitates the identification of bots by analyzing behavioral and content patterns on suspicious accounts. The system employs machine learning algorithms to process large volumes of tweets, and its implementation allows for automated detection of bots in real time. The results obtained show a 75.96% accuracy in detecting bots, validating the effectiveness of the Recurrent Neural Networks used in the model. Additionally, metrics such as retweets and likes were analyzed, which demonstrates the functionality and viability of the proposed approach to combat misinformation on social networks.

Keywords: Analysis of Sentiments, Bots, Neural Networks and Social Networks.

1. Introducción

Hoy en día las personas utilizan cada vez más las redes sociales, ya que son uno de los medios de comunicación preferidos [1], debido a que tienen un impacto social decididamente fuerte [2] como Twitter o X, que es un servicio de red social [3], donde muchos usuarios se convierten en

*Autor para la correspondencia: zandy.castillo@gmail.com
Correo electrónico: zandy.castillo@gmail.com (Sandra Paulina Castillo Cárdenas), fjacobavila@tese.edu.mx (Francisco Jacob Ávila Camacho).

Historial del manuscrito: recibido el 28/6/2023 última versión-revisada recibida el 9/08/2024 aceptado el 22/08/2024, en línea (postprint) desde el 15/11/2024, publicado el 15/11/2024. DOI: https://doi.org/10.5281/zenodo.14194198

portadores de información, mediante este sitio las personas se pueden expresar y compartir opiniones, lamentablemente puede haber difusión de información dañina, a través de diferentes cuentas mediante el intercambio de tweets y retweets [4], publicados por usuarios [5] debido a que la información se propaga a una velocidad extremadamente alta. Las noticias falsas que se difunden pueden influir en las personas e inferir en la toma de decisiones, los usuarios reciben tweets, algunos de los cuales son enviados por bots. La mayoría de los estudios sobre los bots sociales se centran en Twitter [6] ya que se puede acceder fácilmente a sus datos y donde la desinformación ha cobrado un gran protagonismo [7].

Los bots mal intencionados pueden causar daño antes de ser detenidos, la propagación de noticias falsas en las redes sociales origina confusión dentro de la sociedad, por lo tanto, es importante detectarlo para reconocerlos. En este trabajo se hace uso del método de análisis de datos, Redes Neuronales Simples y Redes Neuronales Recurrentes RNR (o en inglés Recurrent Neural, Network RNN) [8], para su estudio e identificación entre otras, así como el procesamiento automático de los textos a través del análisis de sentimientos AS (o SA por sus siglas en inglés, Sentiment Analysis) [9] [10] o minería de opinión [11], usando el dataset del repositorio Cresci, donde todo se inició desde cero.

En este estudio, se propone el diseño de un modelo inteligente para la identificación de bots en Twitter, en el cual se realizó la extracción y análisis de sentimientos de nuestro dataset. Para la visualización de los datos se utilizó la librería Matplolib ya que realiza la interpretación gráfica para facilitar la compresión del comportamiento. El modelo fue entrenado por medio de dos tipos de redes neuronales: una red simple y una red neuronal recurrente, en la cual se implementó una arquitectura LSTM (Long Short Term Memory) para optimizar la capacidad de la red al trabajar con secuencias de datos y gestionar dependencias temporales de los datos, mejorando así el análisis de secuencias en la detección de patrones característicos de cada cuenta.

En la clasificación de los datos se utilizó la API de Twitter, junto con el Dashnoard de Twitter, con el fin de obtener las claves de acceso a nuestras API Key necesarias para acceder a los datos, creando una cuenta en Twitter específicamente con las que se hizo las pruebas, desarrollado internamente, sumando a esto, se instaló del dataset Cresci, esto para hacer la extracción de los tweets de datos para detectar bots, estos datos se prueban en tiempo real, optimizando el proceso para alcanzar la mayor precisión posible.

I. Estado del Arte

A medida que han aparecido nuevos métodos de detección de bots más eficaces y complejos, los bots también han evolucionado para adaptarse, por ello existen diferentes métodos de detección de bots.

Los primeros trabajos que introdujeron el término de análisis de sentimientos fue el presentado por [12], en su publicación definen como encontrar expresiones de sentimientos para un sujeto dado y determinar la polaridad de cada sujeto mencionado en el texto. El análisis de sentimientos también se ha utilizado para la detección de bots, los autores [13] aplicaron el análisis de sentimientos como métodos para la verificación y clasificación de cuentas, usando algoritmos de aprendizaje automático, dando como resultado satisfactorio la precisión para la identificación de bots.

El autor P. Burgos [14] para la detección de usuarios falsos en Twitter, se empleó el DataSet Cresci-2017 y como método Random Forest, definiendolo como una de las mejores técnicas de clasificación supervisadas, para determinar tipos de bots, clasificando la información extraida de los metadatos del usuario empleando información derivada de sus tweets.

Los autores [15] aplicaron un método exploratorio no supervisado donde identifica de manera adaptativa el comportamiento de los bots a medida que evolucionan. La implantación de este método da un 30% en la precisión de manera inteligente en el espacio de funciones, proponiendo el algoritmo BotWall este algoritmo exploración adaptativa de Twitter realiza una detección con una precisión de 90% de bots no descubiertos.

En los trabajos [16], [17] emplearon el conjunto de algoritmos de aprendizaje automático, analizando y etiquetando usuarios manualmente, para la detección de falsos seguidores de Twitter, dando como resultado que los modelos Random Forest y las máquinas de vectores soporte (SVM) logran un 97% de precisión. Wang [18] combina métodos similares, llevando a cabo características basadas en gráficos para identificar bots en Twitter y cuatro clasificadores de aprendizaje automático.

