Text Classification on Political Comments using BERT and SHAP

Political XSA

Mauricio Vargas†
Department of Computational Science
Universidad Nacional de San Agustin
Arequipa, Perú
mvargasfr@unsa.edu.pe

ABSTRACT

Comments written in a variety of social media have an impact on general people and it can shift their political views or increase their political bias. Understanding the political bias and context of a political comment is important to study this social phenomenon. In our research, a BERT model applies to a political comment dataset extracted from X, the comments are categorized into partisan and neutral. After classifying political bias, we use an algorithm that explains the reasons why the model classified any comment as partisan or neutral.

CCS CONCEPTS

• NLP • XSA • BERT

KEYWORDS

Explainability, Political Classification, XAI

ACM Reference format:

Mauricio Vargas. 2025. Text Classification on Political Comments using BERT and SHAP. In *Proceedings of the final work for the course: Thesis Seminar. UNSA, Arequipa, AQP, Perú, 7 pages.*

1 Introducción

El análisis de sentimiento explicables (XSA) ha emergido como un campo crucial en la inteligencia artificial, buscando la transparencia y la comprensibilidad en los modelos de análisis de sentimiento [1]. Los métodos tradicionales tienen una "caja negra", lo que no otorga transparencia acerca de las razones de la clasificación, si bien la clasificación suele ser muy precisa, no explican las razones de la clasificación.

El creciente uso de modelos, muy precisos, complejos pero que no son interpretables en áreas de investigación como de trabajo como los modelos basados en Transformers, nos hace ver la necesidad de aplicar un método de explicabilidad para añadir transparencia al modelo [1].

La política es una parte esencial del individuo, ya que esta afecta desde relaciones interpersonales, pasando por movimientos sociales, hasta decisiones políticas que pueden afectar la vida de millones de individuos, y con el uso continuo y creciente de las redes sociales se puede analizar el descontento, afiliaciones y polarización de los usuarios de dichas redes. Además, podemos aplicar el sentimiento de análisis, clasificación de textos o averiguar inclinación política de personas, candidatos, empresarios y demás personas aplicando modelos de análisis de sentimientos, que además deben ser explicados para poder determinar la visión política de una persona.

El objetivo principal es aplicar un modelo XAI, de tipo post-hoc a un modelo de inteligencia artificial NLP con transformers en comentarios políticos.

2 Trabajos Relacionados

Tun Y. & Khaing M. [4] hicieron un estudio comparativo de diferentes modelos con explicación ad-hoc para saber cuál predice mejor la inclinación política de usuarios de la red Twitter.

Maleszka B. [5], analiza las diferencias entre modelos explicables y modelos de caja negra, indicando tendencias y retos del área. Maleszka B [6], compara múltiples métodos de inteligencias artificiales explicativas, tanto post-hoc como ante-hoc. Diwali A. et al [7], también hace una comparativa de modelos ante-hoc y post-hoc, recurriendo a la literatura de diversas investigaciones y sus resultados, al igual que proponer modelos híbridos de explicación.

En el estudio [8] exploran la evolución de los métodos de análisis de sentimientos y explicación de estos durante el último decenio, estos investigadores, tomaron una gran muestra de literatura del tema para analizar y discutir resultados, haciendo uso de métodos como similitud de palabras claves, palabras clave de la comunidad, etc. Los resultados son que se usa mucho los términos de Machine Learning, NLP y Twitter, y cómo estos se aplican en una gran variedad de tópicos, pasando por tipos de modelos de machine learning, diferentes aplicaciones en finanzas, reseñas de diferentes productos y servicios, al igual que en diferentes lenguajes, y la falta de investigación en dichos lenguajes.

Busra S. et al [20], investigan la importancia y nuevas investigaciones de la aplicación de XAI en los LLMs para la toma de decisiones transparentes y qué modelos de XAI se están aplicando en las investigaciones, usualmente modelos como SHAP y LIME.

Margarita R. et al [9] explora como las redes sociales son una gran base de datos para investigaciones en análisis de

sentimientos y cuáles son las herramientas más comunes, las mejores y las más recientes para hacer los análisis, por último, sugiere que todavía hay muchas oportunidades de investigación en esta área de investigación.

Teja S. et al [10] demuestran que las investigaciones en el área de análisis de sentimientos en redes sociales los modelos más aplicados son lexicón y machine learning, en múltiples áreas como: industria, servicios médicos y seguridad.

3 Objetivos

Se hará un Análisis de Sentimiento Explicable, pero, aplicado a comentarios políticos, para poder saber la razón por la cual se clasifican comentarios como pensamientos de izquierdas o derechas.

