ISSN 1870-4069

Comparación de modelos para la clasificación automática de temáticas en tuits de comunicación pública de la ciencia en español de México

Alec Sánchez-Montero, Gemma Bel-Enguix,
Sergio Luis Ojeda-Trueba

Universidad Nacional Autónoma de México,
México

alecm@comunidad.unam.mx,
{gbele, sojedat}@iingen.unam.mx

Resumen. En el contexto mexicano, los estudios exhaustivos sobre la comunicación pública de la ciencia (CPC) a través de redes sociales son una tarea pendiente hasta la fecha. Como respuesta a este vacío, se propone este trabajo desde la perspectiva del procesamiento del lenguaje natural (PLN). En concreto, el estudio apunta al desarrollo y la evaluación de la clasificación automática de tuits de CPC publicados en México mediante el entrenamiento de distintos modelos de aprendizaje automático, incluidos algoritmos clásicos y modelos basados en transformers. Con base en un corpus etiquetado manualmente, se evalúan y se comparan varios enfoques para identificar y clasificar automáticamente las áreas temáticas en los tuits de CPC. Los resultados muestran que los modelos clásicos como support vector machine mantienen un rendimiento sólido, mientras que los transformers ofrecen alternativas prometedoras para esta tarea.

Palabras clave: Procesamiento del lenguaje natural, clasificación multietiqueta de texto, comunicación pública de la ciencia.

Model Comparison for Automatic Topic Classification in Mexican Spanish Tweets on Public Communication of Science

Abstract. In the Mexican context, comprehensive studies on public communication of science (PCS) through social networks are a pending task to date. In response to this gap, this work is presented from the perspective of natural language processing (NLP). Specifically, the study aims at developing and evaluating the automatic classification of PCS tweets published in Mexico by training different machine learning models, including classical algorithms and models based on transformers. On the basis of a manually labeled corpus, several approaches to automatic identification and classification of thematic areas in PCS tweets are evaluated and compared. The results show that classical models such as support vector machine maintain a robust performance, while transformers offer promising alternatives for this task.

Keywords: Natural language processing, multi-label classification of text, public communication of science.

1. Introducción

En el contexto de la denominada "Cuarta Revolución Industrial", caracterizada por el amplio uso de las tecnologías digitales, redes y plataformas sociales se han reafirmado como espacios virtuales propicio para la comunicación de conocimientos científicos de distintas disciplinas. En estas redes y plataformas, la información científica se destina tanto a expertos como a audiencias no especializadas. En particular, Twitter —ahora 𝕏— se ha convertido en un popular espacio de comunicación digital masiva, donde se comparten conocimientos científicos, se discuten descubrimientos y se promueve el diálogo sobre temas científicos de distintas disciplinas. Este tipo de interacciones entre investigadores, divulgadores, entusiastas de la ciencia y el público general proporciona una conveniente fuente de datos para explorar y comprender los fenómenos relativos a la comunicación de la ciencia, en los cuales el lenguaje natural tiene una función central.

Específicamente, la comunicación pública de la ciencia (CPC) implica un proceso de difundir y divulgar conocimientos científicos hacia el público en general en un canal bidireccional, fuera del ámbito especializado entre pares o expertos. En este sentido, los comunicadores de la ciencia se dirigen a personas que no poseen formación especializada en ciencias para construir un diálogo en torno a los descubrimientos, los avances y los debates científicos. En años recientes, esta actividad se ha expandido más allá de los canales tradicionales, como las revistas de divulgación, hacia los entornos digitales de interacción social. Para el caso de la CPC en Twitter/𝕏, los textos publicados, conocidos como "tuits" o posts, son breves y fragmentados, influenciados por factores como la interactividad de la plataforma y las limitaciones de espacio [1, 2].

El uso de Twitter en México es bastante significativo, puesto que la base de usuarios excede los 17 millones. Esto sitúa a México entre los primeros 10 países con mayor cantidad de usuarios activos a nivel mundial y, además, como el segundo lugar en Latinoamérica, sólo por detrás de Brasil; asimismo, se trata del país hispanohablante con mayor presencia global[18]. Esta plataforma es especialmente útil para la investigación en lingüística y en procesamiento del lenguaje natural (PLN), pues se pueden compilar corpus lingüísticos a partir de la amplia cantidad de datos generados diariamente y, de ese modo, se pueden estudiar distintos aspectos lingüísticos en función de los objetivos que se persigan en la investigación [20].

Desde una perspectiva de PLN y de ciencia de datos, el estudio de la CPC en México a través de Twitter permite analizar distintas facetas del fenómeno, por ejemplo: la interacción entre los usuarios y los comunicadores científicos, la cobertura de la información en función de los temas, las disciplinas o áreas en las que pueden clasificarse los textos, la influencia de comunicadores particulares, la percepción pública de la actividad científica a nivel nacional e internacional, el impacto del conocimiento científico comunicado en los individuos y en la sociedad, entre otras. Sin embargo, pese al elevado número de usuarios en México, en la revisión de la literatura no se han localizado conjuntos organizados de datos, de libre acceso, en formato de tuits del género CPC publicados en México. Este vacío implica una falta de estudios exhaustivos sobre la CPC en México y, en particular, la ausencia de modelos de aprendizaje automático, o Machine Learning, especializados en la clasificación de tuits según su contenido temático.

