UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS
FACULTAD DE INGENIERÍA
ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Presentado por
RUBEN DENKY VALLES MARTINEZ

EFICIENCIA DE UN MODELO PREDICTIVO BASADO EN ANÁLISIS DE SENTIMIENTOS EN LOS RESULTADOS DE LAS ELECCIONES PRESIDENCIALES DEL PERÚ EN EL AÑO 2021

Asesor
FLOR DE LUZ PALOMINO VALDIVIA

Co-asesor
HERWIN ALAYN HUILLCEN BACA

TESIS PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

ANDAHUAYLAS – APURÍMAC – PERÚ

2023

DECLARACIÓN JURADA DE AUTENTICIDAD

Yo, Ruben Denky Valles Martinez, identificado con DNI N° 73456174, de la Escuela Profesional de Ingeniería de sistemas.

Declaro bajo juramento que el Informe Final de Tesis intitulado: *"Eficiencia de un modelo predictivo basado en análisis de sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021"*. Es auténtico y no vulnera los derechos de autor. Además, su contenido es de entera responsabilidad del autor del Informe Final de Tesis, quedando la UNAJMA exenta de toda responsabilidad en caso de atentar contra la Ley de propiedad intelectual y derechos de autor.

Andahuaylas, 20 de junio de 2023.

.....

Firma
N° DNI: 73456174
E-mail: denkyruben@gmail.com
N° Celular: 921265841

APROBACION DEL ASESOR

Quién suscribe:
(ING. FLOR DE LUZ PALOMINO VALDIVIA) por la presente:

CERTIFICA,
Que, el Bachiller en Ingeniería de Sistemas, RUBEN DENKY VALLES MARTINEZ ha culminado satisfactoriamente el informe final de tesis intitulado: "(EFICIENCIA DE UN MODELO PREDICTIVO BASADO EN ANÁLISIS DE SENTIMIENTOS EN LOS RESULTADOS DE LAS ELECCIONES PRESIDENCIALES DEL PERÚ EN EL AÑO 2021)" para optar el Título Profesional de Ingeniero de Sistemas.

Andahuaylas, 13 de febrero del 2023.

(Ing. Flor de Luz Palomino Valdivia)
Asesor

Ruben Denky Valles Martinez
Tesista

ACTA DE SUSTENTACION DE TESIS

En la Av. José María Arguedas del Local Académico SL01 (Ccoyahuacho) en el auditorio de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional José María Arguedas ubicado en el distrito de San Jerónimo de la Provincia de Andahuaylas, siendo las 09:00 horas del día 22 de febrero del año 2023, se reunieron los docentes: Mag. Enrique Edgardo Condor Tinoco, Mag. Ivan Soria Solis, Mag. Neptalí Menejes Palomino, en condición de integrantes del Jurado Evaluador del Informe Final de Tesis intitulado: "EFICIENCIA DE UN MODELO PREDICTIVO BASADO EN ANÁLISIS DE SENTIMIENTOS EN LOS RESULTADOS DE LAS ELECCIONES PRESIDENCIALES DEL PERU EN EL AÑO 2021", cuyo autor es el Bachiller en Ingeniería de Sistemas **RUBEN DENKY VALLES MARTINEZ**, el asesor M.Sc. Flor de Luz Palomino Valdivia y el Coasesor M.Sc. Herwin Alayn Huillcen Baca, con el propósito de proceder a la sustentación y defensa de dicha tesis.

Luego de la sustentación y defensa de la tesis, el Jurado Evaluador **ACORDÓ**: APROBAR por UNANIMIDAD al Bachiller en Ingeniería de Sistemas **RUBEN DENKY VALLES MARTINEZ**, obteniendo la siguiente calificación y mención:

Nota escala vigesimal		Mención
Números	**Letras**	
18	Dieciocho	Excelente

En señal de conformidad, se procedió a la firma de la presente acta en 03 ejemplares.

Mag. Enrique Edgardo Condor Tinoco
Presidente del Jurado Evaluador

Mag. Ivan Soria Solis
Primer Miembro del Jurado Evaluador

Mag. Neptalí Menejes Palomino
Segundo Miembro del Jurado Evaluador

APROBACIÓN DEL JURADO DICTAMINADOR

LA TESIS: **"Eficiencia de un modelo predictivo basado en análisis de sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021"**; para optar el Título Profesional de Ingeniero de Sistemas, ha sido evaluada por el Jurado Dictaminador conformado por:

PRESIDENTE: Mag. Enrique Edgardo Condor Tinoco
PRIMER MIEMBRO: Mag. Iván Soria Solís
SEGUNDO MIEMBRO: Mag. Neptalí Menejes Palomino

Habiendo sido aprobado por **UNANIMIDAD**, en la ciudad de Andahuaylas el día 22 del mes de febrero de 2023.

Andahuaylas, 20 de junio de 2023.

Mag. Enrique Edgardo Condor Tinoco
PRESIDENTE DEL JURADO DICTAMINADOR

Mag. Iván Soria Solís
PRIMER MIEMBRO DEL JURADO DICTAMINADOR

Mag. Neptalí Menejes Palomino
SEGUNDO MIEMBRO DEL JURADO DICTAMINADOR

NOMBRE DEL TRABAJO

INFORME DE TESIS (1).pdf

AUTOR

Ruben Valles

RECUENTO DE PALABRAS

18602 Words

RECUENTO DE CARACTERES

107788 Characters

RECUENTO DE PÁGINAS

102 Pages

TAMAÑO DEL ARCHIVO

2.1MB

FECHA DE ENTREGA

Jun 27, 2023 5:36 PM GMT-5

FECHA DEL INFORME

Jun 27, 2023 5:38 PM GMT-5

22% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada bas

- 22% Base de datos de Internet
- 9% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Cros

UNIVERSIDAD NACIONAL JOSÉ MARÍA ARGUEDAS
Ing. Flor de Luz Palomino Valdivia
DOCENTE

Ruben Denky Valles Martinez
DNI: 73456174
Tesista

Flor de Luz Palomino Valdivia
DNI: 09598250
Asesor

DEDICATORIAS

A Dios, por darme sabiduría y guiarme siempre en los proyectos que tengo.

A mis padres, Pedro y Jesusa, por brindarme su amor, apoyo y motivación incondicional.

AGRADECIMIENTO

A Dios, por darme vida y salud.

A mi familia, por motivarme siempre sin importar la situación o distancia.

A mis asesores, Flor de Luz y Herwin, por su experiencia y apoyo en el desarrollo y culminación de este proyecto.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

En el Perú, algunas de las empresas encuestadoras más populares han perdido credibilidad por parte de la población, debido a sus constantes fallas de predicción y tener pasados relacionados con la corrupción principalmente en relación a contiendas electorales. Debido a esto, muchas personas creen que los sondeos electorales son realizados sin la debida objetividad. Esta situación ha llevado buscar nuevas maneras de obtener predicciones, por lo que el objetivo de este trabajo fue evaluar la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021. La investigación fue de tipo aplicada, retrospectiva y con un enfoque cuantitativo, de nivel predictivo y un diseño preexperimental. El enfoque planteado utiliza el conjunto de datos TASS2016 para realizar el entrenamiento del modelo, y para el procesamiento se usó tweets referentes a las elecciones presidenciales, el cual, contó con una población de 81214 y 15007 tweets para la primera y segunda vuelta electoral y una muestra de 5207 y 1495 tweets respectivamente. Los tweets son clasificados de acuerdo a su polaridad, siendo considerados solamente aquellos que sean positivos y que dentro de su contenido exista una referencia hacia un candidato político. Para la evaluación de modelo planteado se utilizó las medidas de error MAE y RSME. Finalmente, se logró obtener un error MAE de 6.152% muy cercano al resultado de las encuestadoras para la primera vuelta, y para la segunda vuelta un error MAE de 2.574% siendo un resultado considerable al haber acertado en la predicción del ganador de la contienda electoral. El modelo desarrollado es una alternativa de predicción de resultados electorales simplemente algorítmica y de bajo costo.

Palabras clave: Análisis de Sentimientos; Encuesta electoral; Modelo predictivo; Resultados electorales.

ABSTRACT

In Peru, some of the most popular polling companies have lost credibility among the population, due to their constant prediction failures and having past records related to corruption, mainly in relation to electoral contests. Because of this, many people believe that electoral polls are carried out without due objectivity. This situation has led to the search for new ways to obtain predictions, so the objective of this work was to evaluate the efficiency of a predictive model based on Sentiment Analysis in the results of the presidential elections of Peru in the year 2021. The investigation was of applied type, retrospective and with a quantitative approach, predictive level and a pre-experimental design. The proposed approach uses the TASS2016 dataset to perform the training of the model, and for the processing tweets referring to the presidential elections were used, which had a population of 81214 and 15007 tweets for the first and second round and a sample of 5207 and 1495 tweets respectively. The tweets are classified according to their polarity, considering only those that are positive and that within their content there is a reference to a political candidate. For the evaluation of the proposed model, the MAE and RSME error measures were used. Finally, it was possible to obtain a MAE error of 6.152% very close to the result of the pollsters for the first round, and for the second round a MAE error of 2.574%, being a considerable result, having been successful in predicting the winner of the electoral contest. The model developed is a simply algorithmic and low-cost alternative for predicting electoral results.

Keywords: Sentiment Analysis, Electoral Survey, Prediction Model, Election Results.

CHUMASQA

Perú suyupiqa wakin aswan riqsisqa empresakuna encuesta nisqakunan manaña creenapaq hinañachu kanku llaqta runakuna ukhupi, sapa kuti predicción nisqapi mana allinta ruwasqankurayku, hinallataq ñawpaq registrokunayoq kasqankurayku, corrupción nisqawan tupaq, aswantaqa electoral nisqapi atipanakuykunawan. Chayna kasqanraykum achka runakunam piensanku electoral encuestakuna mana allin objetividadniyuq kasqanpi. Kay situacionmi aparqan mosoq ñankunata maskhayta predicciones nisqakuna tarinapaq, chaymi kay llank'aypa objetivonqa karqan Análisis de Sentimiento nisqapi ruwasqa modelo predictivo nisqapa allin ruwayninta chaninchanapaq, Perú suyupi presidencial akllanakuypa ruwasqanpi 2021 watapi, investigacionqa karqan de tipo aplicado, retrospectivo chaymanta enfoque cuantitativo kaqwan, nivel predictivo kaqwan chaymanta huk diseño pre-experimental kaqwan. Chay enfoque propuesto nisqapiqa TASS2016 nisqa conjunto de datos nisqawanmi ruwakun chay modelo nisqapa capacitacionnin, chaymanta procesamiento nisqapaqmi, chay watallapi presidencial akllanakuykunamanta rimaq tweets nisqakuna llamkachisqa, chaypim 81214 runayuq karqa, chaynallataqmi 15007 tweets nisqakuna punta kaq hinaspa iskay kaq rondapaq, hinallataq muestra 5207 hinallataq 1495 tweets nisqakuna. Chay tweets nisqakunan clasificasqa kanku polaridadninkuman hina, chaypin qhawarikun chaykunalla positivo nisqakuna hinallataq contenidonku ukhupi huk referencia candidato político nisqamanta. Chay modelo propuesto nisqa chaninchanapahqa, MAE, RSME pantay tupuykunawanmi churakurqa. Tukuyninpiqa, atikurqam 6,152% pantay MAE nisqa tariyta ancha hichpallapi encuesta ruwaqkunapa ruwasqankuman punta kaq rondapaq, iskay kaq rondapaqñataqmi 2,574% pantay MAE nisqatapas, chaymi ancha allin karqa, chaymi allinta ruwarqaku ganaqmanta willakuypi akllanakuypi atipanakuymanta. Chay modelo ruwasqaqa huk alternativa simple algorítmica hinaspa pisi qullqiyuqmi, akllanakuypa ruwakuyninta willanapaq.

Rimarinamanta: Análisis de Sentimiento nisqa; Akllanakuymanta Tapukuy; Modelo Predictivo nisqa; Akllanakuypi rurasqakuna.

TÉRMINOS Y ABREVIATURAS

ONPE	Oficina Nacional de Procesos Electorales
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
ME	Mean Error
SEPLN	Sociedad Española para el Procesamiento de Lenguaje Natural
TASS	Taller de Análisis Semantico en la SEPLN
LSTM	Long Short-Term Memory
BOW	Bag Of Words
HTBSA	Análisis de Sentimientos Basado en Temas Hibridos
TF-IDF	Term Frecuency – Inverse Document Frecuency
LinReg | LR	Linear Regression
RandForest | RF	Random Forest
NN	Neural Network
SVM	Support Vector Machine
SVC	Support Vector Classification
NB	Naive Bayes
GRU	Gated Recurrent Unit
RNN	Recurrent Neural Network

INTRODUCCION

El análisis de sentimientos ha causado interés en muchos investigadores a nivel mundial ya que diversas aplicaciones y especialmente las redes sociales, páginas web, foros, tiendas virtuales, etc. permiten que las personas vuelquen comentarios de distintos aspectos: político, económico, social, etc. A los cuales se puede realizar un estudio para poder predecir y medir los diferentes sentimientos del grupo social analizado.

En el último proceso electoral presidencial realizado en el Perú ocurrió un caso bastante notable de un cambio brusco en referencia a los sondeos electorales de las encuestadoras con los resultados electorales oficiales obtenidos, poniendo en duda la objetividad con que se elaboran, ya que, según una encuesta publicada en Junio por el Instituto de Estudios Peruanos (2021) reveló que en mayo del mismo año el 59% de los encuestados percibía que los medios de comunicación están favoreciendo a algún candidato, y para junio del mismo año el porcentaje se elevó al 70%.

Según un estudio realizado en Perú por Kantar Ibope Media, publicado por Mendoza (2019), reveló que 1 de cada 10 peruanos utilizó Twitter almenos una vez durante el último mes. Teniendo en cuenta que la cantidad de habitantes en el perú supera los 32 millones hasta el 2020 según el Instituto Nacional de Estadística e Informática (2020), se puede aprovechar el uso la red social Twitter como fuente de recolección de datos para estudiar a la población.

Por lo tanto, este trabajo tiene por objetivo evaluar la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tomando como supuesto que el modelo planteado tenga un resultado similar o aproximado en comparación los sondeos realizados por las encuestadoras.

Para el modelo predictivo se hace uso de datos extraídos de la red social Twitter (tweets), a los cuales, mediante el análisis de sentimientos, se analiza si son positivos, negativos o neutros;

y haciendo uso de solo tweets positivos y que éstas hagan referencia a algún candidato o partido político, se estima el resultado mediante la cantidad de tweets obtenidos por cada uno de ellos.

La razón de la utilización de solamente tweets positivos es debido a que se pretende evaluar la preferencia hacia un candidato, y esta preferencia se contrasta con un comentario positivo. En el caso de los comentarios negativos y neutros, no dan la certeza de prefencia a algún candidato en específico.

Este trabajo de investigación está estructurado en VI capítulos. En el capítulo I "Planteamiento del Problema" se detalla los aspectos principales de la problemática a solucionar. En el capítulo II "Marco Teórico" se muestran algunos trabajos de investigación similares y también los conceptos necesarios para el entendimiento de este trabajo. En el capítulo III "Metodología de la Investigación" se menciona el tipo de investigación y la metodología aplicada en este trabajo. En el capítulo IV "Implementación de la propuesta" se muestra el procedimiento realizado para el entrenamiento del modelo de predicción. En el capítulo V "Experimentos y Resultados" se presenta algunos experimentos y resultados obtenidos según a los objetivos de este trabajo. Y finalmente, en el capítulo VI "Discusión" se realiza una comparación de los resultados obtenidos en este trabajo de investigación con resultados de otros investigadores, también las similitudes y diferencias entre ambos.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1. Descripción del problema

A nivel mundial, según el trabajo publicado por López Gonzáles & De Santiago (2018) todos los países democráticos cuentan con sus propios reglamentos para la realización de elecciones, por lo que cada cierto tiempo pasan por un proceso de elecciones para que la población pueda elegir al líder de su nación. Antes de las elecciones existe un tiempo de campaña electoral donde los candidatos al puesto tratan de muchas maneras convencer a los electores de ser la mejor opción y así poder obtener la mayor cantidad de votos.

Durante las campañas electorales en el Perú, según la Ley Orgánica de Elecciones N°26859 (2021) referente al artículo 191, las encuestadoras y los medios de comunicación pueden difundir las encuestas realizadas hasta el domingo anterior a la fecha programada para las elecciones. Hasta esta fecha muchas encuestadoras hacen la labor de salir a las calles a preguntar a las personas sobre su inclinación política, el problema que surge es que en el Perú muchas de las encuestadoras más relevantes han sido tildadas de "mermeleras" por tener indicios o en algunos casos pruebas de tener un pasado relacionado con la corrupción al haber recibido sobornos para modificar y publicar los resultados de las encuestas a favor de algún candidato, tal como se muestra en el video publicado por Aldis Siu (2021) donde muestra las declaraciones hechas por Vladimiro Montesinos.

