TECHNO REVIEW | 2022 | ISSN 2695-9933
International Technology Science and Society Review / Revista Internacional de Tecnología Ciencia y Sociedad
https://doi.org/10.37467/revtechno.v11.4396

APROBACIÓN DEL PRESIDENTE DE PERÚ BASADO EN ANÁLISIS DE SENTIMIENTOS EN TWITTER

Peruvian President's Approval Rating Based on Sentiment Analysis on Tweet Data

LUIS FERNANDO SOLIS NAVARRO
Universidad Nacional de San Cristóbal de Huamanga, Perú

KEYWORDS

Natural Language Processing
Sentiment Analysis
Artificial Neural Networks
Estimated Approval of politicians

ABSTRACT

The popular acceptance rate is a concept used to explain the increase in popular support for a political figure in a country over a given period. This figure is extracted through requested surveys that reach a certain limited sample of willing citizens and are expensive to conduct.
In this research we have implemented an automatic system for estimating the popular approval of the president of Peru using Twitter data. The method is simple, fast and highly sensitive, and can be quickly extended to other cases of opinion analysis.

PALABRAS CLAVE

Procesamiento del Lenguaje Natural
Análisis de Sentimiento
Redes Neuronales Artificiales
Estimación de Aprobación de Políticos

RESUMEN

La tasa de aceptación popular es un concepto que se utiliza para explicar el aumento del apoyo popular hacia un personaje político, de un país, en un periodo determinado. Esta cifra se extrae a través de encuestas solicitadas que llegan a cierta muestra limitada de ciudadanos dispuestos y además son caras de realizar.
En esta investigación se ha implementado un sistema automático para la estimación de la aprobación popular del presidente del Perú utilizando datos de Twitter. El método es simple, rápido y de alta sensibilidad, pudiendo extenderse rápidamente para otros casos de análisis de opinión.

Recibido: 03/ 04 / 2022
Aceptado: 18/ 06 / 2022

1. Introducción

En el escenario actual, las redes sociales desempeñan un papel cada vez más preponderante en la diseminación de información. Las redes sociales consiste de aplicaciones basadas en internet que permiten la comunicación entre familiares, amigos, colegas y demás círculos de interés (Statista, 2021). Entre las principales redes sociales tenemos a Twitter (Twitter, 2006), YouTube (Google, 2005)y Facebook (Meta Platforms, 2004). Según, Statista Research Department (2022), Diariamente la gente se interconecta a una red social para comunicarse y compartir información. Se estima que cada día se envían como 500 millones de tuits, se publican 350 millones de fotos a Facebook y se cargan más de 700 mil horas de vídeos en YouTube. Asimismo, el tiempo que la gente pasa navegando en internet se ha incrementado durante los últimos años. la gente navega en internet alrededor de 150 minutos por día, principalmente utilizando redes sociales según los reportado en cifra también refleja la proliferación de las redes sociales y el aumento exponencial del número de usuarios.

Según el estudio de Sharma & Ghose (2020), el 79% de los datos en internet son de naturaleza no estructurada, en su mayoría tienen el formato de texto como los tuits. Es decir, a diferencia de los datos estructurados, estas no posee una estructura fija, por ello no se puede almacenar en un sistemas de gestión de base de datos relacional (DBMS por sus siglas en ingles) (Mongodb, 2021). Por otra parte, dada la gran cantidad de datos, su manejo, almacenamiento y diseminación ha dado origen a redes sociales especializadas, por ejemplo, mensajes de texto (Twitter), registros de audio Clubhouse (Paul Davison, 2020) y vídeos (YouTube). Con el objetivo de analizar estos grandes volúmenes de datos, durante la última década, la comunidad científica viene realizando esfuerzos a fin de desarrollar métodos automáticos de extracción de información. En este sentido, el campo de la minería de datos viene realizando trabajos para reconocer patrones en mensajes de texto, dando origen a la disciplina de análisis de sentimientos, cuyo objetivo es extraer información sobre estados psicológicos y mentales del sujeto que escribió un mensaje, tales como su estado emocional, intereses, nivel de instrucción, condición mental, entre otros, siendo la principal fuente de análisis la mensajería publicada en Twitter.

Este trabajo tiene como objetivo analizar los tuits con contenido político redactado por ciudadanos peruanos en Twitter con la finalidad de descubrir su valoración y aceptación respecto al presidente de la república. Por lo cual, desarrollamos un clasificador basado en redes neuronales artificiales, el cual fue entrenado con 3400 tuits recolectados durante agosto y diciembre del 2021. Utilizamos una la red neuronal profunda pre-aprendida Word2Vect *Spanish Billion Words Corpus and Embeddings* (SBW) para construir nuestro vector de características (*Embedding*) en Español (Cardellino, 2016) y *GloVe* (Pennington et al., 2014) para construir nuestro vector de característica para la data traducida al Inglés. Finalmente, para evaluar el modelo, hicimos uso de 200 tuits totalmente desconocidos, por cada mes, obteniendo así una exactitud promedio del 90.61% utilizando una red neuronal convolucional.

2. Fundamentación

2.1. Análisis de sentimiento

El análisis de sentimientos o minería de opiniones es un método muy peculiar en el campo del procesamiento de lenguaje natural (NLP por sus siglas en Inglés) para enseñar a una máquina a extraer emociones de un texto dado (Khurana Batra et al., 2020). Esta técnica tiene como objetivo obtener información de textos que pueden almacenarse en formatos estructurados como hojas de cálculo y documentos HTML o en formatos no estructurados como texto plano (Silva et al., 2022). En función del tópico a investigar, estos textos pueden ser clasificadas en positivas, negativas o neutrales (Liu et al., 2020; Poria et al., 2018) . Existen muchas aplicaciones de esta técnica en el ámbito social, entre las cuales se encuentran la identificación de acoso cibernético y violencia contra la mujer.

2.2 Popularidad de políticos

Es práctica común que los gobiernos y los políticos deseen conocer la opinión de los ciudadanos respecto a sus expresiones, actividades y ejecución de políticas. La opinión de las personas se cuantifica y se presenta como el grado de acuerdo o desacuerdo, o aprobación o desaprobación.