En este contexto, el objetivo de este trabajo es desarrollar y evaluar un modelo de clasificación automática basado en un corpus, conformado por tuits de CPC publicados en México, anotado manualmente mediante un sistema multietiqueta conforme a las áreas temáticas abordadas en cada texto del corpus. La metodología propuesta se basa en un enfoque de aprendizaje automático supervisado, en el cual se entrena un modelo computacional a partir de los ejemplos etiquetados previamente por anotadores humanos. En este caso, se lleva a cabo un análisis contrastivo entre algoritmos clásicos de aprendizaje automático, como Support Vector Machine (SVM) y Random Forest Classifier (RFC), y modelos preentrenados, basados en arquitecturas de aprendizaje profundo tipo transformers, como BERT y RoBERTa, con la finalidad de identificar las características más adecuadas en el modelo para una tarea de clasificación automática multietiqueta.

2. La CPC en Twitter/𝕏

En el ámbito de la comunicación científica, se ha generado ambigüedad terminológica entre conceptos como "divulgación científica", "comunicación pública de la ciencia", "alfabetización científica", "periodismo científico", "participación pública en la ciencia", entre otros. A rasgos muy generales, uno de los propósitos de la CPC es acercar el conocimiento científico a un público amplio y no especializado. Al referirse a una CPC en lugar de una divulgación científica, se acentúa la característica bidireccional en el proceso comunicativo.

Según [16], las actividades de CPC como un campo "multi, inter y transdisciplinario que conjunta saberes provenientes de diversas áreas tales como las ciencias naturales, exactas, de la salud, tecnologías, ingenierías y recientemente sociales y humanísticas, así como el manejo de los distintos medios de comunicación y el conocimiento de los diferentes públicos". Desde este enfoque, las actividades de divulgación científica estarían englobadas en las de CPC, como concepto más general.

Un modelo de comunicación científica, como el de la participación del público, encuentra en Twitter/𝕏 un lugar apropiado para su implementación. En este modelo específico se busca generar un diálogo y un compromiso con el público no científico [12]. Por su parte, trabajos previos han estudiado en esta red social fenómenos como la influencia de las interacciones entre usuarios en el interés y la comprensión del público hacia la ciencia [8], el impacto de determinadas figuras clave en la comunicación científica [7], la relación entre determinados temas científicos y el interés público hacia la ciencia mediante métricas de interacción [10], la diseminación del vocabulario científico [19] o el papel de las comunidades educativas para promocionar la comunicación científica [9].

En años recientes, se ha notado cómo el uso de plataformas digitales como Twitter/𝕏 para la comunicación científica despeja las líneas divisorias entre la comunidad "especializada" y el público general [14]. La investigación sobre la comunicación científica en Twitter/𝕏 es un tema incipiente, aunque cada vez se destaca más su relevancia como escenario para la circulación de la información científica y para la interacción entre científicos y audiencias no científicas.

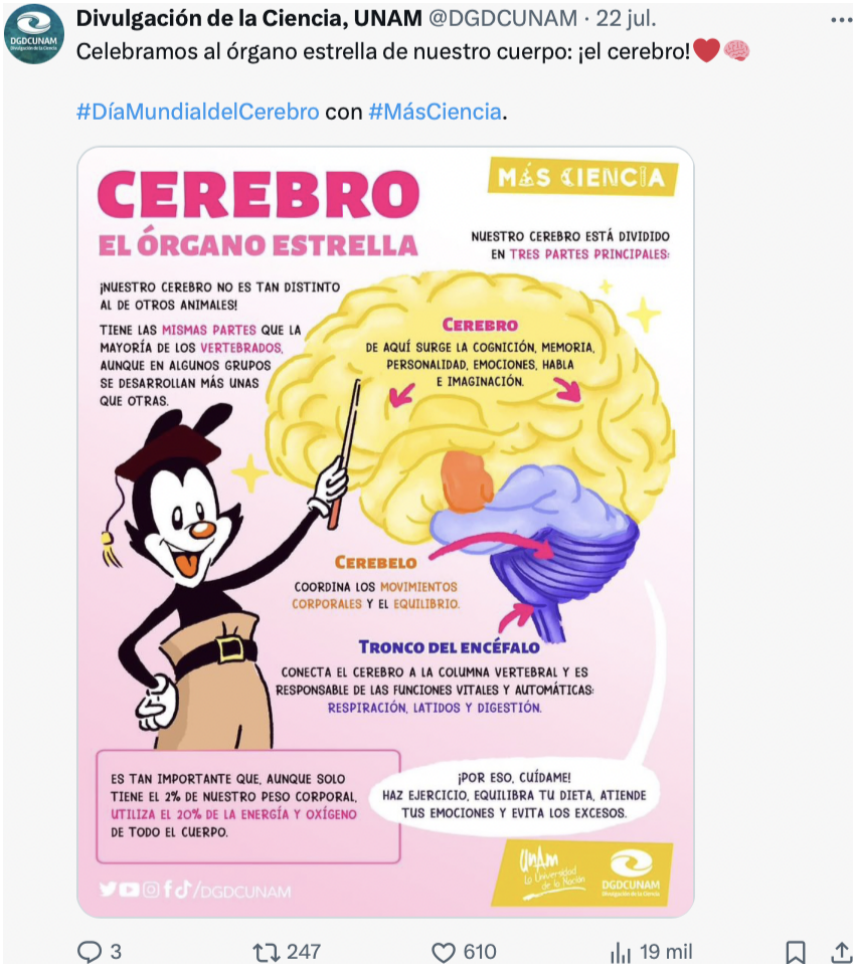

Fig. 1. Ejemplo de tuit de CPC en español mexicano (Fuente: Twitter/𝕏).

