Evaluación automática de preguntas abiertas en el aula
Automatic evaluation of open-ended questions in the classroom

Raúl Cabido, David Concha, Miguel Ángel Rodríguez, Soto Montalvo
raul.cabido@urjc.es, david.concha@urjc.es, miguel.rodriguez@urjc.es, soto.montalvo@urjc.es

[1]Departamento de Informática y Estadística
Universidad Rey Juan Carlos
Madrid, España

Resumen- El *feedback* es un elemento educativo básico en el proceso de enseñanza y aprendizaje. Dentro de esta retroalimentación, el *feedback* inmediato que se da en el aula hace que impacte de forma más directa en el estudiante, intensificando su aprendizaje. Por ello, las soluciones que permiten la interacción entre docentes y estudiantes se utilizan cada vez con más frecuencia en el aula. Por otra parte, el avance de técnicas de Procesamiento del Lenguaje Natural en los últimos años es imparable, teniendo su máxima expresión en los modelos de lenguaje basados en arquitectura de Transformers. En este trabajo se lleva a cabo un estudio sobre cómo integrar modelos de lenguaje en el aula para proporcionar *feedback* en tiempo real a los estudiantes. La propuesta consiste en evaluar de forma automática respuestas a preguntas abiertas en el aula utilizando inteligencia artificial. Los resultados obtenidos son prometedores, logrando los mejores resultados en la evaluación cuando se proporciona la máxima información posible al modelo: la pregunta, la respuesta y la rúbrica de evaluación.

Palabras clave: preguntas abiertas, evaluación automática, modelos de lenguaje, refuerzo del aprendizaje

Abstract- Feedback is a fundamental educational element in teaching and learning processes. Immediate feedback in the classroom has a more direct impact on students, intensifying their learning. Therefore, solutions that allow interaction between teachers and students are increasingly used in the classroom. On the other hand, the advance of Natural Language Processing techniques in recent years is unstoppable, having its maximum expression in language models based on Transformers architecture. In this work, a study is carried out on how to integrate language models in the classroom to provide real-time feedback to students. The proposal consists of automatically evaluating answers to open-ended questions in the classroom using artificial intelligence. The results obtained are promising, achieving the best results in the evaluation when as much information as possible is provided to the model: the question, the answer, and the evaluation rubric.

Keywords: open-ended questions, automatic evaluation, language models, learning reinforcement

1. Introducción

En el aula es cada vez más frecuente el uso de herramientas que permiten la participación de los estudiantes en clase de forma interactiva, y que proporcionan *feedback* en tiempo real. Estas herramientas permiten al docente hacer de las clases un entorno más interactivo, motivando a los estudiantes a sentirse más libres para responder de forma anónima, proporcionando una fotografía más real de sus conocimientos (Pichardo et al., 2021).

Diferentes estudios se han llevado a cabo para ver el efecto de usar este tipo de herramientas como parte del proceso de aprendizaje. Licorish et al. (2018) y Wang y Thair (2020) analizaron el uso de Kahoot!, coincidiendo en el efecto positivo que tiene en el aprendizaje en comparación con el aprendizaje tradicional. Catalina-García y García Galera (2022) utilizaron Wooclap en su asignatura e influyó positivamente en la capacidad de aprendizaje. Gokbulut (2020) analizó el efecto de utilizar Mentimeter y Kahoot! en la enseñanza online, su uso fue positivo y los estudiantes disfrutaron aprendiendo y participaron más en las actividades del aula.

Las preguntas que se suelen plantear con estas herramientas son de tipo test. Son preguntas acotadas y dirigidas que, aunque útiles para tener un *feedback* de los conocimientos de los estudiantes, no permiten al profesorado tener una percepción real de si los conceptos explicados se han asimilado realmente. Estos problemas se evitan formulando preguntas abiertas en el aula, donde cada estudiante puede expresar con sus propias palabras los conocimientos que ha ido adquiriendo. Esto sí permite a los docentes conocer el grado de asimilación de los conceptos explicados. Evaluar a los estudiantes de una asignatura mediante pruebas con preguntas de respuesta corta es beneficioso para su aprendizaje (Chi et al., 1994). Sin embargo, es frecuente que los docentes no las utilicen porque su corrección requiere más tiempo que la corrección de las preguntas de respuesta múltiple. Además, una evaluación en tiempo real de respuestas a preguntas abiertas en el aula permite al profesor tener una percepción más certera de la asimilación de los conceptos. Se pueden ver patrones de respuesta erróneos y comunes en el aula, facilitando su corrección en tiempo real. Así, el profesor puede centrarse en los conceptos no asimilados por el grupo para explicarlos de nuevo, mejorando la asimilación de los conceptos y el aprendizaje de los estudiantes.