Vander Walt & Eloff van der Walt & Eloff [19] detectaron con éxito cuentas falsas creadas por bots, con el uso de modelos de aprendizaje automático supervisado, dando como resultado una predicción de cuentas falsas generadas por humanos con un 49,75% de efectividad.

En 2012 [20] utilizaron un conjunto de datos con usuarios etiquetados manualmente, para identificar características que diferencian a bots, humanos y cyborg en Twitter, el sistema se basó en varios modelos que determina la clase mediante la regla de Bayes obteniendo una precisión del 96%, en el modelo.

Los autores [21] implementaron un algoritmo de aprendizaje automático, tiene como características que incluyen longitud de los nombres, tasa de reenvío, patrones temporales, expresión de sentimientos y variedad de mensajes para la detección de bots. Esta técnica de detección de Twitter es efectiva para detectar bots con una tasa de clasificación errónea del 2,25%.

Los autores Chu, Zi , Gianvecchio, Steven y Wang, Haining [20] diseñaron un sistema de clasificación automatizado, que consta de cuatro partes el componente de antropía (verifica patornes de tiempo de tweet), el componente de aprendizaje automático (comprueba el contenido de spam) y el componente de propiedades de la cuenta (verifica los valores anormales de la cuenta de Twitter) y tomador de decisiones (resume las características identificadas de usuario para

determinarla probabilidad de ser humano, bot o cyborg) para la clasificación de cuentas humanas, bots o cyborg en Twitter, con resultados satisfactorios

En [22] desarrollaron un método de cuatro técnicas para detectar bots, el cual consiste en detección de inconsistencias y modelado de comportamiento, análisis de texto, análisis de red y aprendizaje automático, con resultados satisfactorios.

Por otro lado [23] proponen un marco de referencia y evaluación de detección de bots de Twitter TwiBot-22, con un método basado en gráficos, con resultados satisfactorios.

En 2016 en el proyecto Truthy [24] desarrollaron una técnica supervisada de manera exitosa, para evaluar la probabilidad de que una cuenta sea un bot.

Este trabajo se enfoca en analizar los metadatos del contenido de los twittees en tiempo real para identificar cuentas bots en Twitter. Se diseño y entreno una red neuronal simple y una red neuronal recurrente, para mayor eficacia en los resultados. Nuestro trabajo es un complemento valioso para la investigación existente sobre la detencción de. Bots en Twitter

Los bots de Twitter [25]incluyen:

Spambots: (bots de spam), difundir spam sobre diferentes temas.

Paybots: Ganan dinero ilicitamente. Algunos bots de pago, copian el contenido de los tweets de fuentes respetadas, pagan micro-URL que guian a los usuarios a sitos que pagan al creados del bot, por guiar el tráfico al sitio.

Influencia: Los bots intentan infuir en las conversaciones de Twitter sobre un tema especifico

II. Antecedentes Teoricos

Este apartado se aborda el tema principal de este trabajo, el modelo de RNR, este sistema usa una arquitectura LSTM, una variante de las RNN, la red recurrente simple y análisis de sentimientos. En nuestro modelo se emplean la red neuronal simple, RNR y el modelo neuronal de memoria a corto plazo LSTM.

A. Redes Neuronales Recurrentes

El modelo de Redes Neuronales Recurrentes es muy utilizado en multiltud de dispositivos de uso frecuente en la actualidad.

Las RNR son los modelos de redes de neuronas artificiales (RNA) en el cual las conexiones entre unidades forman un ciclo dirigido, se usan especificamente para el reconocimiento de voz y escritura [26] para porcesar datos de estructura, está conformada por neuronas recurrentes que presentan un bucle de retroalimentación [27]. Puede tomar simultáneamente una secuencia de entradas y producir una secuencia de salidas [28], utilizan su razonamiento de experiencias anteriores, para informar los próximos eventos [29], analizan secuencias temporales, y el procesamiento secuencial datos [30] que cambian con el tiempo. Su principal ventaja es la posibilidad de almacenar una representación de la historia reciente de la secuencia [31].

Las RNR engloban la capacidad de analizar secuencias temporales de datos de tamaño variable y predecir cual será el siguiente valor en la serie, en este caso las RNR toman como entrada una serie temporal de datos y predicen con cierta probabilidad, cuál será el siguiente valor para saber si es una cuenta bot en Twitter o no.

Las RNR hacen uso de información secuencial, para procesar datos. Tienen capacidad para procesar y obtener información de datos secuenciales [32]. Son modelos de redes de neuronas artificiales, en la que las conexiones entre unidades forman un ciclo dirigido, esto es una secuencia en la caminata a lo largo tanto de vértices como bordes. Tiene capa de entrada, capa oculta y capa de salida, como se ilustra en la figura1.

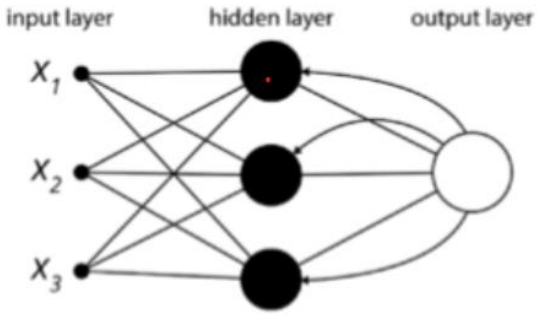

Figura 1: Red Neuronal Recurrente RNR [26].

