MediFlow: Adaptación de un Modelo de Lenguaje para Triaje Automatizado a diferentes especialidades

Erik Perez Sarriegui
eperezs@vicomtech.org

2 de octubre de 2024

1. Introducción

El triaje es un componente esencial en la atención médica de emergencia, ya que facilita la asignación eficiente de recursos y asegura que los pacientes reciban atención adecuada y oportuna en función de la gravedad de su estado [1]. Tradicionalmente, este proceso ha dependido del juicio clínico de los profesionales de la salud, lo cual puede estar sujeto a variaciones subjetivas y limitaciones en la disponibilidad de recursos.

Con los avances recientes en el procesamiento de lenguaje natural (PLN) y el aprendizaje automático, han surgido nuevas oportunidades para mejorar la eficiencia del triaje hospitalario. En particular, los modelos de lenguaje preentrenados, como BERT (Bidirectional Encoder Representations from Transformers) [2], han demostrado ser altamente efectivos en una variedad de tareas de procesamiento del lenguaje natural (NLP).

Este trabajo introduce MediFlow, una adaptación de un modelo de lenguaje diseñada específicamente para automatizar parte del proceso de triaje en los servicios hospitalarios. El objetivo de MediFlow es identificar automáticamente el tipo de especialista al que debe ser referido un paciente, basándose únicamente en la descripción inicial de sus síntomas o motivo de consulta. De esta manera, se busca mejorar la eficiencia operativa del triaje, reducir el tiempo de espera y optimizar la asignación de recursos médicos.

2. Conjunto de Datos

2.1. MedDialog

El conjunto de datos utilizado en este trabajo es MedDialog [6], un recurso a gran escala diseñado para fomentar la investigación y el desarrollo de sistemas de diálogo médico, particularmente en el ámbito de la telemedicina. MedDialog

se divide en dos grandes partes: MedDialog-CN (en chino) y MedDialog-EN (en inglés), ofreciendo una amplia gama de datos que abarcan múltiples especialidades médicas y consultas entre pacientes y médicos.

MedDialog-CN contiene 3,4 millones de conversaciones que equivalen a 11,3 millones de enunciados y 660,2 millones de tokens en caracteres chinos. Estas interacciones cubren un período de diez años, desde 2010 hasta 2020, proporcionando una perspectiva longitudinal sobre las interacciones médicas en China. Las conversaciones abarcan 29 categorías generales y 172 especialidades, incluyendo áreas como medicina interna, pediatría y odontología. Cada consulta en MedDialog-CN incluye una descripción inicial de la condición médica e historial del paciente, seguida de un diálogo de varios turnos entre el paciente y el médico, y en algunos casos concluye con un diagnóstico y recomendaciones de tratamiento.

Por otro lado, MedDialog-EN contiene 0,26 millones de consultas en inglés, que corresponden a 0,51 millones de enunciados y 44,53 millones de tokens. Este subconjunto abarca consultas realizadas entre 2008 y 2020, distribuidas en 51 categorías de comunidades médicas y 96 especialidades, como andrología, cardiología, nefrología y farmacología. Al igual que en MedDialog-CN, cada consulta en inglés presenta una descripción de la condición médica del paciente seguida por un diálogo médico que puede concluir con un diagnóstico o sugerencia de tratamiento.

MedDialog destaca por su escala y cobertura, lo que lo convierte en uno de los conjuntos de datos más grandes en el campo de los diálogos médicos hasta la fecha de su publicación. Su amplia cobertura de especialidades y la diversidad demográfica de los pacientes que componen el conjunto de datos minimizan el sesgo poblacional y aumentan la aplicabilidad de los modelos entrenados en diversas poblaciones. Además, las consultas no solo incluyen el diálogo en sí, sino también información contextual como descripciones médicas detalladas, historiales de pacientes y, en algunos casos, resultados diagnósticos y sugerencias terapéuticas. Esta estructura detallada permite que MedDialog sea utilizado para una amplia variedad de tareas, que van desde la generación de respuestas en diálogos hasta el desarrollo de sistemas de diagnóstico automático.

Los datos de MedDialog consisten en pares de preguntas y respuestas entre pacientes y doctores. Cada conversación se inicia con una descripción de la situación que realiza el paciente, a la cual el médico responde de manera detallada. Además, cada interacción incluye una descripción específica del caso, proporcionando un contexto clínico que enriquece la comprensión del diálogo y facilita su uso en el entrenamiento de modelos de lenguaje orientados al ámbito médico.