En las elecciones generales de primera vuelta en el Perú, según los últimos sondeos publicados por encuestadoras reconocidas como DATUM, CPI e IPSOS, que se muestran en las Figuras 1, 2 y 3, el partido político de mayor aceptación popular y que se suponía que iba pasar a la segunda vuelta era "Acción Popular", representado por el candidato Yonhy Lescano, ya que ocupaba uno de los primeros lugares con un 13.3% de aceptación en promedio, mientras que el partido político del que no se hablaba mucho y que estaba muy por debajo del resto, según las

encuestadoras, fue “Perú Libre” representado por el candidato Pedro Castillo con un 4% de aceptación en promedio.

Figura 1

Último sondeo publicado de la encuestadora DATUM primera vuelta

Nota. Reproducida de Portal Encuestas, Encuestas 1ra Vuelta, 2021 (https://encuestas.com.pe/).

Figura 2

Último sondeo publicado de la encuestadora CPI primera vuelta

Nota. Reproducida de Portal Encuestas, Encuestas 1ra Vuelta, 2021 (https://encuestas.com.pe/).

Figura 3

Último sondeo publicado de la encuestadora IPSOS primera vuelta

Nota. Reproducida de Portal Encuestas, Encuestas 1ra Vuelta, 2021 (https://encuestas.com.pe/).

Al final de las elecciones se pudo comprobar que ésta estadística que las encuestadoras mostraban hasta antes de las elecciones falló en gran manera debido a que, después de las elecciones, los resultados mostraron que el partido “Perú Libre” quedó en primer lugar con un 19.061% de votos, y el partido “Acción Popular” quedó en el quinto lugar con un 9.092% de votos, tal como se muestran en la Tabla 1 y la Figura 4.

Por lo tanto, es una muestra que las encuestas realizadas por estas empresas no son del todo precisas, son poco creíbles y no se puede saber con seguridad si fueron realizadas con la debida objetividad, y esto es una de las razones por la que la población ya no cree en ellas.

Tabla 1

Resultados Oficiales Presidenciales ONPE Primera Vuelta

TOTAL DE VOTOS					
ORGANIZACIONES POLÍTICAS			TOTAL	%VÁLIDOS	%EMITIDOS
		PARTIDO NACIONALISTA PERUANO	230,831	1.603%	1.303%
		EL FRENTE AMPLIO POR JUSTICIA, VIDA Y LIBERTAD	65,300	0.453%	0.369%
		PARTIDO MORADO	325,608	2.261%	1.838%
		PERU PATRIA SEGURA	55,644	0.386%	0.314%
		VICTORIA NACIONAL	814,516	5.656%	4.598%
		ACCION POPULAR	1,306,288	9.071%	7.374%

		AVANZA PAIS - PARTIDO DE INTEGRACION SOCIAL	1,674,201	11.626%	9.451%
		PODEMOS PERU	812,721	5.644%	4.588%
		JUNTOS POR EL PERU	1,132,577	7.865%	6.394%
		PARTIDO POPULAR CRISTIANO - PPC	286,447	1.989%	1.617%
		FUERZA POPULAR	1,930,762	13.407%	10.900%
		UNION POR EL PERU	101,267	0.703%	0.572%
		RENOVACION POPULAR	1,692,279	11.751%	9.553%
		RENACIMIENTO UNIDO NACIONAL	89,376	0.621%	0.505%

		PARTIDO DEMOCRATICO SOMOS PERU	240,234	1.668%	1.356%
		PARTIDO POLITICO NACIONAL PERU LIBRE	2,724,752	18.921%	15.382%
		DEMOCRACIA DIRECTA	50,802	0.353%	0.287%
		ALIANZA PARA EL PROGRESO	867,025	6.021%	4.895%
		TOTAL DE VOTOS VÁLIDOS	14,400,630	100.000%	81.296%
		VOTOS EN BLANCO	2,190,059		12.364%
		VOTOS NULOS	1,123,027		6.340%
		TOTAL DE VOTOS EMITIDOS	17,713,716		100.000%

Nota. Reproducida de Resultados Elecciones Generales y Parlamento Andino, ONPE, 2021 (https://resultadoshistorico.onpe.gob.pe/EG2021/EleccionesPresidenciales/RePres/T).

Figura 4

Estadística de Resultados Elecciones Presidenciales Primera Vuelta

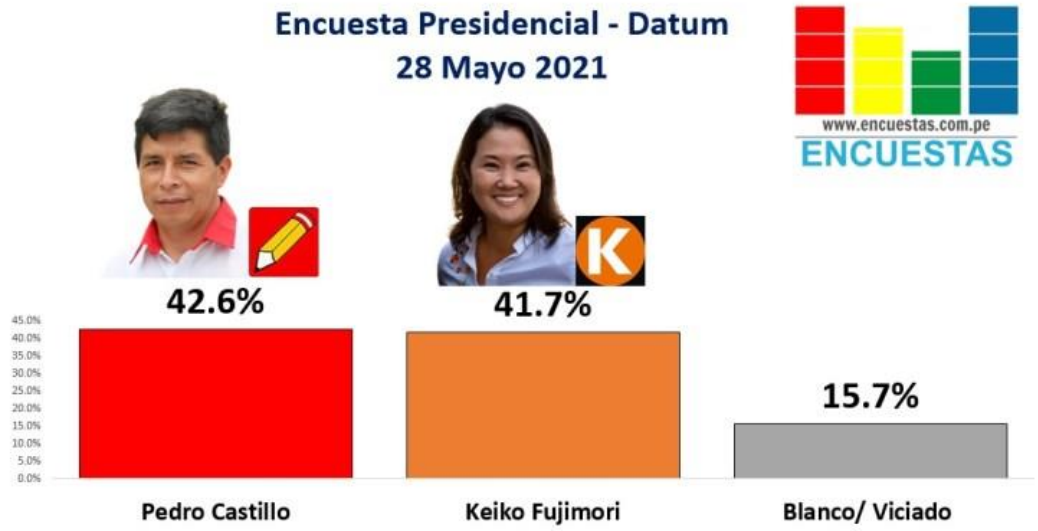
Nota. Reproducida de Resultados Elecciones Generales y Parlamento Andino, ONPE, 2021 (https://resultadoshistorico.onpe.gob.pe/EG2021/EleccionesPresidenciales/RePres/T).

En el caso de la segunda vuelta, las encuestadoras DATUM e IPSOS muestran sondeos más acertados con los resultados oficiales obtenidos, dando como ganador al candidato Pedro Castillo, a diferencia de la encuestadora IDICE que muestra como ganadora a la candidata Keiko Fujimori.

En las figuras 5, 6 y 7 se muestran los resultados de los últimos sondeos electorales publicados de las encuestadoras DATUM, IPSOS e IDICE para las elecciones presidenciales de segunda vuelta, y en la Tabla 2 y la figura 8, se muestran los resultados oficiales obtenidos.

Figura 5

Último sondeo publicado de la encuestadora DATUM segunda vuelta



Nota. Reproducida de Portal Encuestas, Encuestas 2da Vuelta, 2021 (https://encuestas.com.pe/).

Figura 6

Último sondeo publicado de la encuestadora IPSOS segunda vuelta

Nota. Reproducida de Portal Encuestas, Encuestas 2da Vuelta, 2021 (https://encuestas.com.pe/).

Figura 7

Último sondeo publicado de la encuestadora IDICE segunda vuelta

Nota. Reproducida de Portal Encuestas, Encuestas 2da Vuelta, 2021 (https://encuestas.com.pe/).

Tabla 2

Resultados Oficiales Presidenciales ONPE Segunda Vuelta

TOTAL DE VOTOS					
ORGANIZACIONES POLÍTICAS			TOTAL	%VÁLIDOS	%EMITIDOS
		PARTIDO POLITICO NACIONAL PERU LIBRE	8,836,380	50.126%	46.860%
		FUERZA POPULAR	8,792,117	49.874%	46.626%
		TOTAL DE VOTOS V□IDOS	17,628,497	100.000%	93.486%
		VOTOS EN BLANCO	121,489		0.644%
		VOTOS NULOS	1,106,816		5.870%
		TOTAL DE VOTOS EMITIDOS	18,856,802		100.000%

Nota. Reproducida de Resultados Segunda Elección Presidencial, ONPE, 2021

(https://resultadoshistorico.onpe.gob.pe/EG2021/EleccionesPresidenciales/RePres/T).

Figura 8

Estadística de resultados Elecciones Presidenciales Segunda Vuelta

Nota. Reproducida de Resultados Segunda Elección Presidencial, ONPE, 2021 (https://resultadoshistorico.onpe.gob.pe/EG2021/EleccionesPresidenciales/RePres/T).

Tomando en cuenta la enorme cantidad de dinero y recursos humanos que utilizan las encuestadoras para hacer todo el movimiento logístico necesario a nivel nacional para realizar los diferentes sondeos electorales, y que al final la mayoría de estos resulten tener notables errores de predicción, como los mostrados anteriormente, se planteó diseñar un modelo de predicción de resultado electoral basado en análisis de sentimientos, utilizando la opinión (tweets) de los usuarios en publicaciones de la red social Twitter, relacionada con los candidatos a la presidencia, para poder comparar los resultados estadísticos obtenidos mediante el modelo de predicción con los resultados estadísticos de las encuestadoras mencionadas. Luego, hacer un análisis respecto a la eficiencia del modelo predictivo planteado en comparación con las encuestadoras para las elecciones presidenciales del Perú 2021 de primera y segunda vuelta.

1.2. Formulación del problema

1.2.1. Problema General

¿Cuál es la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021?

1.2.2. Problemas Específicos

- ¿Cuál es la eficiencia de un modelo predictivo de análisis de sentimientos de tweets en español?
- ¿Cuál es la eficiencia según el nivel de error RMSE y MAE del modelo predictivo para las elecciones presidenciales en Perú en primera vuelta?
- ¿Cuál es la eficiencia según el nivel de error RMSE y MAE del modelo predictivo para las elecciones presidenciales en Perú en segunda vuelta?
- ¿Cuál es el nivel de error del modelo predictivo en comparación con las encuestadoras IPSOS, CPI y DATUM en primera vuelta?
- ¿Cuál es el nivel de error del modelo predictivo en comparación con las encuestadoras IPSOS, DATUM e IDICE en segunda vuelta?

1.3. Objetivos

1.3.1. Objetivo General

Evaluar la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021.

1.3.2. Objetivos Específicos

- Determinar la eficiencia de un modelo predictivo de análisis de sentimientos de tweets en español.

- Evaluar el RMSE y MAE del modelo predictivo para las elecciones en Perú en primera vuelta.
- Evaluar el RMSE y MAE del modelo predictivo para las elecciones en Perú en segunda vuelta.
- Comparar el nivel de error del modelo predictivo con las encuestadoras IPSOS, CPI y DATUM en primera vuelta.
- Comparar el nivel de error del modelo predictivo con las encuestadoras IPSOS, DATUM e IDICE en segunda vuelta.

1.4. Hipótesis de la Investigación

1.4.1. Hipótesis General

Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un resultado significativo.

1.4.2. Hipótesis Específicas

a) Es posible determinar la eficiencia de un modelo predictivo de análisis de sentimientos de tweets en español.
b) Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un error RMSE y MAE en promedio menor de 10% en primera vuelta.
c) Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un error RMSE y MAE en promedio menor de 5% en segunda vuelta.

d) Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un nivel de error RMSE y MAE en promedio menor en comparación con las encuestadoras IPSOS, CPI y DATUM en primera vuelta.

e) Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un nivel de error RMSE y MAE en promedio menor en comparación con las encuestadoras IPSOS, DATUM e IDICE en segunda vuelta.

1.5. Justificación

A medida que los medios de comunicación y las redes sociales están cada vez más presentes en nuestra vida cotidiana es mucho más probable que nos dejemos influenciar por las noticias, publicaciones y principalmente por las encuestas que muestran.

La importancia de este trabajo es diversa, ya que tratará de solucionar diversos tipos de problemas, por ejemplo, en el aspecto político, la desinformación de los electores a través de las redes sociales, detectar posibles intereses políticos de los medios de comunicación y en la parte social la concientización de la población de la influencia de las publicaciones en las redes sociales principalmente aquellas relacionadas con las encuestas electorales.

Este trabajo de investigación permitirá identificar aquellas encuestas que probabilísticamente estén realizadas sin la debida objetividad, al encontrar la relación existente entre la posición política mostrada en los tweets con los resultados de las encuestas.

Es muy importante conocer de mejor manera la veracidad de las publicaciones que se muestran en las redes sociales, específicamente en las encuestas, porque mucha gente confía en dichas estadísticas para poder tener una inclinación política a favor de algún partido, generalmente

de aquellos que están en los primeros lugares, tal como se mencionó en la investigación hecha por Quispe Auqui & Tomayquispe Mejía (2019) donde, en uno de los resultados que obtuvieron del cuestionario, mencionan que al 58% de personas las encuestas influyeron en su toma de decisión de voto.

De esta manera la presente investigación incentivará, a las encuestadoras, a realizar encuestas con objetividad, de mayor calidad y veracidad y a los investigadores a mejorar o buscar otras alternativas de predicción de resultados electorales. También, los electores tendrán mayor conocimiento e información real reduciendo así el nivel de influencia de las publicaciones de encuestas adulteradas en redes sociales.

Esta propuesta es una alternativa simplemente algorítmica para realizar sondeos electorales, donde no se necesita de dinero ni recursos humanos excesivos.

1.6. Limitaciones

Una limitación fue la bibliografía en español, debido a que este tipo de estudios no son realizados con frecuencia en países hispanohablantes.

Otra limitación muy importante el cual hizo que el proyecto de investigación se retrasara mucho, fue la recolección de datos, debido a la dificultad de descargar tweets antiguos, y algunas librerías que ayudaban como get-old-tweets y tweepy, mediante la actualización de la API de Twitter, fueron restringidas para descargar tweets antiguos.

También, una limitación importante que influyó mucho en los resultados obtenidos por el modelo de predicción planteado fue que las personas que usan Twitter en el Perú la mayoría de ellas viven en las ciudades principales, en cual, fue una desventaja considerable para el candidato *Pedro Castillo* debido a que la mayoría de sus seguidores y votantes en la primera vuelta electoral, viven en la sierra sur y no tienen una cuenta activa de Twitter.

CAPÍTULO II: MARCO TEÓRICO

2.1. Antecedentes de Investigación

2.1.1. Antecedentes Internacionales

Existen algunos trabajos de investigación donde abordaron temas similares al nuestro, por ejemplo, Wang & Gan (2017), hicieron una investigación en Francia donde su objetivo fue proponer un nuevo método para predecir los resultados electorales basado en el análisis de datos de Twitter; Recolectaron datos usando para la búsqueda palabras relacionadas con las elecciones en Francia como los nombres de candidatos, publicados antes de la fecha de votación. Este método, analiza los sentimientos de los tweets para predecir la popularidad de los candidatos clasificándolos en positivos, negativos y neutros basados en un conjunto de palabras; Como resultado supera significativamente al método de Tumasjan, al que toman como referencia. Concluye que el método propuesto considerando los tweets neutrales relacionados con candidatos específicos aumenta la precisión de predicción en el caso de las elecciones francesas. Siendo este uno de los primeros trabajos en considerar el análisis de sentimientos para predecir resultados electorales.

Otro trabajo de investigación desarrollado en Chile por Rodrigues et al. (2018), tuvo como objetivo predecir los resultados de las elecciones presidenciales chilenas en el año 2017 haciendo uso de datos políticos de la red social Twitter, para ello describe la implementación de varios algoritmos de aprendizaje automático para realizar el análisis de sentimientos; Los datos que usaron son tweets generados durante la campaña electoral donde hacían mención a la cuenta o nombre y apellido del candidato, logrando recolectar de entre los 11 candidatos un total de 640224 datos etiquetados entre positivos, negativos y neutros. Los resultados obtenidos mostraron que el

algoritmo “ensemble voting classifier” logró una predicción precisa obteniendo un Error Absoluto Medio del 0.51% para la segunda ronda.

En otro trabajo, Bansal & Srivastava (2018) hicieron un estudio en la ciudad de Uttar Pradesh de la India que tuvo como objetivo introducir una nueva metodología llamada Análisis de sentimientos basado en temas híbridos (HTBSA) con el objetivo de obtener las relaciones de palabras y coocurrencias en tweets para la predicción de elecciones; para este fin, se usó más de 300000 tweets recolectados del 1 al 20 de febrero de 2017. Los resultados que se obtuvieron muestran que el modelo HTBSA propuesto, obtuvo muy buen rendimiento obteniendo un resultado de 8.4 de error absoluto medio, superando así a otras técnicas existentes de predicción de elecciones basado en Twitter como el análisis de sentimientos a nivel de oración el cual obtuvo un resultado de 12 de error absoluto medio.