Según [26], los ciclos dentro de la estructura de la red, se pueden analizar secuencias de tiempo. Frecuentemente se realiza el despliegue de la estructura, para obtener una versión de la red que dependa de una secuencia de entradas, conforme lo visto en la figura 2. Donde los pesos y los sesgos del bloque S son compartidos, la salida es *ht*, las entradas *xi* con $i \in [0; t]$. El número de bloques depende de la longitud de la secuencia a analizar.

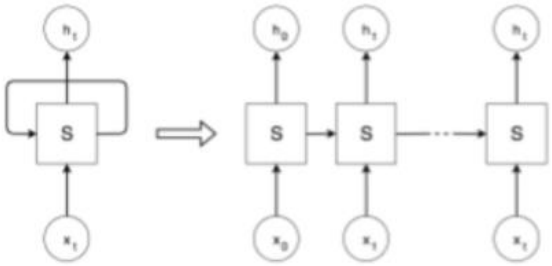

Figura 2: Despliegue de una RNR [26].

Una RNN puede ser clasificada en parcial o totlamente recurrente, las totalmete rrecurentes son las que cada neurona puede estar conectada a cualquier otra y sus conexiones recurrentes son fijas, las parcialmente recurrentes son las que sus conexiones recurrentes son fijas, reconocen o reproducen secuencias las conexiones son hacia adelante, incluyen un conjunto de conexiones retroalimentadas [33], tienen

conexiones recurrentes [34], guardan una copia de los valores anteriores de la capa que contiene los nodos recurrentes y los usa como entrada adicional para el siguiente paso, permitiendo que la red muestre un comportamiento temporal dinámico para una secuencia de tiempo [35].

En el desarrollo de este modelo utilizado para implementar la RNR es secuencial, se compone de dos capas, una oculta con nodos recurrentes y una de salida con uno o dos nodos lineales.

B. Memoria a corto plazo a largo plazo

En 1997 surgen las redes LSTM [36]. Es un tipo de RNN, que contiene celdas especiales que son capaces de aprender dependiendencias a largo plazo. El uso de las LSTM en RNN ayuda a resolver el problema del gradiente de desaparición, puesto que las celdas LSTM permiten que los grandientes fluyan sin cambios [37]. El componente principal de un modelo LSTM es un bloque de memoria, que consta de una o más celdas de memoria, una compuerta de entrada y una compuerta de salida [38]. En la figura 3 se muestran la estructura de una celda LSTM con puertas y funciones de activación.

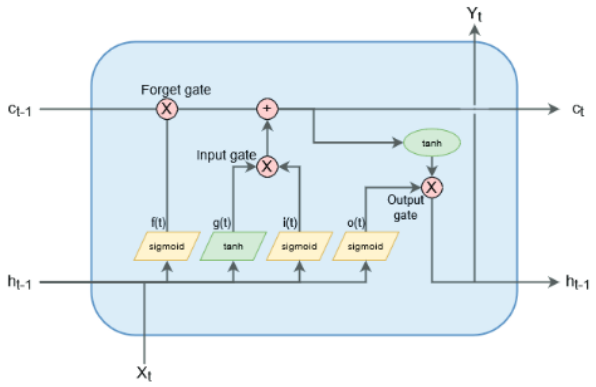

Figura 3: Estructura de una LSTM [39].

2. Materiales y Método

El Proyecto se desarrollo dentro de la plataforma Google Colab, ya que ofrece un entorno de programación basado en la nube, la utilizamos por su integración con Python e instalamos las librerías NLTK, TextBlob y spaCy para realizar el análisis de sentimientos, la librería Pandas se utilizó en el análisis de datos, para el procesamiento de una gran cantidad de tweets a la vez y la identificación de patrones en el comportamiento de cuentas sospechosas se basó en el modelo de red neuronal, utilizando Botometer ya que es un algoritmo basado en Machine Learning que se usa en Inteligencia Artificial para la detección de bots o en este caso humanos.

C. Extracción de datos de Twitter

La extracción de datos se llevó a cabo mediante Tweepy API con el fin de recopilar datos de Twitter incluyendo la autenticación y extracción de tweets, adicionalmente, se utilizó Botometer con el propósito de hacer una evaluación de cuentas sospechosas que pudieran ser bots.

D. Recopilación de datos

Se utilizó el Dataset Cresci, que contiene varios dataset seleccionando el Botometer feedback 2019, este repositorio contiene tweets, con el objetivo de recopilar un listado de usuarios, tanto reales como falsos para etiquetar si es bot o humano, se sacaron todos los metadatos en una tabla para exportarlos en un archivo Excel para hacer el arreglo en una tabla y se sacaron los más importantes lo que nos permitió conocer acerca de la descripción de la cuenta, si la cuenta es de una persona que estuvo retuiteando, compartiendo información y esos datos se colocaron en un arreglo para saber si es un humano o bot, con los siguientes datos: Nombre de la cuenta, de donde proviene la cuenta, descripción, seguidores que tiene, si es una cuenta humana o bot, si tiene verificación o si está protegida la cuenta, con la extracción de estos metadatos se entrenó la red neuronal recurrente.

Para garantizar que los datos estuvieran listos con el fin de ser procesados por el modelo neuronal, se llevó a cabo un proceso de limpieza de datos. Este proceso consistió en la eliminación de información irrelevante que podría afectar la precisión en la clasificación de los tuits, los elementos eliminados incluyen caracteres especiales, direcciones URLs, emojis y menciones a usuarios, etiquetas, que no aportaban valor significativo al análisis del contenido textual, de esta manera, se optimizó la calidad de los datos de entrada, mejorando el desempeño del modelo para su clasificación, después de la limpieza de datos se aplicó el método de extracción de características Tf-idf (es un cálculo estadístico), para transformar los datos cualitativos en representaciones cuantitativas, este proceso permitió convertir el texto en vectores numéricos, asignando un peso a cada término interpretados y procesados de manera eficiente facilitando la clasificación y análisis preciso de la información.

