



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería de Sistemas e Informática

Escuela Profesional de Ingeniería de Sistemas

**Modelo predictivo de Análisis de sentimiento con
respecto a la variación del tipo de cambio del dólar
durante la pandemia en el 2021 en Twitter**

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Joaquín Rodolfo BERNABE POLO

ASESOR

Mg. Yudi Lucero GUZMÁN MONTEZA

Lima, Perú

2025



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

Bernabe, J. (2025). *Modelo predictivo de Análisis de sentimiento con respecto a la variación del tipo de cambio del dólar durante la pandemia en el 2021 en Twitter*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios autor/ asesor

Datos de autor	
Nombres y apellidos	Joaquín Rodolfo Bernabe Polo
Tipo de documento de identidad	DNI
Número de documento de identidad	73456345
URL de ORCID	https://orcid.org/0000-0003-4488-5341
Datos de asesor	
Nombres y apellidos	Yudi Lucero Guzmán Monteza
Tipo de documento de identidad	DNI
Número de documento de identidad	16621460
URL de ORCID	https://orcid.org/0000-0001-5306-5295
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Jorge Leoncio Guerra Guerra
Tipo de documento	DNI
Número de documento de identidad	08473333
Miembro del jurado 1	
Nombres y apellidos	Félix Armando Fermín Pérez
Tipo de documento	DNI
Número de documento de identidad	08736347
Datos de investigación	
Línea de investigación	C.0.5.8. Inteligencia Artificial

Grupo de investigación	No Aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Latitud: -12.0431800 Longitud: -77.0282400
Año o rango de años en que se realizó la investigación	2022 - 2024
URL de disciplinas OCDE	Ingeniería de sistemas y comunicaciones https://purl.org/pe-repo/ocde/ford#2.02.00



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Sistemas

Acta Virtual de Sustentación de Tesis

Siendo las 11:00 horas del día 27 de febrero del año 2025, se reunieron virtualmente los docentes designados como miembros de Jurado de Tesis, presidido por el Mg. Jorge Guerra Guerra, Mg. Félix Armando Fermín Pérez (Miembro) y el Mg. Yudi Guzmán Monteza (Miembro Asesor), usando la plataforma Meet (<https://meet.google.com/yuk-ptbc-yeu>) para la sustentación Virtual de la tesis Intitulada: “ **Modelo predictivo de Análisis de sentimiento con respecto a la variación del tipo de cambio del dólar durante la pandemia en el 2021 en Twitter.**”, del Bachiller: **Joaquín Rodolfo BERNABE POLO**; para obtener el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el Presidente invitó al Bachiller a responder las preguntas formuladas por los Miembros del Jurado.

El Bachiller, en el curso de sus intervenciones demostró dominio del tema, al responder las preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de 14 **CATORCE**

A continuación, el Presidente del Jurado Mg. Jorge Guerra Guerra , declara al Bachiller **Ingeniero de Sistemas.**

Siendo 11.58 horas, se levantó la sesión.

Mg. Jorge Guerra Guerra
Presidente

Mg. Felix A. Fermin Pérez
Miembro

Mg. Yudi Guzmán Monteza
Miembro Asesor



Anexo 1
CERTIFICADO DE SIMILITUD

Yo, **Yudi Guzmán Monteza** en mi condición de asesor acreditado con el Dictamen de la UPG o de la EP N° 000004-2023-EPISI-FISI/UNMSM de la tesis/ monografía/ informe de investigación/trabajo académico, cuyo título es: **Modelo predictivo de análisis de sentimiento con respecto a la variación del tipo de cambio de dólar durante la pandemia en el 2021 en twitter**, presentado por el **bachiller/ magíster/ egresado/ licenciado/ estudiante Bernabe Polo, Joaquin Rodolfo** para optar el grado* de/ **título*/** especialidad en Ingeniería de Sistemas, CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante la plataforma PIDS, el documento evaluado cuenta con el **8.%** de similitud, nivel PERMITIDO para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado*/ título*/ especialidad*

.....

Lima, 14 de julio del 2025

Firma del Asesor

Nombres y Apellidos del asesor: Yudi Lucero Guzmán Monteza

DNI: 16621460

* Según corresponda: bachiller, magíster, doctor; título profesional o título de segunda especialidad

La vigencia del presente certificado corresponde a un período de noventa (90) días desde la fecha de su emisión; siempre que no haya observaciones en el texto durante el proceso de revisión por el jurado informante o surja alguna contingencia.

Resumen

Durante los últimos años las redes sociales se han visto en auge, en donde las personas expresan su opinión sobre diferentes temas entre ellos la variación del dólar, que es un tema importante para la población debido a su influencia en el mercado financiero, debido a esto, en la presente investigación se describe la propuesta de analizar la tendencia de los usuarios de Twitter respecto a la variación del dólar usando algoritmos de machine learning así como datos bancarios obtenidos del BCRP conjuntamente con datos de la red social Twitter para realizar un forecasting del tipo de cambio del dólar.

En la presente investigación se propone una metodología basada en 5 etapas; para la construcción del dataset se utilizó la librería Snsrape de Python, luego se realizó el procesamiento del lenguaje natural de los tweets, seguido del muestreo de los datos en donde se utilizaron particiones porcentuales y validación cruzada, la implementación abarcó una primera subetapa en la que se realiza una clasificación de los tweets obtenidos haciendo uso de los algoritmos de Naive Bayes y Random Forest teniendo un score de 0.901459854, este resultado obtenido junto a información bancaria recopilada sirven como input para la segunda subetapa del forecasting de la variación del dólar, el cual usó los algoritmos de Random Forest Regressor y Support Vector Machine, finalmente usando los errores MSE y MAE se obtiene que el modelo Support Vector Machine muestra un menor error en la predicción del tipo de cambio del dólar teniendo un MAE de 0.009605179, un MSE de 0.000232213 y un F1 Score de 0.92826181.

ABSTRACT

During the last years, social networks have been booming, where people express their opinion on different topics, among them the variation of the dollar, which is an important topic for the population due to its influence on the financial market, due to this, this research describes the proposal to analyze the trend of Twitter users regarding the variation of the dollar using machine learning algorithms and banking data obtained from the BCRP together with data from the social network Twitter to make a forecasting of the dollar exchange rate.

This research proposes a methodology based on 5 stages; the Python Snsrape library was used to build the dataset, then the natural language processing of the tweets was performed, followed by data sampling where percentage partitions and cross validation were used, the implementation included a first sub-stage in which a classification of the tweets obtained using the Naive Bayes and Random Forest algorithms was carried out, with an F1 score of 0.901459854, the result obtained together with the banking information collected serves as input for the second substage of the forecasting of the dollar variation, which used the Random Forest Regressor and Support Vector Machine algorithms, finally using the MSE and MAE errors, it is obtained that the Support Vector Machine model shows a lower error in the prediction of the dollar exchange rate having a MAE of 0.009605179, a MSE of 0.000232213 and a F1 Score of 0.92826181.

Índice

Resumen	2
ABSTRACT	3
Índice	4
Índice de Figuras	9
Índice de Tablas	11
INTRODUCCIÓN	13
Capítulo I: PLANTEAMIENTO METODOLÓGICO	14
1.1 Antecedentes	14
1.2 Definición del Problema	16
1.2.1 Problema	16
1.3 Objetivo	16
1.3.1 Objetivo general	16
1.3.2 Objetivos específicos	16
1.4 Justificación	17
1.5 Organización de la tesis	18
Capítulo II: MARCO TEÓRICO	19
2.1. Machine learning	19
2.1.1 Aprendizaje Supervisado	19
2.1.1.1 Naive Bayes (NB)	19

2.1.1.2 Máquinas de Vectores de Soporte (SVM)	20
2.1.1.3 Random Forest	21
2.1.1.4 Redes Neuronales Convolucionales (CNN)	21
2.1.1.5 Perceptrón multicapa (MLP)	21
2.1.2 Técnicas de vectorización de texto	22
2.1.2.1 Frecuencia de término-frecuencia inversa de documento (TF-IDF)	22
2.1.2.2 BOW	22
2.2 Análisis de Sentimiento	22
2.2.1 Análisis de Sentimiento en redes sociales	23
2.3 Tipo de cambio del dólar	23
2.3.1 Tipo de cambio Nominal	23
2.3.2 Tipo de cambio real	23
2.3.3 Tipo de cambio flotante	24
2.3.4 Tipo de cambio fijo	24
2.3.5 Tipo de cambio en la economía.	24
2.4 Pandemia (COVID-19)	24
2.4.1 Desigualdad Social	24
2.4.2 Salud	25
III: ESTADO DEL ARTE METODOLÓGICO	26
3.1 Metodología	26
Preguntas de investigación	26

Diagramas de Búsqueda	27
Cadena de Búsqueda: “stock market forecasting”	27
Cadena de Búsqueda: “stock market dollar”	27
Cadena de Búsqueda: “sentiment analysis stock market”	28
Cadena de Búsqueda: “sentiment analysis stock market forecasting”	28
Cadena de Búsqueda: “sentiment analysis dollar”	28
3.2 Análisis	31
3.2.1 Métodos para recopilar y construir data sets.	31
Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction	31
Descripción del aporte del autor	31
Bayesian Constitutionalization: Twitter Sentiment Analysis of the Chilean Constitutional Process through Bayesian Network Classifiers	33
Descripción del aporte del autor	33
Stock Market Prediction based on Social Sentiments using Machine Learning	34
Descripción del aporte del autor	34
3.2.2 Métodos de limpieza, preprocesamiento de datos y vectorización de texto.	35
Influence of Pre-processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features	35
Descripción del aporte del autor	35
Exploring the Influence of News Articles on Bitcoin Price with Machine Learning	37
Descripción del aporte del autor	37
3.2.3 Técnicas de machine learning para el procesamiento de datos.	38
Sentiment Analysis with Machine Learning Methods on social media	38
Descripción del aporte del autor	38
Integrating StockTwits with sentiment analysis for better prediction of stock price movement	39
Descripción del aporte del autor	39
Technical analysis and sentiment embeddings for market trend prediction	40
Descripción del aporte del autor	40
3.2.4 Métricas de evaluación	41
Forecasting weekly crude oil using twitter sentiment of U.S. foreign policy and oil companies data	41

Descripción del aporte del autor	41
Predicting the Stock Price Movement by Social Media Analysis	43
Descripción del aporte del autor	43
Foreign Rate Exchange Prediction Using Neural Network and Sentiment Analysis	44
Descripción del aporte del autor	44
Capítulo IV: IMPLEMENTACIÓN	45
4.1 Aporte Teórico	45
4.1.1 Metodología a seguir (Modelo conceptual)	45
4.1.2 Selección y Adaptación de la tecnología	47
Lenguaje de Programación	47
Entorno de desarrollo (IDE)	49
4.2 Aporte Práctico	51
4.2.1 Diseño de la solución	51
4.2.1.1 Arquitectura Tecnológica	52
4.2.2 Implementación de la solución	53
Fase 1: Análisis	53
Construcción de los Dataset	53
Conjunto de datos de Twitter	54
Conjunto de datos Bancarios	55
Fase 2 Preparación:	56
Preprocesamiento del Texto	56
Vectorización del texto	57
Fase 3: Modelado	59
Evaluación de Desempeño - Clasificación	60
Evaluación de desempeño – Regresión	63
4.3. Validación	64
4.3.1 Algoritmo	64

4.3.1.1 Random Forest Classifier	65
4.3.1.2 Naïve Bayes	65
4.3.1.3. Random Forest Regresor	65
4.3.1.4. Support Vector Machine	65
4.3.2 Hardware y Software	65
4.3.3 Instancia de pruebas	66
4.3.4 Resultados	67
Capítulo V. CONCLUSIONES Y TRABAJOS FUTUROS	75
Anexos	77
Anexo 1: Random Forest Classifier	77
Anexo 2: Multinomial Naive Bayes	77
Anexo 3: Random Forest Regressor	78
Anexo 4: Support Vector Machine	78
Anexo 5: Github del código	79
Capítulo VI: Referencias Bibliográficas	80

Índice de Figuras

Figura 1. Variación del dólar entre octubre 2016 y octubre 2021.	14
Figura 3. Naive Bayes.....	20
Figura 4. Termino-Frecuencia.....	22
Figura 5. Frecuencia inversa de documento.....	22
Figura 6. Diagrama de Búsqueda: "stock market forecasting"	27
Figura 7. Diagrama de Búsqueda: "stock market dollar"	27
Figura 8. Diagrama de Búsqueda: "sentiment analysis stock market"	28
Figura 9. Diagrama de Búsqueda: "sentiment analysis stock market forecasting"	28
Figura 10. Diagrama de Búsqueda: "sentiment analysis dollar"	28
Figura 11. Modelo Conceptual.....	47
Figura 12. Lenguaje de programación recomendados.	48
Figura 13. KDnuggets Analytics/Data Science 2019 Software Poll.....	48
Figura 14. Proceso de selección de lenguaje de programación.....	49
Figura 15. Integrated Development Environment by Job Title.....	49
Figura 16. Proceso de selección de IDE.....	50
Figura 17. Fase de Implementación.	52
Figura 18. Arquitectura tecnológica.....	53
Figura 19. Diagrama de barras de los Tweets.	55
Figura 20. Muestra de datos - Dataset Perú.	55
Figura 21 . Muestra de datos (Tipos de cambio del dólar) – BCRP.	56
Figura 22. Preprocesamiento.....	57
Figura 23. Unigramas vs Bigramas, diagrama de barras.	58
Figura 24. Muestra de datos. Tf-Idf Unigramas.....	58
Figura 25. Muestra de datos Tf-idf Bigramas.	59

Figura 26. Random Forest Model – 70% Diagrama de líneas	68
Figura 27. Random Forest Model – 70% Diagrama de Barras	68
Figura 28. Random Forest Model - 75% Gráfica de líneas.....	69
Figura 29. Random Forest Model - 75% Gráfica de Barras	70
Figura 30. Random Forest Model - 80% Gráfica de líneas.....	71
Figura 31. Random Forest Model - 80% Gráfica de líneas.....	72
Figura 32. SVM Model - Gráfica de líneas.....	73
Figura 33. SVM Model - Gráfica de barras	74
Figura 34. Random Forest Classifier - Modelo porcentual 70-30	77
Figura 35. Multinomial Naive Bayes - Modelo porcentual 70-30	77
Figura 36. Random Forest Regressor - Modelo porcentual 70-30.....	78
Figura 37. Support vector machine - Validación cruzada.....	78

Índice de Tablas

Tabla 1.....	29
Tabla 2.....	57
Tabla 3.....	59
Tabla 4.....	60
Tabla 5.....	60
Tabla 6.....	61
Tabla 7.....	61
Tabla 8.....	63
Tabla 9.....	63
Tabla 10.....	64
Tabla 11.....	64
Tabla 12.....	66
Tabla 13.....	67

INTRODUCCIÓN

“Hoy en día, las redes sociales como Twitter, Facebook, Instagram, YouTube, etc. están de moda” (Tiwari, Verma, Garg, & Bansal, 2020). Según Zaidi, (2018), los sitios web como Twitter contienen una gran cantidad de conocimientos que se pueden utilizar para comprender el comportamiento del mercado y según Batra, (2018), el análisis de sentimientos está ganando popularidad en la tecnología de la información puesto que permite determinar los sentimientos, opiniones y emociones de las personas y/o usuarios de estas redes sociales como Twitter.

En este documento se describe la propuesta de realizar un modelo predictivo para el análisis de la tendencia de los usuarios de Twitter respecto a la variación del dólar usando algoritmos de machine learning así como datos bancarios obtenidos del BCRP conjuntamente con datos de la red social Twitter para realizar un forecasting del tipo de cambio del dólar.

Capítulo I: PLANTEAMIENTO METODOLÓGICO

1.1 Antecedentes

La finalidad del Banco Central según la Constitución Política del Perú [Const] (1993) es preservar la estabilidad monetaria. Para esto regulariza la moneda y el crédito del sistema financiero, además de administrar las reservas internacionales, y más funciones que señala su ley orgánica.

Pero el BCRP no es el único influyente en el tipo de cambio. Según Ribeiro (2018), las economías pequeñas y abiertas como la peruana están expuestas de manera constante a diversos choques externos e internos. Y que, ante esta coyuntura, es imprescindible un adecuado manejo de la política cambiaria para asimilar dichos choques, uno de estos choques internos son las elecciones presidenciales, y como se muestra en la Figura 1, la variación del dólar ha sido fluctuante registrando subidas y caídas, pero esta fluctuación se convierte en una tendencia creciente para el año 2020, año en el cual sucedió la pandemia del covid-19.

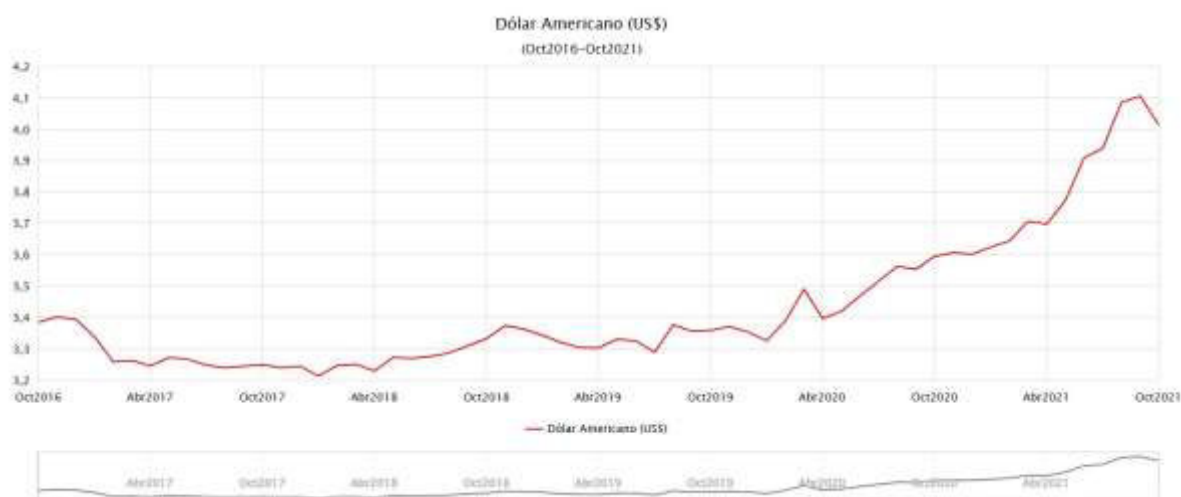


Figura 1. Variación del dólar entre octubre 2016 y octubre 2021. Nota. Tomado de (BCRP Data, 2021).