Por otro lado, Callejas & Velez (2019) hicieron un trabajo con el objetivo de proponer un modelo híbrido mezclando el procesamiento del lenguaje natural, análisis de sentimientos y el análisis de datos numéricos de las encuestas para predecir los resultados de las elecciones presidenciales de primera vuelta en Colombia en el año 2018; Para este fin, utilizaron datos extraídos de las redes sociales de Twitter y Facebook, y datos de encuestas o sondeos de opinión por un periodo de 120 días. Utilizaron LSTM como algoritmo de clasificación de los datos, y RMSE para medir la cantidad de error. Como resultado principal se obtuvo que el RMSE del modelo hibrido propuesto es del 2.47% el cual supera al RMSE de las encuestadoras. Llegando a la conclusión de que las predicciones realizadas basadas en los sentimientos de las personas en plataformas digitales pueden predecir resultados precisos de las elecciones.

Por último, el trabajo de investigación realizado por Khurana et al. (2020), desarrollado en la India, tuvo como objetivo predecir los resultados electorales de Lok-Sabha en el año 2019, para

dicho fin utilizaron un conjunto de datos recolectados tanto de Twitter como de Github los cuales estaban relacionados con dos partidos políticos, logrando asi recolectar un total de 41265 tweets. Para la extracción de características usaron las tecnicas de BOW y tf-idf, haciendo uso de 5 modelos de aprendizaje. El modelo de apredizaje automático Decision Tree haciendo uso de tf-idf obtuvo el mejor resultado en comparación a los demás con una presición del 86.3% aplicados en tweets de India, concluyendo, según la predicción, que el ganador de las elecciones será el partido BJP.

2.2. Bases Teóricas

2.2.1. Modelos Predictivos

Actualmente, los modelos predictivos son muy necesarios e indispensables para muchas organizaciones porque permiten tener una visión del futuro de diferentes situaciones y/o áreas deseadas, son muy necesarios para tomar decisiones importantes.

Según König (2009) señala que "los modelos predictivos se utilizan para encontrar patrones potencialmente valiosos en los datos o para predecir el resultado de algún evento" (pág.3).

Los modelos predictivos son el resultado de convinar datos y matemáticas, normalmente usan tecnicas estadísticas haciendo uso de los datos historicos y actuales para crear un modelo que ayude a predecir el comportamiento futuro (Kalechofsky, 2016, pág.4).

2.2.1.1. Técnicas de Modelos Predictivos Actuales

Existen muchas técnicas para construir modelos predictivos, generalmente se podría clasificar en dos grandes grupos, modelos predictivos de regresión y de clasificación. A continuación, se mencionan algunas de las técnicas más conocidas y utilizadas.

a) Regresión Lineal (RL)

Según Espino Timón (2017) el modelo de regresión lineal “analiza la relación existente entre la variable dependiente o de respuesta y un conjunto de variables independientes o predictoras. Esta relación se expresa como una ecuación que predice la variable de respuesta como una función lineal de los parámetros” (pág.18).

b) Support Vector Machine (SVM)

La máquina de vectores de soporte, o SVM por sus siglas en inglés, es un algoritmo de aprendizaje supervisado que construye modelos predictivos utilizando vectores de n dimensiones llamados hiperplanos que se construyen según la cantidad de clases que se tiene, siendo esta última una cantidad de n+1 clases (Rudd, 2018).

c) K-Nearest Neighbor (KNN)

El k-Vecinos más cercanos, o KNN por sus siglas en inglés, es un método de clasificación no paramétrico que utiliza un valor k, elegido según cada criterio, para buscar a sus vecinos más cercanos formando así una vecindad mediante una votación por mayoría considerando o no la ponderación basada en distancia. (Guo et al.2003)

d) Decisión Tree (DT)

El Árbol de Decisión, o DT por sus siglas en inglés, es un método de clasificación referido como una partición recursiva del espacio de instancia, que utiliza un nodo principal llamado raíz para iniciar el proceso, debajo de este nodo se extienden más nodos llamados nodos internos que cuentan con un borde saliente y, por último, están los nodos denominados hojas que se conocen como nodos terminales o de decisión (Rokach & Maimon, 2005).

e) Naive Bayes (NBC)

Los clasificadores Naive Bayes, o NBC por sus siglas en inglés, se fundamenta en la regla de probabilidad condicional de Bayes, en el cual asume que los predictores son estadísticamente independientes mostrándose así un algoritmo fiable y fácil de interpretar (Espino Timón, 2017).

f) Artificial Neural Network (ANN)

Las redes neuronales artificiales, o ANN por sus siglas en inglés, son métodos de modelado no lineal bastante avanzado, lo que hace que tengan una gran capacidad de modelar funciones complejas, su principal característica es comprender la relación existente entre valores de entrada y salida en el proceso de entrenamiento tanto de supervisados como de no supervisados (Espino Timón, 2017).

g) Recurrent Neural Network (RNN)

Las redes neuronales recurrentes, o RNN por sus siglas en inglés, son un tipo de redes neuronales artificiales utilizada principalmente para el reconocimiento de patrones en una secuencia de datos, por lo que se adapta facilmente para agrupar, clasificar y hacer predicciones de texto, video, voz y series de tiempo (Manaswi, 2018).

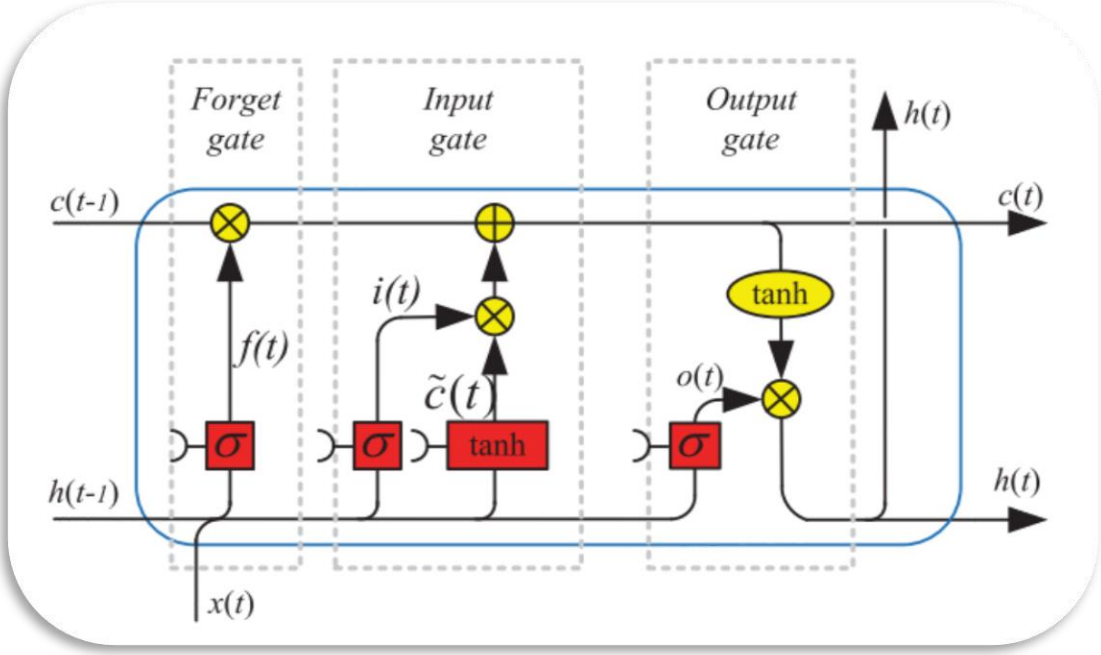
A partir de la modificación de la arquitectura de una RNN, surge Long-Short Term Memory (LSTM), que mejora el entrenamiento al mantener una memoria el cual puede almacenar información durante periodos largos de tiempo.

Esta red LSTM mejora la capacidad de almacenar datos de celda recurrente estándar al agregar una compuerta dentro de la celda. (Yu et al., 2019)

En la Figura 10 se muestra la arquitectura interna de la red LSTM, se puede observar que existen tres compuertas principales: forget gate, Input gate y Output gate; Estas compuertas permiten a la red eliminar datos irrelevantes, actualizarlos y añadir nuevos datos relevantes.

Figura 9

Arquitectura de red LSTM



Nota. Reproducida de ´Architecture of LSTM with a forget gate´, Yu et al., 2019.

2.2.2. Métricas de Evaluación de Modelos Predictivos

Para conocer la calidad de predicción del modelo propuesto en este proyecto, se propuso dos métricas de error distintas como el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE). Estos fueron elegidos debido a que son métricas comúnmente utilizadas en modelos de predicción y cada uno de ellos entrega un resultado distinto con su respectiva interpretación y enfoque.

2.2.2.1. Mean Absolute Error (MAE)

El Error Absoluto Medio, o MAE por sus siglas en inglés, es una medida utilizada para evaluar modelos de predicción. Según Sammut & Webb (2011) menciona la siguiente definición:

El error absoluto medio de un modelo con respecto a un conjunto de prueba es la media de los valores absolutos de los errores de predicción individuales en todas las instancias del

conjunto de prueba. Cada error de predicción es la diferencia entre el valor real y el valor predicho para la instancia. (p.652)

La fórmula para hallar el Error Absoluto Medio es la siguiente:

$$MAE = \frac{\sum_{i=1}^{N} |A_i - F_i|}{N}$$

Donde: A_i es el valor actual, F_i es el valor predicho y N es la cantidad de datos.

2.2.2.2. Root Mean Square Error (RMSE)

La Raíz del Error Cuadrático Medio, o RMSE por sus siglas en inglés, es una medida utilizada para evaluar modelos de predicción. Según Chai & Draxler (2014) dice lo siguiente:

> Si bien el MAE otorga el mismo peso a todos los errores, el RMSE penaliza la varianza ya que otorga más peso a los errores con valores absolutos más grandes que a los errores con valores absolutos más pequeños. Cuando se calculan ambas métricas, el RMSE, por definición, nunca es menor que el MAE (pág.1).

La fórmula para hallar la Raíz del Error Cuadrático Medio es la siguiente:

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(A_i - F_i)^2}{N}}$$

Donde: A_i es el valor actual, F_i es el valor predicho y N es la cantidad de datos.

2.2.3. Análisis de Sentimientos

2.2.3.1. Niveles de Análisis de Sentimientos

El campo de análisis de sentimientos se ha particionado en subcampos más específicos, teniendo estudios principalmente en tres niveles de investigación (Liu, 2012).

a) **Nivel de Documento:** Está enfocado en la clasificación y calificación de un documento completo expresado en sentimientos positivo o negativo, asumiendo que se expresa opiniones sobre una sola entidad (Liu, 2012).

b) **Nivel de Oración:** Está enfocado en la clasificación y calificación de las oraciones de un texto, expresado en sentimiento u opinión positiva, negativa o neutral […] Está estrechamente relacionado con la clasificación de subjetividad (Liu, 2012).

c) **Nivel de Aspecto y Entidad:** Anteriormente llamado nivel de característica debido a que realiza un análisis más detallado […] el nivel de aspecto se centra directamente en la opinión en sí. Una opinión (positiva o negativa) sin un objetivo tiene una utilidad limitada […] Por ende, el objetivo principal de este nivel es descubrir sentimientos sobre las entidades y/o sus aspectos (Liu, 2012).

2.2.3.2. Enfoques del Análisis de Sentimientos

Según Bravo Marquez et al. (2014) al analizar la literatura existente en el campo de análisis de sentimientos, los enfoques actuales se centran principalmente en una dimensión de opinión particular. Por lo tanto, propone una clasificación formada por polaridad, fuerza y emoción.

a) **Polaridad:** Este método enfatiza en la obtención de una variable categórica de una oración, cuyos valores son positivos, negativos o neutrales. Los recursos léxicos encausados a la polaridad están compuestas de una lista de palabras positivas y negativas.

b) **Fuerza:** Este método enfatiza en proporcionar un nivel de intensidad de una oración, según al sentimiento de polaridad, devuelven una puntuación numérica que indican la intensidad o fuerza del sentimiento positivo o negativo. Los recursos léxicos orientados a la fuerza usan las palabras de opinión junto con puntuaciones de intensidad.

c) **Emoción:** Este método enfatiza en la obtención de emociones o estados anímicos de una oración, y clasificarlo en una categoría como alegría, tristeza, ira, entre otras. Los recursos léxicos referentes a la emoción, usan una lista de palabras con diferentes estados anímicos.

2.2.3.3. Análisis de Sentimientos Basados en la Red Social Twitter

Según Alsaeedi & Zubair Khan (2019) existen cuatro enfoques de análisis de sentimientos en la red social Twitter, estas son basados en métodos de aprendizaje supervisado, basado en conjuntos, basados en léxico y el método hibrido.

a) Aprendizaje Supervisado

Son un conjunto de técnicas de métodos de aprendizaje donde se utiliza una colección de datos etiquetados para realizar su entrenamiento y posteriormente generar un modelo para predecir datos no etiquetados. Los algoritmos de aprendizaje supervisado mayormente utilizados en relación con este proyecto son SVM y clasificadores NB.

b) Basado en Conjuntos

Este enfoque consiste en fusionar múltiples métodos de clasificación utilizando las características más relevantes de cada método, con el objetivo de conseguir mejores resultados de predicción acerca de la polaridad de las oraciones.

c) Basados en Léxico

Son un conjunto de metodologías donde su característica principal es el uso de un diccionario de palabras donde, cada uno de estos, guardan relación o son usados normalmente para expresar los distintos sentimientos de polaridad como positivo, negativo o neutral. Este tipo de

método pertenece al campo de los algoritmos no supervisados, debido a que no es necesario tener una etiqueta de polaridad de cada oración.

d) Método hibrido

Son una combinación de las diferentes técnicas de aprendizaje máquina tanto supervisado como no supervisado, tiene como objetivo mejorar los resultados de predicción haciendo uso de las bondades de los distintos algoritmos como por ejemplo los basados en léxico con el de SVM.

2.2.4. Elecciones Generales en el Perú

En el caso de Perú, es un proceso electoral que se realiza cada cinco años con el objeto de elegir al presidente de la República, vicepresidentes, Congresistas de la República y representantes peruanos ante el Parlamento Andino. A continuación, se detalla la función que cumplen cada uno de ellos.

- ***Presidente y vicepresidente.*** Según al enciso 2 del Artículo 118 de la constitución política del Perú el presidente debe representar al estado, en este caso al Perú, dentro y fuera de la república.
- ***Congresistas de la República.*** Según al documento de (Ministerio de Asuntos Exteriores y de Cooperación, 2011) las funciónes de los congresistas son representar, legislar y fiscalizar. Cabe mensionar que los congresistas son elegidos en todas las regiones del Perú en representación de los mismos.
- ***Representantes ante el Parlamento Andino.*** Según el reglamento general del parlamento andino es el órgano deliberante y de control del sistema andino y de integración, son encargados de representar a la comunidad andina.

2.2.4.1. Oficina Nacional de Procesos Electorales (ONPE)

La Oficina Nacional de Procesos Electorales, es un organismo electoral constitucional autónomo que forma parte del estado peruano, es la máxima autoridad encargada de organizar y llevar a cabo procesos electorales. La ONPE cuenta con su portal web en donde se realiza la publicación de todos los resultados electorales, mostrando así el porcentaje de votos obtenido de cada candidato, la cantidad total de votos válidos de cada candidato, cantidad de votos en blanco, cantidad de votos nulos y la cantidad total de votos emitidos. Estos datos se pueden consultar en la siguiente página web https://www.onpe.gob.pe/elecciones/historico-elecciones/.

2.3. Definición de Variables

Variable Dependiente (Predicción de resultados de las elecciones presidenciales del Perú): Son los resultados electorales obtenidos con el modelo de predicción y contabilizados con la metodología propuesta, sobre las elecciones presidenciales de primera y segunda vuelta.

Variable Independiente (Modelo predictivo basado en análisis de sentimientos): Un modelo predictivo hace referencia al uso de un algoritmo haciendo uso de datos históricos o actuales para predecir, en este caso, un resultado probable; se hace mención al análisis de sentimientos debido a que se utilizará como datos al sentimiento expresado en los comentarios realizados por los usuarios.

CAPITULO III: METODOLOGÍA DE LA INVESTIGACIÓN

3.1. Tipo y Nivel de Investigación

3.1.1. Tipo de Investigación

Este trabajo de investigación es de tipo aplicada debido a que busca aportar una posible solución un problema práctico y concreto del entorno social, en este caso en el aspecto político. (Muñoz Rocha, 2015, pág. 142). Según la planificación de los datos es retrospectiva debido a que utiliza datos pasados y secundarios. Según el enfoque de investigación es de tipo cuantitativo, por lo que utiliza datos numéricos y estadísticos. (Muñoz Rocha, 2015, pág. 86)

3.1.2. Nivel de Investigación

Este trabajo de investigación es de nivel predictivo, debido a que pronostica el comportamiento de un hecho social, concretamente en el ámbito político. (Muñoz Rocha, 2015, pág. 85)

3.1.3. Diseño de la Investigación

El diseño de la presente investigación es preexperimental debido a que existe un bajo control de las variables de estudio, la asignación de datos no es aleatoria y no se cuenta con un grupo de control. (Bernal, 2010, pág. 146)

3.2. Operacionalización de Variables

Tabla 3

Operacionalización de Variables

VARIABLE	DIMENSION	INDICADORES	INDICE/ESCALA
Variable independiente: Modelo predictivo basado en Análisis de Sentimientos	Técnicas de Modelos predictivos	Nivel de error RMSE	$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(A_i - F_i)^2}{N}}$
		Nivel de error MAE	$MAE = \frac{\sum_{i=1}^{N} \lvert A_i - F_i \rvert}{N}$
Variable dependiente: Predicción de resultados de las elecciones presidenciales del Perú 2021	Primera vuelta	Resultados predichos	Porcentaje de votos
			Posición obtenida
	Segunda Vuelta	Resultados predichos	Porcentaje de votos
			Posición obtenida

3.3. Población

Para la primera vuelta electoral, la población utilizada está compuesta por 81214 tweets extraídos de la red social Twitter, donde se consideró solo las publicaciones hechas desde el 01 de marzo de 2021 hasta el 29 de marzo de 2021 (desde el inicio del último mes de campaña hasta el inicio del debate electoral de los candidatos).