E. Clasificación de noticias falsas

Se creo un archivo en Excel que incluye noticias recopiladas, conjunto de datos de artículos de noticias falsas y noticias verídicas, noticias de diferentes temas, para evitar el ajuste excesivo de los clasificadores y proporcionar más datos de texto para mejorar el modelo y poder entrenar la red neuronal. El conjunto de datos (noticias) contiene cuatro columnas las cuales son:

- Número (partiendo desde 0).
- Título (Encabezado de la noticia).
- Texto (Contenido de la noticia).
- Etiqueta (0=falso, 1=real).

Con la finalidad de tener una mejor predicción en la

detección de bots se recopilaron 1000 noticias.

F. Entrenamiento de la red

Se creó un archivo de Excel, donde se incluyen noticias recopiladas para poder entrenar la red neuronal. En la creación de la red neural se utilizó la biblioteca para aprendizaje automático Tensor Flow, se instaló e importo las librerías, para poder cargar los datos (el Excel que se creó con las noticias) y se añadió al entorno de Google Colab (es una plataforma para ejecutar código en la nube).

Para el preprocesamiento de los datos se utilizaron dos herramientas de detección de bots, API de Twitter y Botometer, con el fin de revisar la evaluación que hace cada cuenta de manera más gráfica y categorizar algunas cuentas de Twitter que no estaban etiquetadas, nos apoyamos con la plataforma de Botometer, con el propósito de revisar la evaluación que hace cada cuenta. Se creo un arreglo, el cual incluye una lista de cuentas de Twitter que comparten noticias, ingresamos los metadatos, para obtener el nivel de bot de cada cuenta que tiene cada arreglo e importamos Twitter con RapidAPI. Cabe mencionar que Botometer trabaja con la API de Twitter y con RapidAPI, con los metadatos extraídos manualmente se genera el nivel de bot, de cada cuenta que tenemos en el arreglo, esta clasificación permite que el algoritmo de aprendizaje automático clasifique un conjunto de cuentas de prueba. Se crearon variables para poder almacenar la información obtenida de cada metadato, en el archivo de Excel creado que incluye noticias recopiladas, esto para poder entrenar la red neuronal.

Para la creación de la red neuronal, utilizamos la biblioteca para aprendizaje automático Tensorflow, donde se cargaron los datos (nuestro Excel que se creó con las noticias) y se cargó a Google Colab, para ejecutar el entrenamiento y prueba con un 80% del contenido del archivo con noticias. Al finalizar el proceso de prueba se generan matrices para hacer uso de la red.

Con ayuda del repositorio de Cresci, se utilizó el DataSet Botometer feedback, para la implementación de una red neuronal recurrente. Adicionalmente, se instaló la AIP de Botometer para crear un arreglo que incluye una lista de cuentas de Twitter que comparten noticias, se ingresaron los metadatos de estas cuentas que fueron procesados para determinar el nivel de actividad bot de cada una de ellas. En un arreglo se guardaron algunos de las cuentas que se trabajan en Twitter que comprende noticias para su posterior análisis.

Las siguientes secciones describen el marco general para la recopilación de datos y el análisis metodológico de nuestro estudio.

La Metodología desarrollada proporciona un porcentaje de efectividad, detecta con precisión las cuentas de bots en Twitter y la conbinación de una red LSTM, encargada de procesar el texto de los tuitst y una RNR que toma como entrada las palabras procesadas.

La arquitectura del sistema representada en la figura 4, resume de manera integral todo el recorrido de los datos recopilados, para detectar cuentas bot en Twitter.

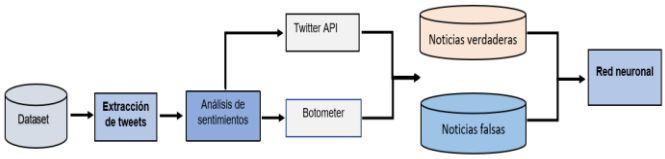

Figura 4 Proceso de reconocimiento de noticias falsas, (fuente propia).

Con ayuda del DataSet se realizó la extracción de los tweets, importamos las librerías nltk y hunspell con el objetivo de clasificar las palabras en categorías como negativas, positivas o neutrales, con el fin de iniciar el análisis de sentimientos, nos apoyamos en la plataforma API Twitter y Botometer. Se importaron los datos de Botometer en un arreglo, que incluye una lista de cuentas de Twitter que comparte una colección de noticias, con esa información se implementó y entreno la red neuronal.

Con la creación de la red neuronal y la instalación e importación de las librerías, se cargaron los datos que se añadieron al entorno de Google Colab, donde se subió a nuestro DataSet para su procesamiento en la aplicación, a partir de este paso, se comenzó a codificar, para trabajar con el DataSet y ver su contenido. Se verifico cada cuenta de Twitter para extraer sus datos y revisar la evaluación que hace de cada cuenta de manera más gráfica.

Se inicio el análisis de sentimientos e importamos la librería nltk y texblob, con nuestro código introducimos el análisis de sentimientos a la tabla de Tweets, en la figura número 7, se presentan los resultados obtenidos de este análisis (los datos que nos arrojó).

