Determination of Political Affinity of Ecuadorian Twitter Users Using Machine Learning Techniques for Authorship Attribution

César Espin-Riofrio, MSc.[1], Jorge L. Charco, MSc.[1], Johanna Zumba Gamboa, MSc.[1], Verónica Mendoza Morán, MSc.[1], Arturo Montejo-Ráez, PhD. [2], Bryan Díaz Campoverde, Ingeniero en Sistemas[1] y Gilson Tomalá Molina, Ingeniero en Sistemas[1]

[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, jorge.charcoa@ug.edu.ec, johanna.zumbag@ug.edu.ec, veronica.mendozam@ug.edu.ec, bryan.diazc@ug.edu.ec, gilson.tomalam@ug.edu.ec

[2]Universidad de Jaén, España, amontejo@ujaen.es

Abstract*– Social networks are a means of wide dissemination of ideas and expression of opinions in various fields, the political issue is no exception, arousing much interest with passionate comments, proclamations, opinions, advertising of a particular candidate or political party. Twitter, as a widely used social network, allows the publication of short messages that can be obtained through some extraction techniques allowing then to be analyzed. Authorship Attribution presents methods that help to determine the author of a certain text, as well as the stylistic characteristics of writing that allow to identify a feeling, affinity to a certain idea, etc. This article aims to investigate through experimentation, the possibility of classifying Ecuadorian Twitter users according to their political affinity through the analysis of short texts published in this network, using Machine Learning (ML) techniques for Authorship Attribution. For this purpose, the political parties with the highest vote in the first round of the 2021 presidential elections in Ecuador are taken as a reference. Classification methods such as Support Vector Machine (SVM) and, from Naive Bayes, Bernoulli and Multinomial are evaluated, comparing them with performance measures to establish which is the most suitable for the proposed task.

***Keywords*-- Authorship Attribution, Machine Learning, Classification Algorithms, Political Affinity, Twitter.**

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.535
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

Determinación de Afinidad Política de Usuarios de Twitter de Ecuador Utilizando Técnicas de Machine Learning para Atribución de Autoría

César Espin-Riofrio, MSc.[1], Jorge L. Charco, MSc.[1], Johanna Zumba Gamboa, MSc.[1], Verónica Mendoza Morán, MSc.[1], Arturo Montejo-Ráez, PhD. [2], Bryan Díaz Campoverde, Ingeniero en Sistemas[1] y Gilson Tomalá Molina, Ingeniero en Sistemas[1]

[1]Universidad de Guayaquil, Ecuador, cesar.espinr@ug.edu.ec, jorge.charcoa@ug.edu.ec, johanna.zumbag@ug.edu.ec, veronica.mendozam@ug.edu.ec, bryan.diazc@ug.edu.ec, gilson.tomalam@ug.edu.ec

[2]Universidad de Jaén, España, amontejo@ujaen.es

***Resumen*– Las redes sociales son un medio de amplia divulgación de ideas y expresión de opiniones en variados ámbitos, el tema político no es excepción despertando mucho interés con apasionados comentarios, proclamas, opiniones, publicidad de determinado candidato o partido político. Twitter, como red social de amplio uso, permite la publicación de mensajes cortos que pueden ser obtenidos mediante algunas técnicas de extracción permitiendo luego ser analizados. La Atribución de Autoría presenta métodos que ayudan a determinar el autor de determinado texto, así como las características estilísticas de escritura que permiten identificar un sentir, afinidad a determinada idea, etc. El presente artículo pretende investigar mediante la experimentación, la posibilidad de clasificar según su afinidad política a usuarios ecuatorianos de Twitter mediante el análisis de textos cortos que publican en dicha red, utilizando técnicas de Machine Learning (ML) para la Atribución de Autoría. Para ello se toma de referencia los partidos políticos de mayor votación en la primera vuelta de las elecciones presidenciales 2021 de Ecuador. Se evalúan métodos de clasificación como Support Vector Machine (SVM) y, de Naive Bayes el Bernoulli y el Multinomial, comparándolos con medidas de rendimiento para establecer cuál es el más idóneo para la tarea propuesta.**

***Palabras claves*-- Atribución de Autoría, Machine Learning, Algoritmos de Clasificación, Afinidad Política, Twitter.**

***Abstract*– Social networks are a means of wide dissemination of ideas and expression of opinions in various fields, the political issue is no exception, arousing much interest with passionate comments, proclamations, opinions, advertising of a particular candidate or political party. Twitter, as a widely used social network, allows the publication of short messages that can be obtained through some extraction techniques allowing then to be analyzed. Authorship Attribution presents methods that help to determine the author of a certain text, as well as the stylistic characteristics of writing that allow to identify a feeling, affinity to a certain idea, etc. This article aims to investigate through experimentation, the possibility of classifying Ecuadorian Twitter users according to their political affinity through the analysis of short texts published in this network, using Machine Learning (ML) techniques for Authorship Attribution. For this purpose, the political parties with the highest vote in the first round of the 2021 presidential elections in Ecuador are taken as a reference. Classification methods such as Support Vector Machine (SVM) and, from Naive Bayes, Bernoulli and Multinomial are evaluated, comparing them with performance measures to establish which is the most suitable for the proposed task.**

***Keywords*-- Authorship Attribution, Machine Learning, Classification Algorithms, Political Affinity, Twitter.**

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2022.1.1.535
ISBN: 978-628-95207-0-5 **ISSN:** 2414-6390