Deseando que estas estimaciones sean características reales de la opinión poblacional, el procedimiento de análisis de la popularidad se realiza utilizando una encuesta sobre una muestra de la población siguiendo procedimientos estadísticos (IPSOS, 2020).

En el caso peruano, se estila presentar mensualmente la tasa de aprobación del presidente de la república, que mide la percepción del ciudadano con relación al trabajo y políticas del gobierno en su conjunto. Este respaldo o rechazo de la población también es percibido como un indicador de la aptitud del gobierno de turno y la estabilidad del estado peruano. Este parámetro es utilizado por los inversionistas extranjeros y emprendedores locales para la toma de decisiones.

Diferentes entidades privadas realizan el análisis de opinión, presentando sus resultados cada mes y realizando proyecciones sobre la curva de tendencia (DATUM, 2021; Medianero Burga, 2014). Tendencias bajistas indican la

debilidad del gobierno y fracaso en la aplicación de políticas, acarreando algunas veces fuga de capitales, recesión, inestabilidad económica y des-inversión extranjera (Rodríguez, C. G. and Tule, 2020).

2.3 Twitter

Twitter es un servicio de mensajería en el que los usuarios publican e interactúan con mensajes conocidos como tuits (Shaghaghi et al., 2021). Los tuits consisten principalmente de opiniones sobre temas de actualidad y puede ser utilizado como una fuente de datos para la de toma de decisiones (Prastyo et al., 2020).

Actualmente, Twitter es la red social más utilizada por los políticos debido a que le permite transmitir mensajes cortos, masivamente, a miles o millones de seguidores (Kydros & Magoulios, 2019).

3. Revisión Bibliográfica

Aunque el comienzo de la inteligencia artificial se remonta a los anteriores ciclos, el procesamiento del lenguaje natural es un tema que surge por primera vez con Alan Turing, en 1950, en su artículo conocido como la prueba de Turing. Desde aquel entonces, la comunidad científica ha intensificado su estudio por el gran aporte que ofrece para solucionar problemas de distinta índole, especialmente aquellas que involucran al lenguaje escrito. Debido al continuo aumento de la computación y el gran desarrollo de algoritmos de aprendizaje automático, se ha producido un gran desarrollo en el campo del procesamiento del lenguaje natural.

Según Kumar et al. (2013), existen una serie de limitantes al trabajar con datos originados en Twitter, uno de los principales problemas es que estos poseen una estructura semántica y gramatical informal. Desde el campo de estudio de Análisis de Sentimientos ha habido un gran interés en estudiar y resolver este problema. Entre las diferentes investigaciones, las más relevantes se resumen a continuación:

En el estudio de Al Shammari (2018), se utilizaron 2 modelos de aprendizaje supervisado (Naive Bayes y voto simple) para clasificar tuits. Cada tuit, después de ser procesado, se descompuso en palabras(token) y se clasificó usando un diccionario con 2014 palabras positivas y 4783 palabras negativas. Para determinar la polaridad de los tuits se utilizó la siguiente ecuación en base al resultado del modelo.

Puntaje = Número de palabras positivas- Número de palabras negativas (1)

Si Puntaje > 0, la oración es positiva.

Si Puntaje < 0, la oración es negativa

Si Puntaje = 0, la oración es neutral

Para la evaluación de cada modelo se realizó un experimento con 1500 tuits etiquetados manualmente por el autor, obteniendo así un 81% de exactitud con el clasificador Naive Bayes y un 74% utilizando el clasificador de voto simple. Una de las limitantes del trabajo es que el resultado del modelo es totalmente dependiente a la cantidad y tipo de palabras que tenga el diccionario.

Con el objetivo de encontrar una respuesta a la siguiente pregunta ¿Se puede utilizar Twitter para observar el efecto de manifestación popular hacia un personaje político?, Shaghaghi et al. (2021), realizó un análisis de sentimientos sobre el nivel de afectividad población al presidente Donald Trump. Se recolectaron tuits empleando la librería Tweepy y la API de Twitter utilizando la palabra clave "Donald Trump". La data recolectada fue etiquetada manualmente por los autores. Utilizaron 2 tipos de arquitecturas: Naive Bayes y LSTM. Como resultado, se obtuvo una precisión del 63% utilizando Naive Bayes y 69% utilizando LSTM.

En el contexto latinoamericano, durante la pandemia del COVID-19. Silva et al. (2022), utilizó el análisis de sentimiento para identificar el sentir de la población brasileña sobre el Sistema Único de Salud (SUS) a través del contenido en Twitter. Además, hicieron comparaciones de sentimientos presentados antes, durante y después de la pandemia. Para ello recolectaron 27500 tuits utilizando la API de Twitter con palabras clave "Saúde" y "SUS" en el periodo diciembre 2019 - octubre 2020. Para clasificar los sentimientos de los tuits utilizaron NRC Sentiment (Mohammad et al., 2012) de la biblioteca Syuzhet de R (Ross Ihaka, 1993). El modelo es capaz de clasificar las oraciones en "Positivo", "Negativo" y "Neutral". Los resultados del modelo se presentaron a través de nubes de palabras por cada uno de los siguientes periodos: diciembre 2019 - febrero 2020, marzo 2020 - abril 2020, mayo 2020 - agosto 2020 y setiembre 2020.

Por otra parte, Gandhi et al. (2021), utilizó el análisis de sentimiento para evaluar la opinión de las personas expresadas en las redes sociales con relación al ámbito político de Indonesia. A diferencia de los trabajos anteriores, se utilizó 2 fuentes de datos distintos (Facebook y Twitter) para extraer un total de 4400 tuits. En primera instancia, se utilizó la arquitectura Word2Vect para construir el vector de características y como clasificador utilizó LSTM. El modelo es capaz de clasificar las oraciones en "Positivo", "Negativo" o "Neutro". Se obtuvo un 85% de exactitud del modelo como resultado.