Para la segunda vuelta electoral, la población utilizada está compuesta por 15007 tweets extraídos de la red social Twitter, donde se consideró solo las publicaciones hechas desde el 01 de

mayo de 2021 hasta el 30 de mayo de 2021 (desde el inicio del último mes de campaña hasta el inicio del debate electoral de los candidatos).

Cabe mencionar que la búsqueda y la obtención de datos se realizó de manera global, dando como parámetro las cuentas oficiales de cada candidato en las elecciones presidenciales Perú 2021.

3.4. Muestra

Para la primera vuelta electoral se contó con 5265 tweets obtenidos mediante el preprocesamiento de datos de la población, donde se escogió aquellos que expresen solamente sentimientos positivos y que tenga la etiqueta de mención a algún candidato participante en las elecciones de primera vuelta dentro de la oración.

Para la segunda vuelta electoral se contó con 1495 tweets obtenidos mediante el preprocesamiento de datos de la población, donde se escogió aquellos que expresen solamente sentimientos positivos y que tenga la etiqueta de mención a algún candidato participante en las elecciones de segunda vuelta dentro de la oración.

3.5. Técnicas e Instrumentos de Recolección de Datos

Para la recolección de datos se utilizó la plataforma web Google Colaboratory, donde se puede ejecutar código Python, en el cual se utilizó la librería denominada “sns-crape” para el raspado de datos de la red social Twitter.

3.6. Plan de Tratamiento de Datos

3.6.1. Técnicas Estadísticas

Los estadísticos que se utilizaron para las diferentes pruebas de hipótesis de la investigación son el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE).

- ✓ Para RMSE, al realizar la comparación del modelo con los distintos resultados de las encuestadoras, el mejor resultado es aquel que tenga menor RMSE. Este tipo de medida penaliza más los errores grandes debido a que la cantidad de error se eleva al cuadrado, por lo que cuanto mayor es el error más grande será el RMSE. Comúnmente, a esta medida se le conoce como la desviación estándar.
- ✓ MAE, tiene un rango que inicia en 0 a más debido a que usa los valores absolutos, esto indicará el tamaño de error que tiene el modelo de predicción en comparación con los resultados de las encuestadoras. Esta medida es la más popular para evaluar modelos de regresión, ya que el resultado es el promedio de todos los errores individuales.

3.6.2. Diseño Estadístico para probar la Hipótesis

Este trabajo de investigación tiene un diseño deductivo-inductivo en vista que los resultados de las métricas de evaluación (RMSE y MAE) contrastan directamente la hipótesis planteada.

CAPITULO IV: IMPLEMENTACIÓN DE LA PROPUESTA

Para la implementación del modelo propuesto de clasificación de sentimientos se utilizó como referencia el pipeline desarrollado por Huillcen et al. (2021) debido a que se adecua perfectamente en los pasos de implementación del modelo.

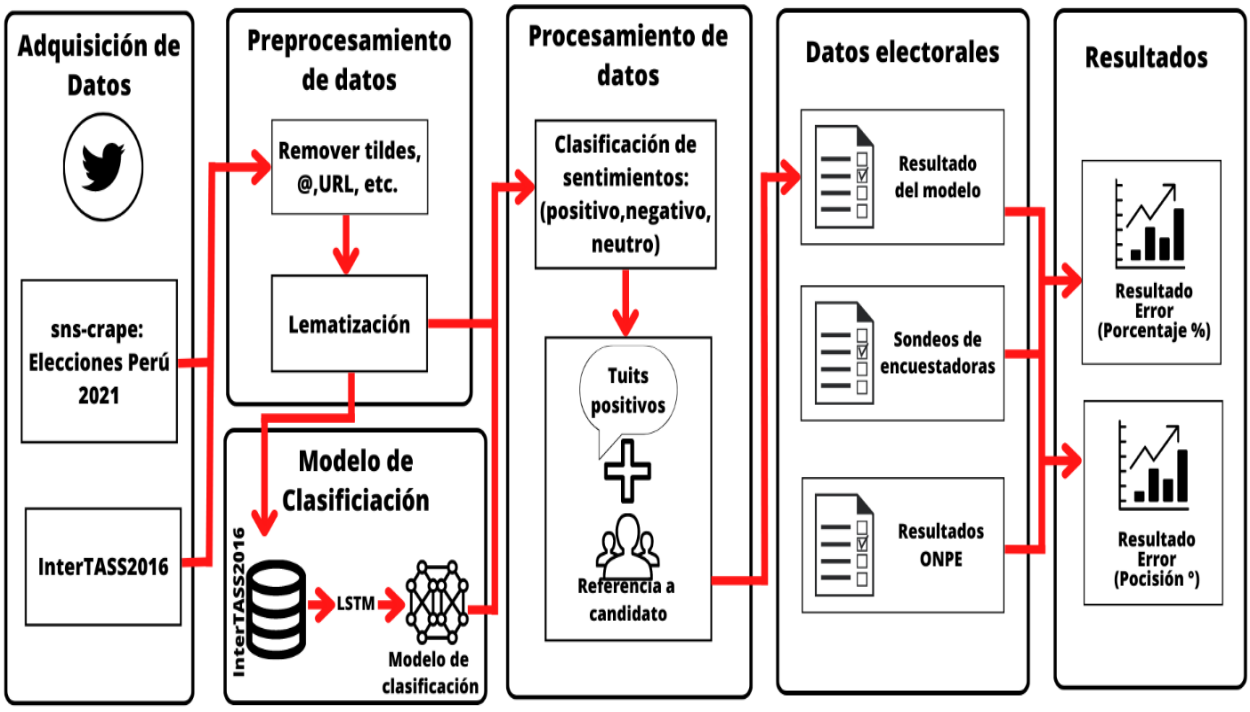
4.1. Pipeline Propuesto

Cabe señalar que no existe una metodología determinada para el desarrollo de trabajos de este tipo, por lo que en cada trabajo es distinto y varía según a los objetivos del mismo.

En la Figura 9 se muestra el pipeline propuesto para el desarrollo general de este trabajo de investigación.

Figura 10

Pipeline propuesto



4.2. Adquisición de Datos

En el desarrollo del trabajo se utilizaron dos conjuntos de datos distintos, uno para el entrenamiento y prueba del modelo de clasificación de sentimientos y el otro netamente para el procesamiento según a los objetivos planteados.

4.2.1. Datos interTASS2016

Para el entrenamiento y prueba del modelo de clasificación de sentimientos se utilizó el conjunto de datos denominada TASS-2016, desarrollada por Martínez Cámara et al. (2016), y ofrecida por la Sociedad Española para Procesamiento del Lenguaje Natural (SEPLN), los cuales se pueden descargar desde el siguiente link: https://github.com/TheDenky/Twitter-Election-Predict-Peru/blob/main/general-train-tagged-3l.xml. así mismo desde su página web: http://www.sepln.org/workshops/tass/2016/tass2016.php.

Se utilizó el corpus general el cual cuenta con 7218 tweets etiquetados y distribuidos como se muestra en la Tabla 4.

Tabla 4

Frecuencia de Polaridad de Datos de Entrenamiento del Modelo

Polaridad	Frecuencia
Positivos	2884
Negativos	2182
Neutros + Sin Polaridad	2152
Total	**7218**

Los datos de entrenamiento del modelo también están divididos por diferentes tópicos, siendo más frecuente los datos relacionados al aspecto político, otros y entretenimiento, tal como se muestra en la Tabla 5.

Tabla 5

Frecuencia de Datos por cada Tópico

Tópico	Frecuencia
Política	2715
Otros	1620
Entretenimiento	1209
Economía	525
Música	412
Fútbol	225
Cine	183
Tecnología	144
Deportes	101
Literatura	84

4.2.2. Datos de Elecciones Perú 2021

Para el procesamiento de los datos se implementó un corpus con tweets referentes a las elecciones presidenciales del Perú en el año 2021 para la primera y segunda vuelta electoral, recolectados mediante la librería ‘sns-crape’ de python desde Google Colab.

Como parámetro de busqueda se utilizó la cuenta oficial de Twitter de cada candidato a la presidencia de la república. En la Tabla 6 se muestra el nombre del candidato y su respectiva cuenta

oficial de Twitter. Cabe señalar que, para la segunda vuelta solo se utilizó las cuentas de los candidatos *Keiko Fujimori* y *Pedro Castillo*.

Tabla 6

Cuenta Oficial de cada Candidato Presidencial

Candidato	Cuenta Oficial Twitter
Ollanta Humala	@Ollanta_HumalaT
Marco Arana	@vozdelatierra
Julio Guzman	@julioguzmanperu
Rafael Santos Normand	@rafaelsantosn
George Forsyth	@George_Forsyth
Jhony Lescano	@yonhy_lescano
Hernando de Soto	@HDeSotoPeru
Daniel Urresti	@DanielUrresti1
Verónica Mendoza	@Vero_Mendoza_F
Alberto Beingolea	@BeingoleaA
Keiko Fujimori	@KeikoFujimori
José Vega	@EquipoVega
Rafael Lopez Aliaga	@rlopezaliaga1
Ciro Gálvez	@Ciro_GalvezH
Daniel Salaverry	@dsalaverryv
Pedro Castillo	@PedroCastilloTe
Andrés Alcantara	@aalacantarape
César Acuña	@CesarAcunaP

Para la primera vuelta electoral, el conjunto de datos recolectado estuvo compuesto por 81214 tweets, a los cuales al aplicar la limpieza y normalización de los datos, se redujo a 11657 tweets, donde se consideró solo las publicaciones hechas desde el 01 de marzo de 2021 hasta el 29 de marzo de 2021 (desde el inicio del último mes de campaña hasta el inicio del debate electoral de los candidatos), estos datos se pueden encontrar en la siguiente link: https://github.com/TheDenky/Twitter-Election-Predict-Peru/blob/main/datosPrimeraCompleto.csv.

Para la segunda vuelta electoral, el conjunto de datos recolectado estuvo compuesto por 15007 tweets, a los cuales al aplicar la limpieza y normalización de los datos, se redujo a 4728 tweets,donde se consideró solo las publicaciones hechas desde el 01 de mayo de 2021 hasta el 30 de mayo de 2021 (desde el inicio del último mes de campaña hasta el inicio del debate electoral de los candidatos), estos datos se pueden encontrar en la siguiente link: https://github.com/TheDenky/Twitter-Election-Predict-Peru/blob/main/datosSegundaCompleto.csv.

4.3. Preprocesamiento de Datos

Para la preparación y el preprocesamiento de datos se tomó como referencia el trabajo realizado por Ruíz (2021), donde cuenta con algunos aspectos imporatantes a considerar para la obtención de buenos resultados de predicción.

En la Tabla 7 se detalla los aspectos considerados en este trabajo para la limpieza y normalización de los datos. Estos aspectos fueron utilizados para los distintos conjuntos de datos, tanto para el entrenamiento con los datos TASS2016 como también para el conjunto de datos de las elecciones presidenciales de Perú.

Tabla 7

Aspectos de Limpieza y Normalización de Datos de Entrenamiento

ASPECTO	DESCRIPCIÓN
Eliminación de tweets de la misma persona	Aquellos tweets que fueron publicados por una misma persona, solo se tomó en cuenta el último tweet publicado y el resto se eliminó.
Eliminación de tweets duplicados	Eliminar tweets repetidos en el conjunto de datos.
Eliminación de tweets sin mención	Eliminar tweets que dentro del contenido no tenga una mención a algún candidato mediante su cuenta oficial (Solo aplica para el conjunto de datos de elecciones de Perú).
Eliminación de signos de puntuación	Al eliminar los signos de puntuación, se reemplazó por espacios en blanco para no afectar el texto por algunas palabras que podrían juntarse. Ejemplos (.;;¡!¿?(){}[]\/”*-).
Eliminación de enlaces, links y hashtags y menciones	Los enlaces, links, hashtags y menciones se reemplazaron por espacios en blanco. Ejemplos (palabras que empiezan por ‘http’, ‘#’ y ‘@’).
Corrección abreviaciones	Se consideró las siguientes abreviaciones por ser los más comunes: ('d': 'de', 'x': 'por', 'xa': 'para', 'as': 'has', 'q': 'que', 'k': 'que', 'dl': 'del', 'xq': 'porqué', 'dr': 'doctor', 'dra': 'doctora', 'sr': 'señor', 'sra': 'señora', 'm': 'me').

Normalización de risas	Para poder normalizar las risas se consideró utilizar las únicas formas de: (‘jajaja’,’jejeje’,’jijiji’,’jojojo’,’jujuju’).
Sentimiento de emojis	Se utilizó el diccionario Emosent para definir el sentimiento expresado por los emojis. Ejemplo: (😉 : sentimiento positivo, 😐 : sentimiento neutro, ☹️ : sentimiento negativo).
Eliminación de stopwords	Se eliminan aquellas palabras que no agregan información en la oración. Pero también se conservan aquellas que si aporten información como: ('no', 'ni', 'poco', 'mucho', 'nada', 'muchos', 'muy', 'nosotros', 'nosotras', 'vosotros', 'vosotras', 'ellos', 'ellas', 'ella', 'él', 'tu', 'tú', 'yo', 'pero', 'hasta', 'contra', 'por').
Convertir todo el texto en minúsculas	Se realiza para que las letras mayúsculas y minúsculas no interfieran en el significado de las palabras, y tener un solo tipo de letra en todos los tweets.
Lematización	Se convierte las diferentes conjugaciones y vocablos de una palabra en su forma canónica. Para este fin se utiliza el componente ‘lemmatizer’ del modelo ‘es_core_news_md’ disponible en Spacy Python. Ejemplo: (bailaste, bailamos, bailo, bailarás = bailar).

4.4. Implementación del Modelo

4.4.1. Características de entrenamiento

Para obtener las características de entrenamiento del modelo, primeramente, se hizo uso de un tokenizador con el que se obtiene la cantidad total de palabras distintas del set de entrenamiento, logrando obtener 7194 palabras, posteriormente, se utilizó la librería 'pad_sequences' para generar una secuencia de la característica de cada tweet, con una longitud máxima de 25 dígitos.

Para la elección de los datos de entrenamiento y prueba se realizó al azar con la librería 'train_test_split' con el parámetro 'random_state' igual a 42. Seguidamente, para el entrenamiento del modelo se utilizó el 85% de datos de la base de datos TASS2016, quedando así el 15% para realizar la prueba o test.

4.4.2. Arquitectura

Para la implementación se utilizó un modelo "sequential" con tres capas: Embedding, Bidireccional LSTM y Dense. En la Tabla 8 se detalla las características del modelo.

Tabla 8

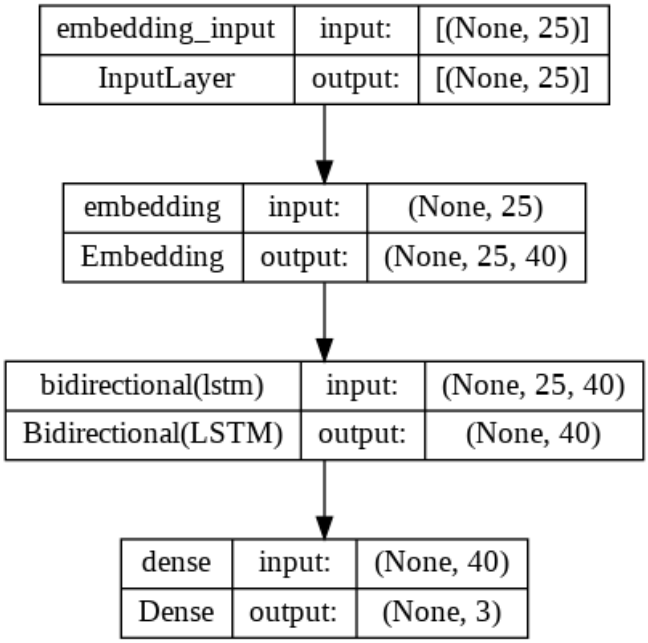
Arquitectura del Modelo

Layer (Type)	Output Shape	Param #
embedding (Embedding)	(None, 25, 40)	508680
bidirectional (Bidirectional)	(None, 40)	9760
dense (Dense)	(None, 3)	123
Total Params: 518,563 Trainable params: 518,563 No trainable params: 0		

Para comprender mejor la arquitectura del modelo, en la Figura 11 se muestra el diagrama de la arquitectura del modelo de predicción construido. Se puede observar que en la capa de entrada inicia con 25 nodos, los cuales son las características extraidas del tweet, y la salida son 3, los cuales pertenecen a la polaridad calculada como positivo, neutro y negativo.