Una evaluación en tiempo real sólo es posible si se puede hacer de forma automática. Existe una creciente línea de investigación denominada ASAG (*Automatic Short Answer Grading*) que evalúa, de forma automática, las respuestas de los estudiantes a preguntas abiertas. Los avances en Procesamiento del Lenguaje Natural y el Aprendizaje Automático de los últimos años han tenido un impacto positivo en este campo (Weegar y Idestam-Almquist, 2023; Ahmed, Joorabchi y Hayes, 2022). La mayoría de las propuestas del estado del arte utilizan conjuntos de datos escritos en inglés orientados a la

DOI: 10.26754/CINAIC.2023.0127

tarea ASAG. Estos conjuntos de datos se utilizan para entrenar arquitecturas neuronales, haciendo muy dependientes las propuestas de ese entrenamiento.

En este trabajo se lleva a cabo un estudio para ver la viabilidad de utilizar modelos de lenguaje (ML en adelante) en español para la evaluación en tiempo real de respuestas cortas a preguntas. En concreto, este estudio no depende de un entrenamiento previo con conjuntos de datos específicos, sino que se utiliza un ML genérico en español para evaluar las respuestas de los estudiantes. Se utiliza la Inteligencia Artificial como asistente del profesor en clase, evaluando respuestas a preguntas abiertas para descubrir qué conceptos no se están entendiendo bien y poder así reforzarlos.

Los ML han evolucionado significativamente en los últimos años. Estos modelos son redes neuronales profundas capaces de generar texto en lenguaje natural, entre otras cosas. Uno de los modelos más conocidos es GPT, actualmente aplicado en varios estudios dentro del ámbito académico. Moore et al. (2022) evaluaron el uso de GPT-3 para medir la calidad de las preguntas de respuesta abierta generadas por los estudiantes y para clasificarlas según la Taxonomía de Bloom. Los resultados que obtuvieron no fueron muy alentadores, ya que el modelo sobreestimaba con frecuencia la calidad de las preguntas y las clasificaba en un nivel incorrecto de la Taxonomía de Bloom. En este trabajo el estudio se ha realizado con GPT-4.

2. Contexto y Descripción

Este estudio se desarrolla en un contexto universitario en asignaturas relacionadas con programación, aunque sería extrapolable a otros entornos educativos. Algunos estudiantes piensan que para programar bien no hay que estudiar ni conocer los conceptos teóricos que soportan el paradigma de programación concreto. Al plantear preguntas sobre conceptos clave, que pueden requerir tanto respuestas en lenguaje natural, como respuestas de código fuente, el abanico de posibilidades para el docente es mayor. La evaluación en tiempo real de las respuestas anónimas en el aula permite reforzar el aprendizaje incidiendo en todo aquello que no se está entendiendo, a la vez que los estudiantes pueden ver su propia realidad en la asignatura. Se persigue mejorar el aprendizaje, capturar la atención de los estudiantes y hacer que sean más participativos.

A. Conjunto de datos

El marco experimental de este trabajo se enmarca en la asignatura "Procesadores Gráficos Avanzados" del Grado en Diseño y Desarrollo de Videojuegos de la Universidad Rey Juan Carlos. Esta asignatura tiene una carga significativa de programación y una carga relativamente alta de conceptos técnicos, lo que habilita la posibilidad de realizar preguntas abiertas de diferente naturaleza.

La evaluación de esta asignatura se realiza desde la plataforma Moodle de la universidad. Para conformar el conjunto de datos, se ha desarrollado una aplicación que permite recuperar de esta plataforma toda la información necesaria para el estudio. Se han recuperado los resultados obtenidos en tres preguntas abiertas realizadas durante la prueba escrita de la convocatoria ordinaria del curso 22/23. Se han recopilado 111 respuestas anonimizadas de 37 estudiantes. Para cada pregunta se tiene la puntuación máxima, la respuesta que el docente establece como correcta y la rúbrica utilizada para la evaluación. Las tres preguntas seleccionadas presentan características muy diferentes:

- Pregunta 1. Pregunta abierta sobre un concepto teórico de la asignatura. Se espera una respuesta cuyo texto defina correctamente el concepto por el que se pregunta.
- Pregunta 2. Pregunta abierta donde se facilita un código de programación incorrecto y se pide al estudiante que identifique y corrija los fallos en el mismo. Se espera una respuesta con el código fuente corregido y el texto con la explicación.
- Pregunta 3. Pregunta abierta donde se plantea un problema de programación y el estudiante debe resolverlo. Se espera una respuesta en forma de código.