Asimismo, la comunidad especializada ha comenzado a centrarse más en esta red social, pues la han reconocido como una oportunidad para investigar, expandir y debatir el conocimiento científico [4, 6, 13]. Según el modelo de participación del público [12], Twitter/𝕏 actúa como una plataforma global para democratizar el conocimiento, aunque, al mismo tiempo, puede contribuir a multiplicar la desinformación y el conocimiento falaz, como se vio a lo largo de la pandemia por COVID-19 [5]. Con respecto al contexto mexicano, se ha observado, en una exploración inicial, una amplia diversidad de los temas comunicados en la plataforma.

Las temáticas destacadas incluyen áreas como la biología, la astronomía y las ciencias de la salud, aunque es común encontrar tuits que integran múltiples áreas científicas o temáticas en un mismo texto. Por otra parte, se debe señalar que los tuits de CPC suelen servir de complemento a otras actividades o enlazar con recursos destinados al público no especializado, ya sean en modalidad física o virtual.

Se incluye la Figura 1 para ilustrar algunas de los principales rasgos de estos tuits, a través de un ejemplo prototípico. Por lo que respecta a los agentes involucrados en la CPC, el conjunto de comunicadores de la ciencia es de carácter heterogéneo. Entre los participantes se puede encontrar tanto instituciones como individuos que comunican los conocimientos científicos. De dichos comunicadores destacan las instituciones de educación superior, así como organizaciones y centros de investigación, publicaciones de divulgación científica, investigadores, estudiantes y divulgadores individuales.

3. Metodología y características del dataset

La metodología de este trabajo se ha desarrollado con base en un pipeline para la clasificación de textos [11]. Este pipeline consiste en las siguientes etapas: 1) recopilación y selección de datos, 2) anotación de los datos, 3) preprocesamiento del texto, 4) extracción de características, 5) selección de las técnicas de clasificación y 6) evaluación. En la primera etapa, se recurrió a la Twitter API v2 para extraer los timelines de una lista de usuarios delimitados mediante la biblioteca Tweepy en un entorno de Python.

Esta lista de usuarios fue el resultado de una investigación para delimitar aquellos perfiles de Twitter relacionados con la CPC en el contexto mexicano. Estos perfiles corresponden a 19 autodenominados "divulgadores", "comunicadores" o "periodistas" científicos, de cuentas individuales e institucionales sobre áreas científicas generales. En otras palabras, las cuentas seleccionadas para este estudio representan una variedad de temáticas científicas generales y no especializadas, que que se corresponde con las áreas del conocimiento divulgadas en el contexto mexicano.

Cabe destacar que estos datos fueron recopilados sin preferencias específicas por un ámbito científico concreto, dado que se ha partido de un escenario con muy poca información cuantitativa y cualitativa respecto al contexto del objeto de estudio. Esta situación trajo consigo una amplia gama de temas en el corpus, desde la astronomía y la física general hasta la genética y la historia de la ciencia, entre otras áreas. En términos de clases del corpus, este tipo de distribución temática al azar puede implicar un desbalance en las clases del corpus.

Sin embargo, se justifica adoptar esta estrategia para reflejar de manera más precisa la distribución natural de etiquetas y clases en el contexto de los tuits de CPC en México, sin manipular los datos para lograr el balance entre clases. Al construir un dataset que refleje la verdadera distribución de clases, se puede identificar áreas específicas donde los modelos pueden tener un rendimiento óptimo y subóptimo, una característica fundamental basada en datos para la investigación de esta emergente área de estudio.

Para constituir un dataset consistente y apropiado para la tarea de clasificación, se delimitó como criterio de homogeneidad fundamental que todos los tuits hubieran sido publicados en español dentro de México, entre enero de 2020 y mayo de 2023, momento en el que se recopilaron de los datos. Tras reunir y estructurar esta información con la Twitter API v2, ocurrieron modificaciones significativas en la plataforma, derivadas del cambio de propietario. La transición del nombre "Twitter" a "𝕏" fue una de ellas. Otro cambio importante, relacionado con el acceso a y la recolección de los datos generados en la plataforma fue la eliminación del acceso gratuito a la API académica,

junto con la introducción de un modelo de pago mensual que comienza en los 100 dólares estadounidenses para acceder a una API con funcionalidades más limitadas, en comparación con la ahora inexistente versión académica. Es relevante considerar esta modificación, puesto que el desarrollo sucesivo de investigaciones relacionadas con esta plataforma podría verse condicionado por limitantes económicas.

Después de eliminar mensajes duplicados, textos de índole personal o de opinión no científica por parte de los autores, el dataset se conformó por 3733 tuits. Este dataset fue tomado como corpus de la investigación para ser etiquetado conforme al contenido temático de cada texto. De acuerdo con la función de tokenización de la biblioteca SpaCy, la cantidad total de tokens en el corpus es de 144,375 y la cantidad de tokens únicos es de 21,830.

Para llevar a cabo la anotación del corpus se seleccionó Argilla, una plataforma de código abierto especializada en el desarrollo de LLM a partir de conjuntos de datos que los usuarios pueden cargar mediante un código de programación a través de un entorno de Python. En esta tarea se buscó identificar y clasificar las áreas temáticas presentes en los tuits del corpus, más allá de procurar que estuvieran balanceadas, con base en una lista de etiquetas predefinidas.

Para anotar cada tuit, se empleó un enfoque de clasificación de texto multietiqueta, es decir, cada tuit podía ser etiquetado con una o más categorías temáticas, contenidas en la lista de etiquetas, según su contenido. Para ello, se optó por la función de "Feedback Dataset" en Argilla, la cual permite la clasificación multietiqueta de registros individuales en un dataset de manera flexible y sencilla.