En la investigación de (Ranjit S. S. S., 2018) se indica que el tipo de cambio sigue la regla de oferta y demanda. Los valores de las divisas tienden a subir con el aumento de la demanda y tienden a caer con la disminución de la demanda y que factores como el mercado

de valores, la situación política, las noticias económicas y las tasas de interés afectan la demanda del tipo de cambio.

Según (Valencia, 2019) igualmente indica que la predicción del mercado de divisas es una tarea compleja, debido a los múltiples factores e incertidumbres que interactúan en los mercados incluidas las condiciones económicas y políticas, así como el comportamiento humano.

Asimismo, la pandemia del COVID-19 ha traído consigo enormes incertidumbres en el suministro mundial y riesgos operativos, como interrupciones en el abastecimiento, escasez de materiales y mano de obra, demanda irracional de los consumidores, etc. El capitalismo mundial fue testigo de la caída de los mercados bursátiles y de la desaparición de billones de dólares en los primeros días del brote de COVID-19 (Chunguang Bai, 2024).

Por lo que, al ser un tema tan relevante en la situación económica en el Perú, las redes sociales como Twitter son un medio bastante frecuente en el cual las personas indican su opinión (tendencia de los usuarios).

Según Alvino (2021), se menciona que un total de 1.55 millones de personas que fueron influenciadas mediante campañas publicitarias, y que para el año 2021, hubo un aumento del 6,9% en los usuarios de Twitter, registrando 100 mil peruanos como nuevos usuarios; asimismo anteriormente ya se han realizado investigaciones en el Perú mediante el uso de la minería de datos para conocer las tendencias de los usuarios de Twitter; un ejemplo de los buenos resultados de su uso se evidencia en la investigación de Alva Segura, (2021) para conocer la tendencia política durante las elecciones congresales del 2020, por lo que realizar una investigación sobre la tendencia del tipo del cambio del dólar durante fenómenos políticos como las elecciones generales es factible, asimismo se tiene la investigación de Ulloa, (2022) que se focalizaron en predecir la tendencia futura del mercado de la Bolsa de Valores de Lima

y teniendo como resultado de una revisión bibliográfica sistemática focalizadas en el Perú, está la única resultante.

1.2 Definición del Problema

1.2.1 Problema

Escasos estudios de modelo predictivos que analicen la predicción del tipo de cambio del dólar a partir de datos del BCRP y de mensajería de Twitter usando técnicas de machine learning durante la pandemia en el 2021 por lo cual se sigue utilizando métodos tradicionales para la predicción del tipo de cambio del dólar en el Perú.

1.3 Objetivo

1.3.1 Objetivo general

Implementar un modelo predictivo para el análisis de sentimiento con respecto al tipo cambio del dólar durante la pandemia en el 2021.

1.3.2 Objetivos específicos

- Construir un dataset bancario y un dataset de los tweets de los usuarios de Twitter respecto a la variación del tipo de cambio del dólar.
- Determinar la estrategia de limpieza y preprocesamiento de datos de los tweets respecto a la variación del tipo de cambio del dólar.
- Implementar un modelo predictivo aplicando técnicas de Machine Learning para el procesamiento de datos.
- Revisar y evaluar el rendimiento de la técnica utilizando las métricas de evaluación, para identificar cuál fue el mejor resultado.

1.4 Justificación

En una entrevista realizada en El Comercio (Lengua, 2021) se ven puntos de vista de diferentes economistas, entre los cuales resalta Jorge Guillén, el cuál con respecto al alza del dólar sostuvo que existen riesgos a nivel crediticios, y que, aquellas personas que sean acreedoras de deudas en dólares necesitarían mayor cantidad de soles para subsanarlas, y eventualmente podrían incurrir en incumplimientos, por otro lado, sostuvo que esta alza del dólar podría beneficiar a otros sectores como el de exportación y minería.

El tipo de cambio sigue siendo uno de los indicadores económicos importantes para cualquier país. Cualquier decisión de inversión por parte de corporaciones multinacionales o gobiernos se basa en los tipos de cambio (Yasir, 2019).

Las redes sociales permiten la participación en una conversación global de públicos variados y también son una fuente de información útil para los modelos predictivos, tal cual como se ejemplifica en la investigación de (Zaidi, 2018), en donde se realiza el análisis de datos de las redes sociales sobre las compañías petroleras para pronosticar los precios del petróleo logrando una precisión de clasificación general superior a la de los modelos existentes.

Y según (Vijh, 2020) con la introducción de la inteligencia artificial y el aumento de las capacidades computacionales, los métodos de predicción programados han demostrado ser más eficientes para predecir los precios, además; sostiene que predecir la tendencia puede resultar de gran utilidad para realizar movimientos bursátiles.

Por lo que al realizar el análisis de los datos de Twitter permitiría conocer la tendencia de los usuarios respecto al tipo de cambio del dólar para posteriormente utilizarla como input para darle más contexto al fenómeno de estudio, y así realizar el forecasting del tipo de cambio de dólar y proporcionar dicha herramienta usando la inteligencia artificial.

Asimismo, el aprendizaje automático no requiere conjeturas previas sobre las relaciones ocultas entre las variables a diferencia de la estadística que deben entender el cómo se reunieron

los datos y diversos factores adicionales que esperarías si hicieras el experimento una y otra vez por lo que no hay una seguridad del 100% de obtener los mismos resultados.

1.5 Organización de la tesis

La presente investigación se desarrolló en 6 capítulos, los cuales están organizados de la siguiente manera:

En el capítulo I se describen los antecedentes del problema, definición o formulación del problema, objetivo general y objetivos específicos, justificación de la solución y la organización de la tesis.

En el capítulo II presenta el marco teórico, que conceptualiza las bases teóricas relevantes vinculadas y utilizadas en esta investigación.

En el capítulo III presenta el estado del arte metodológico, describiendo los trabajos previos desarrollados por otros autores y se describe principalmente la Metodología a seguir (Modelo conceptual).

En el capítulo IV, presenta el desarrollo de la solución planteada y se describe detalladamente cómo se realizó la implementación del modelo predictivo; finalmente, se evalúan los modelos implementando encontrando cual muestra menor error en la predicción del tipo de cambio del dólar.

En el capítulo V, presenta las conclusiones obtenidas luego del desarrollo de la solución, las cuales intentan responder la interrogante formulada acerca del problema, indicando los resultados obtenidos para el modelo y las tecnologías usadas para ello.

En el capítulo VI, presenta las referencias bibliográficas.

Capítulo II: MARCO TEÓRICO

2.1. Machine learning

2.1.1 Aprendizaje Supervisado

2.1.1.1 Naive Bayes (NB)

Este modelo es utilizado en la clasificación textual que incorpora un dataset de entrenamiento de alta dimensión.

Por ejemplo, una verdura puede ser considerado como una beterraga si tiene el color de morado intenso en el exterior, es redonda y tiene aproximadamente 8 centímetros de diámetro. Aun cuando estas características dependieran entre sí o de alguna otra característica, NB estimaría que éstas características aportan de manera independiente a la probabilidad de que esta verdura fuera una beterraga.

Naive Bayes requiere poco espacio de almacenamiento durante las etapas de entrenamiento y clasificación: el mínimo estricto es la memoria necesaria para almacenar las probabilidades previas y condicionales; se puede usar fácilmente como incremental learner, mientras que los otros algoritmos (a excepción de k vecinos más próximos) no pueden.

Según el estudio de Nabipour, Nayyeri, Jabani, Shahab, & Mosavi (2020), se menciona que Naive Bayes tiene fuertes suposiciones de independencia entre las características dado el valor de la variable de clase, así como que puede ser muy rápido en comparación con algoritmos más sofisticados. La separación de las distribuciones de clase significa que cada una puede evaluarse independientemente como una distribución unidimensional. Esto, a su vez, ayuda a aliviar los problemas derivados de la dimensionalidad.

Según Başarslan (2020) los algoritmos bayesianos se encuentran entre las técnicas de clasificación estadística y se basan en el teorema estadístico bayesiano. El clasificador de Bayes

es un modelo predictivo, que es más fácil de aplicar. Sea $X = x_1, x_2, x_3, \dots, x_n$, es el conjunto de muestra, y $C_1, C_2, C_3, \dots, C_k$ es el conjunto de clases.

$$P(X|C_i) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

Figura 2. Naive Bayes. Nota. Tomado de (Başarslan, 2020).

Detallando los elementos de la *Figura 3* tenemos que:

- $P(c)$ es la probabilidad previa de objetivo.
- $P(x|c)$ es la probabilidad del objetivo dada del atributo (predictor).
- $P(x)$ es la probabilidad previa del atributo (predictor).

2.1.1.2 Máquinas de Vectores de Soporte (SVM)

“SVM utiliza un espacio de características de alta dimensión. El objetivo principal de la máquina de vectores soporte es identificar el hiperplano de margen máximo como límite de decisión final.” (Chen, 2019).

“El SVM se puede definir como un método de minería de datos basado en el espacio vectorial que encuentra un límite de decisión entre las dos clases más alejadas de un punto aleatorio en los datos de entrenamiento” (Başarslan, 2020).

Generalmente, SVM tiende a funcionar mucho mejor cuando se trata de características continuas y de múltiples dimensiones. Para SVM, se requiere un tamaño de muestra grande para lograr su máxima precisión de predicción.

Según el estudio de Obthong, (2020), SVM puede proporcionar la solución global óptima y tiene una excelente capacidad de precisión predictiva; también funciona bien en una serie de problemas de clasificación; además de información brindada por estudios análogos realizados sobre su rendimiento, por ejemplo, en el estudio de Al Hamoud (2018) han utilizado Bag of Word (BOW) y TF-IDF en los datos de Twitter para la clasificación de tweets políticos;

los algoritmos de clasificación utilizados fueron SVM y NB teniendo como resultado que SVM habilitado para BOW proporciona la mayor precisión.

2.1.1.3 Random Forest

De acuerdo a Valencia (2019), los RF son metas estimadores que se ajustan a una sucesión de árboles de decisión en muchas submuestras del dataset utilizando un enfoque de conjunto, combinando predictores de árboles y cada árbol necesita de los valores de un vector aleatorio con una distribución uniforme para aquellos árboles que pertenecen al bosque.

Según Obthong (2020) se detalla las ventajas de Random Forest, las cuales son:

- “Método robusto para pronosticar y clasificar problemas ya que su diseño está lleno de varios árboles de decisión y el espacio de características se modela aleatoriamente” (Obthong, 2020).
- “Maneja automáticamente los valores faltantes” (Obthong, 2020).
- “Funciona bien con variables discretas y continuas” (Obthong, 2020).

2.1.1.4 Redes Neuronales Convolucionales (CNN)

“Las redes neuronales convolucionales son una clase de redes neuronales invariantes a la posición de los elementos en la entrada que se inspiran en las células que se encuentran en la corteza visual de los gatos” (Chiruzzo L., 2020).

2.1.1.5 Perceptrón multicapa (MLP)

“Los perceptrones multicapa (MLP) son un tipo de red neuronal que consta de al menos tres capas de nodos. Los MLP pueden utilizar la retro propagación y el aprendizaje supervisado para el entrenamiento” (Valencia, 2019).

2.1.2 Técnicas de vectorización de texto

2.1.2.1 Frecuencia de término-frecuencia inversa de documento (TF-IDF)

“El algoritmo TF-IDF se utiliza en cualquier contenido para sopesar una palabra clave y atribuirle importancia según el número de ocasiones que hace aparición en el documento” (Pimpalkar, 2020). La *Figura 4* y la *Figura 5* nos muestran las fórmulas para hallar el TF e IDF respectivamente según el termino o los números de documentos.

$$TF(t) = \frac{(\text{The amount of occasions the term } t \text{ is seen in the document})}{(\text{Total number of terms in the document})}$$

Figura 3. Termino-Frecuencia. Nota. Tomado de (Pimpalkar, 2020).

$$IDF(t) = \log \frac{(\text{Total number of documents})}{(\text{Number of documents with term } t \text{ in it})}$$
$$TFIDF(t) = TF(t) * IDF(t)$$

Figura 4. Frecuencia inversa de documento. Nota. Tomado de (Pimpalkar, 2020)

2.1.2.2 BOW

Según Pimpalkar (2020) el modelo BOW está asociado sólo con la cuestión de si las palabras reconocidas aparecen en el documento o no. El objetivo es convertir cada documento de texto libre en un vector que se pueda usar como entrada o salida para un modelo de ML.

2.2 Análisis de Sentimiento

Según (Ranjit S. S. S., 2018) el análisis de sentimiento es un proceso para determinar la emoción detrás de una serie de palabras, ya sean actitudes, opciones o emociones.

Los sentimientos se pueden predecir utilizando el conocimiento adquirido a partir de los conjuntos de datos previamente entrenado y usando esta información para probar otro conjunto de datos y obtener sus actitudes, opciones o emociones.

El análisis de sentimientos está ganando popularidad en la tecnología de la información para determinar los sentimientos, opiniones y emociones.

2.2.1 Análisis de Sentimiento en redes sociales

En esta investigación el concepto de análisis de sentimiento se relaciona con la tendencia de los usuarios si están a favor o en contra a un determinado asunto, por lo que para el desarrollo del estudio el sentimiento y la tendencia tienen el mismo significado.

En la investigación de (Valencia, 2019) realizan un análisis de sentimiento sobre los datos sociales de Twitter obteniendo en forma de tweets sin procesar según ciertos criterios de inclusión, mencionando que estos datos pueden usarse para inferir el comportamiento humano futuro y, por lo tanto, podrían usarse para desarrollar estrategias comerciales ventajosa.

En la investigación de (Mankar, 2018) se busca predecir los precios del mercado de valores de varias empresas mediante la realización de análisis de sentimiento de los datos de las redes sociales, como los tweets relacionados con las respectivas empresas. Se detalla como este análisis le ayudará a tomar decisiones de inversión adecuadas con su conocimiento del sentimiento del mercado.

2.3 Tipo de cambio del dólar

El tipo de cambio puede ser definido de diferentes maneras dependiendo del contexto.

2.3.1 Tipo de cambio Nominal

Es el precio de una moneda extranjera en términos de una moneda nacional. Por ejemplo, cuando nos referimos al tipo de cambio sol-dólar, corresponde al número de soles que se requieren para comprar un dólar.

2.3.2 Tipo de cambio real

Es el precio relativo de dos canastas de consumo y sirve para medir el poder adquisitivo de una moneda en el extranjero

2.3.3 Tipo de cambio flotante

Es el tipo de cambio que se mueve libremente, ya que el banco central no interviene nunca en el mercado cambiario.

2.3.4 Tipo de cambio fijo

Es el tipo de cambio estático definido por una entidad reguladora o el estado.

2.3.5 Tipo de cambio en la economía.

La variación del tipo de cambio no solo afecta a las exportaciones e importaciones, sino también a los precios, a los flujos de capital, a las inversiones, al valor de los ahorros y deudas.

Cuando el tipo de cambio cae ocurre una apreciación de la moneda local, la moneda extranjera se hace más barata; requiere menos unidades de la moneda local, y cuando el tipo de cambio sube ocurre una depreciación de la moneda: la moneda extranjera se hace más cara; requiere más unidades de la moneda local (López, 2019).

Asimismo, los bienes nacionales se vuelven relativamente más baratos respecto a los extranjeros, lo que favorece la exportación (importación) de bienes y servicios.

2.4 Pandemia (COVID-19)

La pandemia de COVID-19 provocó una onda expansiva en la economía mundial y desencadenó la mayor crisis económica mundial en más de un siglo. La crisis provocó un aumento drástico de la desigualdad dentro de los países y entre ellos (WorldBank, 2022).

2.4.1 Desigualdad Social

La crisis tuvo un impacto dramático en la pobreza y la desigualdad a nivel mundial. La pobreza mundial aumentó por primera vez en una generación, y las pérdidas desproporcionadas

de ingresos entre las poblaciones desfavorecidas llevaron a un aumento dramático de la desigualdad dentro de los países y entre ellos (WorldBank, 2022).

2.4.2 Salud

Impacta la actividad social, las interacciones y los comportamientos en múltiples niveles. Emocionalmente, las actividades de distanciamiento social y autoaislamiento pueden causar problemas de salud mental, ansiedad y depresión (José, 2020).

III: ESTADO DEL ARTE METODOLÓGICO

3.1 Metodología

Se realizó una adaptación de la metodología de (Barbara Kitchenham, 2013) en la cual se realiza una planificación, ejecución y análisis y resultados del arte. Se contabilizaron un total de 21 documentos de tipo: artículos de revista, trabajos de investigación; en una matriz bibliográfica en Excel. A fin de responder las siguientes preguntas de investigación:

Preguntas de investigación

- ¿Cuáles son los métodos utilizados para recopilar y construir un dataset bancario y un dataset de los tweets de los usuarios de Twitter?
- ¿Cuáles son los métodos de limpieza, preprocesamiento de datos y vectorización de texto?
- ¿Cuáles son las técnicas de machine learning para el procesamiento de datos con mejores resultados?
- ¿Cuáles son las métricas de evaluación de las técnicas de machine learning?

Se utilizaron ScienceDirect y scopus como buscadores para los artículos a través de las cadenas de búsqueda “stock market forecasting”, “stock market dollar”, “sentiment analysis stock market”, “sentiment analysis stock market forecasting”, “sentiment analysis dollar” realizando un filtro de los resultados para años mayores al 2018.