El autor Ansari et al. (2020), lleva a cabo el estudio de los sentimientos de los usuarios de Twitter hacia los principales partidos políticos nacionales que participan en el proceso electoral de la India en el 2019. Se elaboró el modelo de clasificación basado en sentimientos para inferir los resultados de las elecciones. El objetivo fue utilizar arquitecturas de aprendizaje profundo y compararlo con los modelos clásicos de aprendizaje automático. En primera instancia, se recolectaron, 3896 tuits de la plataforma Twitter relacionados a los 2 partidos más

populares en aquel entonces. Como modelos de clasificación, se utilizaron LSTM, Soporte de Maquina de Vectores (SVM por sus siglas en Ingles.), Árbol de decisiones, Regresión Logística y Bosques Aleatorios (Random Forest en Ingles.), obteniendo así una exactitud máxima del 77% con el modelo Bosques Aleatorios.

El autor Khurana Batra et al. (2020), analiza los datos recogidos de twitter para predecir el resultado electoral en la India realizando análisis de sentimiento. Para etiquetar los datos utilizaron la API Valence Aware Dictionary and Estiment Reasoner (VADER), emplearon BOW (Liu et al., 2020) y TF-IDF como técnicas de extracción de características. Utilizaron regresión logística, árbol de decisión, XGBoost, Naive-Bayes y LinearSVC como modelos de aprendizaje automático. El objetivo principal de autor fue llevar un problema de aprendizaje no supervisado a aprendizaje supervisado y aplicar análisis de sentimientos. Como resultado del trabajo, se obtuvo una exactitud del 86% utilizando el algoritmo de Árbol de Decisiones.

Balli et al. (2022), realiza un estudio de sentimiento de varios algoritmos de aprendizaje automático en conjuntos de datos recolectados de Twitter en idioma Turco. Se utilizaron 2 conjuntos de datos diferentes, un conjunto de datos público y otro conjunto de datos creados y etiquetados manualmente por los mismos autores. En la investigación se resalta el uso de los datos originales para el contexto a analizar si queremos obtener mayor exactitud en los resultados. Utilizaron Regresión Logística, Soporte de Maquina de Vectores (SVM por sus siglas en Ingles), bosques aleatorios y el descenso de grandiente estocástico como algoritmos de aprendizaje automático. Además, agregaron LSTM, como arquitectura de aprendizaje profundo, para comparar resultados. Finalmente, el mejor algoritmo, con 87.47% de exactitud, resultó ser el SVM con los datos originales etiquetados por los autores. Por lo tanto, concluyen que los modelos y resultados creados en el artículo muestran que los algoritmos de aprendizaje automático en el idioma turco y el analisis de sentimientos son prometedores y pueden mejorar en el futuro.

Cui et al. (2018), realizó un estudio de cómo utilizar texto con emojis para analizar los sentimientos reflejados por los usuarios en Twitter hacia un tema en particular. El trabajo se centra en usar la arquitectura de una red neuronal convolucional y vectores de incrustación Word2Vect. Como resultado se menciona que al evaluar los tuits que siguen la distribución de sentimiento uniforme, el modelo convolucional funciona notablemente bien en comparación al SVM. Sin embargo, para una clase neutral ambos modelos no son tan buenos como en las otras clases. Además se añade que, la relación entre el sentimiento expresado en el texto y el emoji integrado en el mismo, no siempre tiene cierta consistencia o coherencia.

Maharani y Effendy(2022), realiza un estudio del texto en función a características semánticas que incluyen emoción, sentimiento y el perfil de Twitter disponible públicamente. Predijeron la personalidad basada en el modelo de personalidad Big Five (OCEAN), un modelo enfocado en determinar la personalidad de usuarios en redes sociales. Utilizaron varias técnicas de aprendizaje automático; Naive Bayes (NB), K-vecinos más cercanos y maquina de vectores de soporte (SVM).

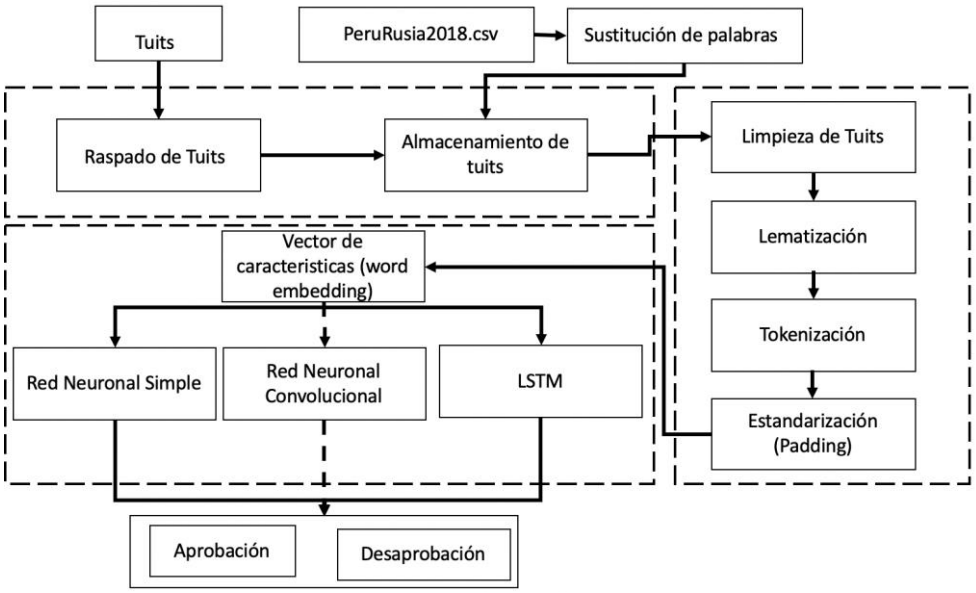
Poornima et al. (2022), utiliza una red neuronal convolucional para analizar la opinión expresada por los usuarios en Twitter. La data utilizada para el entrenamiento y prueba fue recolectada de la plataforma Kaggle, cuya cantidad supera los 1.5 millones de Tweets. El resultado del modelo convolucional alcanzó los 78.64% de exactitud. Como conclusión menciona que una palabra de evaluación que se considera buena en una circunstancia puede considerarse negativa en otra circunstancia.

