ISSN 1870-4069

PLN con Transformers para detección de toxicidad: construcción y evaluación de corpus para la plataforma MisProfesores.com

María Lucía Barrón Estrada, Ramón Zatarain Cabada,
Ramón Alberto Camacho Sapien, Víctor Manuel Bátiz Beltrán

Instituto Tecnológico de Culiacán, Posgrado e Investigación,
México

{lucia.be, ramon.zc, ramon.cs, victor.bb}@culiacan.tecnm.mx

Resumen. El crecimiento de las redes sociales como medios de comunicación ha permitido una interacción más rápida y directa entre los usuarios, pero por otra parte presenta desafíos, como el riesgo de difusión de discursos de odio. Detectar estas publicaciones dañinas tempranamente es crucial. Este artículo presenta una metodología para crear un corpus único de comentarios en español obtenidos de la plataforma MisProfesores.com, abarcando todas las entidades federativas de México. Este proceso resultó en un conjunto de datos de 18,000 muestras no etiquetadas y 853 muestras etiquetadas manualmente. Además de describir el proceso de construcción del corpus, se exponen los resultados de la evaluación de varios modelos entrenados con estos datos, así como su comparación con trabajos previos para la detección de toxicidad existentes en el estado del arte, evidenciando la importancia del desarrollo de corpus en español para tareas específicas.

Palabras clave: Análisis de sentimientos, aprendizaje profundo, BERT, corpus, toxicidad, transformers.

NLP with Transformers for Toxicity Detection: Corpus Construction and Evaluation for MisProfesores.com Platform

Abstract. The growth of social networks as a means of communication has enabled faster and more direct interaction between users, but also presents challenges, such as the risk of spreading hate speech. Detecting these harmful publications early is crucial. This paper presents a methodology to create a unique corpus of Spanish-language comments obtained from the MisProfesores.com platform, covering all Mexican states. This process resulted in a dataset of 18,000 unlabeled samples and 853 manually labeled samples. In addition to describing the corpus construction process, the results of the evaluation of several models trained with these data are presented, as well as their comparison with previous works for toxicity detection existing in the state of the art, evidencing the importance of corpus development in Spanish for specific tasks.

Keywords: Sentiment analysis, deep learning, BERT, corpus, toxicity, transformers.

1. Introducción

El análisis de sentimientos es una subárea del procesamiento de lenguaje natural (PNL) que se enfoca en la identificación automática y la categorización de emociones y sentimientos expresadas dentro de un texto [1]. Este proceso es aplicable a diferentes sectores de la sociedad. Por ejemplo, los medios digitales de comunicación como las redes sociales, en dónde el cambio y expresión de ideas son una actividad recurrente, requieren un monitoreo constante para garantizar la integridad de los usuarios. Esto ha convertido a dichos medios en un sector importante donde aplicar el análisis de sentimientos. Como resultado, el análisis de sentimientos aumentó su popularidad entre las comunidades de investigación en los años recientes [2].

En plataformas como MisProfesores (www.misprofesores.com) en dónde el idioma español es el predominante, usar modelos de aprendizaje para aplicar un análisis de sentimientos representa una tarea complicada debido a la complejidad y diversidad de la lengua. El desarrollo de modelos de aprendizaje automático efectivos para el análisis de sentimientos en textos en español está limitado por los corpus en español existentes. En este artículo, revisaremos el proceso de construcción de un corpus en español a partir de la extracción de los comentarios publicados por usuarios de nacionalidad mexicana en la plataforma MisProfesores.

Este artículo está organizado de la siguiente manera. En la sección 2 se hace un repaso de los trabajos relacionados que hacen referencia a la construcción de corpus en español. En la sección 3, describimos el proceso para la recolección y procesamiento de los datos. La sección 4 muestra la metodología empleada, incluyendo el proceso de construcción de datos, así como los algoritmos y modelos de aprendizaje automático usados para las pruebas. Los resultados de las pruebas están en la sección 5. Y por último en la sección 6 presentamos nuestras conclusiones.

2. Trabajos relacionados

La detección de toxicidad en comentarios de Internet se ha consolidado como un área de interés creciente dentro del campo del PLN. Diversas plataformas que permiten calificar y dejar reseñas sobre docentes se han convertido en focos significativos para el análisis de sentimientos, dado que acumulan una gran variedad de comentarios. Un ejemplo notable de investigación en este ámbito es el estudio realizado por Arceo-Gomez and Campos-Vazquez [3] en donde se llevó a cabo un análisis estadístico exhaustivo sobre aproximadamente 600,000 evaluaciones.