I. INTRODUCCIÓN

En la actualidad, el Machine Learning representa un gran avance en el ámbito tecnológico, puesto que aprende a realizar tareas de forma automática y brinda un nivel de precisión acorde a los datos procesados en un algoritmo; así mismo, trae consigo una variedad de beneficios para el área investigativa ya que puede determinar posibles soluciones a los problemas de distintos ámbitos. [1] manifiesta que usar esta herramienta o aprendizaje automático de manera intensiva genera en grandes cantidades experiencias autónomas, de modo que, proporciona resultados positivos basados en las experiencias obtenidas. La Atribución de Autoría utiliza técnicas de análisis de textos en base a variadas características sintácticas, léxicas y estilométricas, así como de frecuencia de palabras o longitud de frases, atiende tareas como atribuir o verificar la autoría de un texto a determinado autor, o también determinar características de escritura que determinen comportamientos personales, sociales o de grupo. Cabe recalcar que, debido a lo mencionado anteriormente, resulta de vital importancia que la Atribución de Autoría se asocie directamente con las técnicas del Machine Learning.

Falta aún conceder validez a las técnicas científicas de Machine Learning en tareas como la determinación de la afinidad política analizando los mensajes de texto publicados por usuarios de Twitter en Ecuador aparte de que, causa una permanencia de los métodos tradicionales en los procesos electorales.

Se estima de suma importancia tener conocimiento acerca de este tema, puesto que, aparte de que en el contexto ecuatoriano son pocas las investigaciones que comprueben estadísticamente la factibilidad de emplear las técnicas de Machine Learning en la identificación de la inclinación partidaria de los usuarios de Twitter a través de sus publicaciones de texto; clasificar la afinidad política de un usuario de Ecuador permite a los profesionales asentar bases

de información para futuras investigaciones y a la comunidad realizar un análisis para ámbitos políticos o temas relacionados.

El propósito de la investigación propuesta es clasificar usuarios de Twitter de Ecuador utilizando técnicas de Machine Learning como tarea de Atribución de Autoría para determinar su afinidad política, partiendo de establecer el estado del arte sobre la Atribución de Autoría y algoritmos de clasificación, determinando las técnicas más usadas por otros autores en sus contribuciones de investigación sobre el tema. Se recolecta un dataset etiquetado de tweets de cuentas de Ecuador de partidos políticos y personalidades reconocidas como afines a los mismos para su entrenamiento con técnicas de clasificación, posterior se realizan pruebas con datos de usuarios no etiquetados para la predicción de su afinidad política a determinada agrupación política. Así, se evalúan varios métodos de clasificación de Machine Learning con el dataset recolectado para establecer el más idóneo para la tarea propuesta.

Según diversas investigaciones realizadas por distintos autores acerca de la clasificación de texto con la implementación de algoritmos de categorización, se delimita que en trabajos científicos como [2] en su proyecto tienen como objetivo encontrar el mejor algoritmo para la clasificación de texto mediante la combinación de rasgos característicos logrando determinar que el algoritmo SVM es el más idóneo para cumplir con sus objetivos. [3] utiliza el método de clasificación Naive Bayes – Bernoulli en su trabajo con el fin de clasificar las frases que sean ingresadas por personas y determinar el análisis de sentimientos de dicha frase, ya sea frustrado, aburrido, emocionado y comprometido. [4] implementa los algoritmos SVM, Multinomial de Naive Bayes y Maximum Entropy para identificar autores en textos breves o cortos; [5] comprobó que aplicando el algoritmo de Naive Bayes – Bernoulli para la determinación de la afinidad de los usuarios de España mediante el análisis de sus tweets se puede obtener excelente resultados en investigaciones previas; [6] utiliza las redes bayesianas Naive Bayes y árboles de decisión (Decision Trees) en su trabajo para la creación de un sistema de clasificación que permita la identificación y prevención de accidentes laborales; y [7] usó Naive Bayes, SVM, árboles de decisión y bosques aleatorios (Random Forests) en su trabajo con el propósito de extraer las características que describen el estilo de escritura de un autor a partir de un solo documento.

Como se pudo denotar, los métodos de clasificación más usados en los últimos años refieren en: el Naive Bayes, la Máquina de Soporte Vectorial (SVM), el Árbol de Decisión (DT), y las Redes Neuronales; puesto que sus grandes porcentajes de precisión en la clasificación de texto hace que tenga un gran peso de validación científica para futuros trabajos. Cabe destacar que, este proyecto emplea el algoritmo SVM por su impacto y popularidad; y los algoritmos de Naive Bayes por ser uno de los algoritmos con mayor impacto en la clasificación de texto que son Bernoulli y Multinomial para comprobar su efecto en los resultados.

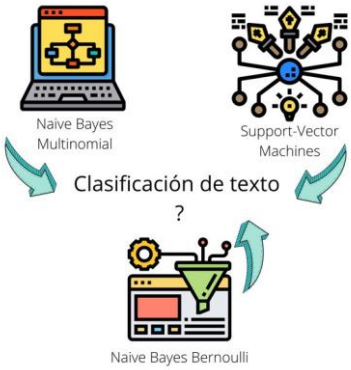

Fig. 1 Algoritmos implementados en la investigación.