A. Modelo de lenguaje

Se ha utilizado GPT-4 como ML en la experimentación. Para poder hacer uso del modelo y evaluar automáticamente las 111 respuestas, se ha desarrollado una aplicación que permite la comunicación con el chatbot de OpenAI chatGPT.

El éxito de los chatbots a la hora de resolver tareas está condicionado a la forma de generar las preguntas. Así, en función de las tareas que se quieran desarrollar existen órdenes (comandos o *prompts*) que permiten obtener el resultado esperado con mayor o menor acierto. En este trabajo se proponen y evalúan diferentes *prompts* con el propósito de encontrar aquellos que permitan obtener mejores resultados en la evaluación de respuestas a preguntas abiertas. Se han tenido en cuenta diferentes aspectos a la hora de construir los *prompts*.

Rango de la nota. La calificación obtenida por chatGPT al actuar como evaluador se suele encontrar en una escala [0, 10]. Con el objetivo de que la evaluación del docente y la del chatbot se encuentren en la misma escala, es importante indicar en la pregunta el rango específico en el que debe encontrarse la nota final. Como ejemplo, algunas de las preguntas evaluadas en nuestro conjunto de datos sólo pueden tomar valores concretos (0, 0.25, 0.5, 0.75, 1.0). Esta información debe ser proporcionada en el comando.

Respuesta correcta. Las respuestas a una misma pregunta pueden variar según el docente y el enfoque seguido en la asignatura.

Rúbrica. La respuesta del docente a una pregunta suele ser muy específica y detallada. Sin embargo, al corregir las respuestas de los alumnos los docentes suelen ser más flexibles. Es común que se permitan expresiones no siempre correctas por parte de los estudiantes y se acepten respuestas que, aunque no sean idénticas a la respuesta esperada, demuestren un entendimiento adecuado del concepto por el que se pregunta. Por este motivo, se decide añadir al ML información de la rúbrica de calificación, para que tenga una guía sobre cómo se evalúan las respuestas y qué aspectos se tendrán en cuenta al asignar una calificación.

Se han creado, por tanto, tres plantillas de *prompts* para la evaluación de preguntas abiertas:

- [Prompt1] Actúa como si fueras un profesor de informática. Dada la pregunta: {PREGUNTA} evalúa con {RANGO_NOTAS} la siguiente respuesta: {RESPUESTA ESTUDIANTE}
- [Prompt2] Actúa como si fueras un profesor de informática. Dada la pregunta: {PREGUNTA} y considerando esta respuesta como correcta: {RESPUESTA CORRECTA} evalúa con {RANGO_NOTAS} la siguiente respuesta: {RESPUESTA ESTUDIANTE}
- [Prompt3] Actúa como si fueras un profesor de informática. Dada la pregunta: {PREGUNTA} y considerando esta respuesta como correcta: {RESPUESTA CORRECTA}. Teniendo en cuenta que se valora lo siguiente: {RUBRICA}, evalúa con {RANGO_NOTAS} la siguiente respuesta: {RESPUESTA ESTUDIANTE}

Los *prompts* se han definido de forma incremental para observar el impacto de las distintas variables en la puntuación final.

3. RESULTADOS

Para medir la precisión del ML en la evaluación, se ha recuperado la nota otorgada por los profesores a las 111 respuestas del conjunto de datos. Utilizando los *prompts* planteados, se ha medido el error cometido por el ML al evaluar cada respuesta. La Tabla 1 muestra el error cometido expresado en forma de media y desviación típica para cada pregunta.

Tabla 1. *Error medio y desviación cometido por el modelo de lenguaje al puntuar respuestas*

	prompt	media	desv.
Pregunta1	1	1.35	1.51
	2	1.28	1.62
	3	1.28	1.73
Pregunta2	1	2.09	1.71
	2	2.09	1.91
	3	0.81	1.18
Pregunta3	1	2.19	1.48
	2	2.53	1.87
	3	1.55	1.26

Los mejores resultados se han obtenido con el Prompt3. Aunque cada pregunta presentaba una puntuación diferente, para una mejor comprensión de los resultados las notas se han normalizado en un rango [0-10]. Los resultados obtenidos revelan la importancia de añadir la rúbrica utilizada en el *prompt* a la hora de conseguir evaluaciones automáticas más precisas. Por otro lado, añadir la respuesta correcta al *prompt* no supone un cambio significativo en la precisión de la evaluación.