Se usará un algoritmo de Análisis de Sentimientos Explicable en comentarios acerca de política, aplicando una arquitectura de redes neuronales tipo transformer y una técnica de explicabilidad local agnóstica, similar a lo aplicado en [3] presentado en la figura 1, donde se ve que modelos con explicabilidad intrínseca y algoritmos de explicabilidad post-hoc pueden ser usados para explicar las razones de la predicción.

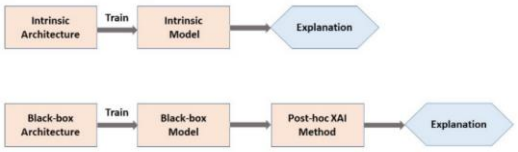

Figura 1: Explicabilidad post-hoc vs ad-hoc

3 Método

El método propuesto sigue las pautas que se presentan en la **figura 2**.

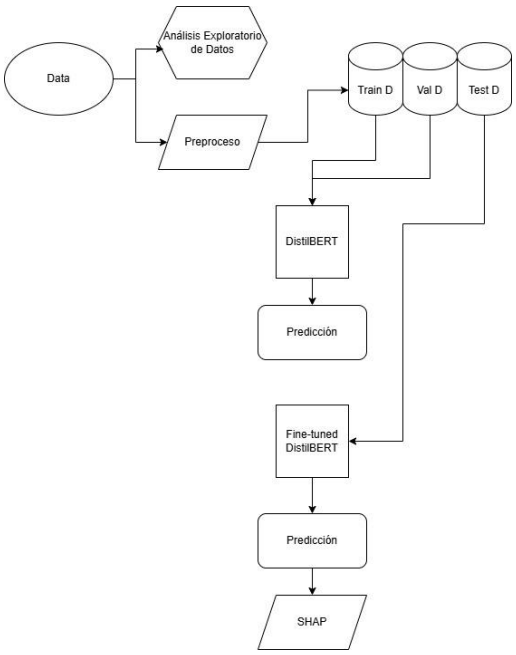

Figura 2: Método propuesto

El método propuesto es bastante común en investigaciones de inteligencia artificial, donde se recopila un dataset, se preprocesa y se entrena un modelo con el dataset, para luego usar un algoritmo post-hoc para que explique las razones de la predicción.

En este caso, elegimos el modelo distilBERT que es un modelo ligero de BERT que da resultados casi idénticos a los de BERT pero que necesita mucho menos tiempo de procesamiento [2].

Obtenemos el dataset de Kaggle, el cual es preprocesado para tener una data que el modelo pueda entender y no contenga caracteres que interfieran en el rendimiento del modelo.

3.1 Definir preguntas de investigación

Las preguntas de investigación aplicadas son:

1. ¿Podemos implementar un algoritmo que use una arquitectura de redes neuronales tipo transformer para realizar una clasificación de textos sobre un conjunto de comentarios políticos obtenidos de una red social?
2. ¿Se evalúan los resultados obtenidos del algoritmo que usa una arquitectura de redes neuronales tipos transformer?
3. ¿Se puede implementar un algoritmo de explicabilidad basado en una técnica local agnóstica?

4 Resultados previos

Para los resultados obtenidos, se usó la librería de HuggingFace: Transformer, de la cual se usa “distil-bert-uncased” que es una versión más pequeña de BERT

producida por la investigación de Chaumond et al. (2020), en la cual encontraron una forma de reducir el modelo BERT en un 40% mientras retiene el 97% de sus cualidades y hacerlo un 60% más rápido.

El proceso es limpiar la data, aplicar un label encoder a las etiquetas, tokenizar los textos, convertir a tensores los textos tokenizados, y crear una red neuronal relativamente pequeña para entrenar el modelo con la data recopilada, luego se evalúa la data en la cual se obtiene una baja "evaluation loss" de 0.4733. A pesar de eso, cuando pasamos a la parte de testing, el modelo lo hace relativamente bien, se entiende que los resultados y las métricas no sean muy buenas debido a la red neuronal entrenada, sin embargo, los resultados son aceptables tal y como se ven en la matriz de confusión (donde 0 representa ser "neutro" y 1 ser "partisano") y las siguientes métricas: Accuracy: 0.770, Precision: 0.723, Recall: 0.737, F1 score: 0.729.

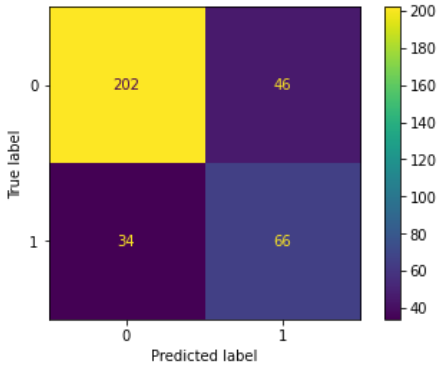

Figura 3: Matriz de Confusion

Para la explicación que es el siguiente paso de nuestra investigación se usa SHAP, ya que las explicaciones de SHAP son individuales, lo que significa que explica la razón del por qué predijo como una determinada clase a un determinado texto, y no una explicación general (para lo cual usaríamos algo como Feature Importance), se agarró solamente una muestra del total, las muestras tomadas tenían como característica que el score o probabilidad, por la cual el modelo eligió una clase determinada, sea alto de 0.9 para arriba, y que se tenga una cantidad igual de cada clase predicha.