Este estudio no solo proporciona perspectivas sobre las interacciones en dichas plataformas, sino que también destaca la presencia de estereotipos de género, subrayando la importancia de las técnicas de PLN para identificar y mitigar sesgos implícitos. Por otra parte, en el trabajo presentado por Kolhatkar et al. [4] se describe el desarrollo de un corpus considerable a partir de comentarios en inglés, recopilados de sitios web de noticias, que incluye cerca de 500,000 muestras.

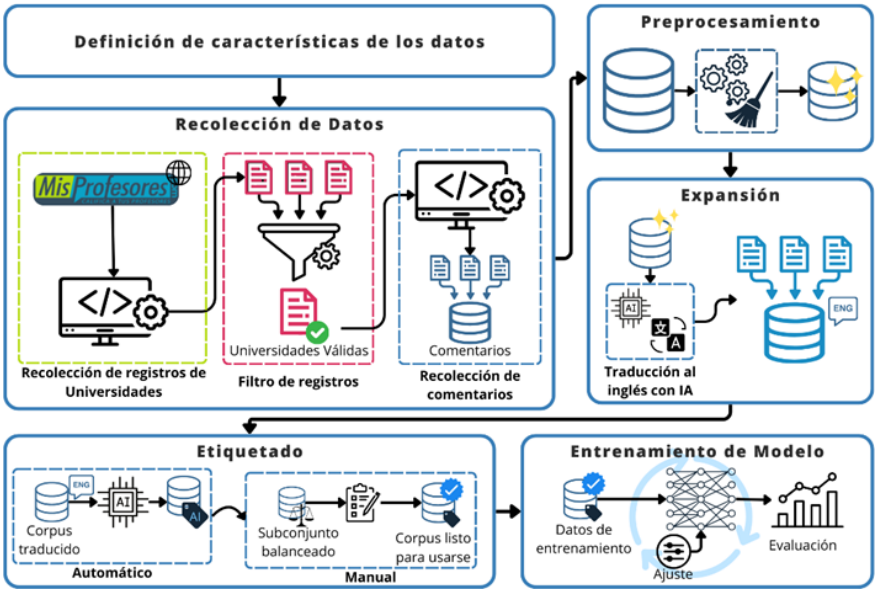

Fig. 1. Diagrama representativo para la metodología.

Este estudio es particularmente relevante porque no solo desarrollaron un corpus amplio, sino que también etiquetaron un subconjunto de aproximadamente 1000 muestras, enfocándose en la clasificación de la toxicidad de los comentarios.

Este enfoque ofrece una base sólida para futuras investigaciones sobre herramientas automatizadas de moderación de contenido. En relación con el uso y la eficacia de los modelos de inteligencia artificial (IA), Nabiilah et al. [5] aportan una valiosa comparativa entre modelos preentrenados en distintos corpus. Los hallazgos indican que aquellos modelos entrenados con corpus en idiomas específicos obtienen mejores resultados en tareas de clasificación de toxicidad en el mismo idioma.

Esto subraya la importancia de considerar las variaciones lingüísticas y culturales al diseñar y entrenar modelos de IA para la moderación de contenido. En contraste con los trabajos previamente mencionados, este estudio presenta un enfoque que amalgama los procedimientos y resultados revisados anteriormente. Aquí, se introduce una metodología integral que abarca áreas como la minería de opiniones, la construcción de conjuntos de datos y el entrenamiento de modelos.

3. Metodología

En la sección de metodología, se detalla un proceso comprensivo que abarca desde la definición inicial de las características deseables del corpus hasta el proceso de entrenamiento de distintos modelos usando un corpus etiquetado manualmente desarrollado en este trabajo. En la Fig.1 se presenta un diagrama que ilustra cada etapa de la metodología.

Fig. 2. Reseña extraída de la plataforma MisProfesores.

3.1. Definición de características de los datos

Como se observa en la Fig. 1, el primer paso para la construcción del corpus es definir las características de los datos a extraer.