Figura 11

Diagrama de la Arquitectura del Modelo



4.4.3. Evaluación

Luego, se analizó si el modelo de predicción entrenado cumple las expectativas de eficiencia y/o precisión. En la tabla 9 se muestra los diversos modelos probados y la precisión que obtuvieron cada uno de ellos, para este fin, se utilizó de base el trabajo de investigación realizado por Ruíz (2021). Como resultado, la red LSTM Bidireccional utilizado entre trabajo obtuvo el mejor resultado de predicción, obteniendo un 61.57% de precisión.

Tabla 9

Precisión de Distintos Modelos de Predicción Probados

Modelo	Precisión
NN GRU layers	0.5991
NN LSTM layers	0.6056
RNN Bidirectional	0.6157
RNN Conv1d	0.3787
RandForest CountVectorizer	0.5491
LinReg CountVectorizer	0.5787
Linear SVC CountVectorizer	0.4843
RandForest TFIDF	0.5704
LinReg TFIDF	0.6093
Linear SVC TFIDF	0.6148

El resultado obtenido con el modelo de predicción planteado fue de 61.57% de precisión, siendo un resultado bueno en comparación con lo que se tiene de literatura para el conjunto de datos interTASS2016 usado en este trabajo. En la Figura 12 muestra los resultados de precisión obtenidos por distintos equipos en competencia, con el mismo conjunto de datos.

Figura 12

Posición y Resultados de Distintos Equipos en Competencia

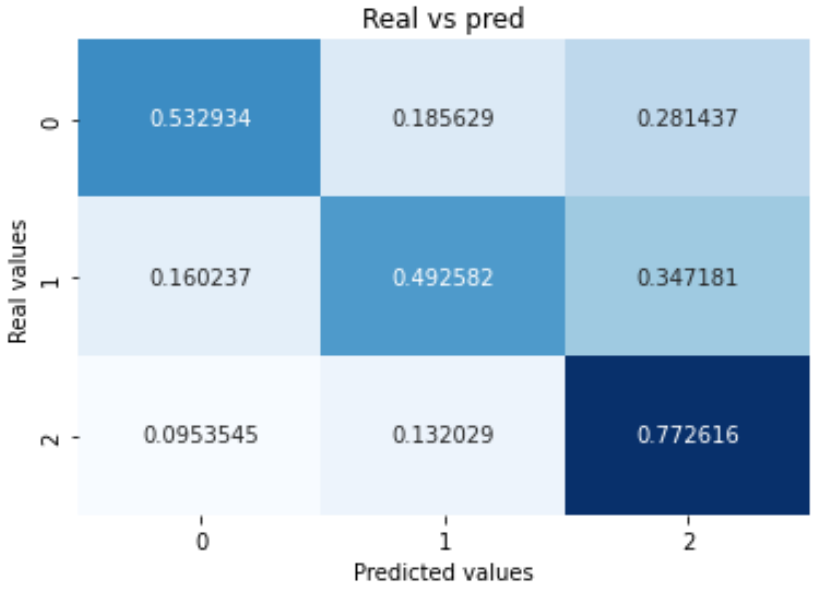
#	△	Team	Members	Score	Entries	Last	Code
1	—	Harvard-cete		0.62255	14	4Y	
2	▲ 3	Spice Girls del Espasio Ecsterior		0.61127	16	4Y	
3	—	Transcriptores		0.60375	26	4Y	
4	▼ 2	Frajumae		0.60000	27	4Y	
5	▼ 1	mahalanoobs		0.59248	36	4Y	
6	▲ 1	Monkeys with Laptops		0.59097	30	4Y	
7	▼ 1	Equipo B		0.57969	48	4Y	
8	—	Matrix Team		0.47969	14	4Y	

Nota. Reproducida del Portal Web Kaggle, Leaderboard de *Clasificación de Opiniones en Tweets en Español*, 2022, (https://www.kaggle.com/competitions/ssatwitter/leaderboard?tab=public).

En la Figura 13 se muestra la matriz de confusión del modelo escogido para la utilización en el presente trabajo. Se puede observar que los tweets positivos reales (fila N°2), son los que mejor predicción tuvieron con un 77.26% de precisión el cual es bueno debido a que se utiliza la mayor cantidad posible de tweets positivos (como se consideró en la metodología); y los de peor predicción fueron los tweets neutros reales (fila N°1) con 49.25% de precisión, este último no afecta en la metodología propuesta ya que son tweets neutros.

Figura 13

Matriz de Confusión del Modelo Entrenado con RNN Bidireccional



4.5. Procesamiento de Datos

Después de haber implementado y evaluado el modelo de predicción de sentimientos, haciendo uso del conjunto de datos recolectados de las elecciones presidenciales de Perú 2021 se hizo la clasificación de sentimientos según a la polaridad. En la Tabla 10 se muestra los resultados de clasificación de polaridad de tweets para la primera vuelta electoral.

Tabla 10

Clasificación de Polaridad de Tweets Primera Vuelta

Polaridad	Frecuencia
Positivo	5265
Negativo	3232
Neutro	3160
Total	**11657**

Utilizando solo los tweets positivos, en la Tabla 11 se muestra a detalle la cantidad de tweets positivos para cada candidato en primera vuelta.

Tabla 11

Tweets Positivos por cada Candidato Primera Vuelta

Candidato	Tweets Positivos
@Ollanta_HumalaT	251
@vozdelatierra	135
@julioguzmanperu	424
@rafaelsantosn	72
@George_Forsyth	218
@yonhy_lescano	189
@HDeSotoPeru	418
@DanielUrresti1	377
@Vero_Mendoza_F	506
@BeingoleaA	339
@KeikoFujimori	271
@rlopezaliaga1	798
@dsalaverryv	229
@PedroCastilloTe	6
@aalacantarape	7
@CesarAcunaP	1025

En la Tabla 12 se muestra los resultados de clasificación de polaridad de tweets para la segunda vuelta electoral.

Tabla 12

Clasificación de Polaridad de Tweets Segunda Vuelta

Polaridad	Frecuencia
Positivo	1495
Negativo	1232
Neutro	2001
Total	**4728**

Utilizando solo los tweets positivos, en la Tabla 13 se muestra a detalle la cantidad de tweets positivos para cada candidato en segunda vuelta.

Tabla 13

Tweets Positivos por cada Candidato Segunda Vuelta

Candidato	**Tweets Positivos**
@PedroCastilloTe	830
@KeikoFujimori	746

Utilizando estos datos, ya se puede pasar a comprobar la eficiencia del modelo predictivo planteado haciendo el comparativo con los sondeos obtenidos por las diversas encuestadoras, utilizando las diversas medidas de error planteados.

CAPITULO V: EXPERIMENTOS Y RESULTADOS

Después de realizar el entrenamiento y prueba del modelo de clasificación planteado, se procede a mostrar los resultados que se obtuvo al aplicar el modelo predictivo con cada uno de los conjuntos de datos de los procesos electorales, tanto para la primera vuelta como para la segunda vuelta.

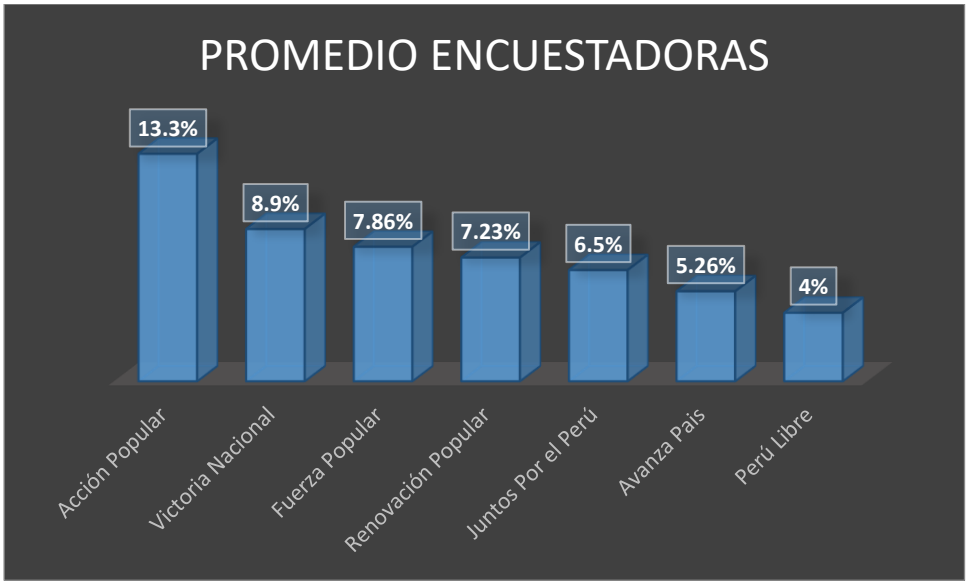
5.1. Elecciones de Primera Vuelta

Primeramente, se muestran los diferentes resultados estadísticos tanto de las encuestadoras como del modelo, así también el resultado oficial obtenidos de las elecciones de primera vuelta.

En la Figura 14, se muestra el gráfico de barras del promedio estadístico de las encuestadoras en la primera vuelta, para ello se utilizó los resultados de DATUM, CPI E IPSOS, mediante la suma de los resultados de las tres encuestadoras divididos entre tres.

Figura 14

Promedio Estadístico de las Encuestadoras DATUM, CPI e IPSOS en Primera Vuelta



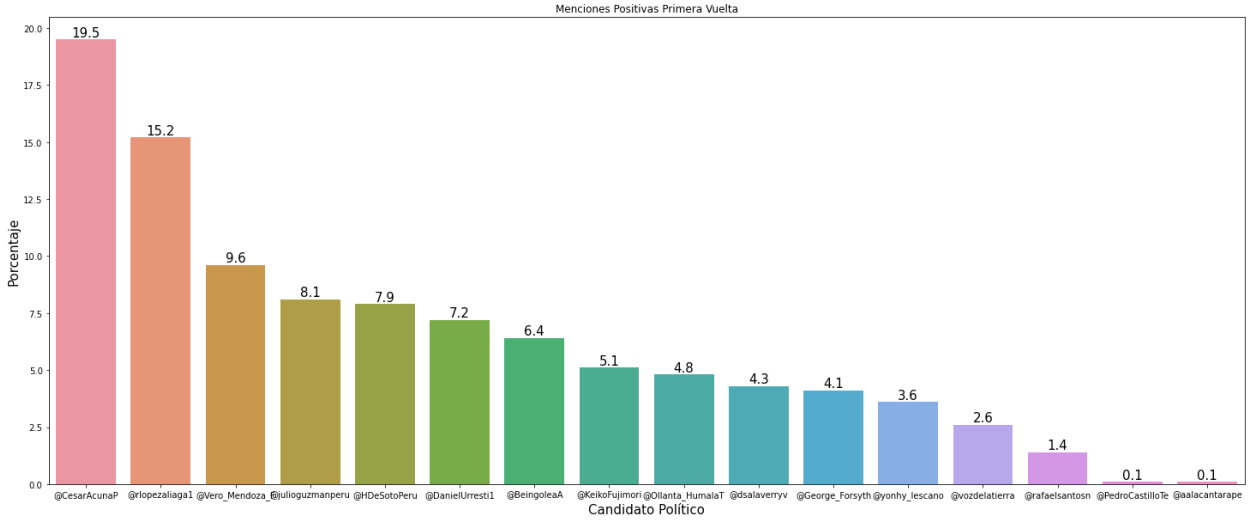
Nota. Adaptada del Portal Encuestas, Encuestas 1ra Vuelta, 2021, *(*https://encuestas.com.pe/).

Se puede observar que el partido Acción Popular tiene la mayor cantidad de aceptación según las tres encuestadoras, seguidamente del partido Victoria Nacional, Fuerza Popular y Renovación Popular; También se puede observar que el partido Perú Libre obtuvo, según dos de las encuestadoras, en promedio el 4% de aceptación, por lo que está ubicado en el séptimo lugar en las intenciones de voto.

Seguidamente, en el gráfico de barras de la Figura 15 se muestran los resultados estadísticos obtenidos mediante el modelo de predicción planteado. Según al resultado del modelo predictivo, el candidato de mayor aceptación popular es César Acuña, con un 19.5% de aceptación, seguido de Rafael López y Verónika Mendoza, esto puede ser debido a algún evento relacionado con el dicho candidato, el cual hizo que se volviera tendencia en la red social.

Figura 15

Resultado Estadístico del Modelo de Predicción Planteado en Primera Vuelta



El 16 de abril, la Oficina Nacional de Procesos Electorales, dio los resultados oficiales de la primera contienda electoral presidencial. En la Figura 16, se muestra el porcentaje real obtenido por cada candidato, los electores hábiles y la cantidad de participación ciudadana.

Figura 16

Resultados Oficiales de las Elecciones de Primera Vuelta

Nota. Reproducida de Resultados Elecciones Generales y Parlamento Andino, ONPE, 2021 (https://resultadoshistorico.onpe.gob.pe/EG2021/EleccionesPresidenciales/RePres/T).

Ahora, tomando en cuenta los resultados de las encuestadoras, el modelo predictivo y los resultados reales obtenidos, se procedió a sacar las distintas medidas de error para cada una de ellas, para luego hacer una comparativa general. Para dicho fin, se utilizó dos puntos de vista o enfoques en los tres casos anteriores, el primero fue tomado según al porcentaje obtenido y el segundo fue tomado según a la posición ocupada por cada candidato electoral.

5.1.1. Resultados Según Porcentaje

5.1.1.1. Resultado General

En la Tabla 14, se muestra el cuadro para analizar los resultados obtenidos del promedio de las encuestadoras, resultados del modelo de predicción y los resultados reales oficiales según el porcentaje.

Tabla 14

Cuadro Comparativo de los Distintos Resultados Según Porcentaje en Primera Vuelta

Candidato Político	Encuesta	Modelo	Real
Yonhy Lescano	13.3%	3.6%	9.071%
George Forsyth	8.9%	4.1%	5.656%
Keiko Fujimori	7.86%	5.1%	13.407%
Rafael López	7.23%	15.2%	11.751%
Verónica Mendoza	6.5%	9.6%	7.865%
Hernando de Soto	5.26%	7.9%	11.626%
Pedro Castillo	4%	0.1%	18.921%

Cabe mencionar que, la cantidad de candidatos considerados son siete debido a que solo se tomó en cuenta aquellos candidatos que aparecen en las encuestas electorales. Además, no se cuenta con los datos estadísticos de las encuestas de los demás candidatos no mencionados, por lo que no se puede realizar las medidas de error de manera total. Seguidamente, se procede a mostrar los resultados de los distintos niveles de error.

En primer lugar, se calculó los niveles de error MAE y RMSE utilizando el promedio de sondeos realizados por las encuestadoras en comparación con los resultados reales y oficiales entregados por la ONPE, tal como se muestra en la Tabla 15.

Se puede observar que, la predicción más acertada que realizó las encuestadoras fue de la candidata *Verónika Mendoza* con un error absoluto de 1.365%; y la predicción peor acertada fue del candidato *Pedro Castillo* con un error absoluto de 14.921%.

Tabla 15

Nivel de Error MAE y RMSE de las Encuestas en Primera Vuelta

Candidatos	Encuestas	Real	Diferencia Absoluta %	Cuadrado de la Diferencia %
Yonhy Lescano	13.3%	9.071%	4.229	17.884441
George Forsyth	8.9%	5.656%	3.244	10.523536
Keiko Fujimori	7.86%	13.407%	5.547	30.769209
Rafael López	7.23%	11.751%	4.521	20.439441
Verónika Mendoza	6.5%	7.865%	1.365	1.863225
Hernando de Soto	5.26%	11.626%	6.366	40.525956
Pedro Castillo	4%	18.921%	14.921	222.636241
		Resultado	**5.742%**	**7.017%**
			MAE	**RMSE**

Del mismo modo, para el caso del modelo predictivo planteado se calculó los niveles de error MAE y RMSE en comparación con los resultados reales y oficiales entregados por la ONPE, y se muestran en la Tabla 16.

Se puede observar que, las predicciones más acertadas que realizó el modelo predictivo fue del candidato *George Forsyth* con un error absoluto de 1.556% y de la candidata *Verónika Mendoza* con un error absoluto de 1.735%; y la predicción peor acertada fue del candidato *Pedro Castillo* con un error absoluto de 18.821%.

Tabla 16

Nivel de Error MAE y RMSE del Modelo Predictivo en Primera Vuelta

Candidatos	Modelo	Real	Diferencia Absoluta %	Cuadrado de la Diferencia %
Yonhy Lescano	3.6%	9.071%	5.471	29.931841
George Forsyth	4.1%	5.656%	1.556	2.421136
Keiko Fujimori	5.1%	13.407%	8.307	69.006249
Rafael López	15.2%	11.751%	3.449	11.895601
Verónica Mendoza	9.6%	7.865%	1.735	3.010225
Hernando de Soto	7.9%	11.626%	3.726	13.883076
Pedro Castillo	0.1%	18.921%	18.821	354.230041
		Resultado	**6.152%**	**8.318%**
			MAE	**RMSE**

En la tabla 17 se muestra un cuadro comparativo entre los distintos tipos de error tanto de las encuestadoras como del modelo predictivo.