	Tweets	SA
0	@JoseExotico Peacock es el streaming de Univer...	-1
1	Encuesta publicada en El Universal refleja que...	0
2	@El_Universal_Mx Quien será el grupo Firme? 🙄	0
3	Tuve el gusto de participar en la #113Asamblea...	1
4	RT @MarkDeReborn: Si se confirma que @Claudias...	0
5	RT @El_Universal_Mx: 🚨 #ÚLTIMAHORA 🚨 Confirmad...	0

Figura 5: Resultado con análisis de sentimientos.

En el código se crearon y las variables para guardar los valores de análisis de sentimiento de cada noticia, clasificando como positiva, negativa o neutral, posteriormente se implementó una función que calcula e imprime el porcentaje que da el análisis de sentimientos.

Para identificar cuentas potencialmente automatizadas, se integró la herramienta Botometer donde se creó una matriz que almacena una lista de cuentas de Twitter (para obtener los datos de Twitter se utilizó Twitter API) dedicadas a compartir noticias, y se ingresaron los metadatos para obtener el nivel del bot de cada cuenta en el arreglo, con los resultados

obtenidos se desarrolló un semaforo de clasificación donde rojo para indicar noticias potencialmente falsas, verde para indicar noticias verdaderas y amarillo para noticias neutrales. Este sistema de clasificación se logró por medio de un arreglo, para que el programa identifique que tipo de color le corresponde a cada una de las cuentas, estas se basan en el análisis del autor y de cada una de las fuentes de las noticias, la implementación de este semáforo como se visualiza en la figura 6.

	Cuenta	Nivel de bot	Semáforo	S.A	C.F	C.A
0	Foro_TV	4.4	Rojo	0	Twitter for iPhone	Veronica Sanchez
1	El_Universal_Mx	3.5	Amarillo	0	Twitter Web App	Veronica Sanchez
2	lajornadaonline	4.0	Amarillo	0	Twitter for iPhone	Veronica Sanchez
3	OVIALCDMX	3.6	Amarillo	0	TweetDeck	Veronica Sanchez
4	infodemiaMex	4.0	Amarillo	0	Twitter for Android	Veronica Sanchez
5	SergioyLupita	3.6	Amarillo	0	Twitter Web App	Veronica Sanchez
6	ElOpinadorTV	1.3	Verde	0	Twitter for iPhone	Veronica Sanchez
7	SASMEX	3.6	Amarillo	1	Twitter Web App	Veronica Sanchez
8	MXvsCORRUPCION	3.4	Amarillo	0	Twitter Web App	Veronica Sanchez
9	FiscaliaCDMX	4.1	Rojo	0	Twitter for Android	Veronica Sanchez
10	AztecaDeportes	4.5	Rojo	0	Twitter for Android	Veronica Sanchez
11	warkentin	1.0	Verde	0	Twitter for Android	Veronica Sanchez

	Cuentas	Nivel de Bot	Verificada
0	@foro_TV	3.8/5	Si
1	@El_universal_Mx	2.8/5	Si
2	@lajornadaonline	3.7/5	Si
3	@OVIALCDMX	3.2/5	Si
4	@infodemiaMex	3/5	Si
5	@SergioyLupita	1.1/5	No
6	@E1OpinadorTV	0.6/5	No
7	@SASMEX	2.9/5	Si

Figura 6: Semáforo de noticias (fuente propia).

G. Clasificación de noticias falsas

Se creo un archivo en Excel que incluye noticias recopiladas, el conjunto de datos de artículos de noticias falsas y noticias verídicas, noticias de diferentes temas para evitar el ajuste excesivo de los clasificadores y proporcionar más datos de texto para mejorar el modelo y poder entrenar la red neuronal.

El conjunto de datos se compone de artículos provenientes de diversas fuentes y temas, para minimizar el sesgo hacia un tema específico, así como mejorar la robustez del modelo de clasificación. La estructura del conjunto de datos se organiza en cuatro columnas clave:

- Número: Un identificador único asignado a cada artículo, comenzado desde 0.
- Título: El encabezado de la noticia.
- Texto: Contenido de la noticia.
- Etiqueta. Es la clasificación binaria donde 0 indica que la noticia es falsa y 1 que es real.

Se combinaron diferentes noticias para obtener información certera en el semáforo, se agregó un porcentaje de confiabilidad a la fuente y el autor, este resultado lo obtenemos con tablas como se muestra la figura 7, donde se encuentra el nombre de cada autor, cada una de las fuentes, si los autores son reales o falsos y envase a eso se obtienen un porcentaje.

3. Resultados

Con la extracción de tweets se crearon variables para poder almacenar la información obtenida de cada metadato y se creó una gráfica partir de los datos obtenidos de los retweets. Donde muestra el comportamiento y comparación de una serie temporal de likes y retweets, destacando la disparidad en la interacción, mientras que los likes permanecen constantes en nivel bajo a lo largo del tiempo y los retweets muestran picos pronunciados, lo que nos indica una mayor participación en comparación de contenido de los likes. Como se puede observar en la figura 7.

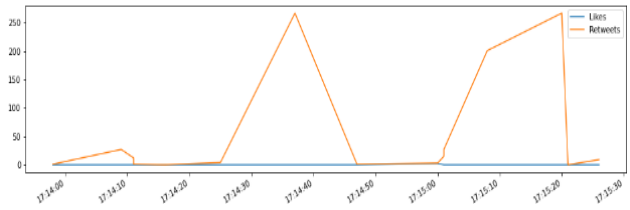

Figura 7: Gráfica de likes y retweets (fuente propia).