A pesar de los esfuerzos para abordar el problema de trabajar con datos no estructurados y en español en las redes sociales, los trabajos dirigidos a detectar comportamientos en lengua española son aún escaso. Por ejemplo, el autor Cuzcano y Ayma (2020), comparó 4 modelos de aprendizaje automático supervisado para detectar el ciberacoso en publicaciones de Twitter escritas en el lenguaje español peruano. Se recolectaron, 10096 tuits utilizando la API de Twitter en español de comentarios peruanos en los periodos agosto 2019 - enero 2020 de usuarios en un rango de 14 y 16 años, estos datos se etiquetaron manualmente con la ayuda de humanos. Se particionaron los datos en una proporción 70% y 30% y se hizo uso del algoritmo TF-IDF para construir el vector de características. Como resultado, se obtuvo un 83% de exactitud utilizando el clasificador Naive Bayes.

4. Metodología

La arquitectura del sistema propuesto se muestra en la Figura 1, el diagrama resume todo el recorrido de la data desde su recolección hasta la ingesta en el modelo de aprendizaje profundo. En primera instancia, se recolectan los mensajes publicados por los usuarios en Twitter relacionados al presidente del Perú. Además, se procede a adaptar al contexto los datos de “PeruRusia2018.csv” utilizando la sustitución de palabras como técnica de aumentación de datos (CHAMBI, 2019). Luego se realiza un pre procesamiento eliminando los datos no deseados y basura, este proceso también se llama limpieza. Después de realizar la limpieza, se analizan utilizando 3 arquitecturas de aprendizaje profundo: red neuronal simple, red neuronal convolucional y la red LSTM. Finalmente, se predice el sentimiento detrás de cada tuit, cuyo valor puede ser “ aprobación“ o “ desaprobación“ .

Figura 1. Arquitectura del sistema propuesto



Fuente: Elaboración propia, 2022.

4.1. Recolección de datos

Hemos construido y hecho público un conjunto de datos que consiste en una colección de 3400 tuits en español a partir de comentarios e interacciones entre usuarios peruanos con relación al presidente de la república en Twitter. Para ello, utilizamos las herramientas de Python; BeautifulSoup y Selenium (Leonard Richardson, 2020). El proceso de recolección se realizó durante el periodo agosto 2021 y diciembre 2021. Además, utilizamos la colección "PeruRusia2018.csv", que consta de 2800 tuits relacionadas a la opinión popular de la selección peruana de futbol en el mundial Rusia 2018, para realizar un segundo experimento.

Tabla 1. Distribución de datos recolectados de Twitter

Nro	Dataset	Fecha de recolección	Cantidad de tuits	Propósito
1	Propio	Desde agosto 2021 hasta diciembre 2021	3400	Modelado
2	PerúRusia2018.csv	Desde octubre 2015 hasta noviembre 2017	2800	
3	Propio	Agosto 2021	100	Predicción
		Septiembre 2021	100	
		Octubre 2021	100	
		Noviembre 2021	100	
		Diciembre 2021	100	
			100	

Fuente: Elaboración propia,

4.2. Pre-procesamiento de datos

El pre-procesamiento de datos es un paso esencial para la construcción de cualquier modelo matemático basado en aprendizaje automático. Según Maharani y Effendy (2022), es la etapa más importante del proceso de análisis de sentimientos y predicción de la personalidad, convertir los datos en bruto en un formato analizable. El resultado del modelo depende qué también se haya realizado este procesado (Harshith, 2019).

4.2.1 Limpieza de tuits

Inicialmente se convirtieron los tuits en minúscula utilizando la función lower() de Python. Se eliminaron los signos de puntuación, caracteres especiales como "#?!()$\&", símbolos numéricos (0-9). Se eliminaron los *stopwords* (palabras vacías). Según (Ferilli et al., 2014), estos por lo general se refieren a palabras más comunes en el idioma y que no poseen un significado individual (no son informativos), para lo cual, utilizamos la librería NLTK (Bird, Steven and Klein, Ewan and Loper, 2009) con su biblioteca de *stopwords* en español.

4.2.2 Lematización

Por razones gramaticales, los documentos utilizan distintas formas de expresar una palabra con el mismo significado. El objetivo de la lematización es reducir las formas flexivas y, a veces, las formas derivadas de una palabra a una forma base común. Por ejemplo: Las palabras sangre, sangrar y sangriento pueden representarse como sangre únicamente (Cambridge University Press, 2008).

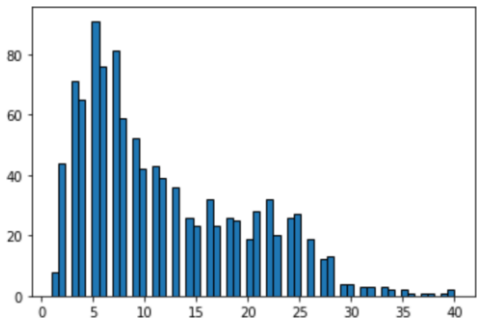
4.2.3 Tokenización

Rai & Borah (2021), define a la tokenización como el mecanismo de dividir o fragmentar la oración en su morfema más pequeño posible llamado palabra o token. Por ejemplo, " Lima es capital de Perú" puede representarse como [" Lima" , " es" , " capital" , " de" , " Perú"].

4.2.4 Estandarización de datos

Una vez realizado el proceso de tokenización, obtuvimos una longitud variable de palabras para cada tuit, es por ello que estandarizamos la longitud de cada tuit teniendo en cuenta la longitud promedio de la misma. En la Figura 2, se muestra la distribución de la cantidad de palabras por tuit, siendo la media 12.4 y la desviación estándar de 9.08. Por tal motivo, se utilizó 12 palabras como umbral de estandarización.

Figura 2. Distribución de palabras en cada tuit



Fuente: Elaboración propia, 2022.

4.3 Modelo de análisis de sentimientos

4.3.1 Vector de características (Word Embedding)

Cada tuit es codificado como un vector 300-dimensional. Para esta codificación utilizamos una red neuronal pre-entrenada Spanish Billion Words Corpus and Embeddings (SBW), con un millón de palabras del Español en su diccionario.

Para realizar la clasificación de los sentimientos de los tuits, optamos por utilizar 3 arquitecturas de aprendizaje profundo: red neuronal simple, red neuronal convolucional y la red LSTM. Para la clasificación de polaridad, se decidió considerar las clases "Positivo" y "Negativo", siendo la clase Positivo asociada a un tuit que representa aprobación hacia el presidente y la clase negativa asociada a una desaprobación del mismo.