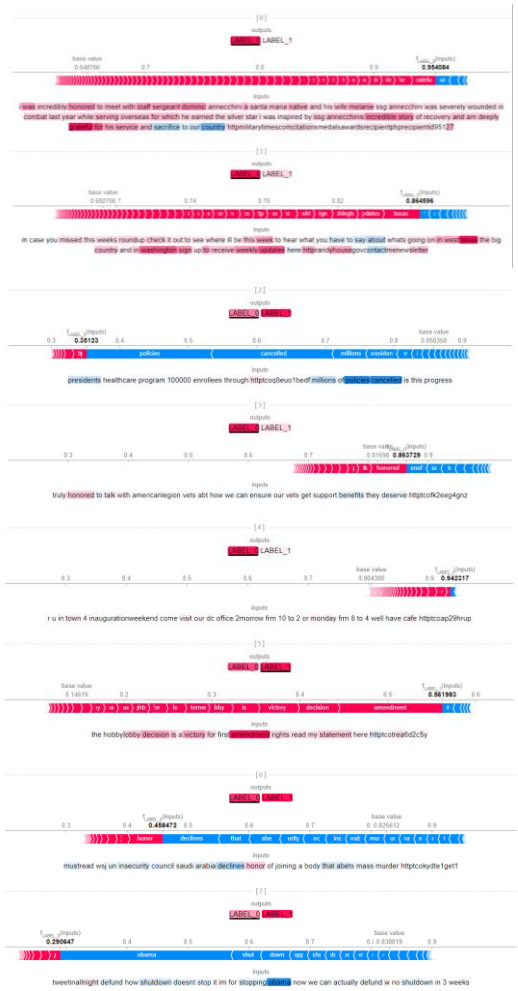

Figura 4: Resultados de SHAP

Si bien, las predicciones son relativamente correctas, las explicaciones de SHAP, no nos dan suficiente información, pero podemos ver que palabras como "honorable", "Texas", "servir" son palabras que se inclinan más a una predicción de tipo "neutra". Por otro lado, la palabra "Obama" es una palabra que pesa a la hora de predecir un comentario de tipo "Partisano", sin embargo, se nota claramente que el comentario va en contra de Obama y lo lógico es pensar que el comentario es "neutro", tal y como está etiquetado en el dataset.

En el análisis exploratorio de datos se encontró que algunas palabras como "president", "american", "health", "care", "government", son frecuentes en ambas clases, sin embargo, tenemos que tomar en consideración que el "neutral" es la clase dominante, casi tres veces más que la clase "partisan". Por ello en el AED el siguiente paso fue dividir las palabras de la clase dominante entre la ratio de diferencia de las dos clases, y encontramos que palabras como "partisan". "Obamacare", "president", "congress", "Bill", "health", "law", "Obama", "government", "tax", "reform",

"economy", "Budget", "legislation", "immigration", "federal", son palabras que tienen más frecuencia en la clase "partisan", esto se puede apreciar en la figura 5, donde las líneas azules representan la frecuencia de palabras de la clase "partisan" y las de color naranja de "neutral". También, debemos considerar que BERT va a entender el contexto en el que se usan estas palabras, sin embargo, esto depende de la cantidad de data que se tenga y de qué tan diferente es el contexto, sin embargo, puede que el modelo tenga un sesgo a la hora de clasificar los textos cuando encuentre estas palabras, porque de todas formas hay una gran diferencia en la frecuencia en la que aparecen estas palabras en cada clase.

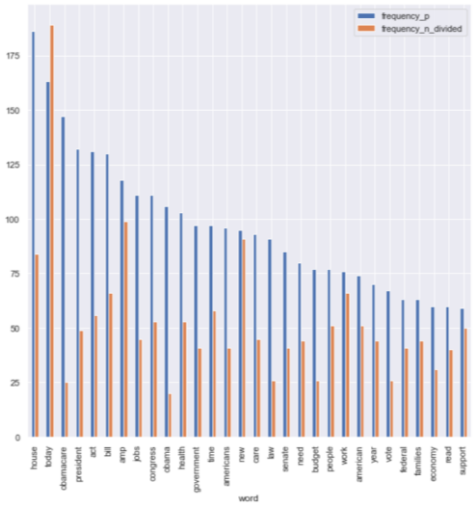

Figura 5: Histograma frecuencia de palabras

5 Discusión

Podemos deducir con el análisis exploratorio de datos que una gran parte de estas explicaciones no convincentes se puede deber a que la calidad de la data es muy baja, al parecer el modelo BERT no está capturando el contexto de los textos, más bien, está usando las palabras como predictores, esto puede deberse a muchas cosas, entre ellas, la baja calidad de la data.