En la siguiente figura 8, se muestra las cuentas que tenemos en nuestro arreglo y su nivel de bot correspondiente.

[72]

	Cuenta	Nivel de bot
0	Foro_TV	4.6/5
1	El_Universal_Mx	3.6/5
2	lajornadaonline	3.7/5
3	OVIALCDMX	4.4/5
4	infodemiaMex	4.0/5
5	SergioyLupita	4.0/5
6	ElOpinadorTV	1.7/5
7	SASMEX	3.8/5
8	MXvsCORRUPCION	3.2/5
9	FiscaliaCDMX	4.2/5
10	AztecaDeportes	1.6/5
11	warkentin	1.4/5
12	MetroCDMX	4.0/5
13	BBCWorld	3.1/5
14	Profeco	3.1/5
15	hdemauleon	1.4/5
16	ricardomraphael	1.8/5
17	TESE_ISC	4.0/5
18	lopezobrador_	3.8/5

Figura 8: Clasificadores (fuente propia).

En la figura 9, la línea azul representa la precisión obtenida en la red. Dando los resultados que obtuvimos del entrenamiento de la red neuronal recurrente.

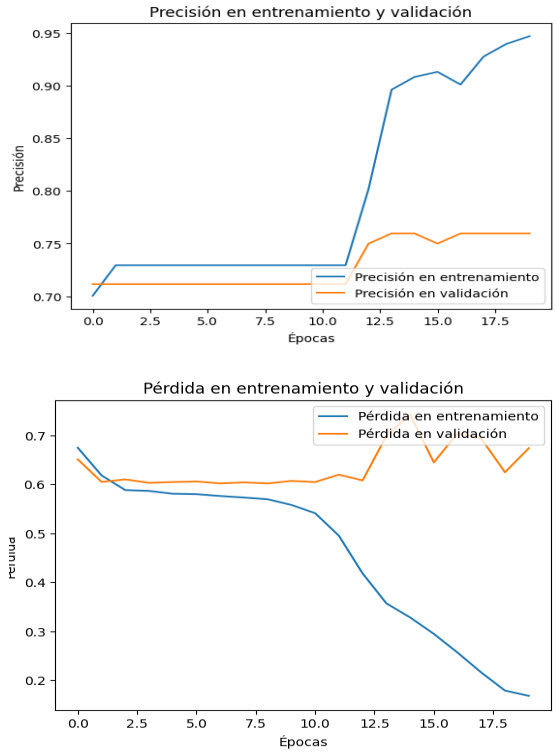

Figura 9: Grafica de presición y perdida

Se logó ejecutar cada componente y funcionalidad de manera correcta. Al final se realizaron las pruebas de calidad de la red neuronal para visualizar la precisión de esta en las cuentas bot y publicaciones falsas.

Los resultados de detección son satisfactorios ya que se puede detectar si es un bot con aproximadamente de 75.96% de efectividad, lo que demuestra su eficiencia en tareas de clasificación.

4. Discusión

La investigación tuvo como propósito identificar y describir aquellas experiencias dentro de la plataforma de Twitter para detectar las noticias falsas y cuentas bot, lo cual influye para alternar la opinión pública frente a un tema en específico que puede convertirse en tendencia.

Se genero un sistema en el cual se detectaron noticias falsas, para que el modelo pudiera entrenar la red neuronal y clasificar las cuentas en bot o humanos.

Los resultados resaltan la importancia del análisis de sentimientos, el aprendizaje profundo y redes neuronales recurrentes en la detección de bots y contenido falso, clasificando las cuentas en bot o humano. Comparado con métodos tradicionales, el modelo implementado ofrece una solución más precisa y adaptable, capaz de manejar grandes volúmenes de datos en tiempo real. Sin embargo, encontramos limitaciones en la variedad de los datos de entrenamiento, lo que sugiere la necesidad de ampliar el dataset.

Este método permitió el desarrollo de un modelo viable para detectar noticias falsas y cuentas bots, creando un conjunto de datos valiosos que pueden utilizarse para futuras investigaciones.

5. Conclusión

El estudio demuestra la viabilidad de utilizar redes neuronales y análisis de sentimientos para detectar bots y noticias falsas en las redes sociales especificamente en Twitter. La precisión obtenida respalda la aplicación del modelo en sistemas de monitoreo de medios digitales, contribuyendo a la mitigación de la desinformación en línea.

El proyecto demuestra el potencial de las redes neuronales para la detección de noticias falsas y cuentas automatizadas en redes sociales. Sin embargo, también destaca las limitaciones actuales, principalmente relacionadas con la dificultad de obtener y manejar datos explorando técnicas que puedan mejorar la presición y reducir el sesgo en este tipo de aplicaciones.

El algoritmo que se desarrolló se aplicó a un conjunto de datos recopilados durante un periodo prolongado, ya que aquello que realmente aportaba valor a la hora de identificar una cuenta como buena o mala, era la tendencia obtenida a partir de las puntuaciones finales de cada usuario a lo largo del tiempo.

Los resultados obtenidos sugieren que las redes neuronales recurrentes y en particular las LSTM, son herramientas poderosas para la identificación de patrones temporales en la actividad de las cuentas lo que es esencial para la detección precisa de bots.

6. Agradecimientos

Agradezco al Consejo Mexiquense de Ciencia y Tecnología COMECYT por su respaldo y financiamiento, que hicieron posible la realización de este trabajo.