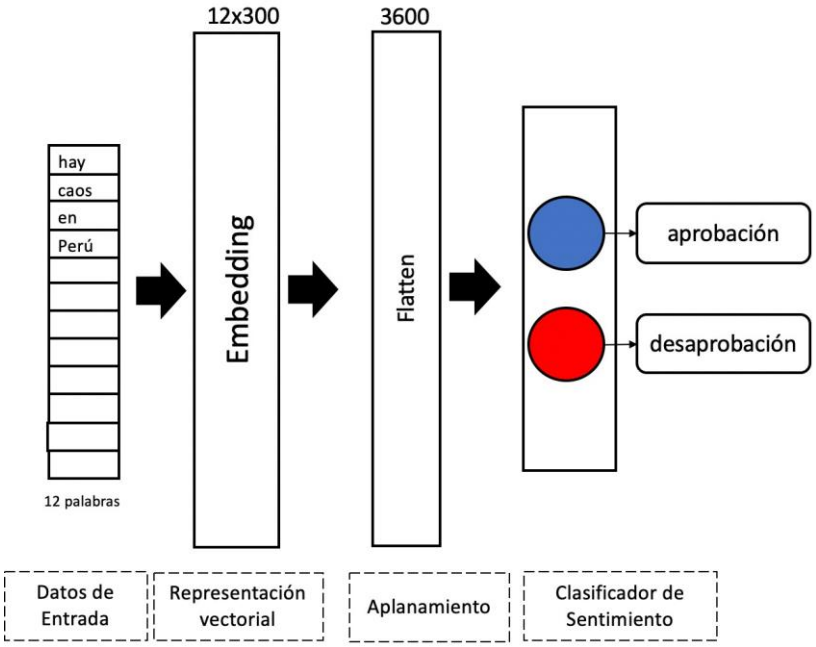
4.3.2 Red neuronal simple

Una red neuronal simple generalmente se compone de capas de entrada, capas ocultas densas y capas de salida, esta ultima puede ser sigmoidea con una neurona para clasificación binaria o una capa de salida softmax con una neurona por clase para clasificación multiclase. Generalmente, antes de esas capas existe una capa de incrustación y una capa de aplanamiento, es decir, el primero utiliza incrustaciones de palabras para transformar un conjunto

de oraciones en una matriz compuesta por vectores de características y el segundo "aplana" los arreglos 2D generados por la capa de incrustación (word embedding) en arreglos 1D que ingresan directamente a una capa densa de clasificación.

Este primer modelo, representado en la **Figura 3**, está compuesto por una capa de entrada y una capa de salida. La primera recibe un vector "plano" de dimensión 3600@1 y se conecta con una única capa de salida sigmoidea que realiza la clasificación.

Figura 3. Arquitectura de la red neuronal simple



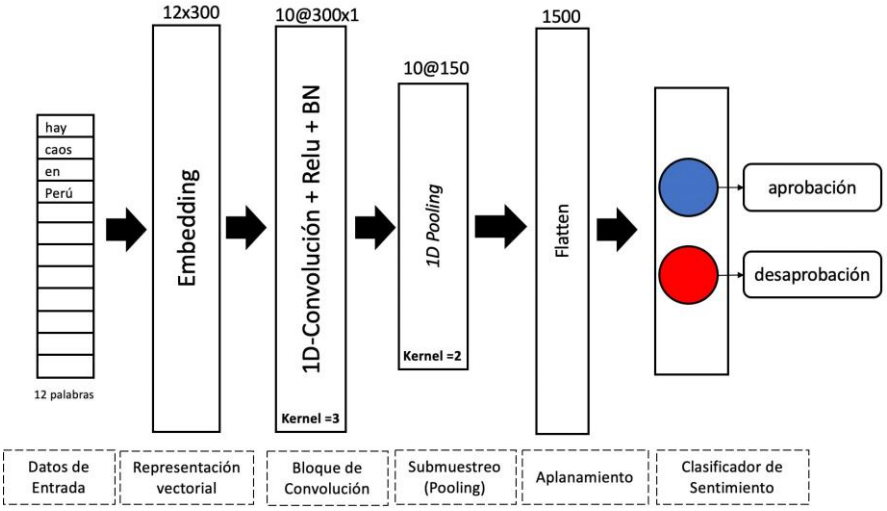
Fuente: Elaboración propia, 2022.

4.3.3. Red neuronal convolucional (Word Embedding)

La red neuronal convolucional ha tenido resultados innovadores e interesantes durante la última década en una variedad de campos relacionados con el reconocimiento de patrones (Albawi et al., 2017). Hoy en día, es ampliamente utilizada en campos como reconocimiento de voz, minería de texto, NLP y análisis de sentimientos.

La **Figura 4** exhibe una arquitectura convolucional para realizar análisis de sentimientos. Inicialmente se tiene una matriz de dimensión 12x300 proveniente de la capa pre-entrenada Word2vect, el resultado ingresa a una capa de convolución con 10 filtros de 300x1 y después a una capa de sub-muestreo (pooling) para reducir la dimensionalidad a 10@150. Finalmente, la salida es enviada al clasificador que consiste de una neurona artificial para predecir la probabilidad de pertenecer el mensaje a la clase aprobación o desaprobación.

Figura 4. Arquitectura de la red neuronal convolucional



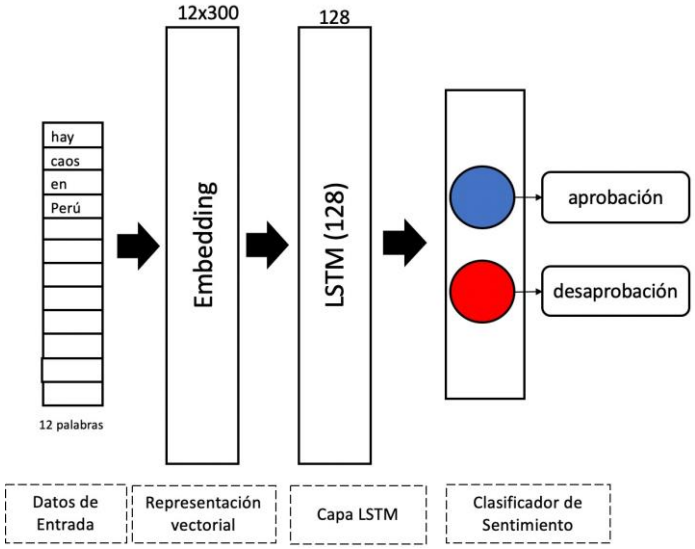
Fuente: Elaboración propia, 2022.