Diagramas de Búsqueda

Cadena de Búsqueda: “stock market forecasting”

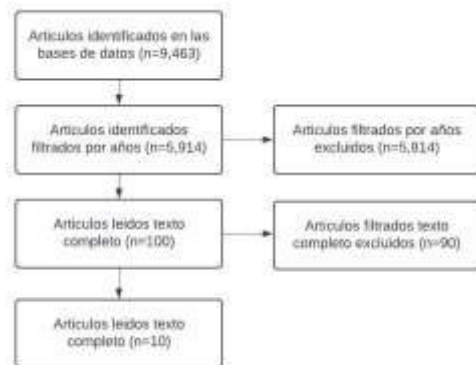


Figura 5. Diagrama de Búsqueda: "stock market forecasting"

Cadena de Búsqueda: “stock market dollar”

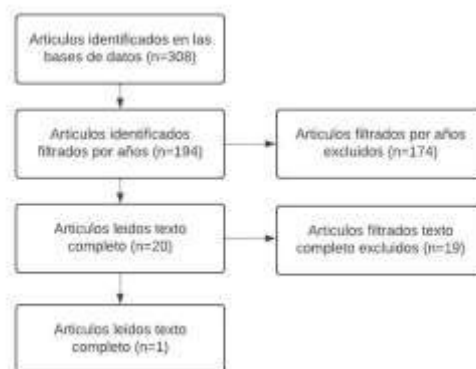


Figura 6. Diagrama de Búsqueda: "stock market dollar"

Cadena de Búsqueda: “sentiment analysis stock market”

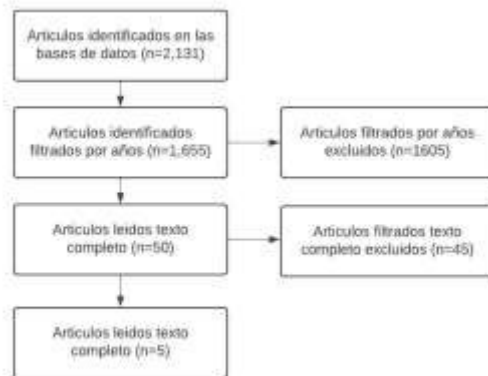


Figura 7. Diagrama de Búsqueda: "sentiment analysis stock market"

Cadena de Búsqueda: “sentiment analysis stock market forecasting”

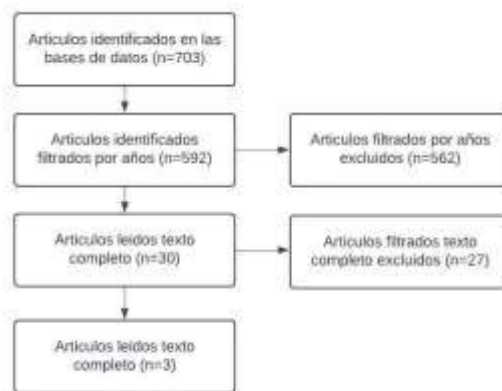


Figura 8. Diagrama de Búsqueda: "sentiment analysis stock market forecasting"

Cadena de Búsqueda: “sentiment analysis dollar”

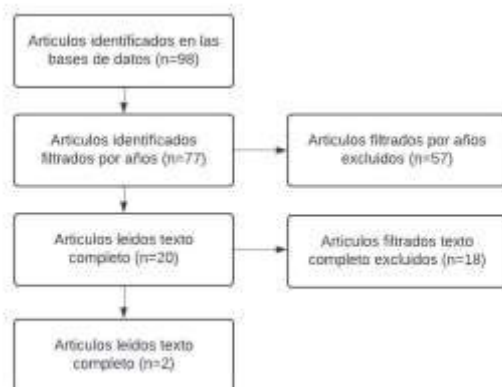


Figura 9. Diagrama de Búsqueda: "sentiment analysis dollar"

Una vez realizada la búsqueda se encontraron más de 200 artículos, entre los cuales se aplicó un filtro para seleccionar se descartaron artículos cuyo centro de investigación quedaba fuera del tema presentado en este trabajo. Se ingresó a los textos para su lectura, una revisión exhaustiva y un análisis textual profundo de la información necesitada en documentos de revisión bajo un formato definido de puntos críticos: Identificación del problema, Metodología y/o Técnica que implementa, Aporte al trabajo de tesis y Resultados obtenidos.

En la *Tabla 1*, se presentará la lista de los documentos utilizados para referencias dentro del presente trabajo que consta de 21 artículos.

Tabla 1

Artículos Científicos

N.	Año	Título
1	2020	“Sentiment Analysis with Machine Learning Methods on social media” (Başarslan, 2020).
2	2019	“Análisis de Sentimiento en el dominio salud: analizando comentarios sobre fármacos” (Segura B. I., 2019).
3	2020	“Sentiment Analysis in Spanish Tweets: Some Experiments with Focus on Neutral Tweets” (Chiruzzo L., 2020).
4	2019	“On the Effect of Word Order on Cross-lingual Sentiment Analysis” (Barnes J., 2019).
5	2019	“Analyzing the Brazilian Financial Market through Portuguese Sentiment Analysis in Social Media” (Carosia, 2020).
6	2020	“Influence of Pre-processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features” (Pimpalkar, 2020).
7	2019	“Foreign Rate Exchange Prediction Using Neural Network and Sentiment Analysis” (Ranjit S. S. S., 2018).

8	2019	“Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning” (Valencia, 2019).
9	2018	“Comparison of algorithms in Foreign Exchange Rate Prediction” (Ranjit S. S. S., 2018).
10	2019	“An Intelligent Event-Sentiment-Based Daily Foreign Exchange Rate Forecasting System” (Yasir, 2019).
11	2018	“Stock Market Prediction based on Social Sentiments using Machine Learning” (Mankar, 2018).
12	2022	“Bayesian Constitutionalization: Twitter Sentiment Analysis of the Chilean Constitutional Process through Bayesian Network Classifiers” (Ruz, 2022).
13	2018	“Integrating StockTwits with sentiment analysis for better prediction of stock price movement” (Batra, 2018).
14	2018	“Forecasting weekly crude oil using twitter sentiment of U.S. foreign policy and oil companies data” (Zaidi, 2018).
15	2019	“Exploring the Influence of News Articles on Bitcoin Price with Machine Learning” (Yao, Xu, & Li, 2019).
16	2019	“Predicting the Stock Price Movement by Social Media Analysis” (Chen, 2019).
17	2018	“Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction” (Das, Behera, Kumar, & Rath, 2018).
18	2020	“Stock Closing Price Prediction using Machine Learning Techniques” (Vijh, 2020).
19	2020	“A Survey on Machine Learning for Stock Price Prediction: Algorithms and Techniques” (Obthong, 2020).
20	2019	“Technical analysis and sentiment embeddings for market trend prediction” (Picasso, 2019).

21	2022	“Stock market prediction using machine learning classifiers and social media, news” (Khan W., 2022)
----	------	--

3.2 Análisis

En el presente apartado se muestra un resumen de las diversas investigaciones referentes a la aplicación del análisis de sentimiento y el uso de técnicas de machine learning resultantes de la revisión bibliográfica realizada.

3.2.1 Métodos para recopilar y construir data sets.

Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction

Autores:	Das, Sushree; Behera, Ranjan Kumar; Kumar, Mukesh; Rath, Santanu Kumar
Año:	2018

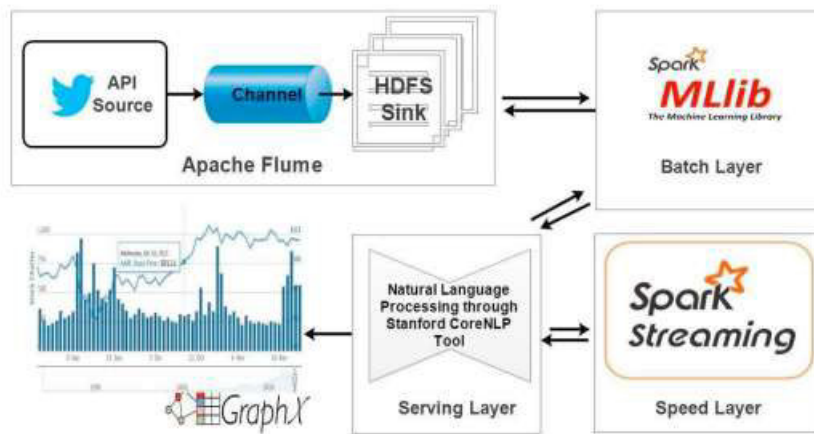
El documento comienza destacando la necesidad de procesar grandes volúmenes de datos en tiempo real, dada la creciente generación de información diaria. Describe el uso del análisis de sentimiento y su relevancia en el contexto financiero, ya que se ha demostrado que la percepción pública puede influir significativamente en el comportamiento del mercado.

Descripción del aporte del autor

Los autores plantean una Arquitectura Lambda, la cuál consta de tres capas principales:

- Capa Batch: Procesa y almacena datos históricos de forma inmutable para crear modelos precomputados.
- Capa de Velocidad: Procesa datos recientes para reducir la latencia y mejorar la respuesta en tiempo real.

- Capa de Servicio: Indexa y organiza los datos procesados para su recuperación eficiente.



Para la ingesta de datos en tiempo real utilizaron la API de Twitter, que luego son almacenados en el sistema de archivos distribuido HDFS mediante Apache Flume. El análisis de sentimiento se realiza mediante el uso de redes neuronales recurrentes (RNN), se utiliza Stanford CoreNLP para clasificar el sentimiento de cada tweet en categorías como "positivo", "negativo" o "neutral". Un conjunto de datos de 560,000 tweets recolectados durante más de un año fue empleado para entrenar y evaluar el modelo.

Se estudian las series temporales de sentimientos asociados a empresas como Google, Microsoft y Apple. Estos datos se correlacionan con las tendencias reales de apertura y cierre de sus precios de acciones. Los resultados muestran que existe una relación significativa entre los sentimientos de los tweets y los movimientos en los precios de las acciones.

El estudio concluye que el análisis de sentimientos de datos de redes sociales como Twitter puede ser una herramienta útil para predecir movimientos en el mercado financiero. Propone explorar modelos híbridos que combinen análisis de sentimientos con tendencias actuales de precios para mejorar la precisión y confiabilidad de las predicciones.

Autores: Ruz, G.A., Henríquez, P.A., Mascareño, A.

Año: 2022

El estudio se sitúa en el contexto del proceso constitucional chileno, iniciado tras las protestas de octubre de 2019, que llevaron a la decisión de redactar una nueva Constitución. Se centra en analizar los tweets de los miembros de la Convención Constitucional (135 de 155 miembros activos en Twitter) durante el período de julio a octubre de 2021.

Descripción del aporte del autor

Se usó la API de búsqueda de Twitter y se procesaron 37,814 tweets que fueron clasificados como positivos o negativos, posteriormente se asignaron puntajes de sentimiento y se identificaron relaciones clave entre las palabras dentro de los tweets.

Se realizó el uso de los clasificadores bayesianos, especialmente el Tree Augmented Naive Bayes (TAN) ajustando estructuras de dependencia entre variables, optimizando la precisión en la clasificación.

Finalmente, se observaron variaciones significativas en el lenguaje utilizado por los diferentes grupos ideológicos y las tendencias en los sentimientos expresados por los miembros reflejaron los momentos clave del proceso constitucional, como la aprobación de reglas generales y la transición hacia el debate del contenido sustantivo. Asimismo, se detalla que el análisis de sentimientos permite capturar matices importantes en las interacciones políticas, revelando tanto patrones colectivos como dinámicas individuales.

Autores: Mankar, T., Hotchandani, T., Madhwani, M., Chidrawar, A., Lifna, C.S.
Año: 2018

El documento destaca la importancia de las redes sociales como reflejo de las opiniones y sentimientos del público hacia eventos o empresas específicas. Estos sentimientos pueden tener un impacto significativo en los precios de las acciones. Por ello, los autores proponen un sistema que analiza sentimientos en Twitter y correlaciona estos datos con valores históricos del mercado bursátil para predecir el movimiento de las acciones.

Descripción del aporte del autor

El proyecto busca facilitar el acceso a inversiones bursátiles para personas sin experiencia financiera, eliminando la necesidad de expertos costosos. La solución es accesible y económica, permitiendo que los usuarios analicen tendencias de mercado y reduzcan riesgos financieros.

En la investigación se realiza la recolección de datos mediante la API de Twitter para recolectar tuits relacionados con empresas específicas, realizando un preprocesamiento para eliminar palabras irrelevantes, URLs, y símbolos. Las etiquetas de los tuits se correlacionan con movimientos históricos de las acciones, obtenidos mediante bibliotecas como Yahoo Finance. Posteriormente se realiza la extracción de características a través de técnicas como unigrams, se identifican palabras clave con mayor valor predictivo. Los datos procesados se usan para entrenar clasificadores como Naïve Bayes y SVM. Finalmente se concluye que el modelo basado en *Support Vector Machine (SVM)* es más eficiente para predecir movimientos del mercado bursátil utilizando sentimientos de redes sociales, el modelo predice la dirección de los precios basándose en datos sentimentales y de mercado.

3.2.2 Métodos de limpieza, preprocesamiento de datos y vectorización de texto.

Influence of Pre-processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features

Autores: Pimpalkar, A., & Raj, R.

Año: 2020

Este trabajo propone un enfoque sistemático para evaluar el impacto de nueve métodos de preprocesamiento, junto con cinco algoritmos de ML. El objetivo es mejorar la clasificación de sentimientos positivos y negativos mediante la aplicación de estas técnicas en dos conjuntos de datos con más de 54,000 tweets y reseñas de películas.

Descripción del aporte del autor

Se destaca que, aunque existen numerosos métodos de preprocesamiento, no hay consenso sobre cuál es el más adecuado. Los resultados varían ampliamente según los datos y métodos seleccionados.

En la investigación se utilizan dos conjuntos de datos, los tweets generales (49,159 registros) y reseñas de películas (4,845 registros), a estos registros se aplicaron técnicas de preprocesamiento de texto y técnicas de vectorización como BOW y TF-IDF para transformar los textos en representaciones numéricas. Finalmente, se implementaron los modelos de SVM, XGBoost, MNB, Decision Trees teniendo como resultados que SVM fue el algoritmo más eficaz para ambos conjuntos de datos, logrando una precisión del 94.91% en tweets y del 81.94% en IMDb tras el preprocesamiento. El estudio concluye que el preprocesamiento de datos es esencial para maximizar la eficiencia de los clasificadores de ML en tareas de análisis de sentimientos.

Este enfoque proporciona una guía clara sobre cómo el preprocesamiento y la selección de características afectan el rendimiento de los modelos de clasificación en dominios desafiantes como las redes sociales.

Autores: Yao, Wenbing; Xu, Ke; Li, Qi
Año: 2019

Las criptomonedas, incluido Bitcoin, son influenciadas por factores como políticas gubernamentales, eventos adversos, aceptación social, actividades ilícitas y sentimientos públicos. Las noticias, como fuente clave de información, reflejan estos factores y pueden influir directamente en las decisiones de compra y venta.

Descripción del aporte del autor

En la investigación se recopila más de 108,000 artículos de noticias sobre Bitcoin recopilados de 10 sitios web especializados en criptomonedas (2012-2019), posteriormente se preprocesaron para eliminar caracteres no relevantes. Se extrajeron características textuales usando métodos tradicionales N-Gram, TF-IDF y Doc2vec y SentiGraph. Además se usaron las fluctuaciones diarias del precio de Bitcoin para etiquetar las noticias como "sube" o "baja".

Los algoritmos utilizados son los de SVM, Random Forest, Perceptrón Multicapa y Regresión Logística, obteniendo finalmente una precisión más alta de 59% con Regresión Logística.

El estudio demuestra que las noticias influyen en el precio del Bitcoin, aunque los resultados sugieren limitaciones inherentes al uso de noticias para predecir precios, la técnica es adaptable a otras tareas de análisis de texto.

3.2.3 Técnicas de machine learning para el procesamiento de datos.

Sentiment Analysis with Machine Learning Methods on social media

Autores: Başırsan, Muhammet & Kayaalp, Fatih

Año: 2020

La proliferación de redes sociales como Twitter ha generado una gran cantidad de datos textuales que reflejan las opiniones de las personas sobre diversos temas

Descripción del aporte del autor

En la investigación se utilizan dos conjuntos de datos, 4,500 tuits etiquetados como positivos (1,220), negativos (1,600) o neutrales (1,680) relacionados con temas de salud, y 1,000 reseñas de películas de IMDb, etiquetadas previamente. Se aplicaron técnicas de preprocesamiento para realizar la limpieza de los datos como la eliminación de símbolos y lematización. Se representó el texto mediante los métodos de TF-IDF y Word2Vec.

Los algoritmos utilizados fueron Naïve Bayes (NB), Super Vector Machine y redes neuronales (ANN). Para la evaluación del rendimiento se usó una validación cruzada $k=5$ y las métricas de Accuracy y F-measure; obteniendo finalmente que ANN superó a los demás algoritmos con una precisión del 86% usando TF-IDF y 87% con Word2Vec, NB tuvo el peor desempeño con precisiones del 72% (TF-IDF) y 72% (Word2Vec) y SVM logró resultados intermedios con un máximo del 84% de precisión usando Word2Vec.

El estudio concluye que ANN es el algoritmo más eficaz para el análisis de sentimientos en redes sociales, seguido de SVM. Word2Vec demostró ser más eficiente que TF-IDF en la representación de texto, especialmente para datos no estructurados como los tuits. NB fue el menos efectivo en ambas pruebas.

El estudio hace énfasis en la importancia de seleccionar una combinación adecuada de representaciones de texto y métodos de clasificación para obtener los mejores resultados.

Autores: Batra, R., Daudpota, S.M

Año: 2020

Los mercados financieros se han analizado basándose en teorías que sostienen que los precios de las acciones son impredecibles y reaccionan únicamente a nueva información. Sin embargo, investigaciones recientes han demostrado que el sentimiento público tiene un impacto significativo en los precios de las acciones

Descripción del aporte del autor

Para la investigación se obtuvo uso de tweets relacionados con Apple(2010-2017) utilizando APIs. Estos tweets incluyen etiquetas predefinidas como "bullish" (alcista) o "bearish" (bajista) teniendo un total de 300,000 tweets; asimismo se obtuvieron datos financieros históricos, como precios de apertura, cierre y volumen, desde Yahoo Finance.

Estos tweets se limpiaron eliminando URLs, emoticonos, y cualquier información irrelevante; además se realizó tokenización para dividir los textos en palabras clave. Para los datos bancarios se interpolaron valores para los días sin actividad bursátil (fines de semana y feriados).