Como resultado, se obtuvo un conjunto de 18 etiquetas que trató de ser lo más exhaustivo y representativo posible: "astronomía y espacio", "matemáticas", "física", "biología", "medicina y salud", "tecnología", "química", "historia de la ciencia", "ingeniería", "computación", "ciencias de la tierra", "materia y energía", "psicología", "invitación a evento o a recursos", "efeméride", "mujeres en la ciencia", "cultura pop" y "otro". La visualización de estas 18 etiquetas en la interfaz de Argilla[1] se muestra en la Figura 2.

Como instrucción para el etiquetado se indicó leer con atención cada uno de los tuits por separado y, con base en la información del texto y los rasgos contextuales de cada registro, seleccionar todas las etiquetas que describieran el contenido de cada tuit. Una vez terminada la tarea de anotación del corpus, se obtuvieron los resultados presentados en el gráfico de la Figura 3. Como puede observarse en el gráfico, cada tuit puede estar asociado con una o varias etiquetas, por la naturaleza multietiqueta del etiquetado, de forma que la suma de las frecuencias no es igual al total de registros del corpus.

Los resultados del etiquetado revelan la distribución de áreas temáticas predominantes en el corpus, la cual no corresponde a un reparto equilibrado de clases, pues algunas etiquetas mantienen una amplia representación en contraste con otras etiquetas con pocos casos identificados. Entre las etiquetas más frecuentes se encuentran "biología" con 1701 instancias, "invitación a evento o a recursos" con 1664, "física" con 1481, "astronomía y espacio" con 1223 y "medicina y salud" con 755. En contraste, las áreas menos representadas en el corpus son "computación" con 98 instancias, "cultura pop" con 78 y "otro" con 26.

[1] argilla.io

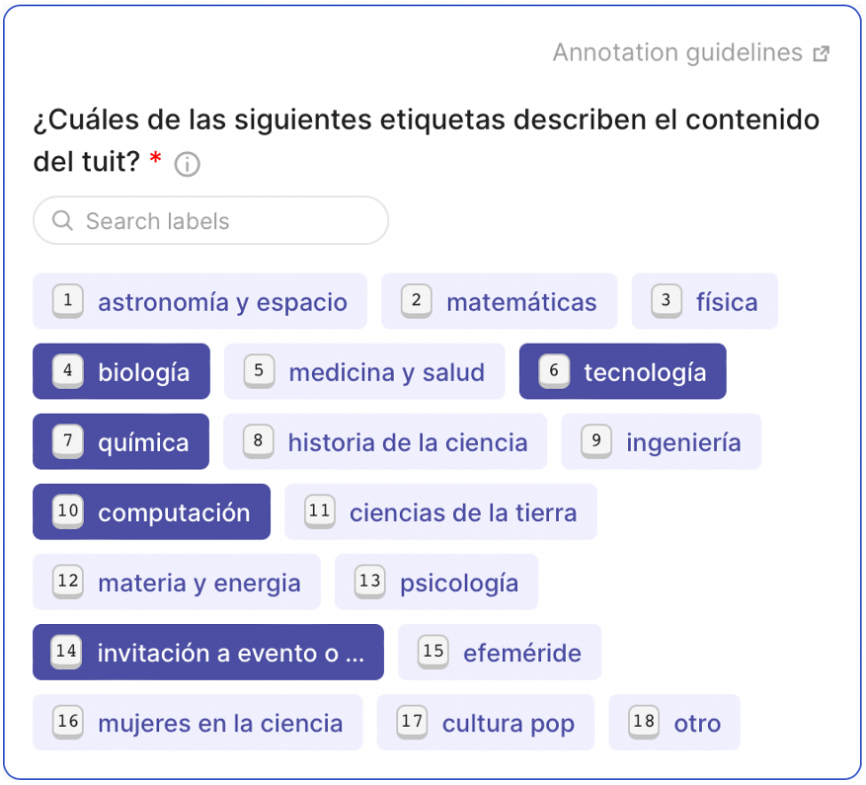

Fig. 2. Visualización de las etiquetas utilizadas para la anotación multietiqueta de las áreas temáticas del corpus en Argilla.

Esta distribución indica la prevalencia de áreas generales (como la biología y la física) y específicas (como la astronomía y temas del espacio) que podrían corresponder a los temas más relevantes en el contexto mexicano de la CPC, al mismo tiempo que las áreas menos representadas podrían ser las menos divulgadas (como el caso de la computación), complemento a otra área temática más general (como el caso de la cultura pop) o no relevante para el estudio de la CPC en Twitter (como la categoría de "otro"). Con este corpus como base para el entrenamiento para el modelo de clasificación automática de tuits, se procedió a las siguientes fases de la metodología.

Se seleccionaron algoritmos clásicos de aprendizaje automático debido a que algunos trabajos previos [15] han demostrado su utilidad para la clasificación de texto con datasets pequeños. A su vez, se buscó contrastar el rendimiento de estos algoritmos en relación con modelos basados en transformers, que representan un estado más avanzado en lo referente a tareas de PLN. Al utilizar los algoritmos clásicos, se consideró necesario preprocesar los textos del corpus. En esta etapa, se recurrió a diversas técnicas de limpieza y normalización de texto, como la eliminación de signos de puntuación, de emoji, de hashtags, de stop words, de hipervínculos, de direcciones de correo electrónico y otros elementos no léxicos, la lematización de palabras y la conversión a minúsculas.

