Polo del Conocimiento

Pol. Con. (Edición núm. 70) Vol 7, No 9
Septiembre 2022, pp. 1457-1474
ISSN: 2550 - 682X
DOI: 10.23857/pc.v7i8

Identificación de ideología política mediante un modelo Transformer para estilometría y Clasificación por votos en Machine Learning

Identification of political ideology through a Transformer model for stylometry and Classification by votes in Machine Learning

Identificação de ideologia política através de um modelo Transformer para estilometria e classificação por votos em Machine Learning

César Espín-Riofrio [I]
cesar.espinr@ug.edu.ec
https://orcid.org/0000-0001-8864-756X

William Ferruzola-Sánchez [II]
william.ferruzolas@ug.edu.ec
https://orcid.org/0000-0002-5388-1337

Abel Aspiazu-Torres [III]
abel.aspiazut@ug.edu.ec
https://orcid.org/0000-0003-1918-1385

Verónica Mendoza-Morán [IV]
veronica.mendozam@ug.edu.ec
https://orcid.org/0000-0001-7520-3505

Correspondencia: cesar.espinr@ug.edu.ec

Ciencias Técnicas y Aplicadas
Artículo de Investigación

*** Recibido:** 23 de julio de 2022 ***Aceptado:** 12 de agosto de 2022 *** Publicado:** 12 de septiembre de 2022

I. Magíster en Sistemas de Información Gerencial, Universidad de Guayaquil, Ecuador.
II. Universidad de Guayaquil, Ecuador.
III. Universidad de Guayaquil, Ecuador.
IV. Magíster Universitario en Software y Sistemas, Universidad de Guayaquil, Ecuador.

Resumen

El objetivo principal de este artículo es la determinación de la inclinación ideológica de usuarios de Twitter en Ecuador. Los datos recopilados se obtuvieron de la plataforma Twitter, estos se almacenaron en Datasets, se procesaron y etiquetaron para alimentar los métodos clasificadores los cuales entrenaron para realizar la predicción de ideología política a través del uso de modelos Transformer y Voting Classifier en Machine Learning, se usará Validación Cruzada para potenciar y evaluar durante el entrenamiento a modelos clasificadores como Logistic Regression, Random Forest, Decision Tree, Multilayer Perceptron y Gradient Boosting. Se ejecutará el modelo Transformer pre-entrenado para el español llamado Roberta-large-bne destinado para la extracción de características estilométricas halladas en textos, además se tendrá características fraseológicas como MeanWordLen, LexicalDiversity, MeanSentenceLen, StdevSentenceLen, MeanParagraphLen, DocumentLen y, de palabras de uso frecuente tomadas del corpus en español llamado CREA, este proceso permitió formar un vector final de características los cuales servirán para el entrenamiento. Se busca clasificar la ideología política en base a textos cortos tomados de Twitter y analizar los resultados de cada clasificador para validar cual es el más adecuado para la tarea de clasificación y predicción, dichos resultados servirán como indicador de factibilidad para estudios similares en un futuro.

Palabras clave: Transformers; Ideología política; Estilometría; Machine Learning.

Abstract

The main objective of this article is the determination of the ideological inclination of Twitter users in Ecuador. The collected data were obtained from the Twitter platform, these were stored in Datasets, processed and labeled to feed the classifier methods which trained to perform the prediction of political ideology through the use of Transformer and Voting Classifier models in Machine Learning, Cross Validation will be used to enhance and evaluate during training classifier models such as Logistic Regression, Random Forest, Decision Tree, Multilayer Perceptron and Gradient Boosting. The pre-trained Transformer model for Spanish called Roberta-large-bne will be executed for the extraction of stylometric features found in texts, in addition to phraseological features such as MeanWordLen, LexicalDiversity, MeanSentenceLen,

StdevSentenceLen, MeanParagraphLen, DocumentLen and frequently used words taken from the Spanish corpus called CREA, this process allowed to form a final vector of features which will be used for training. The aim is to classify political ideology based on short texts taken from Twitter and analyze the results of each classifier to validate which is the most suitable for the classification and prediction task, these results will serve as a feasibility indicator for similar studies in the future.

Keywords: Transformers; Political Ideology; Stylometry; Machine Learning.

Resumo

O objetivo principal deste artigo é determinar a inclinação ideológica dos usuários do Twitter no Equador. Os dados coletados foram obtidos da plataforma Twitter, estes foram armazenados em Datasets, processados e rotulados para alimentar os métodos classificatórios que foram treinados para prever a ideologia política através do uso de modelos Transformer e Voting Classifier em Machine Learning, utilizará Cross Validation para impulsionar e avaliar modelos de classificador como Regressão Logística, Floresta Aleatória, Árvore de Decisão, Perceptron Multicamada e Aumento de Gradiente durante o treinamento. Será executado o modelo Transformer pré-treinado para espanhol chamado Roberta-large-bne, destinado à extração de características estilométricas encontradas em textos, bem como características fraseológicas como MeanWordLen, LexicalDiversity, MeanSentenceLen, StdevSentenceLen, MeanParagraphLen, DocumentLen e, de palavras de uso frequente retiradas do corpus em espanhol denominado CREA, este processo permitiu formar um vetor final de características que serão utilizadas para o treinamento. Busca classificar a ideologia política com base em pequenos textos retirados do Twitter e analisar os resultados de cada classificador para validar qual é o mais adequado para a tarefa de classificação e previsão, esses resultados servirão como indicador de viabilidade para estudos semelhantes no futuro.