Se utilizó SVM como clasificador para categorizar tweets como "bullish" o "bearish". Los datos se dividieron en conjuntos de entrenamiento (80%) y prueba (20%). Los datos de la tendencia se combinaron con el movimiento del precio de las acciones para generar un conjunto de datos consolidado. Finalmente, se creó un modelo para predecir si una acción debe comprarse o venderse obteniendo. La combinación de sentimiento público y datos financieros mejoró significativamente la predicción obteniendo una precisión del 91.2% frente a la precisión obtenida de 75.22% haciendo uso de datos financieros únicamente.

Autores: Picasso, A., Merello, S., Ma, Y., Oneto, L., Cambria, E.
Año: 2019

Descripción del aporte del autor

En la investigación se hacen usos de datos textuales; noticias relacionadas con las empresas, extraídas de agencias como Reuters y Bloomberg y de precios; valores de apertura, cierre, máximos y mínimos de las 20 empresas con mayor capitalización del índice NASDAQ100 (2017-2018).

Estos datos textuales se transformaron en vectores de sentimiento utilizando el Diccionario de Loughran y McDonald y AffectiveSpace. Los datos numéricos (precios) se normalizaron y se calcularon indicadores técnicos, como el índice de fuerza relativa (RSI) y las medias móviles.

Se evaluaron los algoritmos de Random Forest, Support Vector Machine y redes Neuronales Feedforward (NN). Se utilizó una técnica de validación cruzada llamada increasing window cross-validation, obteniendo finalmente que los vectores generados con AffectiveSpace superaron a los del diccionario de Loughran y McDonald, mostrando mayor precisión y equilibrio entre la clasificación de tendencias positivas y negativas. La red neuronal logró resultados más balanceados y precisos en comparación con RF y SVM, especialmente al identificar cambios relevantes en el mercado.

El modelo demuestra la viabilidad de combinar análisis técnico y fundamental con ML para predecir tendencias de mercado. Entre los hallazgos se tienen que la integración de sentimientos mejora la capacidad predictiva.

3.2.4 Métricas de evaluación

Forecasting weekly crude oil using twitter sentiment of U.S. foreign policy and oil companies

data

Autores: Zaidi, A., Oussalah, M.

Año: 2018

El estudio explora la relación entre el sentimiento expresado en Twitter sobre la política exterior de Estados Unidos y las compañías petroleras, y los movimientos en los precios del petróleo.

Descripción del aporte del autor

Para el estudio se identificaron tweets relevantes mediante cuentas de instituciones influyentes en política exterior y grandes compañías petroleras. Se emplearon las herramienta de SentiStrength y Stanford NLP Sentiment, los tweets se evaluaron para determinar la fuerza del sentimiento (positivo o negativo). Posteriormente, se analizaron características como frecuencia de menciones a "petróleo" y a miembros de la OPEP, además se utilizó la prueba de causalidad de Granger para identificar correlaciones significativas.

Los clasificadores utilizados fueron SVM (Support Vector Machine), Naïve Bayes y Redes Neuronales Artificiales (ANN). Los datos fueron etiquetados como "aumento" o "disminución" del precio para resolver un problema de clasificación binaria.

Finalmente los resultados mostraron que las características analizadas tienen un impacto significativo en el precio del petróleo con un retraso de 7 semanas, teniendo que SVM fue el clasificador más efectivo, alcanzando una precisión del 74.29% pero la combinación de los tres modelos mediante un mecanismo de votación mayoritaria elevó la precisión al 76.29%, superando modelos previos.

El estudio que este nuevo enfoque de análisis de sentimiento ofrece una precisión superior a métodos tradicionales y demuestra el potencial de los modelos no lineales para este tipo de predicciones.

Autores: Chen, Sitong & Gao, Tianhong & He, Yuqi & Jin, Yifan.
Año: 2019

La investigación explora cómo el análisis de comentarios en redes sociales puede mejorar la predicción de precios de acciones.

Descripción del aporte del autor

En la investigación se detalla en que los mercados financieros son sistemas complejos y no lineales, y que factores como la tendencia del inversor pueden influir en el comportamiento del mercado.

El conjunto de datos recopiló más de 480,000 comentarios del foro Eastmoney y precios de cierre del índice compuesto de Shanghái. Esta información fue preprocesada eliminando el contenido no relevante, posteriormente se vectorizó el texto mediante la técnica TF-IDF. Se emplearon los algoritmos de SVM (máquinas de soporte vectorial), Regresión logística, Naive Bayes, K-vecinos más cercanos (KNN), Árboles de decisión, Bosques aleatorios (Random Forest) y AdaBoost.

Se utilizó la validación cruzada K-fold para dividir los datos en 5 partes, y las métricas de evaluación de accuracy, precisión y recall; se obtuvo finalmente una precisión mínimo del 75%, con SVM logrando la mejor precisión (88.16%) seguido de Regresión logística y Naive Bayes con accuracy(~88%).

Autores: Ranjit S., Shrestha S., Subedi S., Shakya S.

Año: 2018

Descripción del aporte del autor

En el estudio se utilizan 2 enfoques, el primero es la predicción con redes neuronales artificiales (ANN) y el segundo haciendo uso del análisis de sentimiento.

Para la predicción con ANN se recopiló un histórico de precios del 2013 al 2017 teniendo como principales variables los precios del oro, petróleo, índice NASDAQ y precios de apertura, cierre, máximo y mínimo; estos datos fueron normalizados y segmentados para datos de entrenamiento y prueba; evaluando la precisión para la métrica de error cuadrático medio (RMSE).

Este modelo técnico basado en ANN con 6 nodos ocultos obtuvo un RMSE de 0.0034 (el más bajo) y el modelo fundamental con 8 nodos ocultos alcanzó un RMSE de 0.0044.

Para el enfoque del análisis de sentimientos se recopilaron tweets de operadores clasificados como positivos (1) o negativos (0). Se usó de Naive Bayes y un algoritmo basado en lexicones para el análisis, se realizó la limpieza de los tweets y la extracción de características en unigramas y bigramas; evaluando finalmente los algoritmos con las métricas de precisión, recall, y F-score para clases positivas y negativas, teniendo como resultado una precisión general del modelo de 90.625% y un F-score promedio: ~90.61%.

Se concluye finalmente que las ANN son herramientas efectivas para predecir tasas de cambio con alta precisión y que para el análisis de sentimientos los datos de redes sociales ofrecen información valiosa para predecir comportamientos de mercado.

Capítulo IV: IMPLEMENTACIÓN

4.1 Aporte Teórico

4.1.1 Metodología a seguir (Modelo conceptual)

Detallando el modelo conceptual mostrado en la *Figura 12*, tenemos que:

Se recolectarán los precios de apertura y cierre respecto al tipo de cambio del dólar. Se discretizan los valores numéricos, esto se realiza comparando todos los valores numéricos con el valor del precio de cierre anterior. La salida de esta etapa se utiliza como entrada para el modelo de predicción.

Se recolectarán los datos de los usuarios de Twitter con respecto a la variación del dólar usando la librería de Python, Snsrape. Debido a que los datos presentan inconsistencias, valores perdidos, etc., se realizará un análisis, preprocesamiento y tratamiento de estos.

Preprocesamiento:

- Conversión de texto en minúscula. – Esto se hará a fin de reducir el diccionario de palabras a fin de no reconocer palabras iguales en esencia (diferencia en mayúsculas) como diferentes.
- Eliminación de Hashtag e Identificadores. – Ambas características son muy utilizadas por los usuarios de Twitter, pero no aportan valor al análisis.
- Eliminación de Stopwords. – Estas palabras no aportan valor al análisis.
- Tokenización. – Segmentan los tweets en palabras independientes.
- Stemming. – Tiene como objetivo eliminar afijos, extensiones de inflexión a fin de obtener la raíz de la palabra

Posteriormente se realizará la vectorización de texto utilizando n-gram y Tf-Idf. Con los datos preprocesados, se utilizarán 3 algoritmos:

Machine Learning:

- Support Vector Machine (SVM)
- Naive Bayes (NB)
- Random Forest (RF)

La etapa de entrenamiento y validación de la fase inicial se usará una partición por porcentajes para los algoritmos de NB y RF, la segunda etapa se llevará a cabo bajo una partición por porcentajes para Random Forest y una validación cruzada para el algoritmo de SVM de $K=5$ (K-folds cross validation) que divide el total de los datos en K partes, seleccionando una parte para test y el resto para entrenamiento, realizando este procedimiento K veces.

Posteriormente a la primera fase, se buscará optimizar el rendimiento del modelo, utilizando distintas técnicas de preprocesamiento, etc., y se volverá a evaluar con el método mencionado anteriormente. La salida de esta etapa será input para la segunda fase junto con el corpus de datos bancarios.

Se unirán los resultados de la primera etapa y el dataset bancario según el día. Se usará este dataset mixto para la creación de los nuevos modelos. Finalmente se realizará un forecasting sobre el tipo de cambio del dólar.

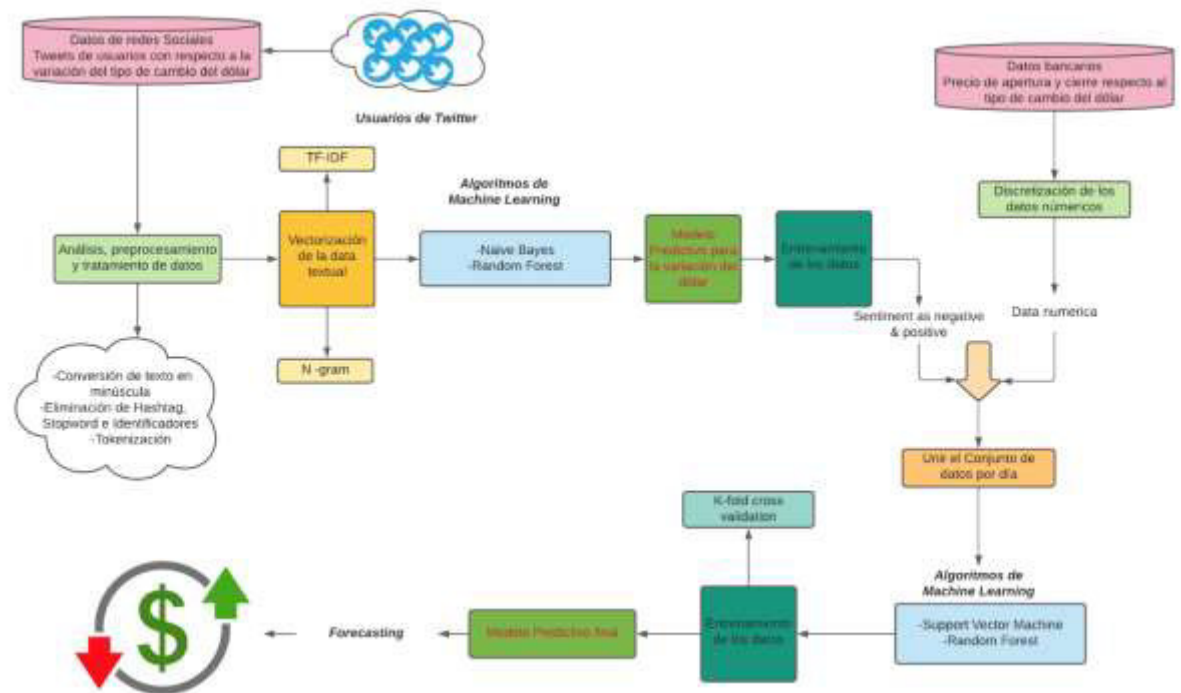
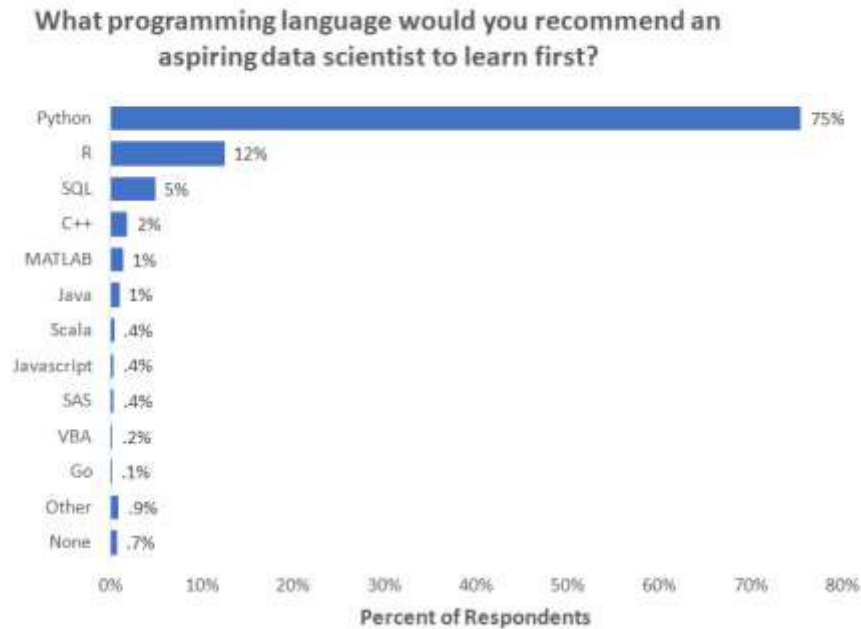


Figura 10. Modelo Conceptual. Nota. Adaptación de (Chen, 2019)

4.1.2 Selección y Adaptación de la tecnología

Lenguaje de Programación

En la *Figura 15* para el desarrollo de la investigación se está evaluando 3 lenguajes top a usar (Python, R y SQL), dichos lenguajes se evaluaron según los rankings top de los lenguajes para la ciencia de datos de Hayes (2019) y Piatetsky (2019), posteriormente se filtran los lenguajes de programación según el dominio personal de los mismos y finalmente se selecciona Python.



Note: Data are from the 2018 Kaggle ML and Data Science Survey. You can learn more about the study here: <http://www.kaggle.com/kaggle/kaggle-survey-2018>. A total of 23859 respondents completed the survey; the percentages in the graph are based on a total of 18788 respondents who provided an answer to this question.



Copyright 2019 Business Over Broadway

Figura 11. Lenguaje de programación recomendados. *Nota.* Tomado de (Hayes, Business Broadway, 2019)

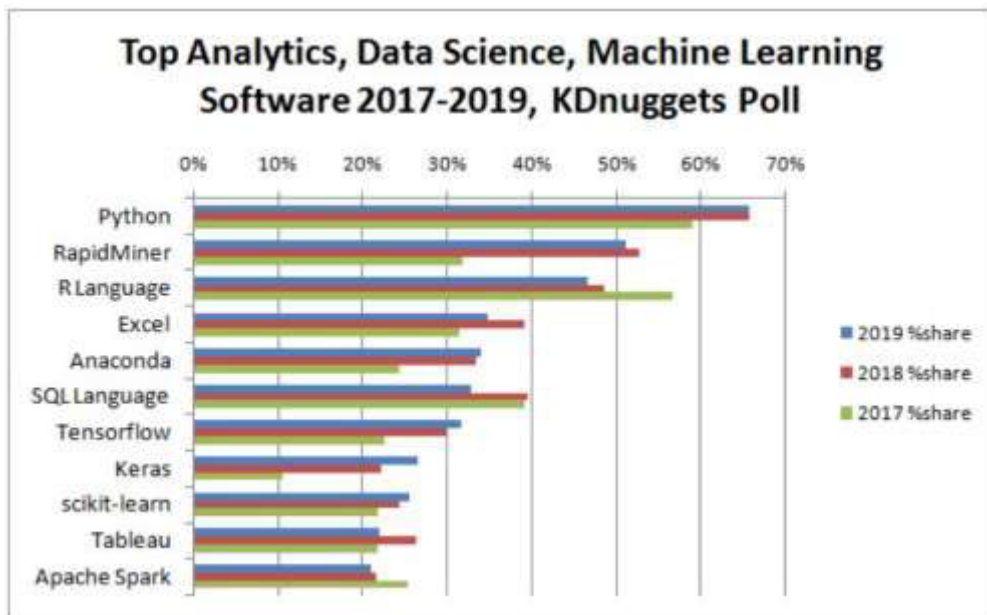


Figura 12. KDnuggets Analytics/Data Science 2019 Software Poll. *Nota.* Tomado de (Piatetsky, 2019)

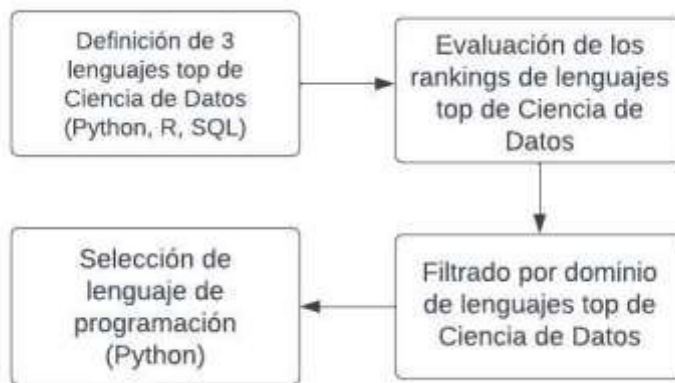


Figura 13. Proceso de selección de lenguaje de programación

Entorno de desarrollo (IDE)

En la Figura 17 para el desarrollo de la investigación se está evaluando 3 IDE top a usar (Visual Studio, Jupyter y Spyder), dichos IDE se evaluaron según los rankings top IDE usados para la ciencia de datos de Nehme (2023) y Hayes (2020) posteriormente se filtraron según el dominio personal de los mismos para finalmente seleccionar Jupyter Notebook.