Se estableció una escala geográfica a nivel nacional como la más adecuada, dado que está limitada por el alcance de la plataforma MisProfesores. Se seleccionó cada estado del país como puntos de interés para la extracción de los datos de dicha plataforma. Esto aseguró abarcar la diversidad lingüística de México. Posteriormente, se definieron las características de las instituciones académicas de las cuales se extrajeron los datos. Dichas instituciones cumplen con las siguientes características: ser universidad pública, escolarizada y con la mayor relevancia a nivel estatal según la cantidad de matrículas registradas. Lo anterior se realizó con base en los datos de la ANUEIS 2023 [6]. Los datos de interés para su obtención fueron las reseñas realizadas por los alumnos hacía los profesores de las instituciones previamente seleccionadas. Cada reseña en la plataforma está compuesta por un comentario/opinión del alumno, la materia, la calificación obtenida con el docente en la materia, la fecha, entre otros (Véase Fig. 2).

3.2. Recolección de datos

Mediante técnicas de extracción de textos de sitios Web, comúnmente conocidas como Web Scrapping se extrajeron los datos del sitio web MisProfesores. Para ello, se utilizaron herramientas como: Selenium y BeautifulSoup, ambas bibliotecas de Python usadas para extraer datos generados de manera dinámica y estática respectivamente. Debido a la flexibilidad que ofrece la plataforma MisProfesores para registrar profesores e instituciones educativas, se implementó un filtro en el algoritmo de extracción. Este filtro únicamente capturaba los registros de las universidades más destacadas en los resultados de la plataforma, como se muestra en la Fig. 3. La selección del filtro se basó en el número de registros de profesores asociados a cada institución educativa.

3.3. Preprocesamiento de los datos

Para el comentario de cada reseña, el texto obtenido fue sometido a un proceso en donde fueron eliminados signos de puntuación innecesarios. Por ejemplo, signos de exclamación o interrogación repetidos al iniciar o terminar una oración. Con este procedimiento, fue reducido el ruido presente en el corpus. También fueron descartadas las reseñas que presentaban un comentario:

Escuela	Ciudad	Estado	Num. de Profs.
Universidad Autónoma de Sinaloa	Culiacán	Sinaloa	117
Universidad autónomas de sinaloa	Culiacán	Sinaloa	1
Universidad Autónoma de Sinaloa	Angostura	Sinaloa	0
UNIVERSIDAD AUTONOMA DE SINALOA	LOS MOCHIS	SINALOA	6
Universidad autónoma de Sinaloa	Guamuchil	Sinaloa	1
Universidad Autónoma de Sinaloa	Culiacan	Sinaloa	16
Universidad Autónoma de Sinaloa	Mazatlán	Sinaloa	34

Fig. 3. Resultados de búsqueda para "Universidad Autónoma de Sinaloa" remarcado en rojo el registro más relevante en base al número de profesores registrados.

Fig. 4. Muestra de reseña con comentario inválido.

- Que solo estuviese compuesto por espacios en blanco o completamente vacío.
- Compuesto solo por caracteres especiales y/o números.
- En espera de revisión o bloqueados por la misma plataforma MisProfesores.

En la Fig. 4 se muestra un ejemplo de una reseña con un comentario inválido, en este el comentario se encuentra en espera de revisión por la plataforma. Posterior a este procedimiento de filtración de los datos, se obtuvo un corpus de 18,000 muestras dónde cada muestra contiene los datos mencionados anteriormente en el apartado 3.1.

3.4. Expansión del conjunto de datos

Para enriquecer el conjunto de datos se aplicó una traducción del español al inglés a cada comentario del corpus. Por la cantidad de muestras presentes en el corpus, la traducción se realizó de manera automática mediante modelos basados en Transformers. Como resultado, se obtuvo una versión en inglés del corpus, la cual amplió los casos de uso de este. Dicha versión en inglés nos sirvió para la etapa de etiquetado automático.

3.5. Etiquetado del conjunto de datos

Por la magnitud del corpus obtenido, se realizó un etiquetado automático a los comentarios en su versión en inglés utilizando varios modelos clasificadores basados

Tabla 1. Comparación entre distintas arquitecturas de modelos usando el Corpus MisProfesores.

Modelo	Exactitud	Recall	F1
EvoMSA BoW	0.8479	0.8212	0.8067
EvoMSA BoW + Text Representation	0.8596	0.8522	0.8262
LSTM	0.7134	0.6714	0.6761
mBERT base	0.8011	0.8011	0.8027
XLM RoBERTa base	0.8245	0.8245	0.8233
BETO base	0.9649	0.9649	0.9645

en Transformers entrenados para analizar sentimientos y clasificar toxicidad en texto. Este procedimiento permite obtener resultados preliminares y conocer el estado del balance de los datos. Con base en este primer etiquetado automático, se extrajo un subconjunto balanceado de muestras tóxicas y no tóxicas. Posteriormente, profesores de nuestra institución lo etiquetaron manualmente.