2.2. Adaptación de MedDialog

MediFlow tiene como objetivo realizar un triaje automático que derive a los pacientes al especialista adecuado basándose exclusivamente en las descripciones de sus síntomas. No obstante, aunque el conjunto de datos MedDialog contiene numerosos ejemplos de diálogos entre pacientes y médicos, no incluye de manera

explícita la información sobre el especialista al que debe derivarse cada paciente. Para resolver esta limitación, implementamos un proceso exhaustivo de preprocesamiento que nos permitió extraer de forma eficiente esta información crítica.

2.2.1. Extracción de especialidades

Al analizar las respuestas de los médicos en MedDialog, observamos que en muchas ocasiones se recomendaba explícitamente a los pacientes visitar a un especialista, según su condición médica. Aprovechamos este patrón para automatizar la extracción de dicha información mediante el uso de expresiones regulares. Estas expresiones nos permitieron identificar de manera precisa y eficiente las menciones de especialistas en las respuestas de los médicos, estableciendo una correspondencia directa entre los síntomas descritos por los pacientes y los especialistas recomendados. Este enfoque nos permitió estructurar el conjunto de datos, obteniendo pares de descripciones y especialidades.

2.2.2. Validación de las especialidades extraídas

Para garantizar que las recomendaciones de especialidades fueran correctas y no el resultado de interpretaciones erróneas del contexto, aplicamos un modelo de lenguaje que validó si los médicos efectivamente estaban recomendando a los pacientes visitar a dichos especialistas. Esta validación fue fundamental para evitar errores contextuales, como recomendaciones negativas o ambiguas (por ejemplo, "No te recomiendo ir a un ginecólogo, sino a un endocrinólogo"). Este paso de filtrado mejoró significativamente la calidad del conjunto de datos, asegurando que las especialidades asignadas correspondieran fielmente a las recomendaciones reales dadas por los médicos.

2.2.3. Conjunto de datos final

Aunque el conjunto de datos extraído incluía una amplia variedad de especialidades médicas, para los fines de este trabajo nos centramos en cuatro áreas clave: Traumatología, Neumología, Salud Mental y Cardiología. Con el fin de mantener un equilibrio en el número de ejemplos por especialidad, aplicamos un proceso de ajuste que consistió en reducir el número de registros en aquellas especialidades con mayor cantidad de datos, para evitar desequilibrios en el entrenamiento del modelo. Como resultado, obtuvimos un conjunto de datos final que consta de 3.820 frases de pacientes, el las que estos describen sus síntomas, con las correspondientes especialidades asociadas. Este conjunto equilibrado es el que se utilizó para entrenar y evaluar nuestro modelo de triaje automático.

3. Entrenamiento

Con el fin de maximizar el rendimiento en la tarea de triaje automático, se entrenaron y evaluaron diversos modelos de lenguaje preentrenados. En este estudio, se seleccionaron tanto las versiones base como grandes de BERT [2],

RoBERTa [3] y XLNet [5]. Estos modelos fueron escogidos debido a su eficacia demostrada en múltiples tareas de procesamiento del lenguaje natural.

En este trabajo se optó por utilizar modelos de lenguaje general en vez de aquellos especializados en el ámbito clínico. Esto se debe a que el lenguaje utilizado por los pacientes al describir sus síntomas es informal y variado, no sigue un formato clínico ni técnico. De este modo, modelos entrenados en lenguaje clínico podrían no ajustarse adecuadamente a la naturaleza de las descripciones proporcionadas por los pacientes. Por tanto, la decisión de usar modelos de propósito general se basa en su capacidad para manejar este tipo de lenguaje.

Todos los modelos se entrenaron utilizando los mismos hiperparámetros. Aunque la optimización de los hiperparámetros podría haber mejorado el rendimiento, se decidió utilizar una configuración estándar para centrarse en identificar el modelo más prometedor en términos generales. Los hiperparámetros empleados se pueden ver en el Cuadro 1.

Cuadro 1: Hiperparámetros empleados

Hiperparámetro	***Valor***
Learning Rate	$2e-5$
Batch Size	4
Training Epochs	3
Weight Decay	0,015
Optimizer	$AdamW$
Train & Test Split	$80\,\% - 20\,\%$

Esta configuración permitió entrenar los modelos de manera eficiente, equilibrando la velocidad de entrenamiento con la capacidad de generalización. Aunque un mayor ajuste de estos parámetros podría conducir a mejoras adicionales.