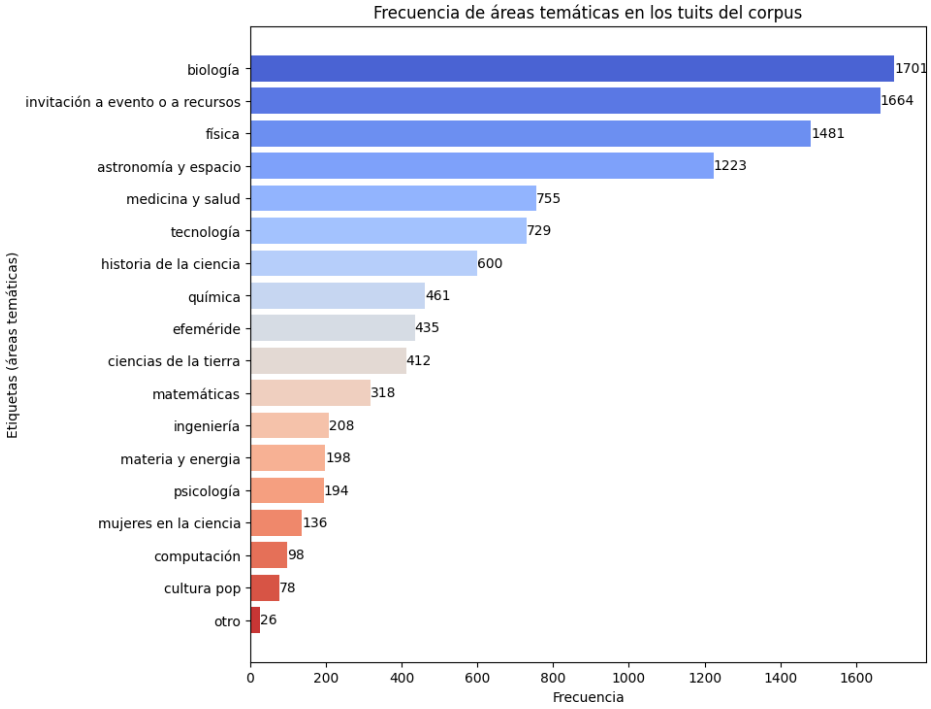

Fig. 3. Distribución de las áreas temáticas en los tuits del corpus.

A continuación, se extrajeron las características de los tuits preprocesados mediante la técnica TF-IDF para obtener representaciones numéricas de las características semánticas y contextuales de los textos, para poder entrenar los algoritmos clásicos de aprendizaje automático seleccionados. Por las características de los modelos basados en arquitecturas transformers, el proceso de entrenamiento y evaluación de los datos se simplificó en comparación con los modelos tradicionales de aprendizaje automático. En otras palabras, como las arquitecturas transformers están diseñadas para trabajar con secuencias de texto de manera eficiente, llevar a cabo una etapa de preprocesamiento exhaustiva se vuelve algo opcional.

En este caso, no se realizó preprocesamiento de los textos y, para la extracción de características, se empleó la biblioteca Transformers de Hugging Face a partir de la representación de los datos en forma de tokens de texto sin procesar, los cuales se proporcionaron como entrada a los modelos para que aprendieran automáticamente las representaciones de características durante el proceso de entrenamiento, como es práctica estándar con este tipo de modelos en PLN.

Por otra parte, ya que se trabajó con un conjunto de datos anotados con 18 etiquetas diferentes, se utilizó un enfoque de codificación de etiquetas conocido como multilabelbinarizer, mediante el uso de la biblioteca Scikit-learn. Esta técnica permite representar adecuadamente las etiquetas multietiqueta en forma de vectores binarios conformados por unos y ceros, donde cada elemento del vector indica la presencia (uno) o ausencia (cero) de una etiqueta particular.

Tabla 1. Resultados de los modelos SVM y RFC en el conjunto de datos de evaluación con base en métricas de precisión, exhaustividad y F1.

Modelo	Precisión	Exhaustividad	F1
SVM	0.89	**0.6**	**0.72**
RFC	**0.9**	0.56	0.69

Este enfoque fue utilizado tanto en los algoritmos clásicos como en los modelos transformers para homogeneizar el entrenamiento del modelo en la clasificación de tuits con múltiples etiquetas. Para la selección de algoritmos clásicos, se optó por dos algoritmos ampliamente utilizados en tareas de clasificación: Support Vector Machine (SVM) y Random Forest Classifier (RFC). En cuanto a los modelos preentrenados basados en arquitecturas transformers, se seleccionaron los siguientes: `distilbert-base-multilingual-cased` [17], `distilroberta-base` [17] y `twitter-xlm-roberta-base-sentiment` [3].

Por lo que respecta a los primeros dos, se trata de versiones ligeras y eficientes de BERT (en el caso de distilbert) y de RoBERTa (en el caso de distilroberta), los cuales han sido preentrenados en en varios idiomas, incluido el español. Finalmente, el modelo `twitter-xlm-roberta-base-sentiment` está específicamente diseñado para tareas de análisis de sentimientos en texto de redes sociales como Twitter y también fue preentrenado en español. Para evaluar estos modelos se emplearon métricas estándar de evaluación de clasificación: precisión, exhaustividad (recall) y F1-score.

4. Entrenamiento y evaluación de los modelos de clasificación automática

Como se mencionó en la sección anterior, para el entrenamiento con los algoritmos clásicos de aprendizaje automático (i.e. SVM y RFC), se llevó a cabo un preprocesamiento de los textos y una extracción de características mediante la técnica de TF-IDF. Asimismo, se empleó la técnica de validación cruzada con 5 folds o iteraciones para evaluar el rendimiento de los modelos con mayor precisión. Esta técnica permite dividir el conjunto de datos para obtener cinco medidas de rendimiento, que después se promedian para obtener una estimación general del desempeño del modelo. La Tabla 1 presenta los resultados detallados de los algoritmos SVM y RFC en el conjunto de datos de evaluación, o test dataset, correspondiente al 20 % del dataset con un random state de 42 para reproducibilidad.