4.3.5. Red LSTM

En las redes neuronales recurrentes, a diferencia de las redes neuronales simples, las conexiones entre neuronas forman un gráfico dirigido que incluye algunas conexiones auto-reflexivas. Además, este tipo de redes pueden modelar dependencias temporales y son especialmente adecuados para la predicción de datos de secuencia como el texto (Parmezan et al., 2019).

Nuestro tercer modelo, inicialmente, integra una matriz de dimensión 12x300 proveniente de la capa pre-entrenada Word2vect, este resultado ingresa a la red LSTM con 128 estados. Finalmente, la salida es enviada al clasificador que consiste de una neurona artificial para predecir la probabilidad de pertenecer el mensaje a la clase aprobación o desaprobación.

Figura 5. Arquitectura de la red LSTM



Fuente: Elaboración propia, 2022.

5. Objetivos

Este trabajo tiene por objetivo estimar la tasa de aprobación del presidente del Perú a partir de datos en Twitter. Para la recuperación de los datos (tuits), utilizamos la técnica de raspado web (web scraping); recolectamos 3400 tuits en español relacionados con el acontecimiento actual asociado al presidente de agosto a diciembre del 2021. Basado en la técnica de análisis de sentimientos, los tuits fueron asociados manualmente a un sentimiento

positivo (aprobación) o negativo (desaprobación); luego construimos un clasificador binario capaz de categorizar los mensajes, utilizando una red neuronal convolucional (RNC). Nuestra RNC consiste de una capa de convolución, una capa de sub-muestreo y dos neuronas; para ser procesado un tuit por la RNC, primeramente, se descompone el mensaje en palabras; luego se codifica cada palabra como un vector 300-dimensional, utilizando una red neuronal (embedding word2vec) pre-entrenada en un millón de palabras de la lengua española.

6. Resultados

Este trabajo analiza las técnicas de extracción de características Word2Vect, combinados con 3 modelos de aprendizaje profundo; red neuronal simple, red neuronal convolucional y la red LSMT. Para una evaluación confiable de nuestra propuesta, los datos fueron separados en dos conjuntos disjuntos llamados conjunto de entrenamiento y conjunto de prueba (test), con la proporción 70:30, habiendo obtenido una exactitud media máxima de 90.61% utilizando la red neuronal convolucional. Los resultados se pueden observar en la Tabla 1. A consecuencia del presente trabajo, Introdujimos una nueva forma de calcular el nivel de aceptación popular al presidente de la república y aplicamos modelos de aprendizaje profundo con técnicas de procesamiento de lenguaje natural como el Análisis de Sentimientos. Para validar mejor el resultado de nuestros modelos, entrenamos el mismo conjunto de datos, pero traducidos a las ingles y utilizando el diccionario de palabras en ingles GLOVE para construir nuestros vectores de características. También utilizamos la librería Googletrans (Han, 2022) de Python para traducir nuestros tuits al inglés. Además, utilizamos la base de datos "PeruRusia2018. csv" modificado para realizar los experimentos. Para cada conjunto de datos, el entrenamiento se realizó durante 7 épocas, con un tamaño de lote igual a 32 y el algoritmo de optimización Adam (Kingma & Ba, 2015). El resumen del experimento se puede observar en la Tabla 1.

Tabla 1. Tabla de resultados

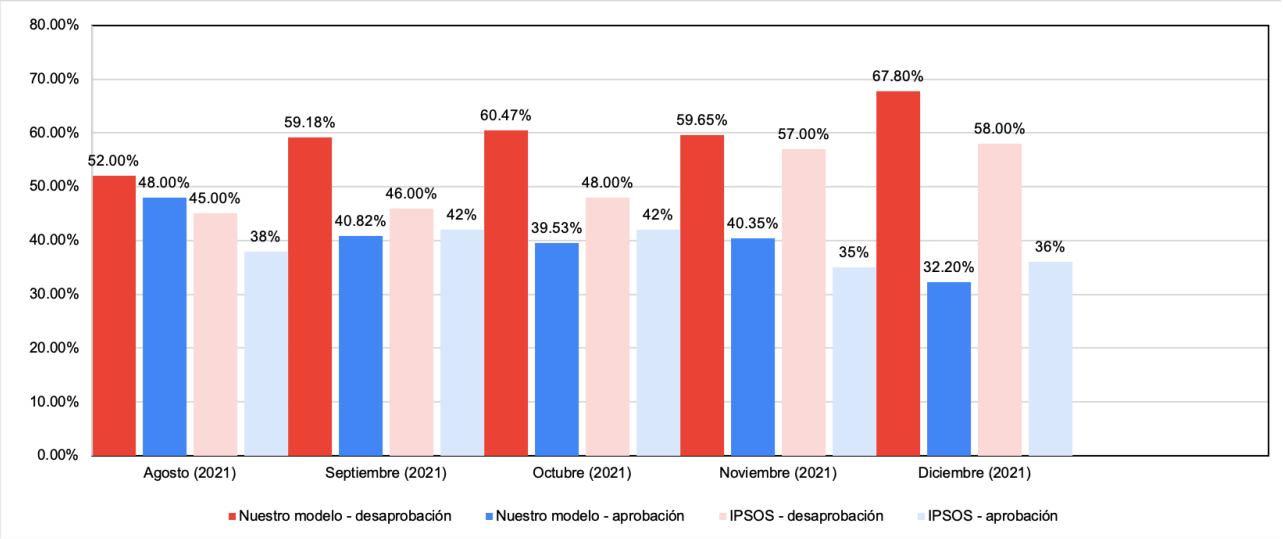
	Español				Inglés			
Tabla de Exactitud (%)	Entrenamiento con datos propios	Prueba con Datos propios desconocidos	Entrenamiento con datos propios más "PeruRusia2028. csv"	Prueba con Datos propios desconocidos	Entrenamiento con datos propios	Prueba con Datos propios desconocidos	Entrenamiento con datos propios más "PeruRusia2018. csv"	Prueba con Datos propios desconocidos
Red Neuronal Simple	86.19%	87.5%	82.98%	70.00%	81.65%	86.50%	78.84%	59.5%
Red Neuronal Convolucional	89.2%	90.61%	84.90%	72.5%	88.99%	89.50%	80.80%	68.00%
LTSM	80.98%	84.50%	81.13%	73.5%	88.68%	86.00%	79.07%	65.5%

Fuente: Elaboración propia, 2022.