Por otra parte, la Atribución de Autoría se basa en detectar, desde un grupo de probables autores, al propietario o autores de un escrito en cuestión por medio de la investigación de ejemplares de documentos escritos por los probables autores [8]. Es la ciencia que estudia los distintivos textuales y aplica el conocimiento lingüístico para obtención de las variables gramaticales; mismas que, analizándolas pueden determinar el autor al que le corresponde la autoría sea tanto de documentos o evidencias de texto escrito en cuestión [9]. Empieza desde el siglo XIX en donde [10] fue el primero en intentar identificar al autor basándose en el estilo de escritura luego [11] y [12] utilizaron la frecuencia relativa y la longitud de la frase para medir la cantidad de palabras, pero sin resultados favorables. Tiempo después [13] implementaron un enfoque estadístico multivariado en los "Federalist Papers" con un clasificador Naive Bayes marcando el inicio de la Atribución de Autoría desde un área computacional, iniciando los enfoques basados en el Machine Learning, que consiste en crear modelos de aprendizaje supervisado y no supervisado para lograr obtener una herramienta de suma importancia para la resolución de tareas relacionadas con la atribución de autoría de manera optimizada [14]. En los de aprendizaje supervisado tenemos los algoritmos de clasificación, [15] demostró que el algoritmo de clasificación SVM (Support Vector Machine) es ideal para la Atribución de Autoría y la clasificación de texto por su capacidad para manejar datos escasos, también tenemos el modelo Naive Bayes, [16] presentaron su método TAN (Tree Augmented Naive Bayes) con una simplicidad computacional para inducir a la clasificación de datos, luego contamos con el algoritmo K-Neighbors en donde [17], mediante una técnica codiciosa, presentaron una clasificación al mayor valor del vecino más cercano con mejoras en el rendimiento con respecto a las otras técnicas. Linear SVC en el que [18] presentó como mejora la clasificación de texto para la Atribución de Autoría con el método SVC. También tenemos a los algoritmos Perceptrón [19] lo aplicaron en los artículos federalistas. Luego tenemos el inicio de la Deep Learning aplicando técnicas redes neuronales [20] obteniendo resultados muy acogedores para la Atribución de Autoría [21], como las redes neuronales convulsiónales presentando una mejora con respecto a

modelos simples basados SVM [22]. Aparte tenemos los algoritmos de regresión que son de aprendizaje supervisado, pero no de la categoría de clasificación, [23] en su investigación de reconocimiento de autoría mediante una implementación de regresión múltiple lograron tener un porcentaje de éxito de 65% que normalmente se obtiene por métodos tradicionales a 72% y también se tiene los algoritmos de clustering o de agrupación donde [24] lo implementa para una revisión e interpretación en varios campos de la clasificación de texto con la finalidad de identificar las áreas importantes y realizar una revisión y análisis. Para mejorar las técnicas o encontrar nuevas alternativas se crearon los talleres PAN "Análisis de Plagio, Identificación de Autoría y Detección de casi Duplicados" que tuvo el origen de su nombre en el primer evento de SIGIR en el año 2007 que son campañas con el objetivo de presentar nuevas áreas e ideas mediante la integración de colaboradores o participantes en los campos relacionados, en el mismo los autores no solo presentan sus documentos, también se da paso a las discusiones productivas antes, durante y después de los talleres, es de ahí donde surgen nuevas metodologías partiendo de las expresiones de cada autor [25]. Así mismo, tenemos los algoritmos de aprendizaje no supervisado en el que [26] presentaron una distribución de párrafos, con algoritmos no supervisados estudiaba los patrones de organismos de longitud a partir de documentos de longitud variable. También se presentan los modelos de representación de lenguaje de Transformer por su popularidad en las tareas del procesamiento del lenguaje natural y teniendo mejoras en el documento de Google "Attention is All You Need" consistiendo en extraer cada palabra para deducir la importancia de todas las demás palabras obtenidas con la misma. Transformers son los modelos de última generación en evaluaciones recientes de traducción automática, por ejemplo. Mantiene dos líneas de investigación, por un lado, Transformer-Big que utiliza redes amplias que ha sido el estándar de factores en el desarrollo del modelo y por otro lado, utiliza una representación del lenguaje más profunda que supera a Transformer-Big [27]. Uno de los modelos Transformer es el de BERT que significa Representación de Codificador Bidireccional de Transformadores. A diferencia de otros modelos de representación de lenguaje, BERT se encuentra diseñado para entrenar previamente representaciones bidireccionales profundas de texto no etiquetado por condicionamiento conjunto en ambos contextos izquierdo y derecho en todas las capas. El modelo BERT, se puede ajustar para una amplia variedad de tareas, como responder preguntas e inferir en el lenguaje. De esta forma, BERT se considera conceptualmente simple pero empíricamente es poderoso [28].

Con respecto a la clasificación automática de textos, ésta se logra a partir de dos partes fundamentales que son el Procesamiento de Lenguaje Natural (PLN) y el aprendizaje automático o Machine Learning, donde por un lado el PLN estudia los inconvenientes inherentes al procesamiento y manipulación de lenguajes naturales, realizando uso de computadoras en el cual pretende conseguir conocimiento sobre el modo en que los humanos comprenden y usan el lenguaje, de tal forma que se logre realizar el desarrollo de herramientas y técnicas para lograr que las computadoras puedan comprenderlo y manipularlo, además sus fundamentos residen en un grupo bastante extenso de disciplinas como pueden ser en ciencias de la información, matemáticas, IA y robótica, entre otros. Por otro lado, sobre el ML y sus técnicas, cubren tareas como el análisis sintáctico y morfológico de los textos, extracción de información, clasificación automática de documentos, agrupación semántica, entre otras [29].

II. Metodología

Esta investigación presenta una metodología experimental en el que se abordan métodos de extracción de tweets y manejo de dataframes para usarlos en el aprendizaje de los algoritmos de clasificación empleados en la resolución de este proyecto. A su vez, se emplea un enfoque bibliográfico-documental en analizar contribuciones científicas de relevancia sobre técnicas de clasificación en Atribución de Autoría y Machine Learning empleadas en trabajos similares de otros autores.

A continuación, se observa en la figura 2 el proceso de desarrollo y ejecución del proyecto, donde se indican las principales etapas para la resolución de este.