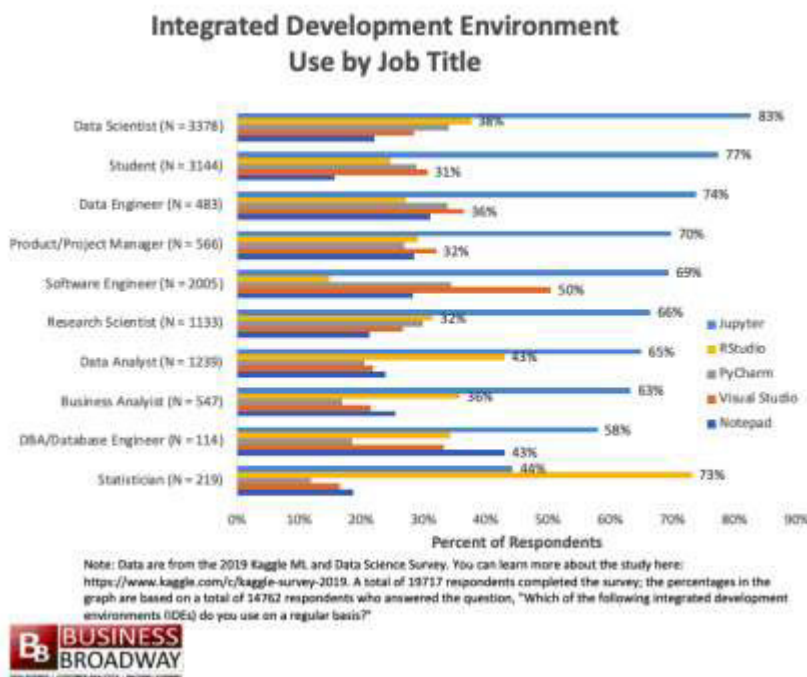


Figura 14. Integrated Development Environment by Job Title. **Nota.** Tomado de (Hayes, Business Broadway, 2020)

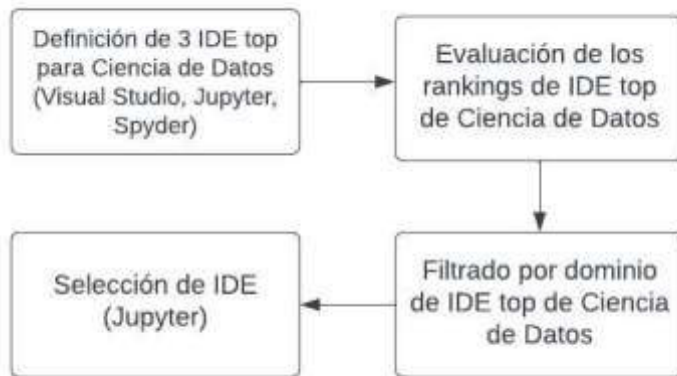


Figura 15. Proceso de selección de IDE

4.2 Aporte Práctico

4.2.1 Diseño de la solución

La implementación del proceso predictivo se implementa por medio de fases, las cuales dependen entre sí de forma secuencial como de forma cíclica con el fin de ajustar la precisión finalmente obtenida.

Detallando la Fase de implementación mostrada en la *Figura 18*, tenemos:

Fase 1: Análisis. – Esta fase comienza con la recopilación de los tweets de los usuarios con respecto a la variación del dólar y posteriormente se reconocen las características generales de los datos realizando un análisis exploratorio.

Fase 2: Preparación. - La fase de preparación comienza con las actividades para el tratamiento del dataset, realizando técnicas de preprocesamiento como: ‘Eliminación de Hashtag’, ‘Stemming’, etc. Asimismo, se realizará la vectorización haciendo uso de Tf-idf y N-gram.

Fase 3: Modelado. - La fase de modelado es donde se aplican las técnicas de machine learning y se realiza su entrenamiento. Debido a la variedad de técnicas también se ajustarán los parámetros a fin de satisfacer los requisitos específicos de formato de datos.

Fase 4: Evaluación. - En la fase de evaluación se realizará la evaluación del modelo a fin de determinar su precisión utilizando data del BCRP.

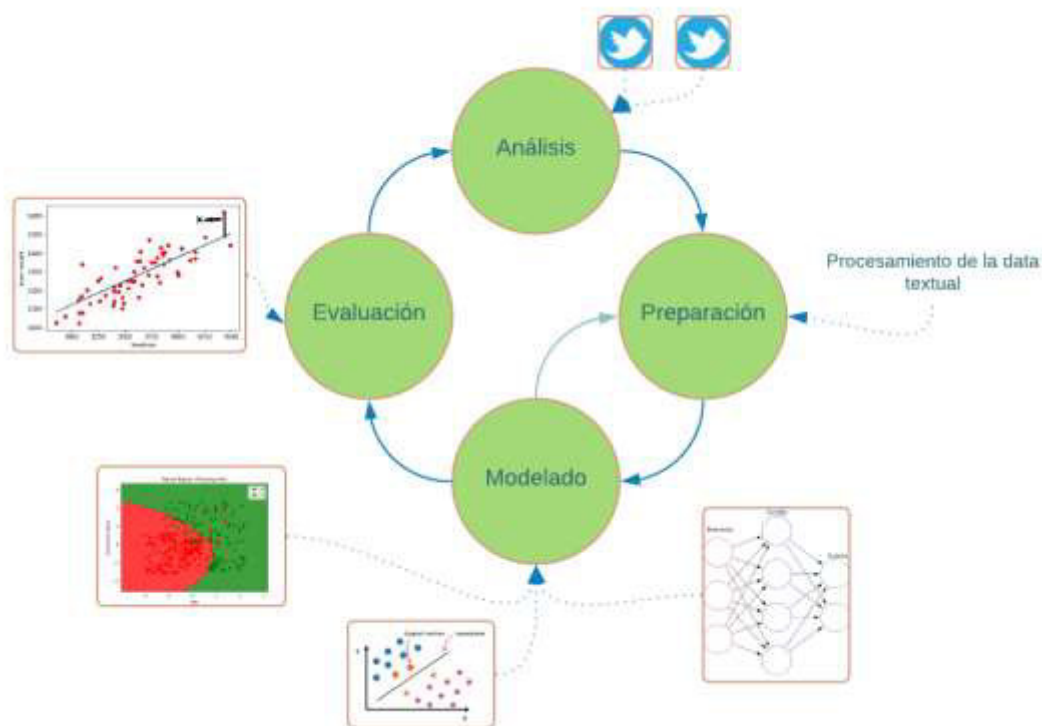


Figura 16. Fase de Implementación. Nota. Adaptación de (Pimpalkar, 2020), (Ranjit S. S. S., 2018)

4.2.1.1 Arquitectura Tecnológica

Detallando la Arquitectura tecnológica mostrada en la *Figura 19*, tenemos:

En la **Capa de Interfaz** se dará esta interacción con el usuario permitiendo ver los resultados de los modelos, gráfica de la tendencia y el forecasting del tipo de cambio del dólar.

En la **Capa de Aplicación** se hará uso principalmente de Python para realizar la recopilación de las fuentes de datos, su preparación que servirá de entrada al modelo, la construcción del modelo y su entrenamiento y posterior evaluación.

En la **Capa de Acceso a datos** es donde se tendrán los dataset a trabajar que están segmentados por geolocalización (Perú).

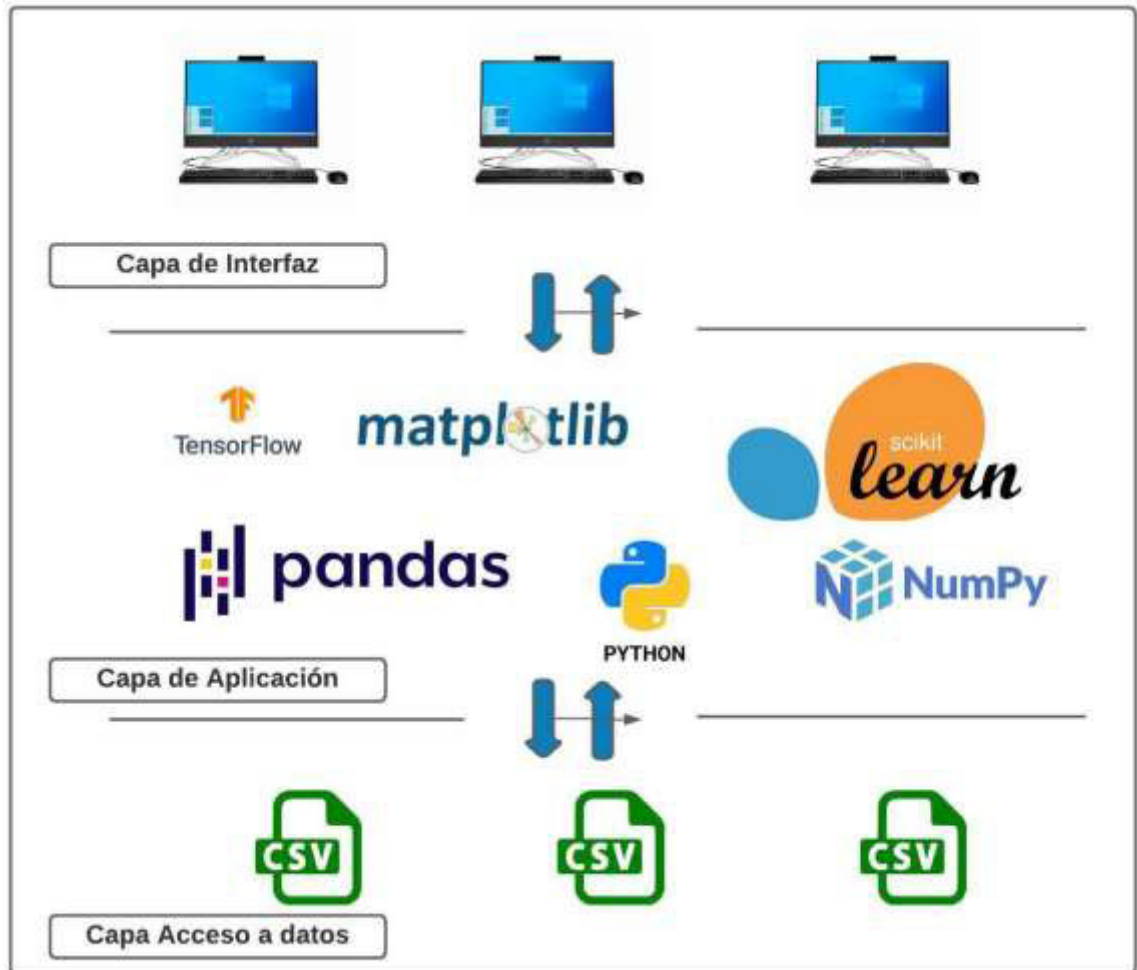


Figura 17. Arquitectura tecnológica.

4.2.2 Implementación de la solución

Fase 1: Análisis

Construcción de los Dataset

En esta investigación sobre la variación del dólar se trabajará con 2 conjuntos de datos(dataset) almacenados en archivos “.csv”.

Conjunto de datos de Twitter

La recopilación del dataset de Twitter se recopila utilizando la librería Snsrape de Python sobre la red social Twitter, considerando el keyword: 'dólar'. Además, el rango temporal de los tweets se limitará del 01 de enero del 2021 al 31 de diciembre del 2021.

Entre las características principales a tratar del dataset están:

'content'. - Proporciona la información textual del tweet del cual se obtendrá la tendencia.

'date'. - Proporciona la fecha del tweet, permitirá agrupar los tweets por día.

'id'. - Proporciona el id del tweet y permitirá eliminar duplicados.

'retweetdTweet'. - Permitirá diferenciar los tweets de los retweets, estos últimos no proporcionan nueva información por lo que serán separados del dataset.

'lang'. - Nos indica el idioma del tweet

'user'. - Proporciona las características del usuario de la cual se obtendrá la subcaracterística 'location', esta será concatenada al '.csv' ('None' en caso de vacíos) como 'localización Usuario' y posteriormente utilizada para filtrar el dataset y tener los tweets recopilados geolocalizados en Perú.

Los tweets para la construcción del modelo están formados por 28 variables y la volumetría como se muestra en la *Figura 20*. La recopilación de los metadatos de Twitter fue mensual, y el análisis de datos fue diario en el rango indicado. Los resultados de Perú fueron filtrados, obteniendo un total de es de 78278 registros para el dataset total.

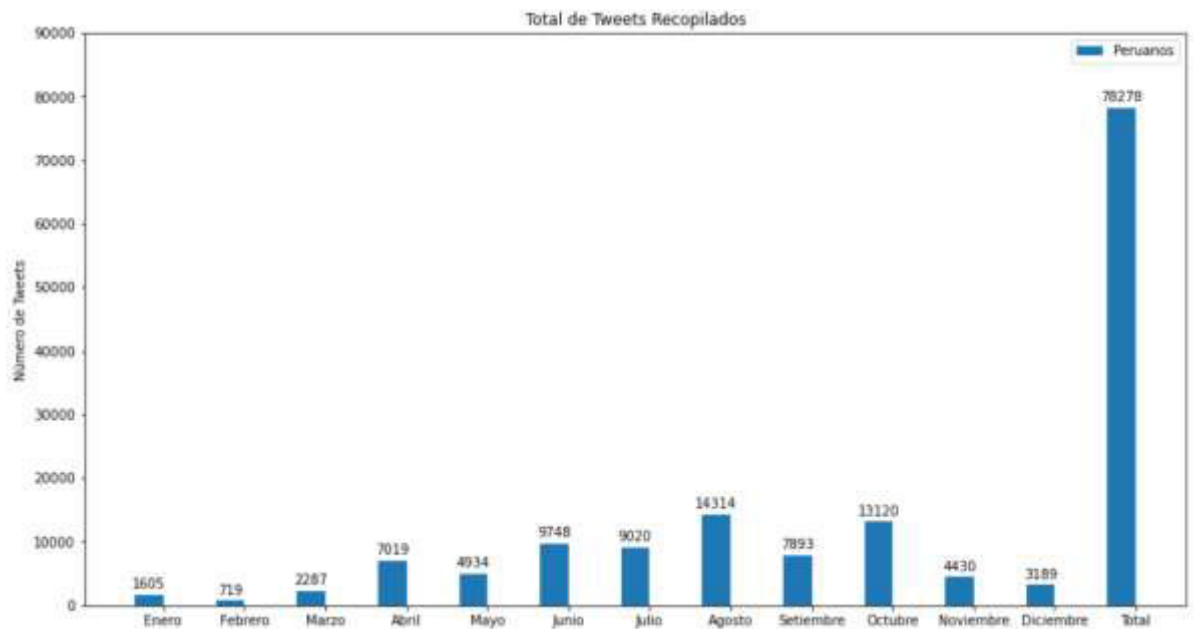


Figura 18. Diagrama de barras de los Tweets.

La muestra de los tweets recopilados se muestra en las siguientes figuras, en la Figura 21 se visualiza los tweets cuya localización de usuario se establece en Perú o Lima que contienen el keyword de ‘dólar’.

date	content	id	user	retweetCount	likeCount	lang	retweetedTweet	localizacionUsuario
2021-01-01 0:	Tipo de cambi	1.3448E+18	{'username': '}	4	0	es	None	PerÃ
2021-01-01 0:	Precio del dÃ	1.3448E+18	{'username': '}	2	0	es	None	Lima, PerÃ
2021-01-01 0:	Tipo de cambi	1.3448E+18	{'username': '}	4	0	es	None	PerÃ
2021-01-01 0:	@VeroLinare:	1.3448E+18	{'username': '}	0	0	es	None	Barranco, Lima Peru
2021-01-01 0:	Tipo de cambi	1.3448E+18	{'username': '}	4	0	es	None	PerÃ
2021-01-01 0:	Precio del dÃ	1.3449E+18	{'username': '}	4	0	es	None	PerÃ
2021-01-01 0:	@JBCPERU Y	1.3449E+18	{'username': '}	0	2	es	None	Lima, Peru
2021-01-01 1:	DÃlar blue: a	1.3449E+18	{'username': '}	4	1	es	None	Lima, Peru
2021-01-01 1:	Peso cierra su	1.3449E+18	{'username': '}	2	0	es	None	Lima, PerÃ
2021-01-01 1:	DolarToday: p	1.345E+18	{'username': '}	4	0	es	None	Lima, Peru
2021-01-01 1:	Venezuela: Pr	1.345E+18	{'username': '}	4	0	es	None	Lima, Peru
2021-01-01 1:	Precio del dÃ	1.345E+18	{'username': '}	4	0	es	None	PerÃ
2021-01-01 1:	Tipo de cambi	1.345E+18	{'username': '}	4	0	es	None	Peru

Figura 19. Muestra de datos - Dataset Perú.

Conjunto de datos Bancarios

El dataset bancario se obtiene de los datos del BCRP para el tipo de cambio del dólar. Los valores numéricos se discretizaron, todos los valores numéricos discretizados se compararon con los precios de apertura y cierre del día anterior para predecir el valor del tipo de cambio del dólar del día siguiente. Se visualiza en la Figura 22 la muestra de datos recopilados del BCRP.

Fecha	Tipo de cambio - TC Interbancario (\$/ por US\$) - Compra	Tipo de cambio - TC Interbancario (\$/ por US\$) - Venta	Tipo de cambio - TC Sistema bancario SBS (\$/ por US\$) - Compra	Tipo de cambio - TC Sistema bancario SBS (\$/ por US\$) - Venta	Tipo de cambio - TC 11:00 AM (\$/ por US\$) - Compra	Tipo de cambio - TC 11:00 AM (\$/ por US\$) - Venta	Tipo de cambio - TC 01:30 PM (\$/ por US\$) - Compra	Tipo de cambio - TC 01:30 PM (\$/ por US\$) - Venta
04Ene21	3.624	3.627	3.624	3.628	3.623	3.624	3.628	3.629
05Ene21	3.628	3.634	3.627	3.631	3.628	3.629	3.628	3.630
06Ene21	3.625	3.627	3.625	3.630	3.627	3.629	3.624	3.626
07Ene21	3.621	3.623	3.620	3.623	3.621	3.622	3.620	3.624
08Ene21	3.611	3.614	3.610	3.615	3.614	3.615	3.606	3.610
11Ene21	3.615	3.617	3.615	3.618	3.619	3.620	3.612	3.613
12Ene21	3.606	3.609	3.606	3.608	3.603	3.606	3.610	3.611
13Ene21	3.613	3.614	3.610	3.615	3.613	3.615	3.614	3.616
14Ene21	3.611	3.614	3.610	3.613	3.608	3.610	3.609	3.611
15Ene21	3.611	3.613	3.610	3.614	3.611	3.612	3.613	3.615
18Ene21	3.611	3.617	3.609	3.617	3.612	3.615	3.615	3.616
19Ene21	3.612	3.613	3.611	3.615	3.612	3.613	3.614	3.615
20Ene21	3.613	3.615	3.612	3.615	3.614	3.615	3.615	3.617
21Ene21	3.619	3.622	3.618	3.622	3.620	3.621	3.623	3.627

Figura 20 . Muestra de datos (Tipos de cambio del dólar) – BCRP. Nota. Tomado de BCRP Data (Banco central de Reserva del Perú, 2021).