Las encuestas, aunque son famosas por tener cierto margen de error con respecto al presidente Pedro Castillo, siguen siendo la mejor métrica para determinar el sentimiento público en torno al desempeño de un presidente. Sin embargo, en lugar de confiar únicamente en un encuestador político, se puede tomar como alternativa utilizar un modelo matemático que me permita determinar la aceptación popular del presidente de una forma simple, práctica y confiable.

En base al resultado de los modelos mencionados, utilizamos la red neuronal convolucional para determinar el porcentaje de aprobación y desaprobación de los comentarios en Twitter sobre el presidente Pedro Castillo Terrones, los cuales fueron organizados por meses. Observando los resultados plasmados en la Figura 6, los sentimientos negativos (desaprobación) fueron mayoritarios y mostraron una tendencia creciente durante el periodo agosto 2021 y diciembre 2021. Además, Comparando nuestras predicciones respecto a las ofrecidas por las encuestadoras nacionales, se puede observar que estas están fuertemente correlacionadas.

Figura 6. Gráfico de barras para la aprobación/desaprobación del presidente del Perú.



Fuente: Elaboración propia

La **Figura 7**, muestra una nube de palabras realizada con Python obtenida a partir de la opinión popular sobre el actual presidente del Perú. La nube proporciona una primera visión del conjunto de datos que nos ayuda a construir nuestro modelo. Además, nos permite visionar en primera instancia la inclinación de la desaprobación popular hacia el presidente del Perú en el periodo agosto 2021 y diciembre 2021.

¡Figura 7. Arquitectura del sistema propuesto

Fuente: Elaboración propia

7.1 Discusión de Resultados

Los resultados no se ven mal reflejados al utilizar un modelo entrenado con datos de distinto contexto y aplicado a otra situación. Uno de los retos de los modelos de análisis de sentimiento, o el procesamiento de lenguaje natural, es generalizar las predicciones para contextos distintos, sin embargo, esto puede resultar un trabajo complejo ya que se necesitaría muestras de datos de todos los contextos a involucrar. Además, según Monhaler y Miranda (2017), la forma de cómo se expresa y percibe el mensaje en el lenguaje español resulta ser relativamente variada

a causa de distintos aspectos, entre las principales son; variedades diacrónicas (históricas), variedades diatópicas (geográficas); variedades diafásicas (funcionales) y variedades diastráticas (socioculturales).

8. Conclusiones

En el presente trabajo, hemos implementado 3 tipos de modelos de aprendizaje profundo; una red neuronal simple, una red neuronal convolucional y una red LSTM. Nuestro mejor modelo (CNN) obtuvo un 91% de sensibilidad y 90.2% de especificidad. Con este resultado, se ha implementado un sistema automático para estimar la aprobación popular del presidente del Perú. Comparando los resultados obtenidos con los informes reportados por encuestadoras nacionales, nuestras estimaciones están fuertemente correlacionadas. Por lo tanto, nuestro método ofrece una alternativa rápida, eficiente y de bajo costo para monitorear y obtener la tasa de aprobación del presidente peruano.

9. Agradecimientos

Quiero agradecer a José Elías Yauri Vidalón por su orientación durante todo el desarrollo de este trabajo.

Referencias

Al Shammari, A. S. (2018). Real-time Twitter Sentiment Analysis using 3-way classifier. 21st Saudi Computer Society National Computer Conference, NCC 2018, 1–3. https://doi.org/10.1109/NCG.2018.8593205

Albawi, S., Mohammed, T. A. y Al-Zawi, S. (2017). Understanding of a convolutional neural network. International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

Ansari, M. Z., Aziz, M. B., Siddiqui, M. O., Mehra, H., y Singh, K. P. (2020). Analysis of Political Sentiment Orientations on Twitter. Procedia Computer Science, 167, 1821–1828. https://doi.org/10.1016/J.PROCS.2020.03.201

Balli, C., Guzel, M. S., Bostanci, E., & Mishra, A. (2022). Sentimental Analysis of Twitter Users from Turkish Content with Natural Language Processing. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/2455160

Bird, S., Klein, E. y Loper, E. (2019, 4 de septiembre). Natural language processing with Python: analyzing text with the natural language toolki. https://www.nltk.org/book/.

Cambridge University Press. (2008). Stemming and lemmatization.

Cardellino, C. (2016). Spanish Billion Words Corpus and Embeddings. https://crscardellino.ar/SBWCE/

Chambi, m. F. (2019). Análisis de opinión del microblogging twitter por la clasificación al mundial de fútbol rusia - 2018 de la selección peruana de fútbol, usando el framework spark.[tesis de pregrado, universidad nacional del antiplano]. http://repositorio.unap.edu.pe/handle/UNAP/13506

Cui, H., Lin, Y., y Utsuro, T. (2018). Sentiment Analysis of Tweets by CNN utilizing Tweets with Emoji as Training Data. Wisdom, August, 1–8. https://sentic.net/wisdom2018cui.pdf

Cuzcano, X. M., & Ayma, V. H. (2020). A comparison of classification models to detect cyberbullying in the Peruvian Spanish language on twitter. International Journal of Advanced Computer Science and Applications, 11(10), 132–138. https://doi.org/10.14569/IJACSA.2020.0111018

Canal N. (2021, October 21). Datum: Aprobación del presidente Pedro Castillo llega al 40 % | Canal N. 21 de Octubre Del 2021. https://canaln.pe/actualidad/pedro-castillo-aprobacion-mandatario-llega-al-40-segun-datum-n440163

Ferilli, S., Esposito, F., y Grieco, D. (2014). Automatic learning of linguistic resources for stopword removal and stemming from text. Procedia Computer Science, 38(C), 116–123. https://doi.org/10.1016/j.procs.2014.10.019