Como puede observarse en estos datos, el modelo RFC tiene una precisión ligeramente mayor (0.9) en comparación con el modelo SVM (0.89). Sin embargo, el modelo SVM tiene una exhaustividad (o recall) más alta (0.6) en comparación con el RFC (0.5). En cuanto al F1-score, medida que combina precisión y exhaustividad, el modelo SVM obtuvo un valor de 0.72, mientras que el RFC obtuvo 0.69. Estos resultados indican un rendimiento muy similar por parte de los algoritmos clásicos de aprendizaje automático seleccionados. Ambos modelos muestran valores de precisión altos, lo cual significa que son capaces de realizar predicciones correctas con una tasa alta.

Tabla 2. Resultados de los modelos transformers con diferentes números de epochs en el conjunto de datos de evaluación.

Modelo	# de epochs	Precisión	Exhaustividad	F1
`distilbert-base- multilingual-cased`	5	**0.83**	**0.60**	**0.64**
	10	**0.83**	0.62	**0.67**
`distilroberta-base`	5	0.74	0.48	0.52
	10	0.73	0.56	0.60
`twitter-xlm- roberta-base-sentiment`	5	0.82	0.59	0.63
	10	0.81	**0.63**	0.65

Por lo que respecta a la exhaustividad y a la puntuación F1, se registran valores por encima del azar. Cabe señalar que estos resultados se refieren al rendimiento general de los modelos y no a las métricas por etiquetas específicas. En cuanto a los modelos basados en arquitecturas transformers, como ya se mencionó, tres modelos preentrenados en textos multilingües fueron seleccionados para la tarea de clasificación automática de texto. Dos de estos modelos corresponden a diferentes versiones del modelo RoBERTa y uno de ellos a una versión más compacta del modelo BERT.

Tanto `distilbert-base-multilingual-cased` como `distilroberta-base` fueron configurados para la tarea de clasificación de secuencias con la biblioteca Transformers, con el problema de una clasificación multietiqueta. Por su parte, el modelo `twitter-xlm-roberta-base-sentiment` fue finamente ajustado para adecuarse a la tarea de clasificación con 18 etiquetas, ya que su arquitectura base comprende la predicción de tres etiquetas, conforme al análisis de sentimientos: positivo, negativo y neutro.

Al igual que se hizo con los modelos de SVM y RFC, para el entrenamiento de los modelos basados en transformers, se dividió el conjunto de datos en una proporción 80-20 (entrenamiento-prueba) con un random state de 42. Además, para las predicciones de los modelos basados en arquitecturas transformers, se definió un umbral de 0.5 para la asignación de etiquetas. Esto significa que una etiqueta se asigna a una instancia de texto si la probabilidad predicha para esa etiqueta es igual o mayor a 0.5. Los resultados de los modelos basados en transformers, en relación con el conjunto de datos de prueba, se muestran en la Tabla 2.

En dicha tabla, se presenta la precisión, la exhaustividad (recall) y la puntuación F1 para cada modelo en torno a un entrenamiento basado en 5 y en 10 epochs. Al comparar estos resultados con los obtenidos por los modelos SVM y RFC, se observa un rendimiento generalmente inferior en términos de precisión, exhaustividad y puntuación F1, lo cual puede indicar que los algoritmos clásicos de aprendizaje automático que se emplearon aquí podrían resultar como opciones oportunas para una tarea de clasificación multietiqueta en un conjunto de datos pequeño, en contraposición a los modelos ligeros basados en transformers.

Dentro de los modelos transformers evaluados, el `distilbert` mostró el desempeño ligeramente más sólido tanto en el entrenamiento con 5 epochs como con 10 epochs. Dicho modelo muestra una precisión y una puntuación F1 más alta en contraste con los otros modelos evaluados.

Tabla 3. Resultados para los modelos SVM y `distilbert-base-multilingual-cased` donde se evalúa si al menos una etiqueta fue predicha correctamente por cada texto.

Modelo	Precisión	Exhaustividad	F1
SVM	**0.91**	**0.65**	**0.76**
`distilbert-base- multilingual-cased (10 epochs)`	0.88	**0.65**	0.73

No obstante, el modelo `twitter-xlm-roberta` también demuestra resultados competitivos, con un rendimiento muy próximo al del `distilbert` en todas las métricas de evaluación e, incluso, con valores superiores en la exhaustividad en un entrenamiento basado en 10 epochs. Por su parte, el modelo `distilroberta-base` presenta un rendimiento notablemente inferior en todas las métricas evaluadas.

Estos hallazgos podrían sugerir al `distilbert` como una opción óptima para esta tarea específica de clasificación de textos multietiqueta, aunque también debería considerarse la alternativa de ajustar los parámetros del modelo `twitter-xlm-roberta-base-sentiment`, el cual ha sido entrenado con textos provenientes de Twitter.

Dado el amplio conjunto de etiquetas y la baja representación de algunas de ellas, como pudo verse en las características del dataset en términos de un desbalance de clases, se decidió llevar a cabo un experimento adicional para evaluar el rendimiento de los modelos tras realizar ciertos ajustes en el conjunto de datos. En este experimento, se excluyeron las etiquetas "cultura pop" y "otro", debido a su escasa presencia y su menor relevancia para la CPC. Al mismo tiempo, se eliminaron los textos que quedaron sin etiquetas después de esta exclusión.