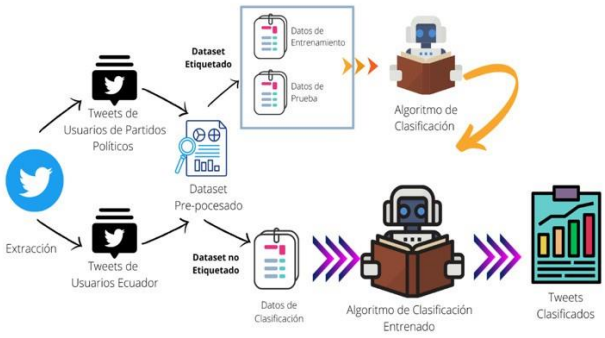

Fig. 2 Arquitectura implementada en la investigación.

A. *Extracción de Tweets*

Se implementó una web scrapping denominada "Snscrape" para la extracción de tweets, y se llevó a cabo la recolección del dataset etiquetado (usuarios principales por cada partido político) y de datos no etiquetados.

Fig. 3 Proceso de extracción de datos.

En la creación del dataset etiquetado, se buscó a usuarios que tengan un alto grado de vinculación con los partidos políticos (CREO, UNES, PACHAKUTIK e IZQUIERDA DEMOCRÁTICA) para el entrenamiento y aprendizaje de los algoritmos Bernoulli, Multinomial y SVM; y en la de datos no etiquetados, se extrajo los tweets de usuarios aleatorios de Ecuador con el fin de poner a prueba el aprendizaje adquirido de los algoritmos. Cabe mencionar que los tweets fueron extraídos desde el 01-12-2020 hasta el 10-02-2022 (primero de diciembre de 2020 se convocó a elecciones, ocho de abril de 2021 culminaron con la segunda vuelta electoral), para recoger tweets durante todo el proceso de elecciones y luego posterior al mismo.

Se empleó las librerías Unicode para llevar los tweets correctamente en formato de texto, dado que el proceso de web scrapping devuelve uno no accesible, y el csv para almacenar los datasets obtenidos en archivos con la misma modalidad.

Respecto a los datos etiquetados, se presenta la cantidad de tweets obtenidos por cada etiqueta y, en base a los datos no etiquetados, se señala el número de publicaciones adquiridas por los usuarios en el contexto ecuatoriano.

TABLA 1
TOTAL DE DATOS ETIQUETADOS Y NO ETIQUETADOS.

Datos Etiquetados		**Datos no Etiquetados**
Partido Político		
UNES:	40,914	
CREO:	24,084	Tweets: **17,726**
IZQDEMOCRATICA:	21,452	
PACHAKUTIK:	16,964	
TOTAL:	**103,414**	

En figura 4 se aprecian los tweets extraídos de los usuarios con alto grado de vinculación de los partidos políticos, en este caso PACHAKUTIK.

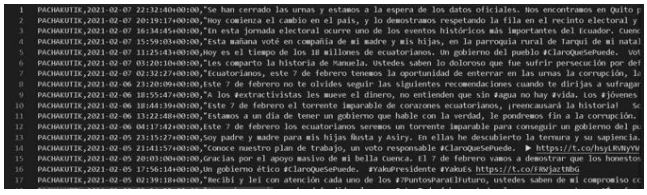

Fig. 4 Tweets extraídos.

B. Pre-Procesamiento

Durante la extracción de tweets, se obtuvieron datos que no son de gran aporte para el aprendizaje de los algoritmos como pueden ser los emojis, links, datos en blanco y retweets, por esa razón, se emplea un pre-procesamiento para obtener datasets estructurados y óptimos, puesto que estos datos redundan en muchos tweets de diferentes usuarios, y de alguna u otra forma terminan afectando a los algoritmos al momento de predecir el partido político de un usuario.

	Partido_Politico	Fecha	Texto
0	CREO	2022-02-08 18:01:45+00:00	Este jueves 10 de febrero, a las 07:00 pm, les...
1	CREO	2022-02-07 22:14:40+00:00	Felicidades a Puebloviejo en sus 176 años de ...
2	CREO	2022-02-07 18:34:35+00:00	Con la presencia de la Gobernadora de Los Rio...
3	CREO	2022-02-04 17:06:40+00:00	⚠ Aviso importante La cuenta .@BancadaAN_EC...
4	CREO	2022-02-04 16:55:54+00:00	4 de febrero día mundial contra el Cáncer ...
...	...	...	...
104324	IZQDEMOCRATICA	2020-12-01 21:49:16+00:00	@pequeandrea @natalylop @SimpliciterPaco Ajaja...
104325	IZQDEMOCRATICA	2020-12-01 20:59:31+00:00	No todas, yo soy Libra 💜 https://t.co/2HyFJoTN9X
104326	IZQDEMOCRATICA	2020-12-01 20:37:57+00:00	@AMPUEROF Andas afiladísimo. Te amo 💜
104327	IZQDEMOCRATICA	2020-12-01 20:36:39+00:00	@soleangusfr @rodolfoasar @MigueIngelZumb1 @Li...
104328	IZQDEMOCRATICA	2020-12-01 20:25:22+00:00	@rodolfoasar @MigueIngelZumb1 @soleangusfr @Li...

104329 rows × 3 columns

Fig. 5 Dataset sin pre-procesar.

En esta fase se aplican procesos relacionados al Procesamiento de Lenguaje Natural (PLN) a los datasets etiquetados y no etiquetados obtenidos en la fase de extracción de tweets en el que se realiza un reconocimiento y verificación de los datos contenidos en los datasets que se consideren no útiles, donde a partir de ello, se procede a la eliminación de la columna Fecha, los Retweets, emojis, datos en blanco y links, todo esto con la finalidad de disminuir las inconsistencias o alteraciones en el aprendizaje de los algoritmos.