Para el caso del Error Absoluto Medio (MAE), los resultados para ambos tipos tienen una diferencia de 0.383%, esto quiere decir que el error es bajo tomando en cuenta la magnitud de error que hubo con los resultados reales.

Para el caso de la Raíz del Error Cuadrático Medio (RMSE), los resultados para ambos tipos tienen una diferencia de 1.301%, la razón por la que es mayor es debido a que al elevar al cuadrado la diferencia absoluta, el error más grande toma mayor importancia y crece

exponencialmente; Esto quiere decir que en algunos candidatos el modelo predictivo tiene errores de mayor proporción a comparación de las encuestadoras.

Tabla 17

Cuadro Comparativo de Error General Según Porcentaje en Primera Vuelta

Error General		
TIPO	MAE	RMSE
Encuestadoras	5.742%	7.017%
Modelo Predictivo	6.152%	8.318%

A continuación, en la tabla 18 se muestra a manera de experimento un cuadro comparativo entre los distintos tipos de error tanto de las encuestadoras como del modelo predictivo, pero sin considerar al candidato *Pedro Castillo*. La decisión de no considerar dicho candidato es debido a que, para ambos tipos, el nivel de error es grande e influye mucho en los resultados finales de error.

En el resultado del Error Absoluto Medio, se puede notar una reducción de error considerable con respecto al caso anterior, también un aspecto importante es que el modelo predictivo obtuvo un **menor error** a comparación de las encuestadoras, que era todo lo contrario en el caso anterior.

En el resultado de la Raíz del Error Cuadrático Medio también existe una reducción de error considerable, pero la diferencia con el error absoluto medio es que, como el caso anterior, las encuestadoras siguen teniendo menor error, aunque la diferencia sea muy mínima con 0.148%. Esto quiere decir que el modelo predictivo sigue teniendo en algunos candidatos errores de mayor proporción a comparación de las encuestadoras.

Tabla 18

Cuadro Comparativo de Error General sin Pedro Castillo en Primera Vuelta

Error General sin Pedro Castillo		
TIPO	MAE	RMSE
Encuestadoras	4.212%	4.509%
Modelo Predictivo	4.041%	4.657%

5.1.1.2. Resultados Individuales

Otro experimento realizado fue obtener los niveles de error, para cada una de las tres encuestadoras tomadas en cuenta, de manera individual. En la Tabla 19 se muestra los resultados de los errores individuales para las encuestadoras CPI, DATUM e IPSOS y del modelo predictivo.

Para el caso de la encuestadora CPI, en el error absoluto medio, hay una diferencia de 0.232%, por lo cual los resultados son cercanos en comparación con el modelo predictivo; en la raíz del error cuadrático medio hay una diferencia de 1.117%, quiere decir que el modelo predictivo tiene algunos datos con errores de mayor proporción que las encuestadoras.

Para el caso de la encuestadora DATUM, en ambas medidas de error, tanto MAE como RMSE, existe una considerable diferencia, por lo tanto, la encuestadora realizó una mejor predicción.

Para el caso de la encuestadora IPSOS, en el error absoluto medio, existe una diferencia de 0.608%, a pesar que el nivel de error es menor que las encuestadoras anteriores, el modelo predictivo obtuvo un mejor resultado; en la raíz del error cuadrático medio hay una diferencia de 0.61 a favor del modelo predictivo obteniendo un **mejor resultado**, también esto quiere decir que la encuestadora IPSOS tuvo errores de mayor proporción a comparación del modelo predictivo.

Tabla 19

Nivel de Error Individual para las Encuestadoras y Modelo Predictivo

Errores Individuales por cada Encuestadora		
TIPO	**MAE**	**RMSE**
CPI	5.920%	7.201%
Modelo Predictivo	6.152%	8.318%
DATUM	5.335%	6.823%
Modelo Predictivo	6.152%	8.318%
IPSOS	4.649%	5.267%
Modelo Predictivo	4.041%	4.657%

5.1.2. Resultados Según Posición

El siguiente enfoque considerado fue obtener los resultados de los niveles de error de las encuestadoras y el modelo predictivo, tomando en cuenta la posición obtenida; Como el caso anterior, también se toma en cuenta para los resultados en general y los resultados de manera individual.

5.1.2.1. Resultado General

En la Tabla 20, se muestra el cuadro para analizar los resultados obtenidos del promedio de las encuestadoras, resultados del modelo de predicción y los resultados reales oficiales según la posición obtenida.

Tabla 20

Cuadro Comparativo de los Distintos Resultados Según Posición en Primera Vuelta

Candidato Político	Encuesta	Modelo	Real
Yonhy Lescano	1°	12°	5°
George Forsyth	2°	11°	8°
Keiko Fujimori	3°	7°	2°
Rafael López	4°	2°	3°
Verónica Mendoza	5°	3°	6°
Hernando de Soto	6°	5°	4°
Pedro Castillo	7°	15°	1°

Seguidamente, en la Tabla 21 se muestra el cuadro comparativo de los resultados obtenidos de los errores MAE y RMSE de las encuestadoras y el modelo predictivo.

En el caso del error absoluto medio, ambos tipos cuentan con un nivel de error bastante grande considerando que se utilizó datos de posición; También existe una diferencia de 1.857% entre ambos tipos, siendo el menor de estos de las encuestadoras, por lo que el modelo predictivo obtuvo resultados no muy prometedores.

Para el caso del error cuadrático medio, los resultados incluso empeoran mucho, en ambos casos existen un nivel de error elevado, pero lo que sobresale más es que el modelo predictivo tiene 6.437% de error, esto significa que en algunos casos tiene errores de proporción muy grandes.

Esto podría deberse a que no se está utilizando los datos de la cantidad total de candidatos en el caso de las encuestadoras ya que no se cuenta con estos, mientras que para el modelo predictivo si se cuenta con esos datos.

Tabla 21

Cuadro Comparativo de Error General Según Posición en Primera Vuelta

Error General Según Posición		
TIPO	MAE	RMSE
Encuestadoras	3.000	3.684
Modelo Predictivo	4.857	6.437

Como en el enfoque mediante porcentaje, también en este caso se consideró la obtención de niveles de error sin considerar al candidato Pedro Castillo, debido a que influye mucho en los resultados obtenidos; En la Tabla 22 se muestra dichos resultados.

Se puede observar que la cantidad de error para ambas medidas disminuyó considerablemente, destacando la del modelo predictivo; Por lo que se puede deducir que efectivamente los datos del candidato *Pedro Castillo* afecta considerablemente en la predicción de resultados de las encuestadoras y del modelo predictivo.

Tabla 22

Cuadro Comparativo de Error General sin Pedro Castillo en Primera Vuelta

Error General sin *Pedro Castillo* Según Posición		
TIPO	MAE	RMSE
Encuestadoras	2.500	3.136
Modelo Predictivo	3.333	3.958

5.1.2.2. Resultados Individuales

También se consideró la obtención de los niveles de error de manera individual según la posición obtenida como un experimento para analizar aquellas donde se cuenta con una mayor cantidad de error; En la Tabla 23 se muestra los resultados individuales.

Para el caso del error absoluto medio, la encuestadora que obtuvo el menor error fue DATUM, teniendo una ventaja grande en comparación al modelo predictivo; En comparación con la encuestadora IPSOS, el modelo predictivo disminuye el nivel de error considerablemente, esto debido a que dicha encuestadora no consideró al candidato *Pedro Castillo*.

En las medidas obtenidas de la raíz del error cuadrático medio se puede observar que el mejor resultado se obtuvo de la encuestadora IPSOS y del modelo predictivo, teniendo una diferencia pequeña; Como fue con el caso MAE es debido a que no considera al candidato *Pedro Castillo* como si lo hicieron CPI y DATUM, por lo que, principalmente afecta al modelo predictivo resultando un error elevado de desviación estándar de 6.437%.

Tabla 23

Nivel de Error Individual Según Posición en Primera Vuelta

Errores Individuales por cada Encuestadora		
TIPO	**MAE**	**RMSE**
CPI	3.286	3.798
Modelo Predictivo	4.857	6.437
DATUM	2.429	3.359
Modelo Predictivo	4.857	6.437
IPSOS	2.571	3.162
Modelo Predictivo	3.143	3.742

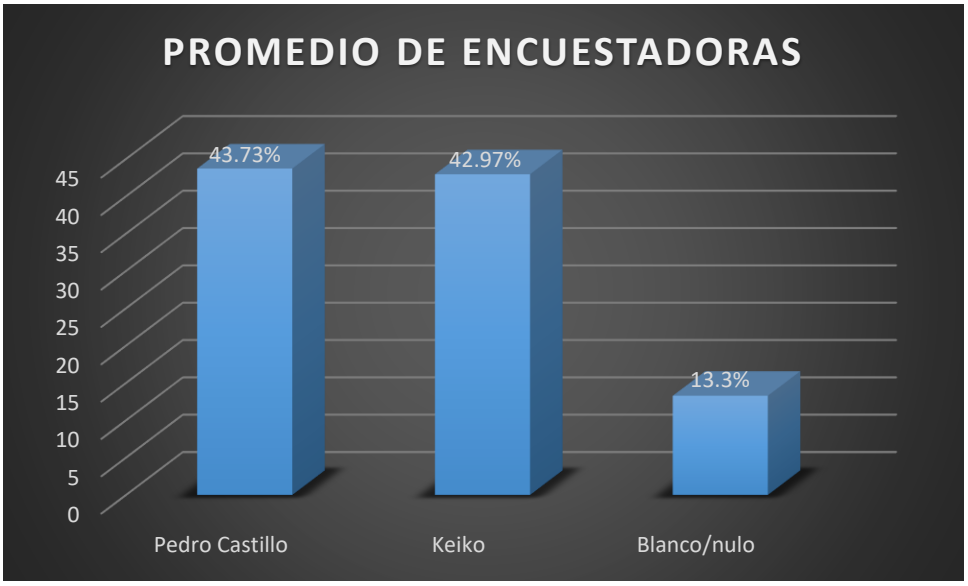
5.2. Elecciones de Segunda Vuelta

Seguidamente, se muestran los diferentes resultados estadísticos tanto de las encuestadoras como del modelo, así también el resultado oficial obtenidos de las elecciones de segunda vuelta.

En la Figura 17, se muestra el gráfico de barras del promedio estadístico de las encuestadoras en la segunda vuelta, para ello se utilizó los resultados de DATUM, IPSOS e IDICE mediante la suma de los resultados de las tres encuestadoras divididos entre tres.

Figura 17

Promedio Estadístico de las Encuestadoras DATUM, IPSOS e IDICE en Segunda Vuelta



Nota. Adaptado del Portal Web Encuestas, Encuestas Segunda Vuelta, 2021, (https://encuestas.com.pe).

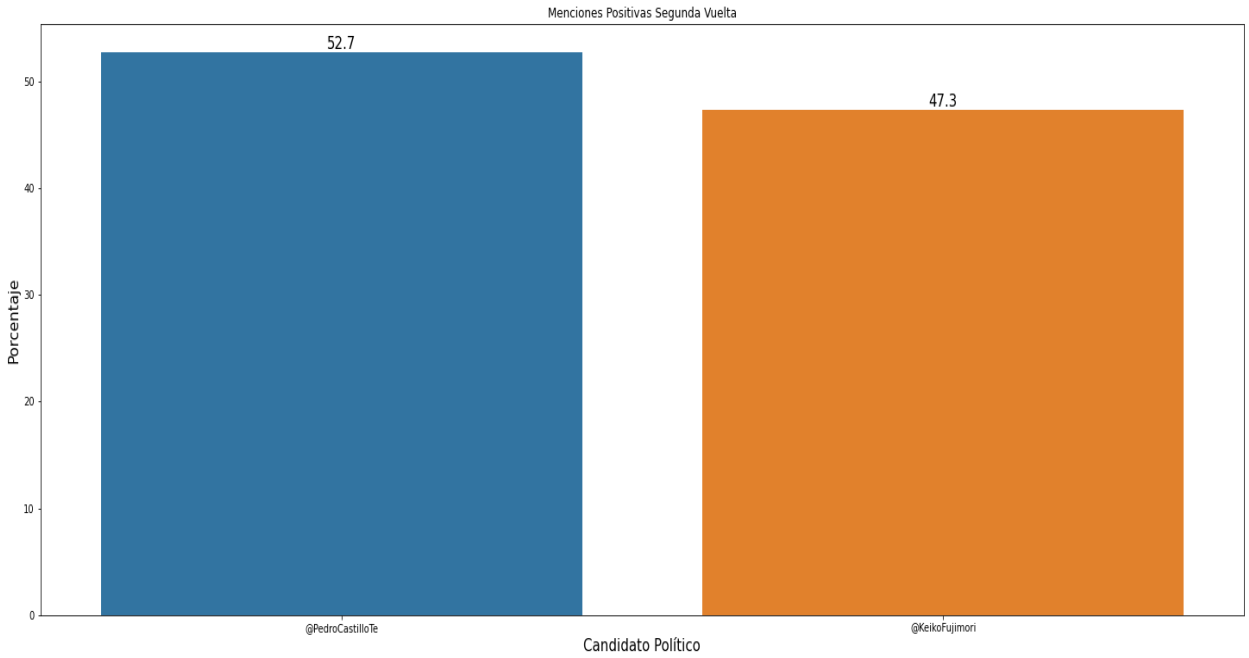
Según las encuestadoras, ambos partidos que pasaron a segunda vuelta, tienen un estimado casi cercano entre ambos, por lo que se puede considerar que existe un empate técnico. En el caso de los votos en blanco o nulos, no se consideró como un indicativo en el modelo predictivo, así que, para comparar los resultados de las encuestadoras con del modelo predictivo, se consideró la suma de las estimaciones de ambos partidos como el 100%.

Por otro lado, con el modelo de predicción que se implementó, en la Figura 18, muestra los resultados que se obtuvo con el modelo predictivo, aplicando la metodología propuesta y sin considerar aquellos votos nulos y blancos.

Como se puede observar, los resultados obtenidos con el modelo de predicción también son casi cercanas entre ellas, según la metodología utilizada en este trabajo, no fue considerado la obtención de otros resultados considerados por las encuestadoras como por ejemplo el porcentaje de votos blancos o nulos, pero tiene la información suficiente para obtener y comparar los resultados de los distintos niveles de error.

Figura 18

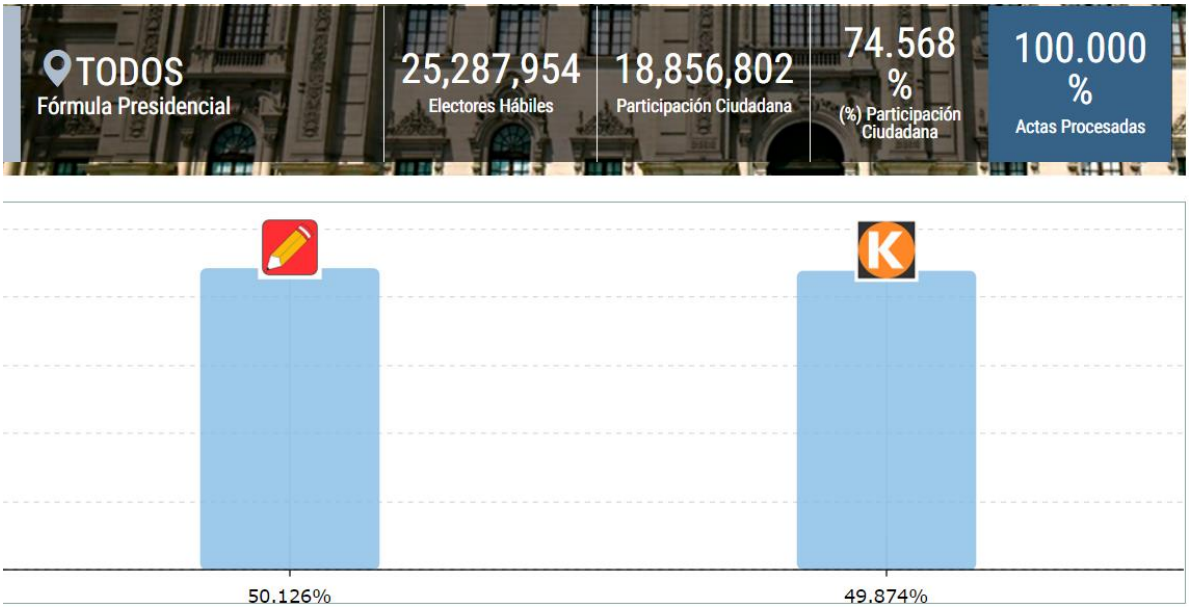
Resultado Estadístico del Modelo Predictivo en Segunda Vuelta



Por último, el 23 de junio, la Oficina Nacional de Procesos Electorales, dio los resultados oficiales de la primera contienda electoral presidencial. En la Figura 19, se muestra el porcentaje real obtenido por cada candidato, los electores hábiles y la cantidad de votos válidos realizados.