Con estos ajustes, el tamaño del dataset disminuyó a 3629 tuits. El objetivo principal de este experimento fue evaluar la capacidad de los modelos para acertar por lo menos una de las etiquetas verdaderas para cada texto, de la lista de 16 etiquetas restantes. Para ello, se seleccionaron los modelos SVM y distilbert, puesto que fueron los que presentaron un mejor rendimiento según las métricas de evaluación utilizadas. La Tabla 3 presenta los resultados de este experimento con base en las mismas métricas de evaluación utilizadas para la primera parte.

De acuerdo con estos resultados, el modelo SVM mantiene el rendimiento más solido para la tarea de clasificación multietiqueta en términos de precisión y puntaje F1. Este modelo alcanzó una precisión destacada de 0.91, es decir, la mayoría de las etiquetas predichas por el modelo fueron correctas en comparación con las etiquetas reales. Además, el F1 de 0.76 indica un buen equilibrio entre la precisión y la exhaustividad, aunque el valor de 0.65 en la exhaustividad se mantiene como el resultado más bajo del modelo.

Por otro lado, el modelo `distilbert-base- multilingual-cased`, entrenado durante 10 epochs, mostró una precisión ligeramente menor (0.88) en comparación con el SVM. Sin embargo, mantuvo la misma exhaustividad de 0.65, de modo que capturó correctamente el mismo porcentaje de etiquetas reales que el SVM. El F1-score para este modelo fue 0.73, bastante cercano al rendimiento del SVM. En general, los resultados de este experimento suponen una leve mejora en la evaluación de la clasificación multietiqueta para la tarea modificada, en función de las etiquetas más relevantes del corpus.

En todos los casos, tanto en la clasificación general como en el experimento con ajustes de clasificación, se observa que la exhaustividad se mantiene como el valor más bajo entre las métricas de evaluación, derivado de la falta de equilibrio entre las clases. Esto podría indicar que los modelos tienen dificultades para recuperar todos los casos positivos en relación con las etiquetas correctas.

Un nivel bajo de exhaustividad significa que el modelo está dejando pasar cierta cantidad de casos positivos no identificados. En el contexto de la clasificación multietiqueta, una baja exhaustividad puede surgir por varias razones, como la complejidad de las relaciones entre las etiquetas, el desbalance de las clases, la calidad del conjunto de datos utilizado para el entrenamiento o la configuración del modelo. Una forma de abordar este problema es analizar las métricas por cada etiqueta particular para identificar dónde está funcionando mejor el modelo.

Al hacer esto, sería posible identificar tanto las clases para las cuales el modelo tiene un buen desempeño como las que presentan desafíos. De cualquier forma, la intención de este trabajo, una vez evaluados los algoritmos, ha sido la de presentar un modelo general de clasificación entrenado con datos que representan la realidad de la CPC en el contexto mexicano. Si bien los valores de exhaustividad se mantienen como los más bajos, es posible realizar en el futuro un análisis pormenorizado de cada clase, con el objetivo de identificar las áreas de mejor rendimiento del modelo y, posteriormente, seguir una ruta de acción más especializada, con relación a los datos de entrenamiento.

Entre estas rutas, se sugiere la de buscar más ejemplos de las etiquetas menos representadas para conseguir un modelo con un rendimiento general sólido en la clasificación de las áreas temáticas de la CPC, o bien la de concentrarse sólo en las áreas que presentan un mejor rendimiento, con el fin de perfeccionar un modelo con un mejor desempeño general para menos clases.

5. Conclusiones y trabajo futuro

En este trabajo, se ha llevado a cabo una tarea de clasificación automática de textos en el contexto de la CPC en Twitter, sobre la base de un sistema multietiqueta de las áreas temáticas identificadas en el corpus de la investigación. Para lograr el objetivo de desarrollar y evaluar un modelo de clasificación automática, se adoptó un enfoque de aprendizaje automático supervisado, a partir de un corpus anotado, conformado por tuits de CPC publicados en México. En la anotación de este corpus se buscó identificar y clasificar las áreas temáticas presentes en los tuits del corpus, con base en un sistema multietiqueta, donde cada tuit podía podía ser etiquetado con una o más de las 18 categorías temáticas predefinidas.

Con base en este corpus anotado, se evaluaron varios modelos de aprendizaje automático, tanto algoritmos clásicos (SVM y RFC) como modelos basados en arquitecturas transformers (`distilbert-base-multilingual-cased`, `distilroberta-base` y `twitter-xlm-roberta-base-sentiment`). Los resultados de la evaluación, en función de las métricas de precisión, exhaustividad y puntuación F1, destacaron el rendimiento del modelo SVM para predecir correctamente las etiquetas asociadas con los textos del corpus.

Asimismo, los modelos ligeros basados en transformers también ofrecieron resultados competitivos, aunque con ciertas variaciones en las métricas de evaluación. De acuerdo con los resultados reportados en este trabajo, se ha resaltado la necesidad de mejorar la exhaustividad de los modelos en general, especialmente en un contexto de clasificación multietiqueta donde algunas clases tienen una representación limitada en el conjunto de datos.

En este sentido, uno de los desafíos consiste en identificar estrategias efectivas para capturar adecuadamente todas las clases durante el entrenamiento del modelo. Como se ha señalado, una de las limitantes con respecto al tipo de datos del corpus se refiere a la restricción impuesta para recopilar datos de Twitter/𝕏 mediante la API, como consecuencia de los cambios efectuados en la plataforma durante el último año. Para futuros trabajos, se propone investigar enfoques avanzados de procesamiento de texto y aprendizaje automático que puedan mejorar el rendimiento en tareas de clasificación multietiqueta en tuits de CPC.