La figura 6 presenta el dataset luego de la fase de pre-procesamiento.

	Partido_Politico	Texto
0	CREO	Este jueves 10 de febrero, a las 07:00 pm, les...
1	CREO	Felicidades a Puebloviejo en sus 176 anos de c...
2	CREO	Con la presencia de la Gobernadora de Los Rios...
3	CREO	Aviso importante La cuenta .@BancadaAN_EC h...
4	CREO	4 de febrero dia mundial contra el Cancer Po...
...	...	...
104324	IZQDEMOCRATICA	@pequeandrea @natalylop @SimpliciterPaco Ajaja...
104325	IZQDEMOCRATICA	No todas, yo soy Libra
104326	IZQDEMOCRATICA	@AMPUEROF Andas afiladisimo. Te amo
104327	IZQDEMOCRATICA	@soleangusfr @rodolfoasar @MigueIngelZumb1 @Li...
104328	IZQDEMOCRATICA	@rodolfoasar @MigueIngelZumb1 @soleangusfr @Li...

103414 rows × 2 columns

Fig. 6 Dataset pre-procesado.

Además, es importante balancear el dataset etiquetado, puesto que, en la fase anterior, las etiquetas de partidos políticos no mantienen una igualdad en cantidad, lo cual podría perjudicar el rendimiento de los algoritmos debido a que tendrán mayores datos para aprender de una etiqueta y menores de otra. Por ejemplo, de la etiqueta UNES se obtuvieron 40,914 tweets, mientras que de la etiqueta PACHAKUTIK se obtuvieron 16,964. A continuación, se muestra el balanceo dado en el dataset etiquetado.

TABLA 2
BALANCEO DEL DATSET CON DATOS ETIQUETADOS.

ETIQUETA	**Tweets**
CREO	16,964
UNES	16,964
PACHAKUTIK	16,964
IZQ. DEMOCRÁTICA	16,964
TOTAL	**67,856**

Con la implementación de la librería RandowUnderSampler se modificó el dataset para lograr obtener el mismo número de datos para cada etiqueta con el fin de establecer una fase de entrenamiento equilibrada y lograr excelentes resultados. El balanceo de datos consistió en definir la etiqueta con la menor cantidad de datos y posteriormente asignar dicha cantidad a las demás etiquetas, en este caso, la etiqueta PACHAKUTIK obtuvo la menor cantidad de datos con 16 964 tweets, por lo tanto, a las demás se les asignó la misma cantidad dando un total de 67856 tweets.

C. Tokenización de los datos

En la tokenización de los datos cada palabra se la convierte en un token, debido a que los modelos de clasificación no reconocen texto, por lo tanto, se debe realizar el proceso de tokenización para que el dataset pueda ser procesado por los modelos de Machine Learning.

D. Fase de entrenamiento (Train) y prueba (Test)

Para preparar los algoritmos se debe contar con dos variables, una para el entrenamiento y otra de prueba para validar lo que aprendió en el entrenamiento, por lo que declaramos nuestras variables asignándoles un 80% de los datos para entrenamiento lo que equivale a 54,284 datos y el 20% de datos de prueba que equivale a 13,572 datos con en una serie de 3 rondas de ejecución.

III. RESULTADOS

Se observa en la figura 7 las métricas obtenidas por las etiquetas correspondiente a cada algoritmo implementado. La métrica de precisión, como lo indica el mismo nombre, refleja la precisión del algoritmo alcanzada durante el aprendizaje. La métrica Recall indica la cantidad de valores positivos, en este caso, las etiquetas de partidos políticos, que el algoritmo pudo identificar correctamente. La métrica F1_score indica el valor obtenido de la combinación de la precisión y Recall Accuracy indica los valores porcentuales de la clasificación correcta de los algoritmos. De esta forma, realizando un análisis de los resultados apreciados, se determina que el algoritmo Naive Bayes con una distribución Multinomial se encuentra más acorde a los requerimientos de la investigación.

		Precisión	Recall	F1-Score	Accuracy
Multinomial	**UNES**	0.79	0.73	0.76	**79.67%**
	CREO	0.89	0.77	0.82	
	PACHAKUTIK	0.81	0.79	0.80	
	IZQDEMOCRATICA	0.66	0.81	0.73	
Bernoulli	**UNES**	0.80	0.7	0.75	**76.50%**
	CREO	0.86	0.78	0.82	
	PACHAKUTIK	0.80	0.77	0.79	
	IZQDEMOCRATICA	0.64	0.81	0.71	
SVM	**UNES**	0.78	0.73	0.76	**77.52%**
	CREO	0.89	0.77	0.82	
	PACHAKUTIK	0.81	0.79	0.80	
	IZQDEMOCRATICA	0.66	0.81	0.73	

Fig. 7 Métricas de cada algoritmo.

Posteriormente se aplica la matriz de confusión normalizada, que consiste en valorar qué tan excelente son los modelos de clasificación, en dicha matriz mientas los valores en diagonal se encuentren más cercanos al valor 1 más óptimo es el algoritmo, con el fin de evaluar sus resultados y el nivel de precisión que refleja cada algoritmo implementado, donde a partir de ello, se puede establecer el algoritmo más óptimo y práctico para determinar la afinidad política de un usuario.