Fase 2 Preparación:

Preprocesamiento del Texto

Recopilado el dataset en el formato “.csv” pasamos a utilizar Python para realizar el preprocesamiento de los datos.

El proceso visualizado en la Figura 23 se describirá a continuación:

Como primer paso realizamos la tokenización de los tweets, el cual consiste en separar el texto palabra por palabra. Una vez obtenidos los tokens se eliminarán los hashtags, menciones de usuario, los signos de puntuación, las “url”, números y emoticones que puedan existir en los tweets recopilados haciendo uso de expresiones regulares. Además, se eliminarán los tokens que formen parte de los stopwords de NLTK para el idioma español, los cuales no generan valor para el análisis.

Finalmente se realizará el Stemming, que es el proceso de regresar la palabra (token) a su raíz o “stem”.

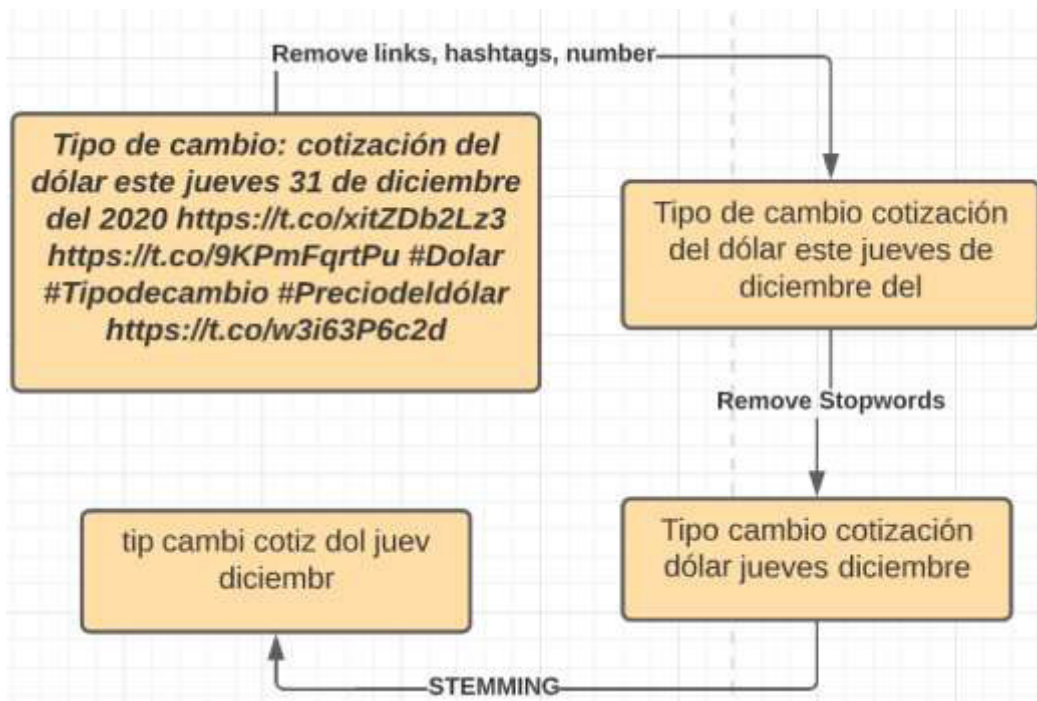


Figura 21. Preprocesamiento. *Nota.* Adaptación de (Batra, 2018), (Chen, 2019)

Vectorización del texto

Una vez culminado el preprocesamiento del texto se utilizarán los tokens, agrupándose según la fecha del tweet teniendo finalmente un conjunto de tokens para cada día y se determinara la tendencia o sentimientos diarios según el total de tweets catalogados como positivos o negativos del día. Estos tokens diarios serán segmentados según el gramo asignado, en esta investigación se harán uso de Unigramas y Bigramas.

Tabla 2

Unigramas vs Bigramas

N-gram	Días Analizados	Columnas
Unigramas	365	20809
Bigramas	365	350218

Para la vectorización de utilizará “TfidfVectorizer” la biblioteca de Python “sklearn” estableciendo como parámetro el rango de ngram a utilizar, Unigrama=1-gram y Bigrama=2-gram. Obteniendo finalmente, como se observa numéricamente en la Tabla 2 y gráficamente

en la *Figura 24*, que al momento de utilizar 2-gram aumenta la cantidad de datos a comparación del uso de 1-gram.

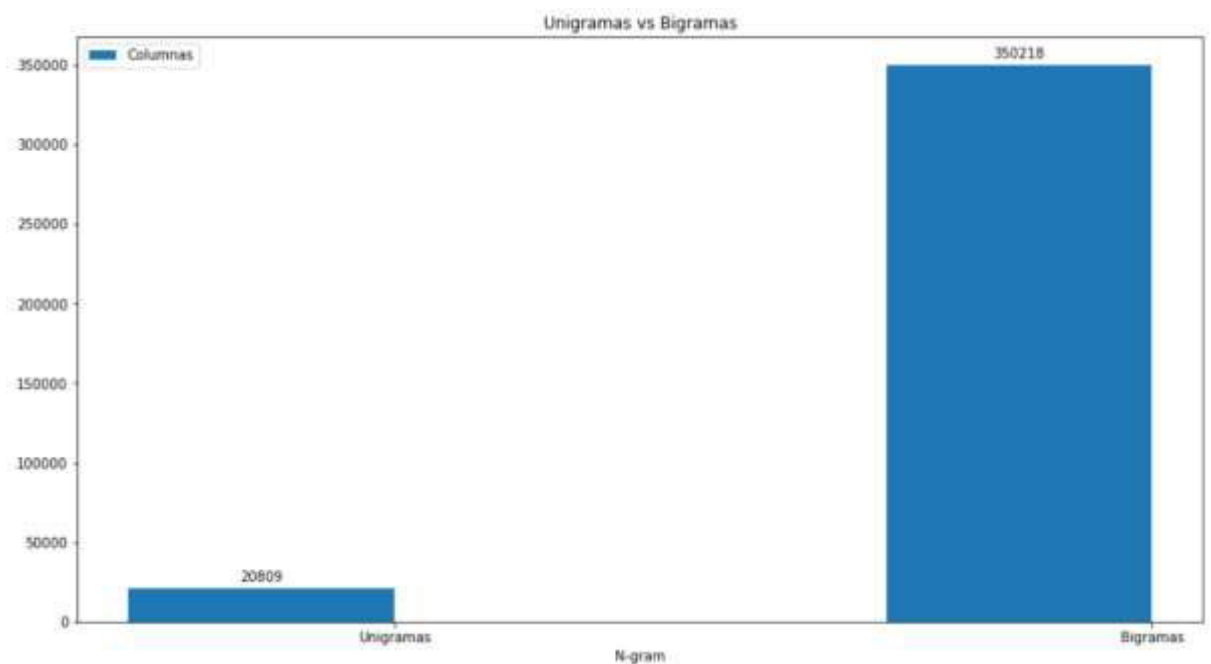


Figura 22. Unigramas vs Bigramas, diagrama de barras.

Posterior al realizado de la vectorización se tendrá un nuevo dataset puesto que se han aginado valores a cada token mediante termino-frecuencia(tf) y la frecuencia inversa de documento(idf), teniendo como output las siguientes características mostradas en la *Figura 25* cuando se usa Unigramas y la *Figura 26* cuando se usa Bigramas.

Dia	abalanz	abancay	abandon	abandonast	abarat	abarc	abarrot	abarroto
1	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0
12	0	0	0.04080037	0	0	0	0	0

Figura 23. Muestra de datos. Tf-Idf Unigramas.

modelo de Random Forest Classifier, y cuyo resultado de la estimación se obtienen como mejores parámetros los siguientes valores:

- N_estimadores: 100
- Max_features: “sqrt”
- Max_depth: None
- Criterion: “gini”

Tabla 4

Parámetros contemplados para RandomForestClassifier

N_estimadores	Max_features	Max_depth	Criterion
100	"sqrt"	None	"gini"
200	"log2"	2	"entropy"
300		4	"log_loss"
400		6	
500		8	

Evaluación de Desempeño - Clasificación

En la primera fase para evaluar el desempeño de los modelos se hará uso de Precisión, recall, f1_score y el accuracy; para cada uno de los dataset resultantes de la vectorización del texto, es decir, utilizando Unigramas y Bigramas. Los resultados se encuentran en las siguientes tablas: Tabla 5, Tabla 6 y Tabla 7.

Tabla 5

Muestreo 70-30

Avarage Micro				
Algoritmos	Precision	Recall	F1_Score	Accuracy
NaiveBayes - Unigram	0.8984375	0.8984375	0.8984375	0.8984375
NaiveBayes - Bigram	0.8984375	0.8984375	0.8984375	0.8984375
RandomForest - Unigram	0.8984375	0.8984375	0.8984375	0.8984375
RandomForest - Bigram	0.8984375	0.8984375	0.8984375	0.8984375

En la Tabla 5 se muestra los resultados usando la partición porcentual de 70% de los datos para entrenar los modelos y 30% para las pruebas de ambos algoritmos con los datos de Unigramas y Bigramas.

Tabla 6

Muestreo 75-25

Algoritmos	Avarage Micro			
	Precision	Recall	F1_Score	Accuracy
NaiveBayes - Unigram	0.901459854	0.901459854	0.901459854	0.901459854
NaiveBayes - Bigram	0.901459854	0.901459854	0.901459854	0.901459854
RandomForest - Unigram	0.901459854	0.901459854	0.901459854	0.901459854
RandomForest - Bigram	0.901459854	0.901459854	0.901459854	0.901459854

En la Tabla 6 se muestra los resultados usando la partición porcentual de 75% de los datos para entrenar los modelos y 25% para las pruebas de ambos algoritmos con los datos de Unigramas y Bigramas.

Tabla 7

Muestreo 80-20

Algoritmos	Avarage Micro			
	Precision	Recall	F1_Score	Accuracy
NaiveBayes - Unigram	0.897260274	0.897260274	0.897260274	0.897260274
NaiveBayes - Bigram	0.897260274	0.897260274	0.897260274	0.897260274
RandomForest - Unigram	0.897260274	0.897260274	0.897260274	0.897260274
RandomForest - Bigram	0.897260274	0.897260274	0.897260274	0.897260274

En la Tabla 7 se muestran los resultados usando la partición porcentual de 80% para entrenar los modelos y 20% para las pruebas de ambos algoritmos con los datos de Unigramas y Bigramas.

Haciendo una comparación de resultados de las Tablas 5, 6 y 7 se puede visualizar que cuando se hacen uso de los dataset unigram y bigram se obtienen resultados iguales para los

modelos según las particiones porcentuales antes mencionadas, pero al comparar los resultados de las distintas particiones porcentuales se encuentra que se obtienen mejores resultados cuando los modelos son entrenados con el 75% de los datos, resaltando aumentando el porcentaje de los datos usados para el entrenamiento no necesariamente se aumenta el accuracy de los modelos, por el contrario, se ven reducidos.

Una vez obtenido los datos de los sentimientos diarios de los tweets se usarán como entrada junto a los datos bancarios obtenidos del BCRP, estos datasets se unirán según el día teniendo como target el promedio de los valores del dólar del día siguiente.

Para esta segunda fase se hará con Support vector machine y RandomForestRegressor realizando un proceso análogo a la fase anterior para Random Forest, en cambio, para SVM se hará uso de la validación cruzada con un $K=5$.

Haciendo uso de “GridSearchCV” hallamos los parámetros para ambos algoritmos, En las Tabla 8 y 9 se muestran los parámetros contemplados para Random Forest y SVM respectivamente.

En la Tabla 8 se muestran los parámetros evaluados para el modelo de Random Forest Regressor, y cuyo resultado de la estimación se obtienen como mejores parámetros los siguientes valores:

Random Forest

- N_estimadores: 100
- Max_features: “sqrt”
- Max_depth: 6
- Criterion: “squared_error”

En la Tabla 9 se muestran los parámetros evaluados para el modelo de SVM y cuyo resultado de la estimación se obtienen como mejores parámetros los siguientes valores:

SVM

- C: 10000
- Epsilon: 0.0000006 (1e-06)
- Gamma: 0.001

Tabla 8

Parámetros contemplados para RandomForestRegressor

N_estimadores	Max_features	Max_depth	Criterion
100	"sqrt"	None	"squared_error"
200	"log2"	2	"absolute_error"
300		4	"poisson"
400		6	
500		8	

Tabla 9

Parámetros contemplados para SVM

C	Épsilon	Gamma
0.0001	0.000001	0.000001
0.01	0.00001	0.00001
0.1	0.0001	0.0001
1	0.001	0.001
10	0.01	0.01
100	0.1	0.1
1000	1	1
10000	10	10
100000	100	100
1000000	1000	1000
10000000	10000	10000

Evaluación de desempeño – Regresión

Una vez obtenido los parámetros para el modelo se pasará a la etapa de entrenamiento como se observa en la Figura 12. El modelo de Random Forest Classifier será entrenado y testeado según las particiones porcentuales mostradas en la *Tabla 3*. Finalmente, para medir el desempeño de los algoritmos se hará uso del error medio cuadrático y el error medio absoluto; los resultados los modelos se encuentran en las siguientes tablas: Tabla 10 y Tabla 11.

Tabla 10

Resultados para Random Forest Regressor

Algoritmo - Muestreo	MAE (Mean Absolute Error)	MSE (Mean Squared Error)
Random Forest - 70%	0.012644025268554689	0.0003230630295872753
Random Forest - 75%	0.01338296475821048	0.0003691462910915809
Random Forest - 80%	0.013848965111304028	0.00040495040295541156

En la Tabla 10 se puede visualizar los resultados alcanzados para el modelo de Random Forest Regressor usando las particiones de 70%-30%, 75%-25% y 80%-20% según los errores (MAE y MSE) para los datos predichos, teniendo como mejor modelo de RF cuando se utiliza la partición 70%-30%.

Tabla 11

SVM model results using cross-validation

Algorithm	Score Mean	MAE (Mean Absolute Error)	MSE (Mean Squared Error)
SVM	0.92826181	0.009605179	0.000232213

En la Tabla 11 se muestra el resultado del algoritmo SVM mediante validación cruzada. Se obtienen 0,009605179 y 0,000232213 para las métricas MAE y MSE, respectivamente. Teniendo así que para los resultados de los modelos de la segunda fase mostrados en la Tabla 10 y Tabla 11 es evidente que en la segunda fase de la investigación se obtienen mejores resultados del porcentaje de corrección con la técnica algorítmica SVM en comparación del modelo de RF.

4.3. Validación

4.3.1 Algoritmo

Para la primera fase de la investigación se han utilizado dos algoritmos, RandomForestClassifier y NaiveBayes (Multinomial), y para la segunda fase de la

investigación se han utilizado otros dos algoritmos, RandomForestRegresor y Support Vector Machine.

4.3.1.1 Random Forest Classifier

Se implementó el modelo Random Forest Classifier haciendo uso de la data Unigram, asimismo el uso de la data Bigram obteniendo el score y métricas de evaluación, ver detalle en el [Anexo 1](#).

4.3.1.2 Naïve Bayes

Se implementó el modelo Multinomial Naive Bayes haciendo uso de la data Unigram, asimismo el uso de la data Bigram obteniendo el score y métricas de evaluación, ver detalle en el [Anexo 2](#).

4.3.1.3. Random Forest Regresor

Se implementó el modelo Random Forest Regresor haciendo uso de la data resultante de la primera fase de la investigación obteniendo el score y errores manejados, ver detalle en el [Anexo 3](#).

4.3.1.4. Support Vector Machine

Se implementó el modelo Support Vector Machine haciendo uso de la data resultante de la primera fase de la investigación obteniendo el score y errores manejados, ver detalle en el [Anexo 4](#).

4.3.2 Hardware y Software

Como parte de las instalaciones previas y las características del equipo requerido necesario mínimamente para lograr el desarrollo del proyecto se tiene a continuación el siguiente listado:

- 1 laptop o computadora con un procesador Intel core7, memoria RAM de 16GB y espacio libre mínimo de 200GB.
- Anaconda (Python 3.8) – Jupyter Notebook.
- Librerías de Python utilizadas en el experimento: Sklearn, Numpy, Pandas, re, String y Nltk.

4.3.3 Instancia de pruebas

Para las pruebas sobre los modelos se ha recopilado la data bancaria como los tweets del mes de enero del 2021. Se realizó el tratamiento de la data bancaria y se determinaron las tendencias del mes utilizando el modelo con los mejores resultados de la primera fase a fin de servir como entrada para el forecasting del dólar.

Tabla 12.

Muestra datos Bancarios - Pruebas

Compra11	Venta11	Compra13	Venta13	PromedioSiguiente
3.9915	3.995	3.975	3.9755	3.98425
3.9915	3.995	3.975	3.9755	3.975
3.997	3.998	3.96	3.96	3.96025
3.96	3.964	3.957	3.96	3.95625
3.95	3.952	3.96	3.963	3.96975
3.971	3.972	3.966	3.97	3.94275
3.954	3.956	3.93	3.931	3.938
3.943	3.945	3.9315	3.9325	3.938
3.943	3.945	3.9315	3.9325	3.93325

Como se muestra en la Tabla 12, los datos bancarios a utilizar para las pruebas se han procesado parcialmente. A estos datos se le procesará una vez más, para obtener así los datos de entrada (Compra, venta y Sentiment o target de la primera fase) y el objetivo (promedio siguiente), los cuales son requeridos por los modelos.

4.3.4 Resultados

Una vez realizado el procesamiento de los datos obtenemos los errores de las predicciones realizadas para los modelos.