Palavras-chave: Transformadores; Ideologia política; Estilometria; aprendizado de máquina

Introducción

En la actualidad, la Inteligencia Artificial (IA) se utiliza para un sin número de tareas y es tan prometedora dado que está impulsando la productividad como nunca, la razón de esto es que esta tecnología permite que las máquinas comprendan y alcancen objetivos específicos con mayor

eficiencia y reduciendo la posibilidad de errores al mínimo. Machine Learning (ML) es un término muy nombrado dentro del campo de la Inteligencia Artificial, y de hecho estos están estrechamente relacionados, sin embargo, no son lo mismo, debido a que el ML es una rama o subcategoría que pertenece a la IA, ahora bien, dentro de esta tecnología, contamos principalmente con 2 tipos de aprendizajes, el supervisado y no supervisado.

La estilometría surge por consecuencia del comienzo de la Atribución de Autoría en el siglo XIX, donde expertos enfocados en el campo lingüístico lograban determinar a qué autor corresponden textos y documentos desconocidos aplicando métodos basados en expertos, clasificando características importantes dentro del texto o el habla de una persona. El primer método propuesto para identificar autores basado únicamente en el estilo de escritura es el método Chi-cuadrado, que consiste en generar una curva para cada archivo en cuestión, reflejando así la relación entre la longitud de palabra y su frecuencia (Mendenhall, 1889). Posteriormente, en el siglo XX, se comenzaron a utilizar métodos estadísticos para determinar la distribución de una auditoría, como el método de frecuencia relativa que permite identificar la autoría en función únicamente del número de apariciones de palabras en un texto (Kingsley Zipf, 1932). Debido al uso de métodos estadísticos y sus problemas al aplicarlos en la identificación de un autor determinado, (Mosteller & Wallace, 2012) lograron adoptar un enfoque de investigación multivariante en los "Federalist Papers" analizó palabras de uso frecuente como 'a', 'y', etc. Luego usaron 30 palabras y un clasificador Naive Bayes para resolverlo, que comienza asignando la autoría desde un área computacional. Posteriormente se ideó un nuevo enfoque basado en el aprendizaje automático, teniendo en cuenta el aprendizaje supervisado, basado en un proceso que permite entrenar a través de las características y etiquetas de un texto dado, para finalmente poder hacer predicciones basadas en características conocidas con anterioridad, donde (Rosenblatt, 1958) demostró que el algoritmo Multilayer Perceptron provee predicciones rápidas después del entrenamiento con datos de grandes longitudes.

Transformer fue popular gracias al documento de Google “Attention is All You Need” (Vaswani et al., 2017), donde han logrado mejoras significativas en el desempeño de varias tareas de aprendizaje en el Procesamiento del Lenguaje Natural (PLN) y la visión por computadora, reemplazando a los modelos neuronales convolucionales y recurrentes. Como indica (Gardner et al., 2018) la estructura de Transformers fue inspirada en la biblioteca pionera tensor2tensor y el código fuente original de BERT, surge del concepto de proporcionar un almacenamiento en

caché fácil para modelos previamente entrenados presentado de AllenNLP, iniciando con el primer modelo transformer denominado Generative Pretrained Transformer, también conocido como GPT, creado por (Openai et al., 2018), Posteriormente GoogleAI creó el modelo Bidirectional Encoder Representations from Transformer o BERT el cual (Devlin et al., 2019) describe como un modelo que interpreta con precisión todos los elementos de una consulta de búsqueda en contexto. OpenAI crea un modelo mejorado perteneciente a la serie GPT denominado GPT-3 donde (Floridi & Chiriatti, 2020) establecen como un modelo de lenguaje autorregresivo que utiliza el aprendizaje profundo para generar textos que imitan la escritura humana. Con el tiempo, la arquitectura de Transformer ha demostrado ser particularmente beneficiosa para la capacitación previa en un gran corpus de documentos, lo que resulta en aumentos significativos en la precisión para tareas posteriores como clasificación de texto, comprensión del idioma, traducción automática, centrado, resolución normal y resumen, entre otros (Wolf et al., 2020).

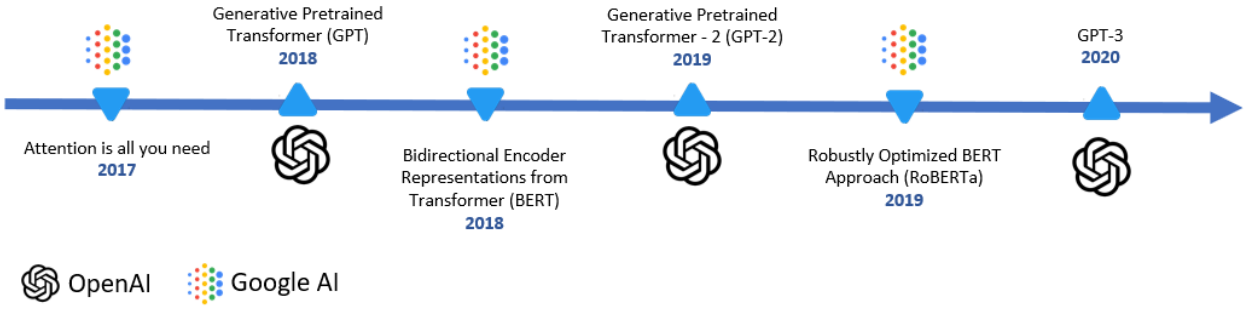

Figura 1: Timeline de modelos Transformers

La presente investigación se orienta hacia el aprendizaje de tipo supervisado, dado que provee técnicas que nos permitirán realizar predicciones en base a datos etiquetados suministrados previamente. La investigación girará en torno a este tipo de implementación de Machine Learning ya que se pretende clasificar a los usuarios de la red social Twitter en Ecuador, de acuerdo con su afinidad política binaria (izquierda y derecha) y multiclase (izquierda moderada, izquierda, derecha moderada, derecha), esperando tener resultados que indiquen qué movimiento político tiene mayor cantidad de partidarios. Por otra parte, es importante mencionar la utilización de un listado exhaustivo de palabras más utilizadas del idioma español, llamado CREA (Corpus de Referencia del Español Actual) referido por la Real Academia Española, este será empleado