En la Figura 1 se muestra la distribución de los errores cometidos al evaluar las tres preguntas utilizando el Prompt3. Para las tres preguntas más del 62% de las respuestas evaluadas automáticamente presentan un error menor o igual a un punto en una escala [0,10]. Destacan los resultados obtenidos para la pregunta 1, donde el 67% de las respuestas evaluadas automáticamente han obtenido la misma puntuación que fue asignada por el profesor. Destacan también los errores de puntuación obtenidos para las preguntas 2 y 3 donde, no mayoritariamente, pero sí en un número reducido de respuestas se han obtenido errores de puntuación de tres, cuatro y cinco puntos. Para estos casos, se ha revisado la justificación facilitada por el ML a cada una de las puntuaciones. La discrepancia en la nota final se debe principalmente a dos factores. Por un lado, las preguntas 2 y 3 guardan relación con desarrollo o escritura de código y aquí el ML se muestra poco permisivo frente a fallos sintácticos en el código. Sin embargo, los docentes a la hora de evaluar no penalizaban este tipo de fallos al no disponer los alumnos durante el tiempo de examen de un entorno de desarrollo para programar. Por otro lado, hay casos donde la calificación del ML parece más adecuada que la propia facilitada por el docente, ayudando a identificar respuestas que habían recibido una puntuación más elevada de lo que correspondía. De este estudio, también se deriva la posibilidad del uso de ML como sistemas de apoyo durante el proceso de evaluación.

En la Figura 2 se muestran las funciones de densidad de probabilidad de las puntuaciones asignadas por el docente y las asignadas por el ML. El modelo utilizado es menos propenso a asignar puntuaciones en los extremos del rango [0,10], siguiendo sus puntuaciones una distribución normal. Determinar si esta normalidad es más adecuada que la propia distribución del docente necesitará de una extensión del estudio realizado para profundizar en la forma de evaluar del docente, así como una ampliación del conjunto de datos empleado.

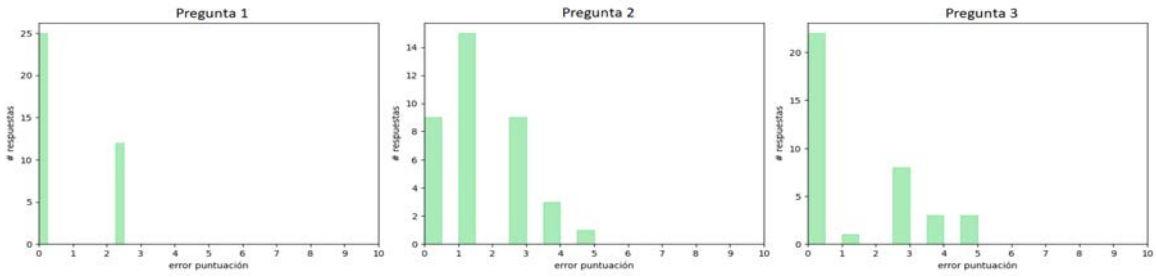

Figura 1. *Error cometido en puntuación en cada pregunta*

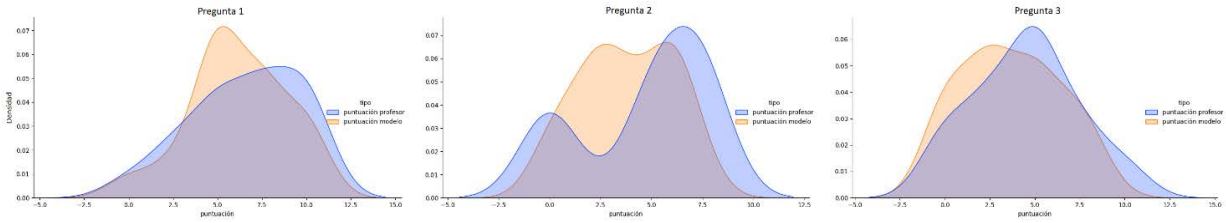

Figura 2. *Función de densidad de probabilidad de las puntuaciones asignadas por chatGPT (naranja) y docente (azul) para las distintas preguntas evaluadas*

El estudio realizado presenta ciertas limitaciones. El conjunto de datos empleado se concentra en una asignatura, sería necesario extenderlo a más asignaturas y distintos tipos de preguntas para garantizar que el modelo es aplicable en otros contextos docentes. Además, la reproducibilidad de los resultados está condicionada por el ritmo de actualización del chatbot de OpenAI (chatGPT). Es de esperar que nuevas actualizaciones traigan consigo mejores resultados, pero el hecho de usar esta tecnología como servicio compromete la reproducibilidad de los resultados experimentales tras una actualización del servicio.