Gandhi, U. D., Malarvizhi Kumar, P., Chandra Babu, G., y Karthick, G. (2021). Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). Wireless Personal Communications, 0123456789. https://doi.org/10.1007/s11277-021-08580-3

Google, L. L. C. (2005). Youtube. https://www.youtube.com/

Han, S. (2022). googletrans · PyPI. https://pypi.org/project/googletrans/

Harshith. (2019). Text Preprocessing in Natural Language Processing. Towardsdatascience. https://towardsdatascience.com/text-preprocessing-in-natural-language-processing-using-python-6113ff5decd8

IPSOS. (2020). Ficha Técnica: Encuesta Nacional Urbana. https://www.ipsos.com/sites/default/files/ct/news/documents/2020-04/opinion_data_-_22_de_abril_del_2020.pdf

Khurana Batra, P., Saxena, A., Shruti, y Goel, C. (2020). Election result prediction using twitter sentiments analysis. PDGC 2020 - 2020 6th International Conference on Parallel, Distributed and Grid Computing, 182–185. https://doi.org/10.1109/PDGC50313.2020.9315789.

Kingma, D. P., y Ba, J. L. (2015). Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1–15. https://arxiv.org/abs/1412.6980.

Kumar, S., Morstatter, F., y Liu, H. (2013). Twitter Data Analytics. SpringerBriefs in Conputer science. https://doi.org/10.1007/978-1-4614-9372-3.

Kydros, D., & Magoulios, G. (2019). Twitter content analysis on Greek political leaders. MIBES Transactions. vol. 13 (1), pp. 30–44.

Leonard Richardson. (2020). Beautiful Soup Documentation. https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Liu, Z., Lin, Y., & Sun, M. (2020). Representation Learning and NLP. Representation Learning for Natural Language Processing, 1–11. https://doi.org/10.1007/978-981-15-5573-2_1

Maharani, W., & Effendy, V. (2022). Big five personality prediction based in Indonesian tweets using machine learning methods. International Journal of Electrical and Computer Engineering, 12(2), 1973–1981. https://doi.org/10.11591/ijece.v12i2.pp1973-1981

Medianero Burga, D. (2014). Metodología de Estudios de Línea de Base. Pensamiento Crítico, 15, 061. https://doi.org/10.15381/pc.v15i0.8994

Meta Inc. (2004). Facebook. https://www.facebook.com/

Mohammad, S. A. I. F. M. M., Urney, P. E. D. T., y Canada, C. (2012). CROWDSOURCING A WORD – EMOTION ASSOCIATION LEXICON. Computational Intelligence. https://onlinelibrary.wiley.com/doi/10.1111/j.1467-8640.2012.00460.x

Mongodb. (2021). What Is Unstructured Data? | MongoDB. https://www.mongodb.com/unstructured-data

Monhaler, Edna Maria; Matias Miranda, A. F. (2017). La diversidad lingüística del español en el mundo contemporáneo: propuestas de actividades didácticas. En Actas Del III Congreso Internacional SICELE. Investigación e Innovación En ELE. Evaluación y Variedad Lingüística Del Español. https://cvc.cervantes.es/ensenanza/biblioteca_ele/sicele/sicele03/006_matiasmonheler.htm

Parmezan, A. R. S., Souza, V. M. A., y Batista, G. E. A. P. A. (2019). 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. Information Sciences, 484, 302–337. https://doi.org/10.1016/j.ins.2019.01.076

Paul Davison, R. S. (2020). Clubhouse. https://www.clubhouse.com/

Pennington, J., Socher, R., y Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. https://nlp.stanford.edu/pubs/glove.pdf

Poornima, A., Nataraj, N., Nithya, R., Nirmala, D., y Divya, P. (2022). Sentiment Analysis of Tweets in Twitter Using CNN. 2022 International Conference on Computer Communication and Informatics, ICCCI 2022, 25–28. https://doi.org/10.1109/ICCCI54379.2022.9740779

Poria, S., Hussain, A., y Cambria, E. (2018). Multimodal Sentiment Analysis (Vol. 8). Springer International Publishing. https://doi.org/10.1007/978-3-319-95020-4

Prastyo, P. H., Sumi, A. S., Dian, A. W., & Permanasari, A. E. (2020). Tweets Responding to the Indonesian Government's Handling of COVID-19: Sentiment Analysis Using SVM with Normalized Poly Kernel. Journal of Information Systems Engineering and Business Intelligence, 6(2), 112. https://doi.org/10.20473/jisebi.6.2.112-122

Rai, A., & Borah, S. (2021). Study of Various Methods for Tokenization. Lecture Notes in Networks and Systems, 137, 193–200. https://doi.org/10.1007/978-981-15-6198-6_18

Rodríguez, C. G. and Tule, L. G. (2019). Honduras 2019: Persistent economic and social instability and institutional weakness. Revista de Ciencia Politica, 40, 379–400. https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-090X2020005000112&lng=en&nrm=iso&tlng=en

Ross Ihaka, R. G. (1993). R: The R Project for Statistical Computing. https://www.r-project.org/

Shaghaghi, N., Calle, A. M., Manuel Zuluaga Fernandez, J., Hussain, M., Kamdar, Y., & Ghosh, S. (2021). Twitter Sentiment Analysis and Political Approval Ratings for Situational Awareness. Proceedings - 2021 IEEE International Conference on Cognitive and Computational Aspects of Situation Management, CogSIMA 2021, 59–65. https://doi.org/10.1109/COGSIMA51574.2021.9475935

Sharma, A., & Ghose, U. (2020). Sentimental Analysis of Twitter Data with respect to General Elections in India. Procedia Computer Science, 173(2019), 325–334. https://doi.org/10.1016/j.procs.2020.06.038

Silva, H., Andrade, E., Araujo, D., & Dantas, J. (2022). Sentiment Analysis of Tweets Related to SUS before and during COVID-19 pandemic. IEEE Latin America Transactions, 20(1), 6–13. https://doi.org/10.1109/TLA.2022.9662168

Statista. (2021). Media usage in an internet minute as of August 2021. Statista; Springer Vienna. https://doi.org/10.1007/s13278-021-00853-w

Twitter. (2006). Twitter. https://twitter.com/