durante el análisis de publicaciones, para determinar la frecuencia de utilización de palabras que componen los textos, un factor fundamental a tomar en cuenta durante el estudio estilométrico. Otro aspecto importante a mencionar es la utilización de la biblioteca desarrollada por Jeff Potter (*Jpotts18 (Jeff Potter) · GitHub*, n.d.), ubicada en un repositorio de Github, esta fue destinada para la extracción de características fraseológicas contenidas en los textos. La importancia de llevar a cabo un estudio de este tipo surge gracias a la necesidad de conocer el favoritismo de la ciudadanía, la inclinación política de nuestra población objetivo, Ecuador. Puesto que tener acceso a esta información, según (Proaño et al., 2018) permite "la toma de decisiones rápidas y acertadas", lo cual a su vez conlleva a una notable mejora en la propuesta de los candidatos, considerando que lo que generalmente buscan determinadas organizaciones o grupos políticos es enterarse del curso que toma la sociedad en tiempos de campaña electoral y de acuerdo a este dato, ofrecer a la población lo que esta desea, lo anteriormente expuesto se da gracias al uso de los sistemas de información que sirven de apoyo con el proceso de toma de decisiones. A su vez, determinar hacía qué partido político se inclina un individuo, comprende un impacto considerable en el ámbito sociopolítico de una nación, ya que estar al tanto de la afinidad política de una persona o usuario, mediante sus publicaciones de texto en la red social Twitter, permitirá predecir elecciones presidenciales o, en su defecto, determinar el partido político con mayor popularidad o cantidad de partidarios en el Ecuador. El presente proyecto de investigación pretende dar cumplimiento a la identificación y clasificación de ideologías políticas binarias y multiclases de usuarios políticos de Twitter en el Ecuador empleando técnicas de estilometría, modelo Transformer ROBERTA-large-bne y métodos de clasificación en ML, es así como se obtendrán resultados de los cuales nos podremos valer para presentar de manera gráfica y posteriormente determinar qué afinidades políticas son más influyentes en las plataformas de red social y en nuestra población.

Metodología

Para este trabajo, se implementó la investigación bibliográfica, ya que se va a recopilar información a partir de materiales publicados en línea, o incluso recursos más habituales, clásicos o tradicionales como libros, periódicos, informes o revistas referentes a investigaciones de clasificación, entrenamiento y predicción de texto.

Pol. Con. (Edición núm. 70) Vol 7, No 9, Septiembre 2022, pp. 1457-1474, ISSN: 2550 - 682X

Para el estudio de este trabajo de investigación se empleó algoritmos de aprendizaje supervisado enfocados en los métodos de clasificación, tales como Logistic Regression propuesto por (Berkson, 1944) el cual define como un método de análisis estadístico que predice resultados binarios, basándose en observaciones previas de un conjunto de datos, (Pranckevičius & Marcinkevičius, 2017) demostraron mediante resultados de clasificación de multiclase que el método Logistic Regression lograba mayor precisión en los resultados en comparación con los métodos de clasificación Naïve Bayes, Random Forest, Decision Tree y Support Vector Machine. Por otro lado, el método Decision Tree introducido por (Quinlan, 1986) define como una forma de análisis de variables múltiples las cuales permiten predecir, explicar, describir o clasificar un resultado. (Charbuty & Abdulazeez, 2021) Demostraron que el algoritmo Decision tree en contraste a otros algoritmos de clasificación crean una colección de reglas eficiente y sencilla de entender realizadas en el área de clasificación de textos. Otro método de clasificación son los Random Forest propuesto por (Laboratories et al., 1995), donde (Shah et al., 2020) evaluaron diferentes algoritmos de clasificación tales como Logistic Regression, Random Forests y K-Nearest Neighbour, teniendo resultados óptimos en el algoritmo Random Forests en la clasificación de texto. Además otro algoritmo de clasificación es el Multilayer Perceptron propuesto por (Rosenblatt, 1958) el cual consta de tres tipos de capas: la capa de entrada, la capa de salida y la capa oculta donde la capa de entrada recibe la señal de entrada para ser procesada, la capa de salida realiza la predicción y la clasificación. Donde (Kamath et al., 2018) aplicó el algoritmo Multilayer Perceptron enfocados en la clasificación de texto, los resultados del algoritmo fueron prometedores tanto en documentos brutos como procesados.

Para dar cumplimiento a la investigación experimental se implementó un método de extracción de tweets para su posterior preprocesamiento, se extrajeron 3 diferentes tipos de características de texto, para el aprendizaje de los métodos de clasificación mencionados anteriormente y su posterior predicción.

El método para seguir en este trabajo de investigación es el siguiente:

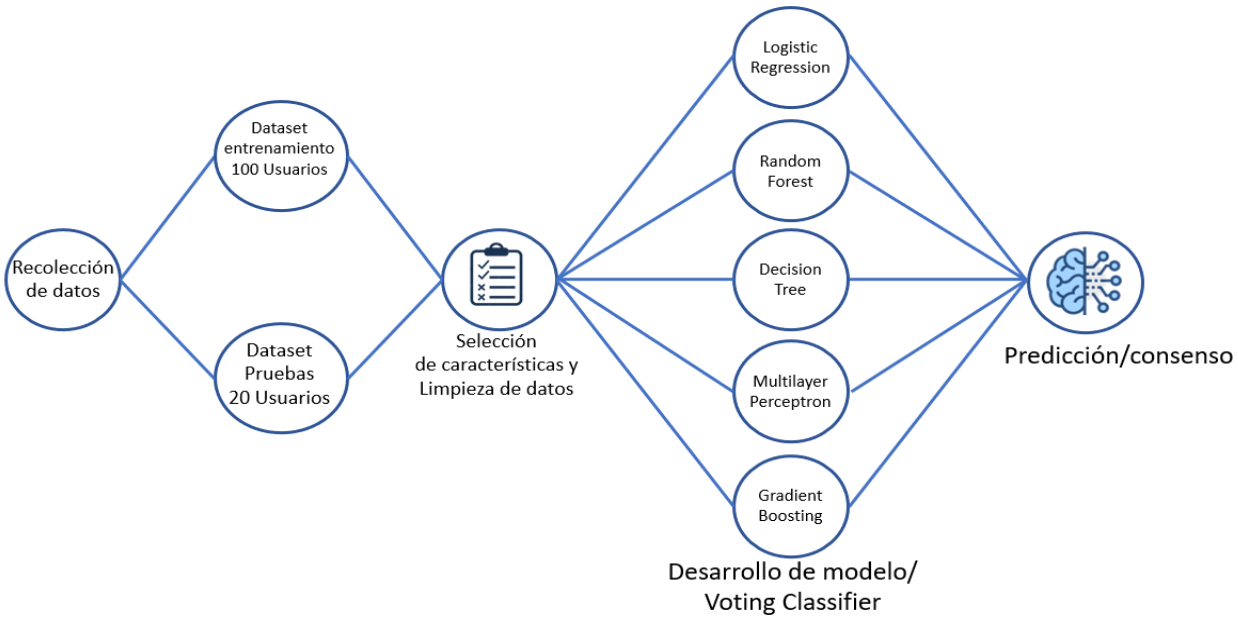

Figura 2: Método implementado en la investigación

Extracción de datos

Para la extracción de datos se utilizó la API Tweepy de Twitter logrando así la extracción de 6000 tweets de 120 usuarios políticos del Ecuador, obteniendo 50 tweets por usuario, formando así dos datasets, uno destinado al entrenamiento constituido por 100 usuarios y 5000 tweets y otro conjunto de datos de prueba con 20 usuarios resultando en 1000 tweets, como se muestra en las Figuras 2 y 3.

Pol. Con. (Edición núm. 70) Vol 7, No 9, Septiembre 2022, pp. 1457-1474, ISSN: 2550 - 682X

Fuente: Para Train/entrenamiento:

index	user	ideologia_binaria	ideologia_multiclase	Tweet
0	@user001	derecha	derecha-moderada	Muy doloroso lo ocurrido en el Cristo del Consuelo. Sobran los motivos para la acción coordinada entre el Municipio y el Gobierno Nacional, al terrorismo se lo vencerá trabajando unidos. Los guayaquileños no merecemos ser víctimas de los berrinches y bochinches entre autoridades. https://t.co/bsfsVmckcl
1	@user001	derecha	derecha-moderada	En Milagro participé de un encuentro con jóvenes rurales, quienes saben que el desafío es acceder a mejores alternativas de formación y lograr mayor productividad para avanzar hacia la transformación rural. Todos merecemos las mismas oportunidades. 🧡 ec https://t.co/e9Du48Y6AB
2	@user001	derecha	derecha-moderada	Desde hace dos años he denunciado el uso fraudulento de mi imagen. Estos anuncios son FALSOS, por favor no caigan y ayúdenme a evitar que más personas sean víctimas de esta estafa. No tengo, no he tenido, ni he recomendado jamás inversiones en Bitcoin. ¡No existe el dinero fácil! https://t.co/SHtPtDsYbU
3	@user001	derecha	derecha-moderada	Una gran noticia para los guayaquileños. @PedroPabloDuart representa una opción de honestidad, arduo trabajo y servicio hacia los ciudadanos. ¡Es tiempo de una nueva generación! https://t.co/h4H3XzcePn
4	@user001	derecha	derecha-moderada	Participé en el encuentro académico de estudiantes y graduados de la @EspochRio donde expuse sobre la necesidad de innovar para lograr cambios reales. En mi paso por Riobamba aproveché para comer el rico hornado de esta tierra maravillosa. 👌 https://t.co/olU9Gqt9wk
5	@user001	derecha	derecha-moderada	Hay quienes se dedican a convertir la política en una cloaca, para ahuyentar a la gente decente, para que nada cambie. Lo triste es que muchas veces logran su propósito y eso debería ser motivación suficiente para no rendirse. Buena lectura ℹ️ https://t.co/VnwSeJLwxt
6	@user001	derecha	derecha-moderada	Disfrutemos su historia, su gastronomía, su cultura y su arte. #VivaGuayaquil 👏🏻👏🏻 https://t.co/1FDEVl720V
7	@user001	derecha	derecha-moderada	Los jóvenes lo que quieren son oportunidades. Estuve con estudiantes de la @unesumoficial, en Jipijapa, intercambiando ideas respecto a alternativas de desarrollo que el Ecuador necesita para progresar y ser más competitivo. ¡La innovación es clave! 👍🏻 https://t.co/rYDZyRkyhI
8	@user001	derecha	derecha-moderada	En Naranjito conversamos sobre la importancia de apostarle a la transformación rural para brindar nuevas y mejores oportunidades al campo. La ruralidad debe dejar de ser sinónimo de pobreza y convertirse en ejemplo de prosperidad. Gracias 'La Comunidad de Karla' por invitarme. 🧡 https://t.co/6hYmf2e5Vc
9	@user001	derecha	derecha-moderada	Mi solidaridad contigo @PedroPabloDuart por los reiterados e injustificados ataques. Hay personas que se molestan por su trabajo solidario y permanente. https://t.co/HREory1WFN

Figura 3: Extracción 5000 tweets de 100 usuarios políticos.