Figura 19

Resultados Oficiales de las Elecciones de Segunda Vuelta



Nota. Reproducida de Resultados Segunda Elección Presidencial, ONPE, 2021 (https://resultadoshistorico.onpe.gob.pe/EG2021/EleccionesPresidenciales/RePres/T).

Ahora, tomando en cuenta los resultados de las encuestadoras, el modelo predictivo y los resultados reales obtenidos, se procedió a sacar las distintas medidas de error para cada una de ellas, para luego hacer una comparativa general. Como para las elecciones de primera vuelta electoral, también se utilizó dos puntos de vista o enfoques, el primero según al porcentaje obtenido y el segundo según a la posición ocupada por cada candidato electoral.

5.2.1. Resultados Según Porcentaje

5.2.1.1. Resultado General

En la Tabla 24, se muestra el cuadro para analizar los resultados obtenidos del promedio de las encuestadoras, resultados del modelo de predicción y los resultados reales oficiales según el porcentaje.

Tabla 24

Cuadro Comparativo de los Distintos Resultados Según Porcentaje en Segunda Vuelta

Candidato Político	Encuesta	Modelo	Real
Pedro Castillo	50.442%	52.700%	50.126%
Keiko Fujimori	49.558%	47.300%	49.874%

En primer lugar, se calculó los niveles de error MAE y RMSE utilizando el promedio de sondeos realizados por las encuestadoras en comparación con los resultados reales y oficiales entregados por la ONPE, tal como se muestra en la Tabla 25.

En el cuadro se observa un aspecto importante de tomar en cuenta, que tanto el error absoluto medio como la raíz del error cuadrático medio tienen resultados iguales, esto es debido a que solo se cuenta con dos datos, por lo tanto, en las siguientes tablas se observará el mismo fenómeno.

Tabla 25

Nivel de Error MAE y RMSE de las Encuestas en Segunda Vuelta

Candidatos	Encuestas	Real	Diferencia Absoluta	Cuadrado de la Diferencia
Pedro Castillo	50.442%	50.126%	0.316%	0.326041%
Keiko Fujimori	49.558%	49.874%	0.316%	0.326041%
		Resultado	**0.316%**	**0.316%**
			MAE	**RMSE**

Del mismo modo, para el caso del modelo predictivo planteado se calculó los niveles de error MAE y RMSE en comparación con los resultados reales y oficiales entregados por la ONPE, y se muestran en la Tabla 26.

Tabla 26

Nivel de Error MAE y RMSE del Modelo Predictivo en Segunda Vuelta

Candidatos	Modelo	Real	Diferencia Absoluta	Cuadrado De La Diferencia
Pedro Castillo	52.700%	50.126%	2.574%	6.625472%
Keiko Fujimori	47.300%	49.874%	2.574%	6.625472%
		Resultado	**2.574%**	**2.574%**
			MAE	**RMSE**

En la tabla 27 se muestra un cuadro comparativo entre los distintos tipos de error tanto de las encuestadoras como del modelo predictivo.

Para en caso de las encuestadoras, las medidas de error tienen un valor de 0.316% por lo que es un valor de error bastante bajo, uno de los factores que influyó fue que la encuestadora IDICE consideró en su predicción que la candidata *Keiko Fujimori* obtendría mayor cantidad de aceptación popular que el candidato Pedro Castillo, por lo que favoreció en el promedio de las encuestadoras; Por esta razón, la mejor manera de analizar el resultado general es sacar el promedio de los niveles de error de las encuestadoras de manera individual.

Para el modelo predictivo se obtuvo un error de 2.574%, es una medida de error bastante aceptable ya que se trata solo de dos candidatos, también es una medida obtenida con un conjunto de datos reducido.

Tabla 27

Cuadro Comparativo de Error General Según Porcentaje en Segunda Vuelta

Error General		
TIPO	MAE	RMSE
Encuestadoras	0.316%	0.316%
Modelo	2.574%	2.574%

5.2.1.2. Resultados Individuales

A continuación, en la Tabla 28 se muestra los resultados de los errores de manera individual para las encuestadoras DATUM, IDICE e IPSOS y del modelo predictivo.

Para analizar de mejor manera los resultados, se consideró otra medida de error denominada Error Medio (ME), debido a que las demás no consideran los errores negativos, el cual es importante en este caso.

Se puede observar que la encuestadora DATUM obtuvo el mejor resultado en comparación de las demás encuestadoras, por otra parte, la encuestadora IPSOS obtuvo el resultado más alejado, pero esto no quiere decir que fue la peor, ya que si se observa el error ME de la encuestadora IDICE, tiene un valor negativo, esto quiere decir que aunque la medida de error es más baja que la encuestadora IPSOS y el modelo predictivo, el resultado es peor y opuesto a ambos ya que predijo que la ganadora de la contienda electoral era la candidata *Keiko Fujimori*. En el caso del modelo predictivo, el resultado es aceptable ya que es cercano al error obtenida por una de las encuestadoras.

Por lo anterior, el promedio general de las medidas de error de las encuestadoras es **1.509%**, este es el resultado de sumar los errores individuales y dividirlos entre tres.

Tabla 28

Nivel de Error Individual Según Porcentaje en Segunda Vuelta

Errores Individuales por cada Encuestadora			
TIPO	**MAE**	**RMSE**	**ME**
DATUM	0.408%	0.408%	0.408%
IDICE	1.736%	1.735%	-1.735%
IPSOS	2.383%	2.383%	2.383%
MODELO	2.574%	2.574%	2.574%

5.2.2. Resultados Según Posición

El siguiente enfoque, como en el caso de primera vuelta, fue obtener los resultados de los niveles de error de las encuestadoras y el modelo predictivo, tomando en cuenta la posición obtenida; También se tomó en cuenta para los resultados en general y los resultados de manera individual.

5.2.2.1. Resultado General

Para obtener el resultado general, se consideró los datos mostrados en la tabla 29, el cual muestra las predicciones realizadas de cada encuestadora y del modelo predictivo según la posición, para tomar en cuenta en el resultado general.

Tabla 29

Datos de Posición Individual en Segunda Vuelta

DATOS DE POSICIÓN INDIVIDUALES		
TIPO	**Pedro Castillo**	**Keiko Fujimori**
DATUM	1°	2°
IDICE	2°	1°
IPSOS	1°	2°
MODELO	1°	2°

En la Tabla 30, se muestra el cuadro de los resultados obtenidos, tomando en cuenta la Tabla 23, para analizar los resultados según la posición y de manera general.

Tabla 30

Cuadro Comparativo de los Distintos Resultados Según Posición en Segunda Vuelta

CANDIDATO POLÍTICO	ENCUESTA	MODELO	REAL
Pedro Castillo	1°	1°	1°
Keiko Fujimori	2°	2°	2°

En primer lugar, se calculó los niveles de error según posición de las encuestadoras en comparación con resultados reales, como se muestra en la Tabla 31.

Tabla 31

Nivel de Error de las Encuestas Según Posición en Segunda Vuelta

Candidatos	Encuestas	Real	Diferencia Absoluta	Cuadrado de la Diferencia
Pedro Castillo	1°	1°	0	0
Keiko Fujimori	2°	2°	0	0
		Resultado	**0.0%**	**0.0%**
			MAE	**RMSE**

Seguidamente, se calculó los niveles de error según posición del modelo predictivo en comparación con resultados reales, como se muestra en la Tabla 32.

Tabla 32

Nivel de Error del Modelo Predictivo Según Posición en Segunda Vuelta

Candidatos	Modelo	Real	Diferencia Absoluta	Cuadrado De La Diferencia
Pedro Castillo	1°	1°	0	0
Keiko Fujimori	2°	2°	0	0
		Resultado	**0.0%**	**0.0%**
			MAE	**RMSE**

En la tabla 33 se muestra un cuadro comparativo entre los distintos tipos de error, según la posición, tanto de las encuestadoras como del modelo predictivo.

Se puede observar que los niveles de error para ambos tipos es 0.0%, esto es debido a que en ambos casos la predicción según la posición fue acertada ya que dieron como ganador al candidato *Pedro Castillo*.

Tabla 33

Cuadro Comparativo de Error General Según Posición en Segunda Vuelta

Error General		
TIPO	MAE	RMSE
Encuestadoras	0.0%	0.0%
Modelo Predictivo	0.0%	0.0%

5.2.2.2. Resultados Individuales

Finalmente, en la Tabla 34 se muestran los resultados individuales de los niveles de error. Según estos resultados, se observa que las encuestadoras DATUM e IPSOS tuvieron un error de 0.0 al igual que el modelo predictivo, esto significa que acertaron en la predicción del ganador de la contienda electoral. En el caso de la encuestadora IDICE, obtuvo un error de 1.0, esto significa que tuvo un fallo del 100% en la predicción del resultado según la posición.

Tabla 34

Nivel de Error Individual Según Posición en Segunda Vuelta

Errores Individuales por cada Encuestadora		
TIPO	**MAE**	**RMSE**
DATUM	0.0	0.0
IDICE	1.0	1.0
IPSOS	0.0	0.0
MODELO	0.0	0.0

5.3. Contrastación de Hipótesis

5.3.1. Hipótesis Específicas

Hipótesis Específica 1

H1: Es posible determinar la eficiencia de un modelo predictivo de análisis de sentimientos de tweets en español.

Según la Tabla 9, el mejor resultado de la eficiencia del modelo predictivo es **61.57%** de exactitud, utilizando Redes Neuronales Recurrentes con capa Bidireccional o LSTM. Haciendo una comparativa del score obtenido por los distintos equipos de competencia mostrados en la Figura 12, se puede confirmar que el modelo predictivo implementado en este trabajo obtuvo un resultado de clasificación de sentimientos eficiente. Por lo tanto, la hipótesis H1 es verdadera.

Hipótesis Específica 2

H2: Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un error RMSE y MAE en promedio menor de 10% en primera vuelta.

En la Tabla 17 muestra que el modelo predictivo obtuvo **8.318% y 6.152%** de error RMSE y MAE respectivamente, pero incluso este resultado se puede mejorar aún más al dejar de considerar aquel dato de mayor proporción de error, tal como se muestra en la Tabla 18, obteniendo **4.657% y 4.041%** de error RMSE y MAE respectivamente. Por todo lo anterior, la hipótesis H2 es verdadera.

Hipótesis Específica 3

H3: Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un error RMSE y MAE en promedio menor de 5% en segunda vuelta.

En la Tabla 26 se muestra el nivel de error del modelo predictivo en segunda vuelta, obteniendo un error RMSE y MAE de **2.574%**. Por lo tanto, la hipótesis H3 es verdadera.

Hipótesis Específica 4

H4: Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un error RMSE y MAE en promedio menor en comparación con las encuestadoras IPSOS, CPI y DATUM en primera vuelta.

Según los resultados mostrados en la Tabla 18, al eliminar el dato de mayor proporción de error, se consiguió obtener un resultado MAE y RMSE de **4.041% y 4.657%** para el modelo predictivo y **4.212% y 4.509%** para las encuestadoras respectivamente. Por lo tanto, la hipótesis H4 es verdadera.

Hipótesis Específica 5

H5: Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un error RMSE y MAE en promedio menor en comparación con las encuestadoras IPSOS, DATUM e IDICE en segunda vuelta.

En la tabla 28 se muestra los errores individuales de las encuestadoras y el modelo predictivo, resultando como promedio de las encuestadoras un error MAE y RMSE de **1.509%** y **2.574%** del modelo predictivo, todo esto según el enfoque por porcentaje de aceptación. En el caso del otro enfoque por posición, en la Tabla 34 se muestra que el modelo predictivo obtuvo un error MAE y RMSE de **0.0%**, y en el caso de la encuestadora IDICE tuvo un error MAE y RMSE de **1.0**, por lo que **en promedio** las encuestadoras tuvieron un error mayor que el modelo predictivo. Por lo tanto, la hipótesis H5 es verdadera.

5.3.2. Hipótesis General

H0: Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un resultado significativo.

Al observar las Tablas 17 y 21 de los resultados de primera vuelta, se puede ver que las encuestadoras tienen un menor nivel de error a comparación del modelo predictivo, pero la diferencia entre ambos tipos no es mucha, por lo tanto, el resultado es significativo; Por otro lado, en las tablas 27 y 33 de los resultados de segunda vuelta, se puede notar que los niveles de error tanto de las encuestadoras como del modelo predictivo son muy bajas, y en el enfoque por posición el más confiable fue el modelo predictivo, por lo que fue un resultado significativo. Por lo tanto, la hipótesis H0 es verdadera.

CAPITULO V: DISCUSIÓN

En la implementación un clasificador de sentimientos mediante aprendizaje supervisado, haciendo uso de datos de la red social Twitter, realizado en Irlanda, consiguieron una precisión efectiva del 65% en la predicción de polaridad (Bermingham & Smeaton, 2011 citado en Arcila Calderón et al., 2017). En otro trabajo realizado en el mismo país, haciendo uso de 2634 tweets se implementó un clasificador de sentimientos de polaridad que además podía etiquetar el sarcasmo, logrando un 61.6% de precisión (Bakliwal et al., 2013 citado en Arcila Calderón et al., 2017). En el presente trabajo de investigación se logró obtener 61.57% de precisión, el cual, comparando con las investigaciones anteriores tiene un resultado aproximado y pudiendose considerar como bueno para realizar el análisis de sentimientos; cabe mensionar que para obtener mejores resultados además de aumentar la cantidad de datos etiquetados se debería mejorar la calidad de los mismos.

En los resultados obtenidos de Rodrigues et al. (2018) donde analizaron las elecciones en Chile, pudieron obtener en total 1380932 tweets para primera vuelta, teniendo mucha diferencia en comparacion con este trabajo que utilizó solamente 81214 tweets para la primera vuelta. Con estos datos, el trabajo hecho en Chile pudo obtener como mejor resultado de su modelo un error MAE de 6.35%, un error mayor si comparamos con el presente trabajo de investigación, que obtuvo un error MAE de 6.152%. En el trabajo realizado por Bansal & Srivastava (2018) utilizando un nuevo método Hibrido (HTBSA) para predecir las elecciones legislativas de estado de Uttar Pradesh, en el país de la India, pudieron obtener un MAE de 8.4% de error, un resultado mayor en comparación con el método propuesto en este trabajo. Si bien es cierto, hay muchos aspectos distintos que influyen para hacer una comparativa entre ambos, uno de ellos es el idioma, el cual en el trabajo hecho en la India utilizaron el idioma Hindi a comparación de este que utiliza el español, Otro aspecto es la cantidad de tweets que utilizaron, al ser una ciudad muy poblada, se

contó con 300000 tweets para realizar el análisis, una cantidad bastante considerable a diferencia de este trabajo que contó con 81214 tutis, y por último, la metodología aplicada fue por un modelo hibrido y en el caso de este trabajo fue la utilización de un modelo LSTM para realizar la clasificación de sentimientos.

Para el caso de la segunda vuelta electoral, en los resultados obtenidos de Rodrigues et al. (2018) donde analizaron las elecciones en Chile, recolectaron un total de 422169 tweets para la segunda vuelta, teniendo mucha diferencia en comparacion con este trabajo que utilizó 15007 tweets para la segunda vuelta. Con estos datos en la investigación de Chile se pudo obtener como mejor resultado de su modelo un error MAE de 0.51% en segunda vuelta, a comparación de este trabajo de investigación, que obtuvo un error MAE de 2.574%, por lo tanto, el modelo planteado por el otro investigador, tuvo mayor efectividad, pudiendose tratar por tener mayor cantidad de tweets recolectados de cada candidato en las elecciones de segunda vuelta.

En la investigación realizada por Callejas Cuervo & Velez Guerrero (2019) analizaron las eleccciones en Colombia, solo en primera vuelta, mediante el uso de un modelo hibrido propuesto, pudieron obtener un resultado de RMSE de 2.47% de error, siendo este resultado mejor que las encuestadoras; En el modelo propuesto en este trabajo se obtuvo 6.152% de error en primera vuelta, teniendo un resultado cercano al de las firmas encuestadoras, pero al hacer algunos ajustes del modelo como dejar de considerar al candidato *Pedro Castillo,* se obtuvo un error RMSE de 4.041%, superando al promedio de las firmas encuestadoras. La razón de que el resultado del error en el caso de las elecciones de Perú sea mayor, se puede tratar debido a que hubo mayor cantidad de candidatos electorales, de las cuales se consideró 16 de ellos, a diferencia que en el modelo utilizado en las elecciones de Colombia donde se consideró solamente 5 candidatos.

Finalmente, para el caso de realizar la comparativa de resultados en segunda vuelta con algunas firmas encuestadoras, en los trabajos tomados como referencia para los antecedentes de investigación todas se enfocan principalmente en la primera vuelta electoral, por lo tanto, no hay datos para realizar una comparativa entre ellos; Pero al observar los resultados de este trabajo, donde se obtuvo 2.574% de error y el de las encuestadoras un error promedio de 1.509% en la segunda vuelta electoral, se puede inferir que al tener que realizar un pronóstico para solamente dos candidatos, aunque el nivel de error del modelo predictivo disminuyó considerablemente en segunda vuelta, es mucho más efectiva los resultados obtenidos por las firmas encuestadoras.