7. Referencias

[1] [E. Van Der Walt and J. Eloff, “Using Machine Learning to Detect Fake Identities: Bots vs Humans,” *IEEE Access*, vol. 6, pp. 6540–6549, Jan. 2018, doi: 10.1109/ACCESS.2018.2796018.

[2] S. Sharma and V. Kumar Sharma, “Análisis de Delitos Cibernéticos en Redes Sociales,” *BSSS Journal of Computer*, May 2020, doi: 10.51767/jc1104.

[3] M. Parselis, “Función e innovación social: el caso Twitter Social function and social innovation: the Twitter case,” 2014. [Online]. Available: http://www.lanacion.com.ar/

[4] G. Udge, M. Mohite, S. Bendre, Y. Birnagal, and D. Wankhede, “Statistical Analysis for Twitter Spam Detection,” *Int J Sci Res Sci Eng Technol*, pp. 624–629, May 2019, doi: 10.32628/ijsrset1962170.

[5] Mr. G. Kumar D, “Spam Detection in Twitter,” *Int J Res Appl Sci Eng Technol*, vol. 8, no. 7, pp. 783–787, Jul. 2020, doi: 10.22214/ijraset.2020.30337.

[6] K. C. Yang, E. Ferrara, and F. Menczer, “Botometer 101: práctica de bots sociales para científicos sociales,” *J Comput Soc Sci*, vol. 5, no. 2, pp. 1511–1528, nov. 2022, doi: 10.1007/s42001-022-00177-5.

[7] M. Esther and R. Martínez, “Distinción de bots y humanos en Twitter con Inteligencia Artificial.”

[8] I. Grau, G. Nápoles, I. Bonet, and M. M. García, "Backpropagation through Time Algorithm for Training Recurrent Neural Networks using Variable Length Instances," vol. 17, no. 1, pp. 15–24, 2013.

[9] C. Henríquez, P. Ferran, L.-F. Hurtado, and J. Guzmán, "Análisis de sentimientos a nivel de aspecto usando ontologías y aprendizaje automático," España, pp. 49–56, 2017. Accessed: Feb. 06, 2023. [Online]. Available: http://www.redalyc.org/articulo.oa?id=515754427005

[10] J. Carlos and S. Sande, "Análisis de sentimientos en Twitter."

[11] M. T. Khan, M. Durrani, A. Ali, I. Inayat, S. Khalid, and K. H. Khan, "Análisis de sentimientos y el lenguaje natural complejo," *Complex Adaptive Systems Modeling*, vol. 4, no. 1. Springer, Dec. 01, 2016. doi: 10.1186/s40294-016-0016-9.

[12] T. Nasukawa and J. Yi, "Análisis de sentimiento: captura de la favorabilidad mediante el procesamiento del lenguaje natural," in *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*, Association for Computing Machinery, Inc, Oct. 2003, pp. 70–77. doi: 10.1145/945645.945658.

[13] J. P. Dickerson, V. Kagan, and V. S. Subrahmanian, "Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?" Aug. 2014. doi: 10.1109/ASONAM.2014.6921650.

[14] P. Burgos Gonzalo, "Análisis y detección de bots en Twitter," 2021. [Online]. Available: https://repositorio.uam.es/handle/10486/698001

[15] A. Minnich, N. Chavoshi, D. Koutra, and A. Mueen, "BotWalk: exploración adaptativa eficiente de las redes de bots de Twitter," in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017*, Association for Computing Machinery, Inc, Jul. 2017, pp. 467–474. doi: 10.1145/3110025.3110163.

[16] S. Cresci, R. di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fama en venta: detección eficiente de seguidores falsos en Twitter," Sep. 2015, doi: 10.1016/j.dss.2015.09.003.

[17] S. Cresci, A. Spognardi, M. Petrocchi, M. Tesconi, and R. di Pietro, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *26th International World Wide Web Conference 2017, WWW 2017 Companion*, International World Wide Web Conferences Steering Committee, 2017, pp. 963–972. doi: 10.1145/3041021.3055135.

[18] C. Yang, R. C. Harkreader, and G. Gu, "Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers," Jul. 2013. Accessed: Jan. 16, 2023. [Online]. Available: https://doi.org/10.1007/978-3-642-23644-0_17

[19] E. van der Walt and J. Eloff, "Uso del aprendizaje automático para detectar identidades falsas: bots vs humanos," *IEEE Access*, vol. 6, pp. 6540–6549, Jan. 2018, doi: 10.1109/ACCESS.2018.2796018.

[20] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans Dependable Secure Comput*, vol. 9, no. 6, pp. 811–824, 2012, doi: 10.1109/TDSC.2012.75.

[21] F. G. Efthimion1, S. Payne1, and N. Profers2, "Técnicas supervisadas de detección de bots de aprendizaje automático para identificar bots de redes sociales en Twitter," 2018. [Online]. Available: https://scholar.smu.edu/datasciencereview/vol1/iss2/5

[22] V. Subrahmanian, "El DESAFÍO DEL BOT DE TWITTER DE DARPA," vol. 49, no. 6, pp. 38–46, 2016, doi: 10.1109/MC.2016.183.

[23] S. Feng *et al.*, "TwiBot-22: Towards Graph-Based Twitter Bot Detection." [Online]. Available: https://twibot22.github.io/.

[24] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun ACM*, vol. 59, no. 7, pp. 96–104, Jul. 2016, doi: 10.1145/2818717.

[25] V. Subrahmanian *et al.*, "THE DARPA TWITTER BOT CHALLENGE," Pacific Social, Jun. 2016. doi: 10.1109/MC.2016.183.

[26] J. Pérez Guerrero Sevilla, J. de, T. por, and R. Pino Mejías, "REDES RECURRENTES."