Fuente: Para Test/prueba:

user	ideologia_binaria	ideologia_multiclase	Tweet
@user001	derecha	derecha	Una semana más trabajando con pasión y entrega, para construir el #Guayaquil que soñamos. Siempre junto a ustedes, escuchando y actuando. 💪🏻 #NuevasIdeas #ConIlusión https://t.co/VvppP6EjLc
@user001	derecha	derecha	Me invitaron a formar parte de 'Yo Leo', un festival que incentiva la lectura entre estudiantes. Me llena de esperanza saber que existen iniciativas como esta y más aún que los guayaquileños aportan con empeño para que se lleven a cabo con éxito. 📚🙌🗣️ #ElVerdaderoCambio https://t.co/OynaBSwKBZ
@user001	derecha	derecha	Compartimos un momento ameno con los vecinos del Guasmo Sur 📍. Junto a nuestros graduados del curso de enfermería llegamos con dosis de Vitamina C y Complejo B, que ayudan a subir las defensas y fortalecer el cuerpo. Estamos agradecidos por su cariño, estaremos por más sectores. https://t.co/QUtqiiFbaP
@user001	derecha	derecha	Tengo certeza plena que juntos podemos construir el Guayaquil que tanto deseamos 🙏🌞🌳 #LlegóLaHoraDelCambio https://t.co/LCb7nTmVPK
@user001	derecha	derecha	Luis Tomalá, ciudadano guayaquileño radicado en la Isla Trinitaria. Hoy, a sus 68 años colabora en la comunidad reparando ventiladores 💪🏻 Un orgullo presenciar el compromiso que tienen los guayaquileños con su gente 🙏🏻🙌 #HacerMásHablarMenos https://t.co/wNiaoDLZpQ
@user001	derecha	derecha	☝🏻Primero los ciudadanos. 🗣️Si el crimen se organiza para hacer el mal, organicémonos los buenos para hacer el bien. Somos más. Somos más fuertes. ☝🏻Los derechos humanos de los guayaquileños serán respetados. ¿De los delincuentes? veremos luego. https://t.co/GSqeGcq1tc
@user001	derecha	derecha	Construimos un Guayaquil seguro para las nuevas generaciones. 🚒 Cada vez llegamos a más lugares vulnerables. 🏳️ Los pequeños pasos forman grandes cambios 💪🏻 #guayaquil https://t.co/WjskS9vdXv
@user001	derecha	derecha	🗣️"Quien no espera vencer, ya está vencido." Gran frase de José Joaquin de Olmedo, primer alcalde de #Guayaquil https://t.co/Fw5EAY4wUT
@user001	derecha	derecha	Hay que volver a sentirnos orgullosos de ser guayaquileños y llevar nuestra ciudad a estar entre las mejores de latinoamérica y del mundo. Como ciudadanos merecemos obras de CALIDAD. Recuperemos nuestra ciudad juntos. 🤝🏻 ¡Saldremos adelante Guayaquil! 💪🏻 https://t.co/6KTZQbxL8g
@user001	derecha	derecha	El único enemigo a vencer debe ser la delincuencia, narcotráfico e inseguridad. La gente quiere que se resuelvan los problemas, no importa quien lo haga. Lo que importa son los resultados. #guayaquil https://t.co/l3WX5fflkXw

Figura 4: Extracción 1000 tweets de 20 usuarios políticos.

Preprocesamiento de datos

Para ejecutar el respectivo preprocesamiento de datos, primero se llevará a cabo la limpieza de los tweets extraídos tanto para el dataset train, como también para el dataset test, para ello se eliminará enlaces, retweets, emojis y caracteres especiales, y posteriormente se deberá agrupar

los 50 tweets de cada usuario en un nuevo dataset. En las Figuras 4 y 5 se muestran los dataset preprocesados.

Para train:

	user	ideologia_binaria	ideologia_multiclase	Tweet
0	@user001	derecha	derecha-moderada	Muy doloroso lo ocurrido en el Cristo del Cons...
1	@user002	derecha	derecha	Gracias presidente @LassoGuillermo por la opor...
2	@user003	izquierda	izquierda-moderada	Sólo para entendidos…¡Manabí de mis amores!#Ba...
3	@user004	izquierda	izquierda-moderada	Veto total. [SEP] CONTRA LA CORRUPCIÓN EN COTO...
4	@user005	izquierda	izquierda	Yo Silvia Lorena Vera Calderón, ciudadana ecua...
...	...	...	...	...
95	@user096	derecha	derecha-moderada	Que buena noticia señor Presidente. Que Dios l...
96	@user097	derecha	derecha	Estamos próximos a recibir a nuestra querida "...
97	@user098	derecha	derecha	Seguimos celebrando los 487 años de creación d...
98	@user099	izquierda	izquierda-moderada	Nuestra actuación será siempre en defensa de l...
99	@user100	derecha	derecha	Un gran abrazo al cielo amigo @CesarMongeO [S...

Figura 5: Preprocesamiento de 5000 tweets para 100 usuarios políticos.

Para test:

	user	ideologia_binaria	ideologia_multiclase	Tweet
0	@user001	derecha	derecha	Una semana más trabajando con pasión y entrega...
1	@user002	izquierda	izquierda	Mensaje a la militancia @ID12Ecuador Estamos o...
2	@user003	izquierda	izquierda	Los jóvenes son importantes para @ID12Ecuador ...
3	@user004	izquierda	izquierda-moderada	Un saludo caluroso a quienes tienen la gran la...
4	@user005	derecha	derecha-moderada	El #deporte es una manifestación cultural que ...
5	@user006	izquierda	izquierda-moderada	#AgendaLegislativa Nuestra #Minga continua, co...
6	@user007	izquierda	izquierda	#EnMedios Viernes, 19 de agostoRadio La Nueva ...
7	@user008	derecha	derecha-moderada	#LoorLagoAgrio| Saludamos a nuestro querido ca...
8	@user009	derecha	derecha-moderada	Escuchar a los ciudadanos de mi provincia es c...
9	@user010	derecha	derecha	En estos últimos días he sido blanco de divers...
10	@user011	derecha	derecha-moderada	La ignorancia (y la corrupción) es atrevida. P...
11	@user012	derecha	derecha-moderada	Escogimos la Democracia aquí se debate y se vo...
12	@user013	derecha	derecha	Esta tarde, visité a los niños y niñas que par...
13	@user014	derecha	derecha-moderada	El martes 11 de mayo, informé sobre las accion...
14	@user015	derecha	derecha	Te invito a mi entrevista el día de mañana. [S...