Tabla 13

Resultados Obtenidos de las pruebas

Algoritmo	MAE (Mean Absolute Error)	MSE (Mean Squared Error)
Random Forest - 70%	0.014454047379031597	0.000307344792752
Random Forest - 75%	0.017775811491925288	0.00046453315948733
Random Forest - 80%	0.014989979838719098	0.000336327482781309
SVM	0.009219240037769527	0.0001580308508933825

En la Tabla 13 se pueden visualizar los resultados de las pruebas en las que se evidencia como el modelo con mejores resultados a SVM, a fin de una mejor visualización de los resultados se han desarrollado una serie de gráficas lineales y de barras.

En la Figura 27 se visualiza en el diagrama de líneas que usando los valores estimados(predichos) para el valor real o Target(objetivo) cuando se usa el modelo de Random Forest para la partición del 70% de los datos, se aproximan en gran medida a los valores reales del promedio del valor del dólar del día siguiente demostrando que el modelo genera una predicción balanceada.

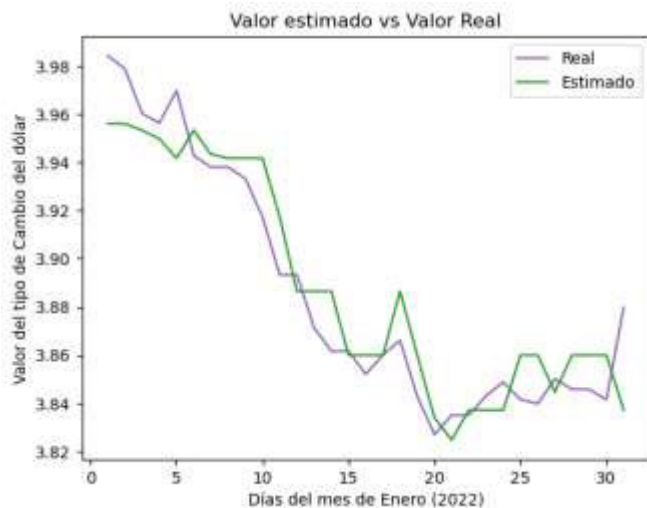


Figura 25. Random Forest Model – 70% Diagrama de líneas

En la Figura 28 se visualiza en el diagrama de barras que usando los valores estimados(predichos) para el valor real o Target(objetivo) cuando se usa el modelo de Random Forest para la partición del 70% de los datos, se aproximan en gran medida a los valores reales del promedio del valor del dólar del día siguiente demostrando que el modelo genera una predicción balanceada.

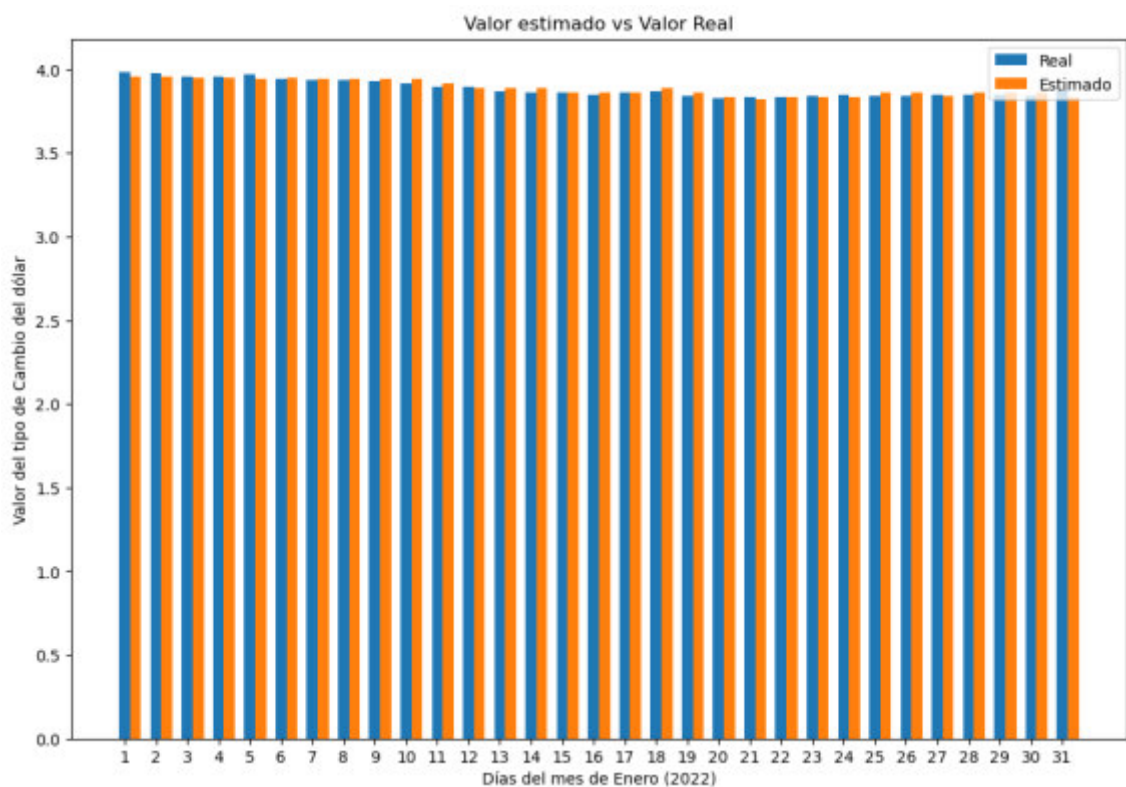


Figura 26. Random Forest Model – 70% Diagrama de Barras

En la Figura 29 se visualiza en el diagrama de líneas que usando los valores estimados(predichos) para el valor real o Target(objetivo) cuando se usa el modelo de Random Forest para la partición del 75% de los datos, se aproximan en gran medida a los valores reales del promedio del valor del dólar del día siguiente demostrando que el modelo genera una predicción balanceada.

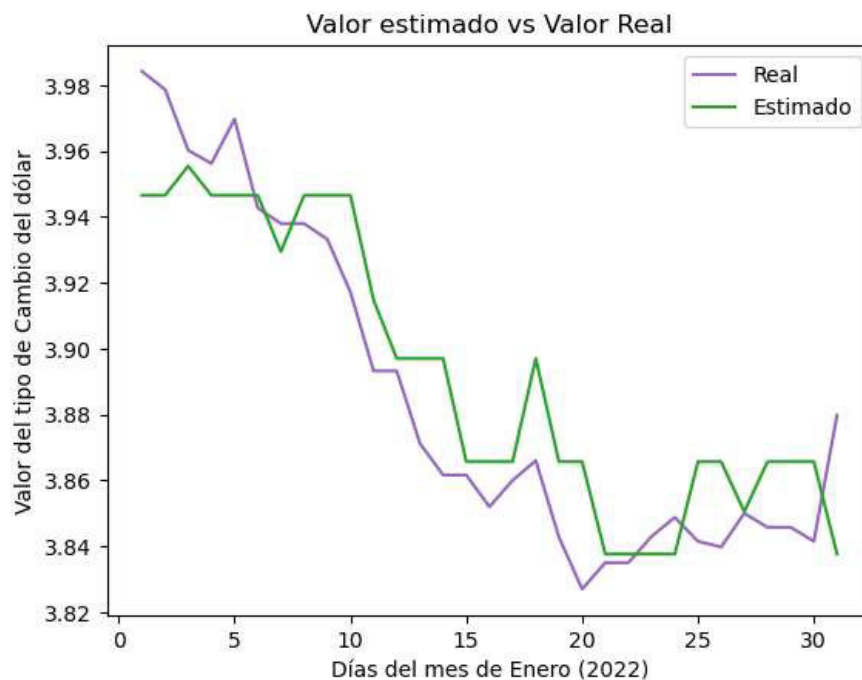


Figura 27. Random Forest Model - 75% Gráfica de líneas

En la Figura 30 se visualiza en el diagrama de barras que usando los valores estimados(predichos) para el valor real o Target(objetivo) cuando se usa el modelo de Random Forest para la partición del 75% de los datos, se aproximan en gran medida a los valores reales

del promedio del valor del dólar del día siguiente demostrando que el modelo genera una predicción balanceada.

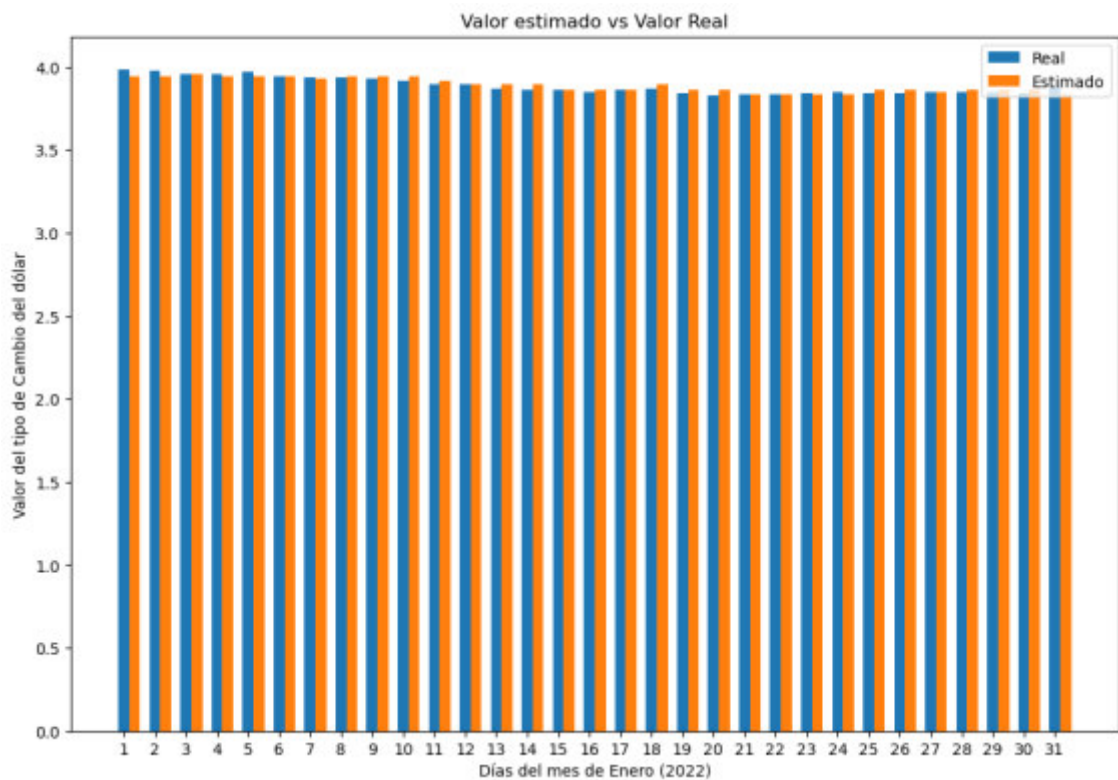


Figura 28. Random Forest Model - 75% Gráfica de Barras

En la Figura 31 se visualiza en el diagrama de líneas que usando los valores estimados(predichos) para el valor real o Target(objetivo) cuando se usa el modelo de Random Forest para la partición del 80% de los datos, se aproximan en gran medida a los valores reales

del promedio del valor del dólar del día siguiente demostrando que el modelo genera una predicción balanceada.

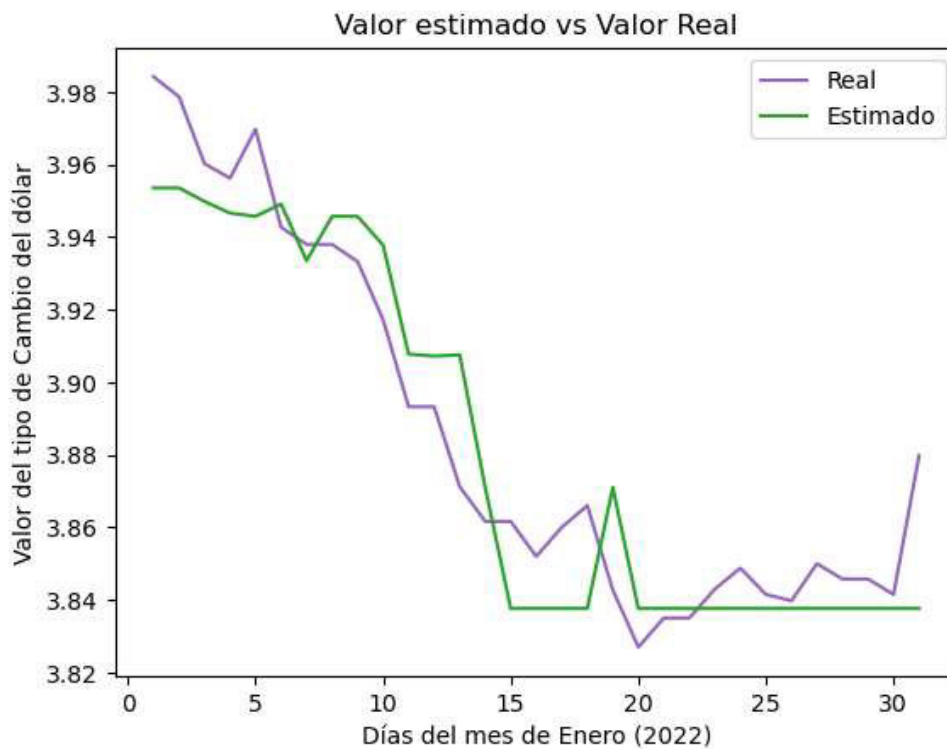


Figura 29. Random Forest Model - 80% Gráfica de líneas

En la Figura 32 se visualiza en el diagrama de barras que usando los valores estimados(predichos) para el valor real o Target(objetivo) cuando se usa el modelo de Random Forest para la partición del 80% de los datos, se aproximan en gran medida a los valores reales del promedio del valor del dólar del día siguiente demostrando que el modelo genera una predicción balanceada.

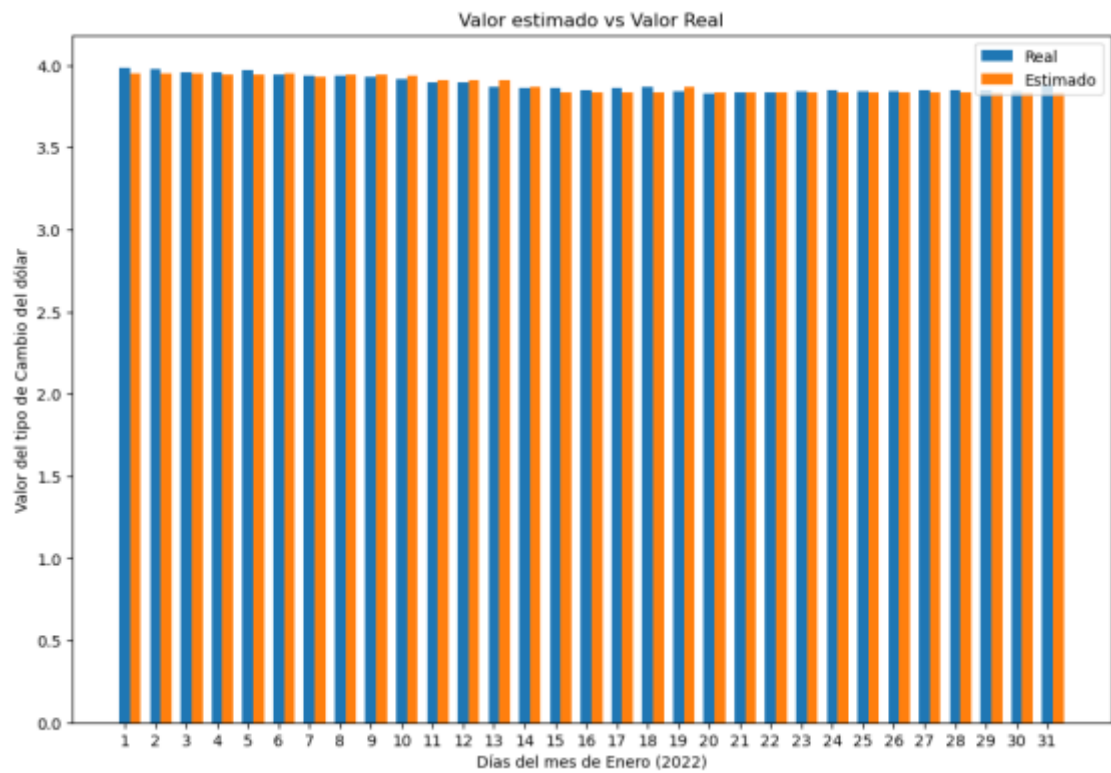


Figura 30. Random Forest Model - 80% Gráfica de líneas

En la Figura 33 se visualiza en el diagrama de líneas que usando los valores estimados(predichos) para el valor real o Target(objetivo) cuando se usa el modelo de support

vector machine no se distancian en gran medida de los valores reales del promedio siguiente.

De hecho, el error en la predicción es menor.

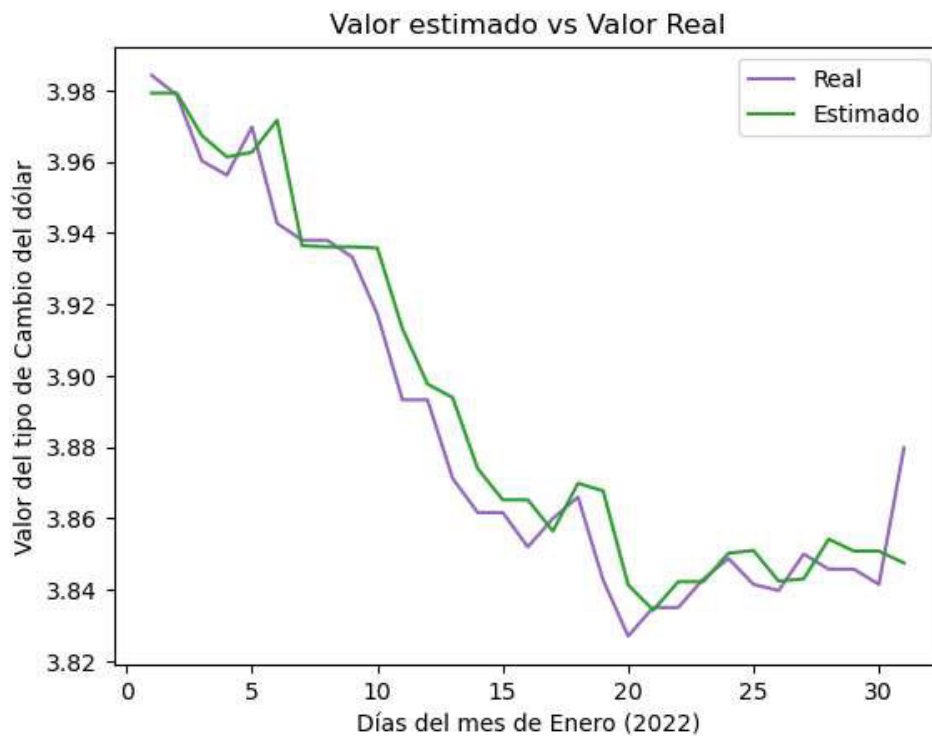


Figura 31. SVM Model - Gráfica de líneas

En la Figura 34 se visualiza en el diagrama de barras que usando los valores estimados(predichos) para el valor real o Target(objetivo) cuando se usa el modelo de Support Vector Machine no se distancian en gran medida de los valores reales del promedio siguiente. De hecho, el error en la predicción es menor.