Para el entrenamiento y evaluación de los modelos, se utilizó el framework *transformers* de Hugging Face [4], empleando específicamente la clase *Trainer*. Esta herramienta facilitó el proceso al integrar de manera eficiente la optimización, evaluación y registro de métricas, lo que permitió gestionar los modelos y sus resultados de forma estandarizada. Esto aceleró el ciclo de experimentación y simplificó la comparación entre ellos.

Los entrenamientos se realizaron sobre una Nvidia P100, y durante este, se realizó una evaluación periódica del desempeño de los modelos para evitar el sobreajuste. Finalmente, se utilizó un conjunto de prueba separado para evaluar la capacidad de generalización de los modelos, lo que permitió identificar cuál de ellos alcanzaba la mayor precisión en la tarea de triaje basada en la descripción de síntomas de los pacientes.

4. Resultados

En esta sección se presentan los resultados obtenidos al evaluar los diferentes modelos en la tarea de triaje automático. Las métricas empleadas para la

evaluación fueron Accuracy, F1-Score y Precision, cada una de ellas aportando una visión complementaria sobre el rendimiento de los modelos.

Accuracy (precisión global) mide la proporción de predicciones correctas sobre el total de ejemplos evaluados. Aunque esta métrica es fácil de interpretar, puede ser insuficiente en conjuntos de datos desbalanceados, ya que no refleja correctamente los errores en las clases minoritarias.

El F1-Score combina la precisión y el recall (sensibilidad) en una sola métrica, calculando la media armónica entre ambos. El F1-Score es especialmente útil cuando las clases están desbalanceadas (aunque no sea el caso), ya que equilibra los falsos positivos y falsos negativos, ofreciendo una visión más robusta del rendimiento en tales casos.

Por último, la Precision (precisión por clase) mide la proporción de ejemplos clasificados como positivos que efectivamente son positivos. Un valor alto de precisión indica que el modelo comete pocos errores al predecir la clase positiva, lo que es importante en escenarios donde los falsos positivos son costosos.

En el Cuadro 2 se muestran el rendimiento de los modelos evaluados.

Cuadro 2: Rendimiento de los diferentes modelos fundacionales.

Modelo	**Accuracy**	**F1-Score**	**Precision**
BERT (base)	88.2 %	0.88	0.88
BERT (large)	88.2 %	0.88	0.88
RoBERTa (base)	89 %	0.89	0.89
RoBERTa (large)	86.8 %	0.86	0.87
XLNet (base)	87.8 %	0.88	0.89
XLNet (large)	**89.3 %**	**0.89**	**0.90**

Los resultados obtenidos revelan una tendencia general en la que las versiones grandes de los modelos tienden a desempeñar de manera inferior en comparación con sus contrapartes pequeñas, lo que sugiere la posibilidad de un overfitting en estas versiones grandes. En cuanto al rendimiento, la versión grande de XLNet es la que destaca, aunque no hay diferencias significativas con otros modelos como RoBERTa base.

Este rendimiento cercano al 90 % sugiere que el conjunto de datos empleado para el entrenamiento de este modelo está adecuadamente adaptado del conjunto de datos general y que la metodología utilizada para la creación del dataset está bien formulada. Esta metodología de minería de datos a partir de un dataset no estructurado puede tener un gran valor para entrenar otros modelos a partir de un conjunto de datos que no contenga de manera explícita los datos que se requierenmismo conjunto, ya que permite adaptar el tipo de dato necesario sin necesidad de generarlo de manera sintética.

Referencias

[1] C Brandao-de-Resende et al. "A machine learning system to optimise triage in an adult ophthalmic emergency department: a model development and validation study". En: *EClinicalMedicine* 66 (2023), pág. 100847.

[2] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". En: *arXiv* (2019). arXiv: 1810.04805.

[3] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL]. URL: https://arxiv.org/abs/1907.11692.

[4] Thomas Wolf et al. "Transformers: State-of-the-Art Natural Language Processing". En: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, oct. de 2020, págs. 38-45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[5] Zhilin Yang et al. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. 2020. arXiv: 1906.08237 [cs.CL]. URL: https://arxiv.org/abs/1906.08237.

[6] Guangtao Zeng et al. "MedDialog: Large-scale medical dialogue datasets". En: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, págs. 9241-9250.