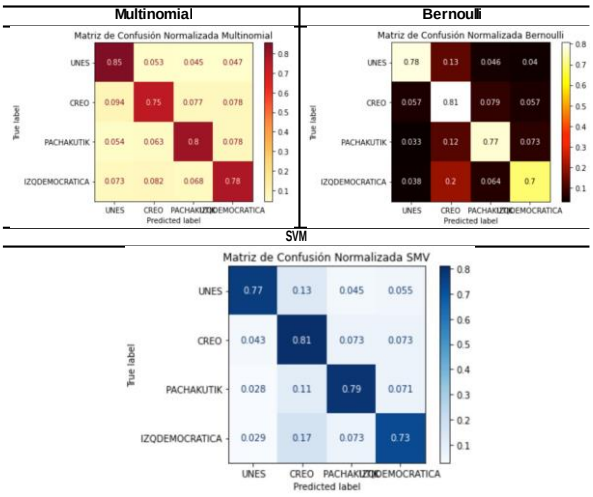

Fig. 8 Matriz de confusión normalizada.

Una vez realizado las fases correspondientes a los algoritmos de clasificación implementados en esta investigación que son Bernoulli, Multinomial y SVM, se procede a clasificar los textos correspondientes del dataset de datos no etiquetados para prueba, es decir, de los textos de usuarios generales de Ecuador, con cada algoritmo

mencionado con el fin de obtener los resultados y poder determinar el algoritmo con mayor precisión en la clasificación de afinidad política. A continuación, se muestran los resultados obtenidos de cada algoritmo.

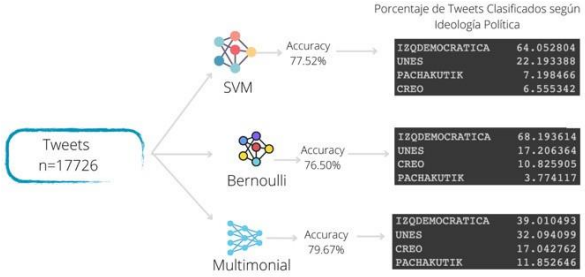

Fig 9 Resultados de los algoritmos.

Figura 9 muestra los resultados de precisión y los porcentajes de clasificación sobre los partidos políticos de los algoritmos implementados. Se observa los valores Accuracy, es decir, el valor porcentual de precisión de cada algoritmo de clasificación usado, donde SVM obtiene un porcentaje de 77.52%, Bernoulli obtiene el 76.50% y Multinomial obtiene el 79.67%. Por otro lado, se observa también los valores porcentuales clasificados por cada partido político o etiqueta que fueron asignadas según la clasificación de los algoritmos, donde el algoritmo SVM en su clasificación refleja que el 64.05% de los datos no etiquetados son partidarios de Izquierda Democrática, el 22.19% son partidarios de UNES, el 7.19% de PACHAKUTIK y el 6.55% de CREO. El algoritmo Bernoulli en su clasificación refleja que el 68.19% son partidarios de Izquierda democrática, el 17.20% son partidarios de UNES, el 10.82% de CREO y el 3.77% de PACHAKUTIK. Por último, el algoritmo Multinomial en su clasificación refleja que el 39.01% son partidarios de IZQUIERDA DEMOCRÁTICA, el 32.09% son partidarios de UNES, el 17.04% son de CREO y el 11.85% son de PACHAKUTIK.

Refiriéndonos a los resultados del algoritmo Multinomial, se hizo una prueba omitiendo la etiqueta IZQUIERDA DEMOCRÁTICA, los resultados cambiarían poniendo a UNES con el 32.01%, CREO con el 17.04% y PACHAKUTIK con el 11.85%, donde los resultados otorgados por el algoritmo comparándolos con los resultados de las elecciones presidenciales en primera vuelta, que se encuentra en la figura 10, mantienen un porcentaje de similitud con las cifras dadas en esa primera fase de la contienda electoral.

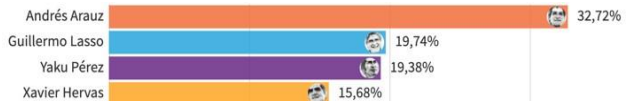

Fig. 7 Resultados de Elecciones Presidenciales 2021 - Primera Vuelta.

Con todo lo expuesto, se establece que el algoritmo con mayor grado de precisión en la clasificación de texto y que mejores resultados a arrojados, es el algoritmo de clasificación Multinomial con una precisión del 79.67%.

IV. Discusión

Con respecto a la recolección de datos, fueron verificados que fueron escritos y publicados en las fechas en el que se estableció, lo que otorga confianza y calidad al dataset. Existen otros medios por el que se pueden obtener mensajes de Twitter por ejemplo con su API gratuita que permite acceder a las publicaciones de los usuarios, por publicaciones en línea o por una determinada localidad, sin embargo, se seleccionó el método de extracción indicado porque se tenía acceso a obtener las publicaciones según lo establecido en las fechas, obteniendo la confiabilidad de la información de su procedencia y al acceso gratuito de los datos.

En la fase de preprocesamiento de la información se determinó diferentes criterios de limpieza, que consisten en la eliminación de Retweets, link de los Tweets, datos vacíos y los emojis que no presentan valor en el reconocimiento de entrenamiento para los algoritmos en materia de análisis de opinión, realizar este procedimiento no presenta ninguna alteración al dataset por el buen manejo y sintonización de la información, pero cabe decir, que para otros trabajos dichos datos serán de gran aporte.

En la etapa de transformación de los datos de texto para ingresar a los algoritmos se utilizó el procedimiento Bag of Word (BoW). Adicionalmente, con respeto a los datos de entrenamiento se aplicó un balanceo de los datos, para tener un mismo número de Tweets de cada etiqueta.