Para llevar a cabo el etiquetado manual, se establecieron diversos criterios a considerar. Por ejemplo, se instruyó al equipo a leer minuciosamente cada comentario con el fin de identificar posibles expresiones sarcásticas. Además, se enfatizó la importancia de no basar exclusivamente la etiqueta del comentario en palabras altisonantes, sino evaluar el contexto en el que se utilizan. Como resultado, se obtuvo un corpus etiquetado manualmente conformado por 853 muestras. Cada muestra cuenta con un texto en español presente en el comentario de la reseña original y una etiqueta binaria. En la etiqueta binaria, el número "1" representa que el texto de la muestra es tóxico y el número "0" significa la ausencia de toxicidad en el texto.

3.6. Entrenamiento de modelo para clasificación de texto

Para desarrollar un modelo capaz de clasificar comentarios en la plataforma MisProfesores como tóxicos o no tóxicos, se procedió al entrenamiento con el subconjunto de 853 muestras etiquetadas manualmente. Dado el enfoque del estudio en español, se optó por utilizar modelos de aprendizaje automático (ML) como máquinas de soporte vectorial y aprendizaje profundo (DL) como modelos neuronales con arquitectura LSTM y modelos basados en arquitecturas de Transformers. Todos los modelos fueron entrenados a 10 épocas, un tamaño de lote de 16 y una caída de peso de 0.01. Definiendo tales hiperparámetros, aseguramos el entrenamiento justo entre los modelos. Los procesos de entrenamiento y evaluación se realizaron en la nube, usando un mismo hardware para cada modelo.

3.7. Experimentos y resultados

Usando el subconjunto de comentarios y sus correspondientes etiquetas, mencionado previamente, se entrenaron distintos modelos de ML y DL. Los modelos de ML utilizados fueron bolsa de palabras y una combinación de bolsa de palabras con representación de texto ambos modelos construidos con apoyo de la biblioteca EvoMSA [7].

En cuanto a los modelos de DL, se entrenó una red LSTM básica, conocida por su efectividad en tareas de procesamiento de lenguaje natural. También se entrenaron 3

Tabla 1. Resultados de distintos modelos de clasificación de toxicidad en texto.

Modelo	Exactitud	Recall	F1
BETO-MP	0.9649	0.9649	0.9645
TextDetox XLMR	0.7894	0.7894	0.7942
dehateBERT	0.6783	0.6783	0.6031

modelos basados en Transformers. El primero que fue mBERT, es un modelo basado en BERT [8] preentrenado con un gran corpus en distintos idiomas. El segundo modelo fue XLM-RoBERTa, un modelo variante de RoBERTa [9] preentrenado en un corpus multilenguaje.

El tercer modelo utilizado fue BETO [10], que, aunque también comparte su arquitectura con BERT, este fue preentrenado con un corpus exclusivamente en español. La tabla 1 nos muestra que el mejor modelo para esta tarea fue BETO base, obteniendo resultados sobresalientas en cada métrica, por lo que se tomó este como referencia para las siguientes comparaciones. A partir de este punto y para efectos prácticos, nos referiremos al modelo seleccionado BETO, como BETO-MP.

Posterior a la selección del modelo, se comparó este con otros modelos que forman parte del estado del arte en clasificación binaria de toxicidad. El primero, un modelo XLM-RoBERTa entrenado con un corpus presentado en [11] para la tarea de clasificación binaria de toxicidad el cual es una compilación de diferentes corpus en varios lenguajes, incluyendo el español. El segundo, "dehateBERT" [12] un modelo basado en BERT multilenguaje el cual fue entrenado con un corpus en español de toxicidad. Los modelos se evaluaron con el 20% restante de muestras para pruebas.

Los resultados de la evaluación de cada modelo se presentan en la Tabla 2. Como se muestra en la tabla anterior, el modelo BETO-MP obtuvo resultados superiores con respecto a los otros modelos presentados en otros trabajos en la tarea de clasificación de comentarios tóxicos en la plataforma MisProfesores.

4. Conclusiones

Este estudio representa un avance significativo en la construcción de corpus en español para el análisis de sentimientos, específicamente en el contexto de comentarios en la plataforma MisProfesores. A través de un meticuloso proceso que incluyó la recolección, preprocesamiento, limpieza, y etiquetado de datos, se desarrolló un corpus único que refleja la diversidad y complejidad del español hablado en México. Este corpus no solo es relevante por su tamaño, con 18,000 muestras no etiquetadas y 853 muestras etiquetadas manualmente (Ambos corpus disponibles en el sitio[1]), sino también por su enfoque en capturar la riqueza lingüística y cultural específica de este contexto.