4. Conclusiones

En este trabajo se ha llevado a cabo un estudio para ver la viabilidad de utilizar modelos de lenguaje para evaluar en tiempo real respuestas a preguntas abiertas en el aula. Se ha tratado de encontrar la forma más adecuada de convertir los nuevos modelos de lenguaje en evaluadores expertos. Para ello, se han explorado diferentes *prompts* combinados con distintas entradas a un modelo: pregunta, respuesta correcta y rúbrica.

Los resultados obtenidos son prometedores e indican que puede ser viable utilizar un ML como asistente para evaluar las respuestas a preguntas abiertas. Además, se pueden emplear no sólo como asistente en el aula en tiempo real, sino también como asistente para las correcciones que debe hacer el docente en sus pruebas de evaluación continua.

La propuesta es aplicable a otros contextos, no sólo en el ámbito universitario para un gran número de asignaturas, independientemente de si son más o menos teóricas, sino también para otros niveles educativos u otros ámbitos relacionados con el aprendizaje. A la vista de los resultados obtenidos en este primer estudio de viabilidad, la mayor eficacia en la evaluación se consigue proporcionando al modelo la pregunta, una respuesta válida y la rúbrica de evaluación. La aplicación de la investigación propuesta en este trabajo no exige un fuerte conocimiento tecnológico por parte del docente. Se ha utilizado chatGPT para simplificar la adopción de modelos de lenguaje en el aula.

Como trabajo futuro se propone introducir mejoras en dos aspectos: (1) búsqueda de nuevos *prompts* que mejoren la precisión del modelo de lenguaje a la hora de evaluar preguntas abiertas; (2) generación de un conjunto de datos diverso que contemple diferentes asignaturas y distintos tipos de preguntas abiertas, sirviendo de marco comparativo para los modelos de lenguaje presentes y futuros.

Agradecimientos

Este trabajo ha sido soportado por el proyecto PLN4PA, concedido en la Convocatoria de Proyectos de Innovación Educativa de la Universidad Rey Juan Carlos 2022-23.

Referencias

Ahmed, A., Joorabchi, A., Hayes, M.J. (2022). *On Deep Learning Approaches to Automated Assessment: Strategies.* Proceedings of the 14th International Conference on Computer Supported Education, vol. 2.

Catalina-García, B. y García Galera, M.C. (2022*). Innovation and hi-tech tools in journalism education. The Wooclap case*. Doxa Comunicación, 34, pp. 19-32.

Chi, M.T., De Leeuw, N., Chiu, M.H., y LaVancher, C. (1994). *Eliciting self-explanations improves understanding*. Cognitive Science 18(3), 439–477.

Gokbulut, B. (2020). *The effect of Mentimeter and Kahoot applications on university students' e-learning*. World Journal on Educational Technology Current Issues, 12.

Licorish, S.A., Owen, H.E., Daniel, B., y George, J.L. (2018). *Students' perception of Kahoot!'s influence on teaching and learning*. Research and Practice in Technology Enhanced Learning 13, 9.

Moore, S., Nguyen, H.A., Bier, N., Domadia, T., y Stamper, J. (2022). *Assessing the Quality of Student-Generated Short Answer Questions Using GPT-3*. EC-TEL 2022.

Pichardo, J.I., López-Medina, E.F., Mancha-Cáceres, O., González-Enríquez, I., Hernández-Melián, A., Blázquez-Rodríguez, M., Jiménez, V., Logares, M., Carabantes-Alarcon, D., Ramos-Toro, M., Isorna, E., Cornejo-Valle, M., y Borrás-Gené, O. (2021). *Students and Teachers Using Mentimeter: Technological Innovation to Face the Challenges of the COVID-19 Pandemic and Post-Pandemic in Higher Education*. Education Sciences, 11(11): 667.

Wang, A.I., Tahir, R. (2020). The effect of using Kahoot! for learning-A literature review. Computers & Education, 149.

Weegar, R., Idestam-Almquist, P. (2023). *Reducing Workload in Short Answer Grading Using Machine Learning*. International Journal of Artificial Intelligence in Education