No obstante, en la actualidad se presentan diversas técnicas y algoritmos de clasificación, para este trabajo se eligieron los algoritmos Multinomial y Bernoulli de Naive Bayes y SVM por ser los que mayores impactos han presentado en la clasificación de texto, lo que conlleva a ser los algoritmos ideales para la determinación de la afinidad política de los usuarios de Twitter en Ecuador inclusive para tener datos estadísticos de cómo se presenta la ciudadanía a las elecciones y verificar los candidatos tendencia.

Algunos factores se pueden considerar y tomar en cuenta para nuevas investigaciones relacionadas, como las fechas en que los mensajes han sido extraídos que en el presente caso se hizo en época de elecciones para los datos etiquetados; la cantidad de tweets extraídos, SVM por ejemplo en un método de mucha aceptación aplicado con volúmenes altos de datos; el público objeto de estudio, ya que en Ecuador la red social Twitter es más usada en estratos sociales medio altos y seguramente influye en un análisis de preferencias o afinidad política; se podría nombrar también otras técnicas de pre-procesamiento y adecuación de los datos, etc.

Los métodos de agrupamiento o clustering deberían también probarse en siguientes o futuros trabajos, dichos algoritmos agrupan ítems en grupos con características similares y se utilizan en trabajos similares al presente.

Durante la investigación se presentaron problemas como la geolocalización inactiva de los usuarios, las publicaciones

de los tweets en formato de imagen, el desconocimiento de las técnicas automáticas, el uso incorrecto de los algoritmos de Machine Learning, y los grandes volúmenes de datos para la tarea de clasificación; que a su vez ocasionan una constante limitación durante la obtención de datos, una reducción de los datos de entrenamiento para el procesamiento de información, un menor rendimiento y automatización en los procesos relacionados al tema planteado, una dificultad de procesamiento y una serie de inconvenientes con los resultados esperados.

Es así como, se podrían obtener mejores resultados incrementado la cantidad de datos de aprendizaje para los algoritmos de clasificación, dado que crece el valor de precisión en ellos. Además, que los algoritmos seguirán actualizándose con nuevas técnicas de Machine Learning que permitirán seguramente mejores resultados de clasificación en trabajos futuros, como bien pueden ser trabajos relacionados al tema expuesto, la determinación de afinidad política de un usuario a partir de un texto. Puesto así, que el presente trabajo pueda motivar a otros investigadores a profundizar más sobre la afinidad política con técnicas de Machine Learning para Atribución de Autoría.

V. Conclusiones

El algoritmo Multinomial de Naive Bayes logró presentar los mejores resultados de la clasificación de mensajes cortos de Twitter para la determinación de afinidad política de los usuarios de Ecuador, frente a SVM y Bernoulli también presentados en este trabajo. Multinomial en valores porcentuales de la siguiente manera: IZQUIERDA DEMOCRÁTICA con un porcentaje de 38%, UNES con 32%, CREO 16% y PACHAKUTIK 12%, el algoritmo delimita por la capacidad de lo aprendido, lo que nos indica que el candidato de Izquierda Democrática es el que mayor apego obtuvo en la red social Twitter, lo que confirma lo propuesto por los medios de comunicación en aquella época, que el candidato de Izquierda Democrática tuvo una gran acogida en las redes sociales. Pero, al no considerar por un momento al partido Izquierda Democrática, es decir, experimentando solo con los datos de UNES, CREO y PACHAKUTIK, los resultados se aproximan a los reales de las elecciones presidenciales primera vuelta 2021, formando una gran iniciativa para futuras investigaciones aplicando criterios que el investigador considere necesarios para la implementación de un algoritmo con un mayor porcentaje de precisión para la tarea encomendada.

Además, los modelos propuestos fueron validados por el porcentaje de precisión y la matriz de confusión, lo que se puede concluir que el modelo Multinomial es confiable y óptimo de acuerdo con los resultados presentados en comparación a los algoritmos SVM y Bernoulli. Es importante mencionar que estas y otras técnicas de Machine Learning para tareas de clasificación pueden dar a futuro mejores resultados en este tipo de investigaciones.

References

[1] J. Francisco, R. Veliz, M. Abelardo, and A. Ramírez, "Estado del arte del aprendizaje automático relacionado con la lógica difusa," *Repos. Inst. - UNAC*, 2019, Accessed: Jan. 21, 2022. [Online]. Available: http://repositorio.unac.edu.pe/handle/20.500.12952/5580.

[2] A. Neocleous and A. Loizides, "Machine Learning and Feature Selection for Authorship Attribution:The Case of Mill, Taylor Mill and Taylor, in the Nineteenth Century," *IEEE Access*, vol. 9, pp. 7143–7151, Jan. 2021, doi: 10.1109/ACCESS.2020.3047583.

[3] M. L. Barrón Estrada, R. Zatarain Cabada, S. L. Ramírez Ávila, R. Oramas Bustillos, and M. G. Guerrero, "Use of Emotion Analyzer in Intelligent Educational Systems," *Res. Comput. Sci.*, vol. 147, no. 6, pp. 179–188, 2018.

[4] B. Vijayakumar and M. M. M. Fuad, "A New Method to Identify Short-Text Authors Using Combinations of Machine Learning and Natural Language Processing Techniques," *Procedia Comput. Sci.*, vol. 159, pp. 428–436, Jan. 2019, doi: 10.1016/J.PROCS.2019.09.197.

[5] B. López Medel, "Estudio de afinidad política en redes sociales a través de Machine Learning," 2019.

[6] R. Mosquera *et al.*, "Predicción de la accidentalidad laboral en la industria de pulpa y papel usando algoritmos de clasificación," *Inf. tecnológica*, vol. 32, no. 1, pp. 133–142, 2021, doi: 10.4067/S0718-07642021000100133.