Los resultados de nuestro modelo superan otros modelos entrenados con otros corpus para la detección de toxicidad, evidenciando la importancia de la construcción y especialización de un corpus para esta tarea específica con el fin de mejorar la precisión en la detección de comentarios tóxicos en plataformas sociales académicas. Estos

[1] catalabs.mx/datasets/misprofesores/

hallazgos no solo contribuyen al campo académico del análisis de sentimientos y la IA, sino que también ofrecen aplicaciones prácticas para plataformas educativas en línea, ayudando a crear ambientes de aprendizaje más seguros y positivos.

Al detectar y manejar de manera proactiva los discursos de odio, se puede fomentar un intercambio de ideas más respetuoso y constructivo, crucial para el desarrollo educativo y social. Finalmente, este estudio subraya la necesidad de continuar expandiendo y refinando los corpus en idiomas distintos al inglés, adaptándolos a contextos específicos para mejorar la efectividad de los modelos de análisis de sentimientos. En el trabajo futuro se recomienda aumentar la cantidad de muestras etiquetadas manualmente y explorar otras plataformas y contextos, con el objetivo de desarrollar modelos más robustos y versátiles. La construcción de estos recursos no solo beneficia la investigación académica, también tienen un impacto directo en la sociedad, promoviendo entornos digitales más inclusivos y respetuosos.

Referencias

1. Tan, K.L., Lee, C.P., and Lim, K.M.: A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. Applied Sciences, vol. 13, no. 7, pp. 4550 (2023). DOI: 10.3390/ app13074550.
2. Wankhade, M., Rao, A.C.S., and Kulkarni, C.: A Survey on Sentiment Analysis Methods, Applications, and Challenges. Artificial Intelligence Review, vol. 55, no. 7, pp. 5731–5780 (2022). DOI: 10.1007/s10462-022-10144-1.
3. Arceo-Gomez, E.O., Campos-Vazquez, R.M.: Gender Stereotypes: The Case of misprofesores.com in Mexico. Economics of Education Review, vol. 72, pp. 55–65 (2019). DOI: 10.1016/j.econedurev.2019.05.007.
4. Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M.: The Sfu Opinion and Comments Corpus: A Corpus for the Analysis of Online News Comments. Corpus Pragmatics, vol. 4, no. 2, pp. 155–190 (2020). DOI: 10.1007/s41701-019-00065-w.
5. Nabiilah, G.Z., Prasetyo, S.Y., Izdihar, Z.N., and Girsang, A.S.: Bert Base Model for Toxic Comment Analysis on Indonesian Social Media. Procedia Computer Science, vol. 216, pp. 714–721 (2023). DOI: 10.1016/j.procs.2022.12.188.
6. Asociación Nacional de Universidades e Instituciones de Educación Superior: Información estadística de educación superior, anuarios estadísticos de educación superior www.anuies.mx/informacion-y-servicios/informacion-estadistica-de-educacion-superior/ anuario-estadistico-de-educacion-superior (2024).
7. Graff, M., Miranda-Jimenez, S., Tellez, E.S., and Moctezuma, D.: Evomsa: A Multilingual Evolutionary Approach for Sentiment Analysis. In: IEEE Computational Intelligence Magazine, vol. 15, no. 1, pp. 76–88 (2020). DOI: 10.1109/mci.2019.2954668.
8. Devlin, J., Chang, M., Lee, K., and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019). DOI: 10.18653/v1/n19-1423.
9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451 (2020). DOI: 10.18653/v1/2020.acl-main.747.
10. Cañete, J., Chaperon, G., Fuentes, R., Ho, J., Kang, H., and Pérez, J.: Spanish Pretrained BERT Model and Evaluation Data (2023). DOI: 10.48550/arXiv.2308.02976.
11. PAN: Multilingual Text Detoxification (TextDetox). http://pan.webis.de/clef24/pan24-web/text-detoxification.html#task (2024).

12. Aluru, S.S., Mathew, B., Saha, P., and Mukherjee, A.: Deep Learning Models for Multilingual Hate Speech Detection. In: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, pp. 1–16 (2020). DOI: 10.48550/arXiv.2004.06465.