[27] V. E. Garcia-Moreno, U. Nacional, S. Marcos, and C. R. Alvarez-Caicedo, "Análisis de sentimientos en la predicción de resultados de elecciones presidenciales," 2021, doi: 10.15381/risi. v14i1.21866.

[28] A. Géron, "Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems," 2017. [Online]. Available: http://oreilly.com/safari

[29] D. E. Maqueda Bojorquez, "DE REDES NEURONALES RECURRENTES A MODELOS DE LENGUAJE: LA EVOLUCIÓN DEL PLN EN LA GENERACIÓN DE TEXTOS." [Online]. Available: https://www.ties.unam.mx/

[30] L. Deng and D. Yu, "Deep learning Methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3–4. Now Publishers Inc, pp. 197–387, 2013. doi: 10.1561/2000000039.

[31] J. Antonio, P. Ortiz, D. Por, M. L. Forcada, and J. C. Rubio, "MODELOS PREDICTIVOS BASADOSEN REDES NEURONALES RECURRENTESDE TIEMPO DISCRETO," 2002.

[32] D. de Trabajo and C. Arana, "UNIVERSIDAD DEL CEMA Buenos Aires Argentina Serie," 2021. [Online]. Available: www.cema.edu.ar/publicaciones/doc_trabajo.html

[33] I. Bonet Cruz, S. Salazar Martinez, A. R. Abed, G. Abalo, and M. M. G. Lorenzo, "Redes neuronales recurrentes para el análisis de secuencias," *Revista cubana de ciencias informáticas*, cuba, pp. 48–57, 2007.

[34] Paul J. Werbos, "Backpropagation Througt Time: What It Does and How to Do It," *Proceedings of the IEEE*, vol. 78, pp. 1550–1560, Oct. 1990.

[35] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and Machine Learning forecasting methods: Concerns and ways forward," *PLoS One*, vol. 13, no. 3, Mar. 2018, doi: 10.1371/journal.pone.0194889.

[36] Sepp Hochreiter, "Long Short-Term Memory".

[37] D. Banks, M. Jordan, R. Kannan, C. Ré, R. J. Tibshirani, and L. Wasserman, "Springer Series in the Data Sciences Series Editors." [Online]. Available: http://www.springer.com/series/13852

[38] J. Antonio, P. Ortiz, D. Por, M. L. Forcada, and J. C. Rubio, "MODELOS PREDICTIVOS BASADOSEN REDES NEURONALES RECURRENTESDE TIEMPO DISCRETO," 2002.

[39] Juan Lulián Cea Morán, "Redes neuronales recurrentes para la generación automática de música," 2020.

[40] D. de Trabajo and C. Arana, "REDES NEURONALES RECURRENTES: ANÁLISIS DE LOSMODELOS ESPECIALIZADOS EN DATOS SECUENCIALES," 2021. [Online]. Available: www.cema.edu.ar/publicaciones/doc_trabajo.html

[41] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied Soft Computing Journal*, vol. 90, May 2020, doi: 10.1016/j.asoc.2020.106181.

[42] J. de Lucio, "Estimación adelantada del crecimiento regional mediante redes neuronales LSTM," *Investigaciones Regionales*, vol. 2021, no. 49, pp. 45–64, 2021, doi: 10.38191/iirr-jorr.21.007.

[43] A. Rafael Sabino Parmezan, V. M. A Souza, and G. E. A P A Batista, "Supplementary Material for Evaluation of Statistical and Machine Learning Models for Time Series Prediction: Identifying the State-of-the-art and the Best Conditions for the Use of Each Model," 2019.

[44] R. Montañés, R. Aznar, and R. del Hoyo, "Aplicación de un modelo híbrido de aprendizaje profundo para el Análisis de Sentimiento en Twitter," pp. 51–56, 2018.

[45] X. B. Olabe, "Redes neuronales artificiales y sus aplicaciones." Bilbao, p. 4.

[46] L. F. S. Navarro, "APROBACIÓN DEL PRESIDENTE DE PERÚ BASADO EN ANÁLISIS DESENTIMIENTOS EN TWITTER," *TECHNO Review. International Technology, Science and Society Review / Revista Internacional de Tecnología, Ciencia y Sociedad*, vol. 11, 2022, doi: 10.37467/revtechno.v11.4396.

[47] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in twitter," *Sensors (Switzerland)*, vol. 19, no. 21, Nov. 2019, doi: 10.3390/s19214654.

[48] X. Liu, "A big data approach to examining social bots on Twitter," *Journal of Services Marketing*, vol. 33, no. 4, pp. 369–379, Sep. 2019, doi: 10.1108/JSM-02-2018-0049.

[49] L. U. I. de la R. (UNIR), Luis de la F. V. P. U. I. de la R. Ernesto del Valle Martín, "Sentiment analysis methods for politics and hate".

[50] I. Latin and A. Transactions, "Sentiment Analysis of Tweets Related to SUS Before and During COVID-19 Pandemic," 2022.

[51] M. Angel Rosales Quiroga, D. Vilariño Ayala, D. Pinto, M. Tovar, and B. Beltrán, "Análisis de sentimientos basado en aspectos: un modelo para identificar la polaridad de críticas de usuario," 2016. [Online]. Available: http://www.lke.buap.mx/

[52] W. Mckinney, "Data Structures for Statistical Computing in Python," 2010.

[53] D. Cao, "Cloud Computing Based Plant Classifiers and Their Real-Life Research Applications," [Online]. Available: www.slayte.com