[7] B. González, F. Tapia, and S. Salas Hernández, "New approach to feature extraction in authorship attribution," *Int. J. Comb. Optim. Probl. Informatics*, 2021, Accessed: Feb. 24, 2022. [Online]. Available: https://web.s.ebscohost.com/abstract?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=20071558&AN=152879397&h=nWz1n1HDbx9KSJ1QjbsfT140fWkKqx3qZVaftUle0MmrayPZuJJ%2FY5ErvCFxSi%2BtOzVkTYgN6vvdHQs5hQwGpA%3D%3D&crl=c&resultNs=AdminWebAuth&resultLocal=ErrCrlNotAuth&crlhashurl=login.aspx%3Fdirect%3Dtrue%26profile%3Dehost%26scope%3Dsite%26authtype%3Dcrawler%26jrnl%3D20071558%26AN%3D152879397.

[8] J. P. Posadas Durán, "Detección automática de plagio usando información sintáctica," Instituto Politécnico Nacional Centro de Investigación en Computación, Ciudad de México, 2016.

[9] M. F. Abad-García, "El plagio y las revistas depredadoras como amenaza a la integridad científica," *An. Pediatría*, vol. 90, no. 1, pp. 57.e1-57.e8, Jan. 2019, doi: 10.1016/J.ANPEDI.2018.11.003.

[10] T. C. Mendenhall, "THE CHARACTERISTIC CURVES OF COMPOSITION," *Science*, vol. 9, no. 214S, pp. 237–246, 1887, doi:10.1126/SCIENCE.NS-9.214S.237.

[11] G. Zipf Kingsley, "Selected Studies of the Principle of Relative Frequency in Language," *Sel. Stud. Princ. Relat. Freq. Lang.*, Dec. 1932, doi: 10.4159/HARVARD.9780674434929/.

[12] G. U. Yule, "ON SENTENCE-LENGTH AS A STATISTICAL CHARACTERISTIC OF STYLE IN PROSE: WITH APPLICATION TO TWO CASES OF DISPUTED AUTHORSHIP," *undefined*, vol. 30, no. 3/4, p. 363, Jan. 1939, doi: 10.2307/2332655.

[13] F. Mosteller and D. L. Wallace, "Inference and disputed authorship : the Federalist," p. 303, 1964.

[14] R. Sarwar and S. Nutanong, "The Key Factors and Their Influence in Authorship Attribution," *Res. Comput. Sci.*, vol. 110, no. 1, pp. 139–150, Dec. 2016, doi: 10.13053/RCS-110-1-12.

[15] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn. 1995 203*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[16] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian Network Classifiers," *Mach. Learn. 1997 292*, vol. 29, no. 2, pp. 131–163, 1997, doi: 10.1023/A:1007465528199.

[17] E. H. S. Han, G. Karypis, and V. Kumar, " Text Categorization Using Weight Adjusted K-Nearest Neighbor Classification," *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.*, vol. 2035, pp. 53–65, 2001, doi: 10.1007/3-540-45357-1_9.

[18] M. Sassano, " Virtual Examples for Text Classification with Support Vector Machines," 2003, Accessed: Feb. 24, 2022. [Online]. Available: http://ls6-www.informatik.uni-dortmund.de/ir/.

[19] F. J. Tweedie, S. Singh, and D. I. Holmes, " Neural network applications in stylometry: The Federalist Papers," *Comput. Humanit. 1996 301*, vol. 30, no. 1, pp. 1–10, 1996, doi: 10.1007/BF00054024.

[20] D. E. Rumelhart, G. E. Hinton, and R. . Williams, " Learning Internal Representations by Error Propagation," *Readings Cogn. Sci. A Perspect. from Psychol. Artif. Intell.*, pp. 319–362, Oct. 1986, doi: 10.1016/B978-1-4832-1446-7.50035-2.

[21] F. Püschel and N. Zárate, " DETERMINACIÓN DE AUTORÍA POR MEDIO DE REDES DE PALABRAS," 2016.

[22] P. Shrestha, S. Sierra, F. González, P. Rosso, M. Montes, and T. Solorio, " Convolutional Neural Networks for Authorship Attribution of Short Texts," *Assoc. Comput. Linguist.*, vol. 2, pp. 669–674, 2017.

[23] E. Stamatatos, " A survey of modern authorship attribution methods," *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, Mar. 2009, doi: 10.1002/ASI.21001.

[24] A. Kyriakopoulou, " Text Classification Aided by Clustering: a Literature Review," *Tools Artif. Intell.*, Aug. 2008, doi: 10.5772/6083.

[25] A. Trotman, S. Geva, and J. Kamps, " Report on the SIGIR 2007 workshop on focused retrieval," *ACM SIGIR Forum*, vol. 41, no. 2, pp. 97–103, Dec. 2007, doi: 10.1145/1328964.1328981.

[26] Q. Le and T. Mikolov, " Distributed Representations of Sentences and Documents," *31st Int. Conf. Mach. Learn. ICML 2014*, vol. 4, pp. 2931–2939, May 2014, Accessed: Feb. 24, 2022. [Online]. Available: https://arxiv.org/abs/1405.4053v2.

[27] Q. Wang *et al.*, " Learning Deep Transformer Models for Machine Translation," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 1810–1822, Jun. 2019, doi: 10.18653/v1/p19-1176.

[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, " BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Mar. 03, 2022. [Online]. Available: https://arxiv.org/abs/1810.04805v2.

[29] S. A. Pérez Vera, " 'Análisis y Clasificación de Textos con Técnicas Semi Supervisadas Aplicado a Área Atención al Cliente,'" pp. 1–47, 2017.