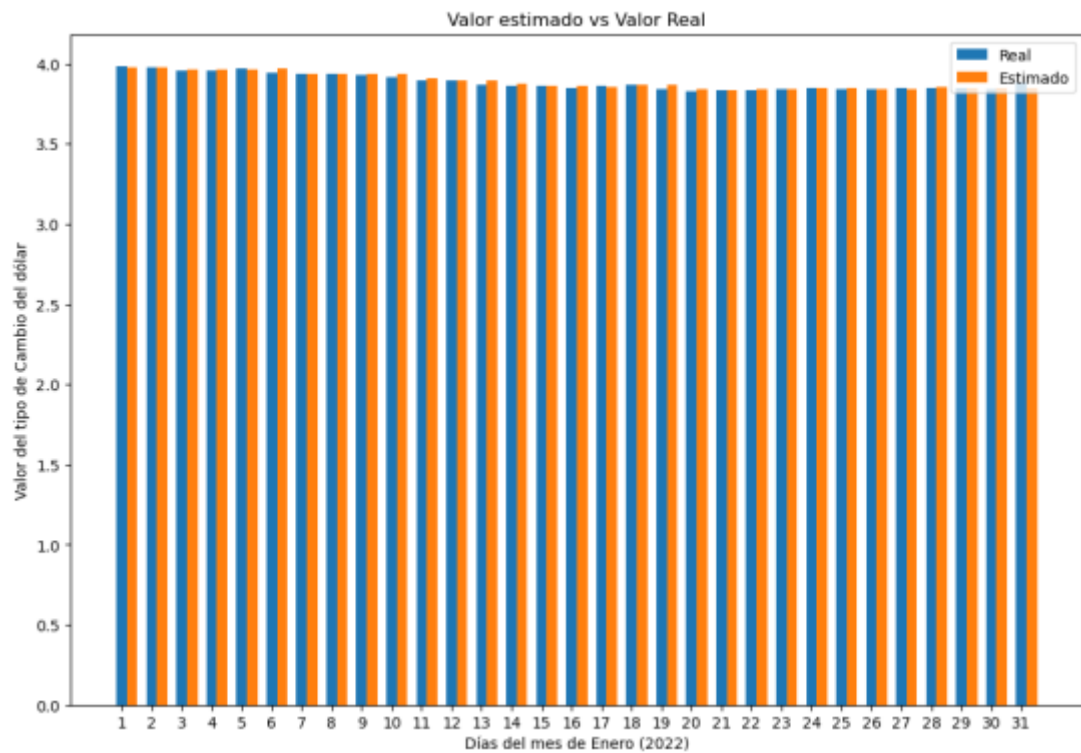


Figura 32. SVM Model - Gráfica de barras

Finalizando las pruebas de los modelos con la nueva data procesada se puede comprobar que el modelo de Support Vector Machine logra un menor error en los valores predichos en contraste a los valores reales del tipo de cambio del dólar.

Capítulo V. CONCLUSIONES Y TRABAJOS FUTUROS

Se extrajeron datos para la construcción de 2 data sets para la investigación, un data set de tweets con respecto a la variación del tipo de cambio del dólar utilizando la librería de Python Snsrape para la extracción de datos y posteriormente se realizó un filtro según la localización para Perú; asimismo un data set bancario de la data abierta del BCRP teniendo como característica principal los precios de apertura y cierre.

Se logró realizar el procesamiento del lenguaje natural de los tweets y se usó TF-IDF y n-gram para la vectorización del texto, además se estandarizó la data bancaria con respecto a los días cuya información no estaba disponible.

Se logró identificar en la primera fase de la investigación la tendencia de los usuarios de Twitter usando los modelos de Random Forest (RF) y Naive Bayes (NB), este resultado se utilizó como input para la segunda fase, en la cuál se logró predecir el tipo de cambio del dólar usando los algoritmos de Random Forest y Support Vector Machine. (SVM).

Se logró medir los rendimientos de los algoritmos implementados, para la primera fase se utilizaron las métricas de precision, recall, F1 score y accuracy, obteniendo finalmente un F1 score de 0.901459854 usando las particiones porcentuales de 75-25 para ambos modelos de RF y NB (Ver Tabla 7); para la segunda fase se midió la tasa de error en la predicción según el MAE (error medio absoluto) y el MSE (error medio cuadrático), obteniendo un MAE de 0.01264 y MSE de 0.000323 aproximadamente y un MAE de 0.009605 y MSE de 0.000232213 para los algoritmos de RF y SVM respectivamente, evidenciando que en la investigación SVM es el más óptimo (Ver Tabla 10 y Tabla 11).

Se debe tener en cuenta que los algoritmos realizados se basan en los datos del 2021 y es importante mencionar que en una siguiente fase de investigación se puede aplicar métodos de extracción de datos de Streaming de tal manera de que se pueda determinar la tendencia de

los usuarios con respecto a la variación del tipo de cambio en tiempo real, lo cual ameritaría a una estructura con mayores recursos computacionales.

ANEXOS

Anexo 1: Random Forest Classifier

En la siguiente imagen se visualiza la creación del modelo, el entrenamiento y la obtención del score para la partición porcentual 70-30, estos pasos son análogos a los realizados para las particiones porcentuales de 75-25 y 80-20 sobre la data unigram, así como para las particiones porcentuales utilizando los bigramas.

```
arbol70=RandomForestClassifier(criterion='gini', max_depth=None, max_features='sqrt')
X_trainarbol70, X_testarbol70, y_trainarbol70, y_testarbol70 = train_test_split(Xbayes, ybayes, test_size=0.7, random_state=0)
arbol70.fit(X_trainarbol70, y_trainarbol70)
y_predarbol70=arbol70.predict(X_testarbol70)
f1scoreArbol70=f1_score(y_testarbol70,y_predarbol70, average='micro')
metricasArbol70=precision_recall_fscore_support(y_testarbol70,y_predarbol70, average='micro')
```

Figura 33. Random Forest Classifier - Modelo porcentual 70-30

Anexo 2: Multinomial Naive Bayes

En la siguiente imagen se visualiza la creación del modelo, el entrenamiento y la obtención del score para la partición porcentual 70-30, estos pasos son análogos a los realizados para las particiones porcentuales de 75-25 y 80-20 sobre la data unigram, así como para las particiones porcentuales utilizando los bigramas.

```
X_train, X_test, y_train, y_test = train_test_split(Xbayes, ybayes, test_size=0.7, random_state=0)
gnb=MultinomialNB()
gnb.fit(X_train,y_train)
y_pred=gnb.predict(X_test)
f1score=f1_score(y_test,y_pred, average='micro')
metricas=precision_recall_fscore_support(y_test,y_pred,average='micro')
```

Figura 34. Multinomial Naive Bayes - Modelo porcentual 70-30

Anexo 3: Random Forest Regressor

En la siguiente imagen se visualiza la creación del modelo, el entrenamiento y la obtención del score para la partición porcentual 70-30, estos pasos son análogos a los realizados para las particiones porcentuales de 75-25 y 80-20 sobre la data resultante de la primera fase de la investigación.

```
arbolRegre=RandomForestRegressor(n_estimators=400,criterion="absolute_error",bootstrap=False, max_depth=4, max_features="log2")
X_trainarbolRegre, X_testarbolRegre, y_trainarbolRegre, y_testarbolRegre = train_test_split(X, y, test_size=0.7, random_state=0)
arbolRegre.fit(X_trainarbolRegre, y_trainarbolRegre)
y_predarbolRegre=arbolRegre.predict(X_testarbolRegre)
maeRegre=metrics.mean_absolute_error(y_testarbolRegre,y_predarbolRegre)
mseRegre=metrics.mean_squared_error(y_testarbolRegre,y_predarbolRegre)
```

Figura 35. Random Forest Regressor - Modelo porcentual 70-30

Anexo 4: Support Vector Machine

En la siguiente imagen se visualiza la creación del modelo, el entrenamiento y la obtención del score haciendo uso de la validación cruzada sobre la data resultante de la primera fase de la investigación.

```
svrModel=SVR(C=10000, epsilon=0.000001, gamma=0.001)

svrModel.fit (X, y)

predictSvrModel = cross_val_predict(svrModel, X, y, cv=5)

scoreSvrModel = cross_val_score(svrModel, X, y, cv=5)
```

Figura 36. Support vector machine - Validación cruzada

Anexo 5: Github del código

<https://github.com/JoaquinBernabe/PredictiveModel>

Capítulo VI: REFERENCIAS BIBLIOGRÁFICAS

- Al Hamoud, A. A. (2018). Classifying political tweets using naïve bayes and support vector machines. *Lecture Notes in Computer Science*, 736-744.
- Alva Segura, D. (2021). *Análisis del Sentimiento político en twitter durante las elecciones congresales 2020 en el Perú*. Recuperado el 10 de Octubre de 2021, de Universidad Internacional de la Rioja: <https://reunir.unir.net/handle/123456789/11364>
- Alvino, C. (07 de Mayo de 2021). *Estadísticas de la situación digital de Perú en el 2020-2021 | Branch*. Obtenido de Branch: <https://branch.com.co/marketing-digital/estadisticas-de-la-situacion-digital-de-peru-en-el-2020-2021/>
- Banco Central de Reserva del Perú. (01 de Julio de 2021). *Banco Central de Reserva del Perú*. Obtenido de NOTAS DE ESTUDIOS DEL BCRP: <https://www.bcrp.gob.pe/docs/Publicaciones/Notas-Estudios/2021/nota-de-estudios-46-2021.pdf>
- Banco central de Reserva del Perú. (10 de octubre de 2021). *BCRP - Series diarias*. Obtenido de BCRP Data: <https://estadisticas.bcrp.gob.pe/estadisticas/series/diarias/resultados/PD04637PD-PD04638PD-PD04639PD-PD04640PD-PD04643PD-PD04644PD-PD04645PD-PD04646PD/html/2021-01-01/2021-01-31/>
- Barbara Kitchenham, P. B. (2013). A systematic review of systematic review process research in software engineering,. *Information and Software Technology*, 55(12), 2049-2075.
- Barnes J., A. À. (2019). On the Effect of Word Order on Cross-lingual Sentiment Analysis. *Procesamiento del Lenguaje Natural*, 23-30.

- Başarslan, M. &. (2020). Sentiment Analysis with Machine Learning Methods on Social Media. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 5-15.
- Batra, R. D. (2018). Integrating StockTwits with sentiment analysis for better prediction of stock price movement. *iCoMET*, 1-5.
- BCRP Data. (10 de Diciembre de 2021). *Dólar Americano (US\$)*. Obtenido de BCRP Data: <https://estadisticas.bcrp.gob.pe/estadisticas/series/mensuales/resultados/PN01234PM/html/2016-10/2021-10/>
- Carosia, A. C. (2020). Analyzing the Brazilian Financial Market through Portuguese Sentiment Analysis in Social Media. *Applied Artificial Intelligence*, 1-19.
- Chen, S. &. (2019). Predicting the Stock Price Movement by Social Media Analysis. *Journal of Data Analysis and Information Processing*, 295-305.
- Chiruzzo L., R. A. (2020). Sentiment Analysis in Spanish Tweets: Some Experiments with Focus on Neutral Tweets. *Procesamiento del Lenguaje Natural*, 109-116.
- Chunguang Bai, Q. Z. (2024). Do blockchain capabilities help overcome supply and operational risks: Insights from firm market returns during COVID-19. *Omega*. Obtenido de <https://doi.org/10.1016/j.omega.2024.103049>
- Constitución Política del Perú [Const]. (1993). *Art. 84 [Título III]*. Editora Perú. Recuperado el 10 de Octubre de 2021, de Constitución Política del Perú: https://cdn.www.gob.pe/uploads/document/file/198518/Constitucion_Politica_del_Peru_1993.pdf
- Das, S., Behera, R. K., Kumar, M., & Rath, S. K. (1 de 2018). Real-Time Sentiment Analysis of Twitter Streaming data for Stock Prediction. *132*, págs. 956-964. Elsevier.
- Hayes, B. (13 de Junio de 2019). *Business Broadway*. Obtenido de Programming Languages Most Used and Recommended by Data Scientists:

<https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/>

Hayes, B. (14 de Julio de 2020). *Business Broadway*. Obtenido de Most Popular Integrated Development Environments (IDEs) Used by Data Scientists:

<https://businessoverbroadway.com/2020/07/14/most-popular-integrated-development-environments-ides-used-by-data-scientists/>

José, H. R. (2020). Impacto de la COVID-19 sobre la salud mental de las personas.

Medicentro Electrónica.

Khan W., G. M. (2022). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing*, 13(7), 3433 - 3456. doi:10.1007/s12652-020-01839-w

Lengua, C. (04 de Marzo de 2021). *Tipo de cambio en máximo histórico: tres economistas explican las causas y consecuencias de esta alza*. Obtenido de El Comercio:

<https://elcomercio.pe/economia/tipo-de-cambio-en-maximo-historico-tres-economistas-explican-las-causas-y-consecuencias-de-esta-alza-ncze-noticia/?ref=ecr>

López, P. (2019). *BBVA*. Obtenido de ¿Qué es el tipo de cambio y cómo evitar que afecte a las finanzas?

Mankar, T. H. (2018). Stock Market Prediction based on Social Sentiments using Machine Learning. *2018 International Conference on Smart City and Emerging Technology*.

Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; A Comparative Analysis. *IEEE Access*, 8, 150199-150212.

Naciones Unidas Perú. (2020). Obtenido de Informe: El impacto del COVID-19 en América Latina y el Caribe.

- Nehme, A. (Febrero de 2023). *Datacamp*. Obtenido de 6 Best Python IDEs for Data Science in 2023: <https://www.datacamp.com/tutorial/data-science-python-ide>
- Obthong, M. T. (2020). A survey on machine learning for stock price prediction: Algorithms and techniques. *2nd International Conference on Finance, Economics, Management and IT Business* (págs. 63-71). Virtual, Online: FEMIB.
- Olmo, G. D. (2022). *BBC*. Obtenido de Qué es la "flotación sucia", el sistema con el que Perú mantiene estable su moneda (y cómo actúa ahora frente la inflación).
- Piatetsky, G. (30 de Mayo de 2019). *KDnuggets*. Obtenido de Python leads the 11 top Data Science, Machine Learning platforms: Trends and Analysis: <https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html>
- Picasso, A. M. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications*, 60-70.
- Pimpalkar, A. &. (2020). Influence of Pre-Processing Strategies on the Performance of ML Classifiers Exploiting TF-IDF and BOW Features. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, 49-68.
- Ranjit S., S. S. (2018). Comparison of algorithms in Foreign Exchange Rate Prediction. *International Conference on Computing, Communication and Security* (págs. 9-13). Kathmandu: IEEE.
- Ranjit S., S. S. (2018). Foreign Rate Exchange Prediction Using Neural Network and Sentiment Analysis. *IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking* (págs. 1173-1177). Greater Noida: IEEE.
- Ribeiro, J. (2018). THE ROLE OF EXCHANGE RATE IN PERU: A SHOCK ABSORBER OR A SOURCE OF SHOCKS? *Revista de Análisis Económico*, 33(2), 79-89.
doi:10.4067/S0718-88702018000200079

- Ruz, G. H. (2022). Bayesian Constitutionalization: Twitter Sentiment Analysis of the Chilean Constitutional Process through Bayesian Network Classifiers. *Mathematics*.
- Segura B. I., M. F. (2019). Análisis de Sentimiento en el dominio salud: analizando comentarios sobre fármacos. *Procesamiento del Lenguaje Natural*, 15-22.
- Tiwari, S., Verma, A., Garg, P., & Bansal, D. (2020). Social Media Sentiment Analysis On Twitter Datasets. *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, (págs. 925-927). Coimbatore, India. doi: 10.1109/ICACCS48705.2020.9074208.
- Ulloa, A. E. (2022). Predicting Daily Trends in the Lima Stock Exchange General Index Using Economic Indicators and Financial News Sentiments. *Communications in Computer and Information Science*. 1577, págs. 34-59. Springer. Obtenido de https://doi.org/10.1007/978-3-031-04447-2_3
- Valencia, F. G.-E.-A. (2019). Price Movement Prediction of Cryptocurrencies Using Sentiment Analysis and Machine Learning. *Entropy*, 21(6). doi:10.3390/e21060589
- Vijh, M. C. (2020). Stock Closing Price Prediction using Machine Learning Techniques. *Procedia Computer Science*, 599-606.
- WorldBank. (2022). *WorldbankGroup*. Obtenido de Informe sobre el desarrollo mundial 2022.
- Yao, W., Xu, K., & Li, Q. (6 de 2019). Exploring the Influence of News Articles on Bitcoin Price with Machine Learning. *Proceedings - IEEE Symposium on Computers and Communications, 2019-June*.
- Yasir, M. D. (2019). An intelligent event-sentiment-based daily foreign exchange rate forecasting system. *Applied Sciences*.
- Zaidi, A. O. (2018). Forecasting weekly crude oil using twitter sentiment of U.S. foreign policy and oil companies data. *2018 IEEE 19th International Conference on*

Information Reuse and Integration for Data Science (págs. 201-208). Salt Lake City:
IEEE.