Esto podría incluir el uso de modelos más grandes o las versiones base de los modelos abordados en este trabajo, así como el ajuste de los hiperparámetros para capturar relaciones complejas entre etiquetas y características del texto. Además, podrían realizarse análisis detallados por cada etiqueta del corpus para identificar las clases con mejores resultados desarrollar estrategias de mejora para las clases con valores de evaluación más bajos. En todo caso, se debe considerar que este estudio representa una de las primeras aproximaciones desde el PLN y el aprendizaje automático a la clasificación de tuits relacionados con la CPC en México.

Los resultados de este trabajo deben intepretarse como una exploración inicial para la clasificación multietiqueta de texto en tuits de CPC. A pesar de que los modelos basados en arquitecturas transformers constituyen un avance significativo en PLN, los algoritmos tradicionales como el SVM siguen mostrando un rendimiento competitivo en la tarea de clasificación multietiqueta automática en datasets pequeños con clases múltiples desbalanceadas. Al abordar la clasificación de tuits de CPC mediante distintos modelos de aprendizaje automático y de aprendizaje profundo, este trabajo establece importantes fundamentos para futuras investigaciones en esta área emergente.

Referencias

1. Aguilar-Tello, V., Angulo-Giraldo, M.: La divulgación científica en twitter durante la pandemia por la COVID 19. Revista Aportes de la Comunicación y la Cultura, vol. 32, no. 32 (2022)
2. Barajas-Galindo, D. E., Rodríguez-Carnero, M. G.: La divulgación científica en los tiempos de twitter. Endocrinología, Diabetes y Nutrición, vol. 67, no. 5, pp. 295–296 (2020) doi: 10.1016/j.endinu.2020.03.001
3. Barbieri, F., Espinosa-Anke, L., Camacho-Collados, J.: XLM-T: Multilingual language models in twitter for sentiment analysis and beyond. In: Proceedings of the 13th Language Resources and Evaluation Conference, European Language Resources Association, pp. 258–266 (2022)
4. Cheplygina, V., Hermans, F., Albers, C., Bielczyk, N., Smeets, I.: Ten simple rules for getting started on twitter as a scientist. PLoS Computational Biology, vol. 16, no. 2, pp. e1007513 (2020) doi: 10.1371/journal.pcbi.1007513

5. Claassen, G.: The viral spreading of pseudoscientific and quackery health messages on twitter - finding a communication vaccine. Current Allergy and Clinical Immunology, vol. 34, no. 1, pp. 18–22 (2021) doi: 10.10520/ejc-caci-v34-n1-a4
6. Daneshjou, R., Shmuylovich, L., Grada, A., Horsley, V.: Research techniques made simple: Scientific communication using twitter. Journal of Investigative Dermatology, vol. 141, no. 7, pp. 1615–1621 (2021) doi: 10.1016/j.jid.2021.03.026
7. Denia, E.: The impact of science communication on twitter: The case of Neil deGrasse Tyson. Comunicar, vol. 28, no. 65, pp. 21–30 (2020) doi: 10.3916/C65-2020-02
8. Denia, E.: Twitter como objeto de investigación en comunicación de la ciencia. Revista Mediterránea de Comunicación, vol. 12, no. 1, pp. 289 (2021) doi: 10.14198/MEDCOM000006
9. Déchène, M., Lesperance, K., Ziernwald, L., Holzberger, D.: From research to retweets—exploring the role of educational twitter (X) communities in promoting science communication and evidence-based teaching. Education Sciences, vol. 14, no. 2, pp. 196 (2024) doi: 10.3390/educsci14020196
10. Guenther, L., Wilhelm, C., Oschatz, C., Brück, J.: Science communication on twitter: Measuring indicators of engagement and their links to user interaction in communication scholars' tweet content. Public Understanding of Science, vol. 32, no. 7, pp. 860–869 (2023) doi: 10.1177/09636625231166552
11. Kowsari, K., Jafari-Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., Brown, D.: Text classification algorithms: A survey. Information, vol. 10, no. 4, pp. 150 (2019) doi: 10.3390/info10040150
12. Lewenstein, B. V.: Models of public communication of science and technology (2003) hdl.handle.net/1813/58743
13. Milbourne, S.: How to use twitter as a scientist (2022) www.letpub.com/How-to-Use-Twitter-as-a-Scientist
14. Peters, H. P., Dunwoody, S., Allgaier, J., Lo, Y. Y., Brossard, D.: Public communication of science 2.0. EMBO Reports, vol. 15, no. 7, pp. 749–753 (2014) doi: 10.15252/embr.201438979
15. Riekert, M., Riekert, M., Klein, A.: Simple baseline machine learning text classifiers for small datasets. SN Computer Science, vol. 2, no. 3, pp. 178 (2021) doi: 10.1007/s42979-021-00480-4
16. Sanchez-Mora, M. C.: Hacia una taxonomía de las actividades de comunicación pública de la ciencia. Journal of Science Communication, pp. 1–9 (2016) ru.ameyalli.dgdc.unam.mx/handle/123456789/73
17. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In: The 5th EMC2 - Energy Efficient Machine Learning and Cognitive Computing Co-located with the 33rd Conference on Neural Information Processing Systems, pp. 1–5 (2020) doi: 10.48550/arXiv.1910.01108
18. Statista Research Department: Countries with most x/twitter users 2023 (2023) www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
19. Sundström, G.: Science communication on twitter an analysis of vocabulary and content (2021) www.diva-portal.org/smash/get/diva2:1603996/FULLTEXT01.pdf
20. Zappavigna, M.: The discourse of Twitter and social media. Continuum International Publishing Group (2012) doi: 10.5040/9781472541642