Figura 6: Preprocesamiento de 1000 tweets para 20 usuarios políticos.

Extracción de características

Para la extracción de características de los tweets se usaron 3 tipos de técnicas: Fraseológicas (MeanWordLen, LexicalDiversity, MeanSentenceLen, StdevSentenceLen, MeanParagraphLen, DocumentLen), dada por la librería creada por Jeff Potter ubicada en un repositorio de Github, sumado al uso de palabras frecuentes tomada del Corpus de Referencia del Español Actual (CREA), y por último el modelo de Transformer RoBERTa-large-bne. De las cuales se van a crear 3 vectores con sus características respectivas. Estos vectores se van a unir en un solo vector el cual se normaliza con el método MinMaxScaler, para crear un vector final y poder realizar el entrenamiento mediante los métodos clasificadores de Machine Learning

```
# Union vectores de características en un solo vector
real_x_train = [lengths_vector,words_vector,transformer_vector]
# Normalizacion del vector
datalv = np.stack(real_x_train[0])
datawv = np.stack(real_x_train[1])
datamt = np.squeeze(np.stack(real_x_train[2]), axis=1)
# Normalización del vector
scalerLV = MinMaxScaler().fit_transform(datalv)
scalerWV = MinMaxScaler().fit_transform(datawv)
scalerMT = MinMaxScaler().fit_transform(datamt)
# Vector Final
x_train_numpy = np.concatenate([scalerLV, scalerWV, scalerMT], axis=1)
```

Figura 7: Vector final normalizado con características fraseológicas, CREA y Transformer.

Entrenamiento de métodos clasificadores

Para el entrenamiento de los métodos clasificadores mediante el dataset de train, se aplicó la librería Voting Classifier (hard voting classifier) de Scikit-Learn, para potenciar la clasificación a manera de método de conjunto heterogéneo para lograr un mejor rendimiento predictivo, además se obtuvieron métricas de evaluación utilizando Cross Validation. Los algoritmos usados para alimentar el clasificador por votos fueron Logistic Regression, Random Forest, Decision Tree, Multilayer Perceptron y Gradient Boosting.

```
clf1 = LogisticRegression(random_state=1)
clf2 = RandomForestClassifier(n_estimators=50, random_state=1)
clf3 = DecisionTreeClassifier(random_state=45)
clf4 = MLPClassifier(max_iter=1000, random_state=45)
clf5 = GradientBoostingClassifier(n_estimators=100, learning_rate=1.0, max_depth=1, random_state=0)

# vamos a almacenar un baseline por caracteristica
baselines = {}

# como vemos, esta tarea se trata de cuatro caracteristicas: 2 demograficas y dos psicográficas. por lo tanto vamos a
# entrenar modelos diferentes y separados para cada tarea
for label in ['ideologia_binaria', 'ideologia_multiclase']:

  # obtenemos un clasificador de referencia
  baselines[label] = VotingClassifier(estimators=[('lr', clf1), ('rf', clf2), ('dt', clf3), ('mlp', clf4), ('gb', clf5)], voting='hard')

  # entrenamos el baseline para la etiqueta actual
  baselines[label].fit (x_train_numpy, dataframes['train'][label])


  metrics = ['accuracy', 'precision', 'recall', 'f1']
  print("para -->", label)
  for clf, labl in zip([clf1, clf2, clf3, clf4, clf5, baselines[label]], ['lr', 'rf', 'dt', 'mlp', 'gb', 'Ensemble']):
    #scores = cross_val_score(clf, x_train_numpy, dataframes['train'][label], scoring='f1_weighted', cv=5)
    scores = cross_val_score(clf, x_train_numpy, dataframes['train'][label], scoring='recall', cv=10)
    # results.loc[alg,:] = [scores['test_'+m].mean() for m in metrics]
    print("recall: %0.2f (+/- %0.2f) [%s]" % (scores.mean(), scores.std(), labl))
    #print("f1_weighted: %0.2f (+/- %0.2f) [%s]" % (scores.mean(), scores.std(), labl))
```

Figura 8: Ejecución de Voting Classifier, Cross Validation y métodos de la librería sklearn empleados en esta investigación.

Predicción

Para la predicción mediante el dataset de train, se utilizó el método predict, tal como lo muestra la siguiente imagen.

```
for label in ['ideologia_binaria', 'ideologia_multiclase']:

  # obtenemos las predicciones
  y_pred = baselines[label].predict (X_test)

  f1_scores[label] = f1_score(dataframes['test'][label], y_pred, average='macro')
```

Figura 9: Predicción a través del método predict

Resultados

Luego de realizar la valoración de resultados, quedó evidenciado que el algoritmo Gradient Boosting tuvo un alto desempeño con respecto a la etiqueta de ideología binaria en comparación con los otros algoritmos utilizados en este trabajo de investigación respecto de la determinación de inclinación política-ideológica de los usuarios en Twitter, este algoritmo produjo el más alto

resultado con un 60% de accuracy en el entrenamiento. Por otra parte, el algoritmo Multilayer Perceptron tuvo un mayor nivel con respecto a los demás algoritmos en la etiqueta de ideología multiclase, reflejando un porcentaje de 37%, los algoritmos empleados, en conjunto con sus resultados, se detallan a continuación en las siguientes tablas.