Por otro lado, las métricas son medianamente buenas. Con lo analizado se necesitaría continuar con más ensayos y ajustes de los hiperparámetros así como buscar un dataset de mejor calidad.

5.1 Oportunidades de investigación

El estudio no aplica ninguna forma de métrica para evaluar modelos de explicabilidad post-hoc, por lo tanto, es una laguna que se puede estudiar o verificar si existe y aplicarla a estudios que apliquen este tipo de XIA o algoritmos de explicabilidad en análisis de sentimientos.

5.2 Limitaciones

Una de las limitaciones es la falta de accesibilidad a artículos de pago. La baja calidad de la data no permite validar el modelo y sus resultados correctamente. El trabajo se centra únicamente en comentarios de X.

6 Conclusiones

Con los resultados podemos concluir que se pueden hacer estudios en evaluaciones de modelos post-hoc o aplicar evaluaciones a estudios de XSA con algoritmos post-hocs, así como investigar o aplicar XSA a comentarios políticos, especialmente a comentarios políticos en español, donde hay una clara falta de investigación en el área.

ACKNOWLEDGMENTS

A las bases de datos de literatura académica y al docente del curso.

REFERENCES

[1] Arrieta, A.B.; Díaz-Rodríguez, N.; Ser, J.D.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. Inf. Fusion 2020, 58, 82–115. [Google Scholar] [CrossRef].

[2] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[3] A. Diwali et al., "Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis" in IEEE Transactions on Affective Computing, vol. 15, no. 03, pp. 837-846, July-Sept. 2024, doi: 10.1109/TAFFC.2023.3296373.

[4] Y. M. Tun and P. H. Myint, "Comparative study for text document classification using different machine learning algorithms,"International Journal of Computer (IJC), vol. 33, no. 1, pp. 19–25, 2019 (PDF) A large-scale sentiment analysis using political tweets. Available from: https://www.researchgate.net/publication/376118499_A_large-scale_sentiment_analysis_using_political_tweets.

[5] Maleszka, B. (2023). A Survey of Explainable Artificial Intelligence Approaches for Sentiment Analysis. In: Nguyen, N.T., *et al.* Intelligent Information and Database Systems. ACIIDS 2023. Lecture Notes in Computer Science(), vol 13996. Springer, Singapore. https://doi.org/10.1007/978-981-99-5837-5_5

[6] Shaun George Rajesh, Smriti Vipin Madangarli, Gauri Santosh Pisharady & Rolla Subrahmanyam. (2025) Enhancement of Virtual Assistants Through Multimodal AI for Emotion Recognition. *IEEE Access* 13, pages 102159-102179.

[7] A. Diwali, K. Saeedi, K. Dashtipour, M. Gogate, E. Cambria and A. Hussain, "Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis," in *IEEE Transactions on Affective Computing*, vol. 15, no. 3, pp. 837-846, July-Sept. 2024, doi: 10.1109/TAFFC.2023.3296373.

[8] Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho & Erik Cambria. Survey on sentiment analysis: evolution of research methods and topics. 06 de Enero del 2023.

[9] Mabokela, K. R., Primus, M., & Celik, T. (2024). Explainable Pre-Trained Language Models for Sentiment Analysis in Low-Resourced Languages. *Big Data and Cognitive Computing*, *8*(11), 160. https://doi.org/10.3390/bdcc8110160

[10] Abdelwahab, Y., Kholief, M., & Sedky, A. A. H. (2022). Justifying Arabic Text Sentiment Analysis Using Explainable AI (XAI): LASIK Surgeries Case Study. *Information*, *13*(11), 536. https://doi.org/10.3390/info13110536

[9] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. CuencaJiménez. "A review on sentiment analysis from social media platforms". In: Expert Systems with Applications 223 (2023), p. 119862. issn: 0957-4174. https://doi.org/10.1016/j.eswa.2023.119862.

[10] T. V. Sai Obulapuram, J. A. Sai Reddy Yendreddy, L. S. Kotari, V. T. Tirumalasetty, T. Santhi Sri and S. S. Imambi, "A Review: Sentiment Analysis Methods and their use in Social Media Platforms," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA),

Uttarakhand, India, 2023, pp. 499-504, doi: 10.1109/ICIDCA56705.2023.10100189.