CONCLUSIONES

1. En este trabajo de investigación, se evaluó la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tomando como medida los niveles de error RMSE y MAE para los procesos electorales de primera y segunda vuelta; Logrando obtener buenos resultados con errores cercanos al de las encuestadoras y mostrar una buena alternativa de predicción de resultados electorales simplemente algorítmica y de bajo costo.
2. Se logró determinar la eficiencia de un modelo predictivo de análisis de sentimientos de tweets en español, obteniendo un 61.57% de exactitud de los datos, el cual pertenece dentro de los mejores resultados obtenidos por otros equipos con el conjunto de datos utilizado.
3. Se logró evaluar el RMSE y MAE del modelo predictivo para las elecciones en Perú en primera vuelta, obteniendo un 8.318% y 6.152% de error respectivamente, siendo estos resultados menores al 10%, mejorando aún más el resultado al considerar otro factor, como la eliminación del candidato con mayor error que en este caso fue el candidato *Pedro Castillo*, y otro enfoque que fue la evaluación según la posición obtenida de los candidatos.
4. Se logró valuar el RMSE y MAE del modelo predictivo para las elecciones en Perú en segunda vuelta, obteniendo un 2.574% de error en ambos casos, siendo este resultado menor al 5%, por lo tanto, bastante eficiente y considerable en la predicción de resultados electorales en segunda vuelta.
5. Se comparó el nivel de error del modelo predictivo con las encuestadoras IPSOS, CPI y DATUM en primera vuelta, obteniendo que las encuestadoras CPI y DATUM obtuvieron mejores resultados que el modelo predictivo, pero cercanos entre ellos, y en el caso de la

encuestadora IPSOS obtuvo un error MAE de 4.649% un error mayor al del modelo predictivo que obtuvo 4.041%, siendo este resultado el menor de todos.

6. Se comparó el nivel de error del modelo predictivo con las encuestadoras IPSOS, DATUM e IDICE en segunda vuelta, donde el modelo de predicción obtuvo 2.574% de error y las encuestadoras DATUM e IPSOS errores de 0.408% y 2.83% respectivamente, siendo mejores que el modelo de predicción, en el caso de la encuestadora IDICE obtuvo un error medio negativo de -1.735%, por lo que el modelo de predicción obtuvo el mejor resultado.

RECOMENDACIONES

- Se recomienda implementar el modelo de predicción realizando un estudio “pre-hoc” ya que para este trabajo de investigación se usó el estudio “post-hoc” que quiere decir que se realizó después de ocurrido los procesos electorales, pero a futuro, se puede realizar un estudio longitudinal analizando los datos antes y después de realizarse las elecciones.
- Agregar mayor cantidad de datos para el entrenamiento del modelo predictivo de análisis de sentimientos con el fin obtener mejores resultados de clasificación de tweets, principalmente datos de fuentes confiables y mayormente relacionados al aspecto político.
- Disminuir el nivel de error promedio aceptable al 5% en las elecciones de primera vuelta, esto con el objetivo de buscar nuevas maneras o metodologías que permitan mejorar la obtención de resultados de predicción con datos más cercanos al real.
- Considerar la obtención del porcentaje aproximando de votos en blanco y nulos del modelo predictivo en las elecciones de segunda vuelta, debido a que las encuestadoras si consideran ese dato y puede influir en el aumento o disminución del nivel de error de la predicción.
- Mejorar el análisis de los datos agregando la segmentación de la población, tomando en consideración los departamentos o ciudades donde la mayor cantidad posible de electores utilicen la red social Twitter.
- Agregar la mayor cantidad de encuestadoras para tener una visión global del nivel de error de las encuestadoras tanto en conjunto como de manera individual en cada uno de los procesos electorales.

REFERENCIAS BIBLIOGRÁFICAS

Bohorquez Lopez, V. W., Mendez Lazarte, C., Altube, L., & Santana, E. (2019). Identificación del sentimiento expresado usando redes sociales en un contexto político. *Twenty-fifth Americas Conference on Information Systems*.

Aldis Siu. (30 de Marzo de 2021). Las #Encuestadoras #Corruptas en el #Perú - #DATUM ,#CPI , #IPSOS - Vladimiro Montesinos [video]. Youtube. Obtenido de https://www.youtube.com/watch?v=tw3gczUIbtI

Alsaeedi, A., & Zubair Khan, M. (2019). A Study on Sentiment Analysis Techniques of Twitter. *(IJACSA) International Journal of Advanced Computer Science and Applications, 10*(2), 361-374.

Arcila Calderón, C., Ortega Mohedano, F., Jiménez Amores, J., & Trullenque, S. (2017). Análisis supervisado de sentimientos políticos en español: clasificación en tiempo real de tweets basada en aprendizaje automático. *El profesional de la información*, 973-982.

Bansal, B., & Srivastava, S. (2018). On predicting elections with hybrid topic based sentiment analysis of tweets. *3rd International Conference on Computer Science and Computational Intelligence 2018.*, 346-353.

Bernal, C. A. (2010). *Metodología de la Investigación* (Tercera ed.). Colombia: Pearson.

Bravo Marquez, F., Mendoza, M., & Poblete, B. (5 de Junio de 2014). Meta-level sentiment models for big social data analysis. *KNOSYS 2857*, 14.

Callejas Cuervo , M., & Velez Guerrero, M. A. (2019). PREDICCIÓN ELECTORAL USANDO UN MODELO HÍBRIDO BASADO EN ANÁLISIS SENTIMENTAL: ELECCIONES PRESIDENCIALES DE COLOMBIA. *Revista Politécnica ISSN*(30), 94-104.

Catacora Flores, N. L. (Junio de 2021). El problema-justificación-objetivos: Material de curso. Andahuaylas, Apurímac, Perú. Obtenido de https://drive.google.com/file/d/14pCGkg5_I2k8AeF56qEtaNycMwRJFLUw/view

Catacora Flores, N. L. (Junio de 2021). Modelo de Plan de proyecto de tesis: un caso aplicado. Andahuaylas, Apurímac, Perú. Obtenido de https://drive.google.com/file/d/1HAeYLlSuH_VXPtmVsjOc41x6oPbhIKcY/view

Cedano, D., Risco, C., Insil, R., Rodrígues, C., & Ticlacuri, C. (2018). Sistema de encuesta virtual sobre los políticos en Perú basado en redes sociales. *PERSPECTIV@S. Revista de Tecnología e Información*(15), 61-69.

Chai, T., & Draxler, R. (30 de Junio de 2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geosci. Model Dev*.

Congreso de la República del Perú. (2021, 30 de Octubre). *Ley Orgánica de Elecciones N°26859.* Diario Oficial El Peruano. Obtenido de https://diariooficial.elperuano.pe/pdf/0000/ley-organica-de-elecciones-ley%20n26859.pdf

Espino Timón, C. (2017). "Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso. *(Trabajo de Fin de Grado).* Universidad Abierta de Cataluña, Cataluña, España.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. *Springer*, 986–996.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la Investigación* (Sexta ed.). México: McGRAW-HILL / INTERAMERICANA EDITORES, S.A.

Huillcen Baca, H. A., Palomino Valdivia, F., Ponce Atencio, Y., J. Ibarra, M., A. Cruz, M., & Huillcen Baca, M. E. (2021). CovidStream: Interactive Visualization of Emotions Evolution Associated with Covid-19. *Annual International Conference on Information Management and Big Data*, 540-551.

Instituto de Estudios Peruanos (IEP). (2021). *IEP Informe de Opinión – Junio II 2021*. Recuperado el Noviembre de 2022, de Encuesta telefónica a celulares a nivel nacional: https://iep.org.pe/wp-content/uploads/2021/06/Informe-IEP-OP-junio-II-2021-completo.pdf

Instituto Nacional de Estadística e Informática. (Julio de 2020). *INEI*. Obtenido de Estado de la población peruana 2020: https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1743/Libro.pdf

Kalechofsky, H. (Septiembre de 2016). *A Simple Framework for Building Predictive Models.* Obtenido de A Little Data Science Business Guide: https://www.msquared.com/wp-content/uploads/2017/01/A-Simple-Framework-for-Building-Predictive-Models.pdf

Khurana Batra, P., Saxena, A., Shruti, & Goel, C. (2020). Election Result Prediction Using Twitter Sentiments Analysis. *2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 182-185.

König, R. (2009). Predictive Techniques and Methods for Decision Support in Situations with Poor Data Quality. *Digitala Vetenskapliga Arkivet*. Obtenido de http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A876907&dswid=9531

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Morgan & Claypool publishers*.

López Gonzáles, J. L., & De Santiago, M. (2018). Significado y Función del derecho de sufragio en la actividad electoral de un estádo democrático. Madrid, España. Obtenido de https://repository.ucatolica.edu.co/bitstream/10983/16350/1/Significado%20y%20funci%c3%b3n%20del%20derecho%20de%20sufragio%20en%20la%20actividad%20electoral%20de%20un%20estado%20democr%c3%a1tico.pdf

Manaswi, N. (2018). RNN and LSTM. En *Deep Learning with Applications Using Python* (págs. 115-126). doi:https://doi.org/10.1007/978-1-4842-3516-4_9

Martínez Cámara, E., García Cumbreras, M., Villena Román, J., & García Morera, J. (2016). TASS 2015 - The The Evolution of the Spanish Opinion Mining Systems. *Procesamiento del Lenguaje Natural, 56.* Obtenido de http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/issue/view/218

Mendoza, S. (21 de Marzo de 2019). *Asociación Nacional de Anunciantes ANDA*. Recuperado el Enero de 2023, de Estudio revela el uso de Twitter en el Perú: https://andaperu.pe/estudio-revela-el-uso-de-twitter-en-el-peru/

Ministerio de Asuntos Exteriores y de Cooperación. (Junio de 2011). *Congreso de la República*. Obtenido de Conociendo el congreso y sus funciones: https://www.congreso.gob.pe/Docs/participacion/parlamento-escolar/files/separata_uso_video.pdf

Muñoz Rocha, C. I. (2015). *Metodología de la Investigación* (Primera ed.). México: Progreso S.A.

Navarro Delgado, J. E. (2020). ANÁLISIS DE INFORMACIÓN PÚBLICA DE CARÁCTER POLÍTICO PROVENIENTE DE REDES SOCIALES UTILIZANDO CIENCIA DE DATOS Y TÉCNICAS ESTADÍSTICAS DE REGRESIÓN. *(Trabajo de Graduación).* Universidad de San Carlos de Guatemala, Guatemala.

Ñaupas Paitán, H., Valdivia Dueñas, M. R., Palacios Vilela, J. J., & Romero Delgado, H. E. (2018). *Metodología de la Investigación* (Quinta ed.). Bogotá, Colombia: Ediciones de la U.

Quispe Auqui, B. E., & Tomayquispe Mejía, S. M. (2019). *Los sondeos electorales y su influencia en la decisión de voto de los universitarios en las elecciones municipales a la alcaldía de Lima, 2018.* [Trabajo de Investigación. Universidad Tecnológica del Perú].

Rodrigues, S., Allende-Cid, H., Palma, W., Alfaro, R., Gonzales, C., Elortegui, C., & Santander, P. (2018). Forecasting the Chilean Electoral Year: Using Twitter to Predict the Presidential Elections of 2017. *Springer International Publishing AG, part of Springer Nature 2018*, 298-314.

Rokach, L., & Maimon, O. (2005). DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK. *Springer*. doi:doi:10.1007/0-387-25465-x_9

Rudd, J. (1 de Agosto de 2018). Application of Support Vector Machine Modeling and Graph Theory Metrics for Disease Classification. *IOS Press*. doi:10.3233/MAS-180444

Ruíz P, A. (25 de Enero de 2021). *Spanish Twitter Sentiment Analyzer*. Obtenido de Github.com: https://github.com/alvaroruizp/spanish-twitter-sentiment-analyzer

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of Machine Learning*. doi:https://doi.org/10.1007/978-0-387-30164-8_525

Swamidass, p. M. (2000). *Encyclopedia of Production and Manufacturing Management*. doi:https://doi.org/10.1007/1-4020-0612-8_580

Villena Román, J., Lana Serrano, S., Martínez Cámara, E., & Gonzales Cristóbal, J. C. (Marzo de 2013). TASS - Workshop on Sentiment Analysis at SEPLN. *Sociedad Española para el Procesamiento del Lenguaje Natural*, 37-44.

Wang, L., & Gan, J. Q. (2017). Prediction of the 2017 French Election Based on Twitter Data Analysis. *IEEE*, 89-93.

Wirth, R., & Hipp, J. (Abril de 2000). CRISP-DM: Towards a Standard Process Model for Data. *4th international conference on the practical applications of knowledge discovery and data mining*, 29-39.

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Computation*, 1-36. doi:10.1162/neco_a_01199

ANEXOS

Anexo 1: Código Fuente del Proyecto

Para ver detalladamente el código fuente del proyecto, revisar el siguiente link:

https://github.com/TheDenky/Twitter-Election-Predict-Peru.

Anexo 2: Matriz de Consistencia

Tabla 35

Matriz de Consistencia

PROBLEMA	OBJETIVOS	HIPOTESIS	VARIABLES E INDICADORES	METODOLOGÍA
Problema General: ¿Cuál es la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021? **Problemas Específicos:** ¿Cuál es la eficiencia de un modelo predictivo de análisis de sentimientos de tweets en español? ¿Cuál es la eficiencia según el nivel de error RMSE y MAE del modelo predictivo para las elecciones presidenciales en Perú en primera vuelta? ¿Cuál es la eficiencia según el nivel de error RMSE y MAE del modelo predictivo para las elecciones presidenciales en Perú en segunda vuelta? ¿Cuál es el nivel de error del modelo predictivo en comparación con las encuestadoras IPSOS, CPI y DATUM en primera vuelta? ¿Cuál es el nivel de error del modelo predictivo en comparación con las encuestadoras IPSOS, DATUM e IDICE en segunda vuelta?	**Objetivo General:** Evaluar la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021. **Objetivos Específicos:** Determinar la eficiencia de un modelo predictivo de análisis de sentimientos de tweets en español. Evaluar el RMSE y MAE del modelo predictivo para las elecciones en Perú en primera vuelta. Evaluar el RMSE y MAE del modelo predictivo para las elecciones en Perú en segunda vuelta. Comparar el nivel de error del modelo predictivo con las encuestadoras IPSOS, CPI y DATUM en primera vuelta. Comparar el nivel de error del modelo predictivo con las encuestadoras IPSOS, DATUM e IDICE en segunda vuelta.	**Hipótesis General:** Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un resultado significativo **Hipótesis Específicas:** Es posible determinar la eficiencia de un modelo predictivo de análisis de sentimientos de tweets en español. Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un error RMSE y MAE en promedio menor de 10% en primera vuelta. Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un error RMSE y MAE en promedio menor de 5% en segunda vuelta. Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un nivel de error RMSE y MAE en promedio menor en comparación con las encuestadoras IPSOS, CPI y DATUM en primera vuelta. Es posible que la eficiencia de un modelo predictivo basado en Análisis de Sentimientos en los resultados de las elecciones presidenciales del Perú en el año 2021, tenga un nivel de error RMSE y MAE en promedio menor en comparación con las encuestadoras IPSOS, DATUM e IDICE en segunda vuelta.	**Variable dependiente:** Predicción de resultados de las elecciones presidenciales del Perú 2021. **Variable Independiente:** Modelo predictivo basado en Análisis de Sentimientos	**Tipo de Investigación:** Este trabajo de investigación es de tipo aplicada debido a que busca aportar una posible solución un problema práctico y concreto del entorno social, en este caso en el aspecto político. (Muñoz Rocha, 2015, pág. 142). Según la planificación de los datos es retrospectiva debido a que utiliza datos pasados y secundarios. Según el enfoque de investigación es de tipo cuantitativo, por lo que utiliza datos numéricos y estadísticos. (Muñoz Rocha, 2015, pág. 86) **Nivel de Investigación:** Este trabajo de investigación es de nivel predictivo, debido a que pronostica el comportamiento de un hecho social, concretamente en el ámbito político (Muñoz Rocha, 2015, pág. 85). **Diseño de Investigación:** El diseño de la presente investigación es preexperimental debido a que existe un bajo control de las variables de estudio, la asignación de datos no es aleatoria y no se cuenta con un grupo de control (Bernal, 2010, pág. 146). **Población y Muestra:** Se contó con una población de 81214 y 15007 tweets para la primera y segunda vuelta electoral y una muestra de 5207 y 1495 tweets respectivamente. Donde para la primera vuelta se consideró solo las publicaciones hechas desde el 01 de marzo de 2021 hasta el 29 de marzo de 2021 y para la segunda vuelta solo las publicaciones hechas desde el 01 de mayo de 2021 hasta el 30 de mayo de 2021.