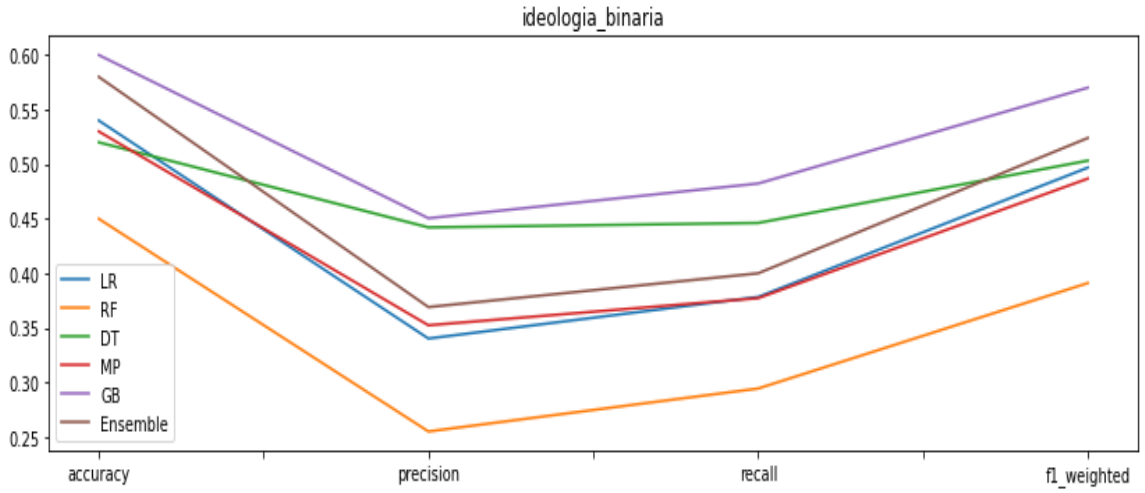

Figura 10: Métricas de los algoritmos clasificadores para ideología binaria.

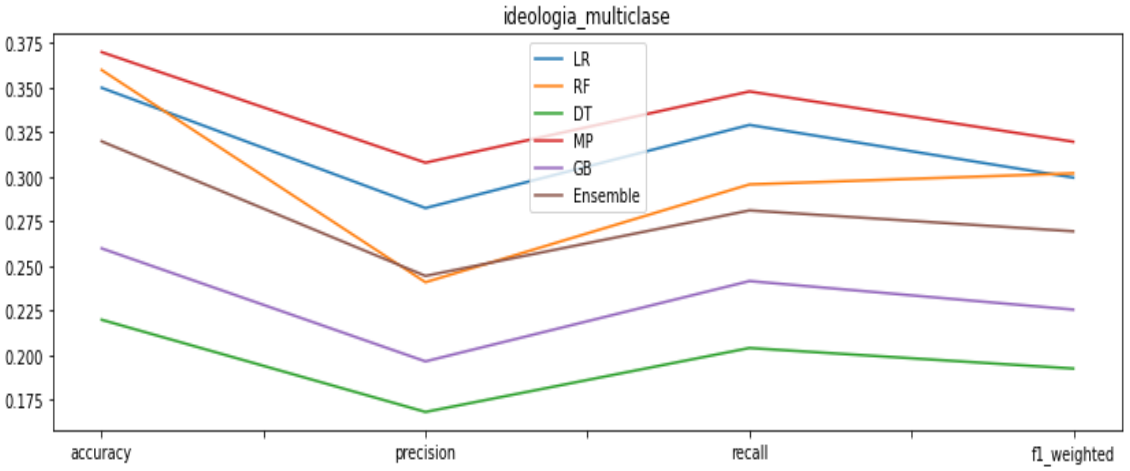

Figura 11: Métricas de los algoritmos clasificadores para ideología multiclase.

Finalizado el entrenamiento de nuestro dataset train, se realizaron pruebas con el dataset test de los cuales obtuvimos los siguientes resultados de predicción:

ideologia_binaria

	precision	recall	f1-score
derecha	0.857143	0.800000	0.827586
izquierda	0.500000	0.600000	0.545455
accuracy			0.750000

Figura 12: Resultados de predicción para ideología binaria.

Donde la precisión obtenida por parte de la etiqueta de ideología binaria en el dataset test fue un 75%.

ideologia_multiclase

	precision	recall	f1-score
derecha	0.000000	0.000000	0.000000
derecha-moderada	0.285714	0.200000	0.235294
izquierda	0.166667	0.500000	0.250000
izquierda-moderada	0.200000	0.333333	0.250000
accuracy			0.200000

Figura 13: Resultados de predicción para ideología multiclase.

Donde la precisión obtenida por la etiqueta de ideología multiclase en el dataset test fue un 20%. Terminando así el proceso de entrenamiento y predicción para posteriormente crear un archivo dataframe con las predicciones hechas por el algoritmo y compararlo con el dataset test original, como se muestra en la siguiente figura.

Pol. Con. (Edición núm. 70) Vol 7, No 9, Septiembre 2022, pp. 1457-1474, ISSN: 2550 - 682X

username	ideologia_binaria	ideologia_multiclase
@user001	derecha	derecha-moderada
@user002	derecha	derecha-moderada
@user003	derecha	izquierda
@user004	izquierda	izquierda
@user005	izquierda	izquierda
@user006	izquierda	izquierda-moderada
@user007	izquierda	derecha-moderada
@user008	derecha	derecha
@user009	izquierda	izquierda-moderada
@user010	derecha	derecha-moderada
@user011	derecha	derecha-moderada
@user012	derecha	derecha-moderada
@user013	derecha	izquierda
@user014	derecha	izquierda-moderada
@user015	derecha	izquierda
@user016	derecha	derecha
@user017	izquierda	izquierda
@user018	derecha	derecha-moderada
@user019	derecha	izquierda-moderada
@user020	derecha	izquierda-moderada

username	ideologia_binaria	ideologia_multiclase
@user001	derecha	derecha
@user002	izquierda	izquierda
@user003	izquierda	izquierda
@user004	izquierda	izquierda-moderada
@user005	derecha	derecha-moderada
@user006	izquierda	izquierda-moderada
@user007	izquierda	izquierda-moderada
@user008	derecha	derecha-moderada
@user009	derecha	derecha-moderada
@user010	derecha	derecha-moderada
@user011	derecha	derecha-moderada
@user012	derecha	derecha
@user013	derecha	derecha-moderada
@user014	derecha	derecha-moderada
@user015	derecha	derecha-moderada
@user016	derecha	derecha-moderada
@user017	derecha	derecha
@user018	derecha	derecha
@user019	derecha	derecha
@user020	derecha	derecha-moderada

Figura 14: Dataset Test ubicado en el lado izquierdo, la Predicción del lado derecho.

Los resultados expuestos anteriormente, demuestran que los algoritmos adoptados en este trabajo de investigación logran cumplir el funcionamiento por el cual fueron elegidos en un principio, la cual consiste en la clasificación de Tweets y posterior predicción. Lograron resultados muy favorables alcanzando un rendimiento óptimo para los algoritmos de clasificación especialmente en ideología binaria, sino además se consiguió corroborar la hipótesis planteada inicialmente la cual plantea la posibilidad de revelar la ideología política por la cual se inclina una persona de acuerdo con el análisis de sus tweets.

Discusión

Para poder predecir la afinidad política de las personas o usuarios, mediante sus tweets es necesario tener presente que las características estilométricas de cada persona son únicas, y por

ende, se deberá llevar a cabo un estudio exhaustivo de dichas características, es ahí donde la implementación de técnicas estilométricas para Machine Learning nos facilitan el trabajo para obtener dichas características, por tanto se decidió emplear el modelo Transformer RoBERTa-large-bne, características fraseológicas y de palabras de uso frecuente, para así poder enriquecer las características que servirán como suministro a los modelos de clasificación propuestos en este proyecto de investigación, permitiendo así potenciar la predicción de la afinidad política de los usuarios para obtener mejores resultados.

Teniendo en cuenta también que, si se logra entrenar mediante un dataset más extenso, podrían mejorar significativamente los resultados de la predicción, esto se debe a que el nivel de precisión de los modelos a entrenar es directamente proporcional a la cantidad de datos y características estilométricas extraídas de los tweets/textos, esto por lo tanto reflejará una notable variación positiva en los resultados. Por otra parte, se ha usado 3 tipos de características estilométricas, pero esto no significa que sean la única o mejor opción, es muy probable que se obtengan mejores resultados con otros tipos de características que beneficien en el entrenamiento y sean más precisos en la predicción.

Conclusiones

Con el análisis de contribuciones científicas relacionadas al estado del arte de los modelos Transformer y métodos de clasificación de Machine Learning para estilometría, se determinó que los algoritmos clasificadores Logistic Regression, Decision Tree, Multilayer Perceptron, Gradient Booster y Random Forest tuvieron un impacto positivo en cuanto a los resultados para la predicción política mediante la clasificación de texto, dando como el mejor resultado para ideología binaria al clasificador Gradient Boosting con un 60%, seguido de Logistic Regression con un 54%, Multilayer Perceptron con un 53%, Decision Tree con un 52%, y por último Random Forests con un 45%, en cuanto a la ideología multiclase Multilayer Perceptron fue el más óptimo con un 37%, seguido de Random Forests con un 36%, Logistic Regression con un 35%, Gradient Boosting con un 26%, y por último Decision Tree con un 22%, evidenciando así que los clasificadores Gradient Boosting y Multilayer Perceptron fueron los más recomendable y óptimo al momento de predecir la afinidad política mediante tweets. Además, estos resultados pueden ser mejorados enriqueciendo los tweets del dataset entrenado. Importante de señalar es que al tratarse de tema político y por los resultados obtenidos se hace evidente experimentar con

otras características estilométricas como pudiera ser el uso de un lexicón de palabras de uso político, seguramente se obtendrían mejores resultados en especial en ideología multiclase.

Referencias

1. Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. *Journal of the American Statistical Association*, *39*(227), 357–365. https://doi.org/10.1080/01621459.1944.10500699
2. Charbuty, B., & Abdulazeez, A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*, *2*(01), 20–28. https://doi.org/10.38094/jastt20165
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (Vol. 1).
4. Floridi, L., & Chiriatti, M. (2020). *GPT-3: Its Nature, Scope, Limits, and Consequences*. *30*, 681–694. https://doi.org/10.1007/s11023-020-09548-1
5. *jpotts18 (Jeff Potter) · GitHub*. (n.d.). Retrieved August 25, 2022, from https://github.com/jpotts18
6. Kamath, C. N., Bukhari, S. S., & Dengel, A. (2018). Comparative study between traditional machine learning and deep learning approaches for text classification. *Proceedings of the ACM Symposium on Document Engineering 2018, DocEng 2018*. https://doi.org/10.1145/3209280.3209526
7. Kingsley Zipf, G. (1932). Selected Studies of the Principle of Relative Frequency in Language. *Selected Studies of the Principle of Relative Frequency in Language*. https://doi.org/10.4159/HARVARD.9780674434929/HTML
8. Laboratories, T. B., Avenue, M., & Murray, U. H. (1995). *Random Decision Forests*.
9. Mosteller, F., & Wallace, D. L. (2012). Inference in an Authorship Problem. *Http://Dx.Doi.Org/10.1080/01621459.1963.10500849*, *58*(302), 275–309. https://doi.org/10.1080/01621459.1963.10500849
10. Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for

Text Reviews Classification. *Baltic Journal of Modern Computing*, *5*(2), 221–232. https://doi.org/10.22364/bjmc.2017.5.2.05

11. Proaño, M., Orellana, S., & Martillo, I. (2018). Los sistemas de información y su importancia en la transformación digital de la empresa actual. *Espacios*, *39*(45), 3–7.
12. Quinlan, J. R. (1986). Induction of Decision Trees. In *Machine Learning* (Vol. 1).
13. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.
14. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augmented Human Research*, *5*(1). https://doi.org/10.1007/s41133-020-00032-0