

Aplicación de la evaluación automática de respuestas y el análisis de sentimientos en entornos de aprendizaje con enfoque constructivista.

David Aguirre Franco
Doctorado en tecnologías de
la información y la comunicación.
Universidad de Granada.
Granada, España
Email: daguirref@correo.ugr.es

Director: Juan Luis Castro Peña
Departamento de Ciencias de la
Computación e Inteligencia Artificial.
Universidad de Granada.
Granada, España
Email: jcastro@ugr.es

Resumen—Dentro de la línea de investigación: Aplicaciones de las TIC, se plantea como propuesta de investigación predoctoral una metodología que permita aplicar técnicas de procesamiento del lenguaje natural (NLP) en contextos educativos. En concreto, se propone la integración de técnicas para evaluación automática de respuestas a preguntas abiertas y de análisis de sentimientos como soporte para la retroalimentación a estudiantes en el marco del proceso de enseñanza - aprendizaje. Se espera que dentro de los aportes se encuentren: *un método para evaluación de bondad de la respuestas basados en embeddings de texto unificado al análisis de sentimientos para dar soporte a la retroalimentación automática de respuestas; la integración del modelo en un sistema de gestión de aprendizaje (LMS); el dimensionamiento de los recursos necesarios para la implementación del modelo en un contexto educativo real y la integración de la metodología en un marco pedagógico de enfoque constructivista.* Este plan de investigación fue aprobado el doce de enero de 2024 por la Escuela Internacional de Postgrado de la Universidad de Granada. Alineado con el plan de investigación, en este documento se hace énfasis en las actividades asociadas al primer objetivo específico de la investigación, que se corresponde con el primer aporte listado. En la sección IV se amplían algunos detalles al respecto.

I. INTRODUCCIÓN

Durante la última década las tecnologías de inteligencia artificial (IA) han ganado participación en múltiples aspectos de la vida diaria y sus aplicaciones son casi ubicuas. El ámbito educativo no es la excepción. Los principales temas reportados en la literatura de este ámbito pueden agruparse en tres ejes temáticos: educación, enseñanza y aprendizaje [1]. En el primer eje, se reportan temáticas asociadas con educación a distancia, online y personalizada. En el segundo eje, temáticas asociadas a sistemas tutores inteligentes y agentes para la enseñanza. Finalmente, en el tercer eje, temas asociados a procesos, entornos y desempeño en el aprendizaje.

Zawacki-Richter y colaboradores [2] exploran las aplicaciones de IA en educación (IAEd) a través de una revisión sistemática de 146 documentos elegidos entre un corpus de 2656 publicaciones, dentro de sus principales hallazgos se encuentra *la carencia de teorías del aprendizaje en las investigaciones en este ámbito.* De otra parte, Paterson y colaboradores [3]

enfatan el rol central de la retroalimentación en los procesos de aprendizaje. En este sentido, señalan que la retroalimentación no sólo es un parámetro usado para medir la calidad de la enseñanza, sino que también soporta las mediaciones que contribuyen al desarrollo de los estudiantes, en línea con los postulados de Wood, Bruner y Vygotsky [4]. A través de una revisión sistemática los autores consolidan evidencia que les permite concluir que los estudiantes valoran y necesitan una retroalimentación multimodal y personalizada en la que haya un balance entre comentarios positivos y oportunidades de mejora, *que sea directa, oportuna y clara*, de manera que puedan incorporarla en su proceso de aprendizaje. Asimismo, es importante resaltar que la evidencia [5] señala que la retroalimentación formativa promueve en los estudiantes una percepción de mayor utilidad sobre la evaluación, mejorando su motivación, auto regulación y desarrollo cognitivo y meta-cognitivo.

I-A. Hipótesis y justificación

La evaluación de la calidad de las respuestas y la detección de sentimientos, soportadas en la implementación exitosa de modelos de NLP en un entorno de aprendizaje, tienen el potencial de mejorar significativamente la retroalimentación, interacción y personalización de la experiencia del estudiante, enriqueciendo así el proceso de aprendizaje.

La implementación de estos modelos involucra tareas de procesamiento de lenguaje natural, como la evaluación de respuestas y la identificación de sentimientos en texto. Para enriquecer la experiencia pedagógica, es fundamental dotar a los modelos de capacidad explicativa y mejora continua, basada en la retroalimentación de los usuarios.

Como elemento de soporte y justificación, en la revisión de la literatura se documentó la disponibilidad de medios técnicos asociados a NLP e IA para la evaluación de respuestas, el análisis de sentimientos y la mejora continua de los modelos basados en la interacción con los usuarios. Asimismo, se documentó una oportunidad de investigación interdisciplinar que involucre aspectos tecnológicos y pedagógicos. También

se documentó el rol clave de la retroalimentación en los procesos educativos. A continuación se citan algunos autores relevantes, de manera sucinta.

De acuerdo con Cetina y colaboradores [6] el termino aprendizaje por refuerzo (RL) es usado para designar a una familia de algoritmos empleada para resolver problemas en los que se requiere una secuencia repetitiva de decisiones. El RL se diferencia de los métodos de aprendizaje supervisado y no supervisado en que su proceso de aprendizaje surge de la interacción del algoritmo (agente) con el entorno. A través de esta interacción el algoritmo descubre la secuencia de decisiones más adecuada para cumplir con un objetivo particular, asociado a la optimización de una función de costo que usualmente se denomina recompensa. El aprendizaje por refuerzo (RL) ha sido ampliamente empleado en diferentes tareas asociadas al NLP [6]. Dentro de estas tareas, se presentan sus aplicaciones en análisis de sintaxis, comprensión de lenguaje natural y sistemas conversacionales. Resulta de especial interés el uso de RL para la mejora en la extracción de información mediante la incorporación de datos externos con la que se logra reducir la ambigüedad en la interpretación de textos, diseñando la recompensa para maximizar la precisión de la extracción.

La evolución de las redes neuronales artificiales (ANN) hacia las redes de aprendizaje profundo (DL) ha llevado al desarrollo de los grandes modelos de lenguaje (LLM) basados en Transformers [7]. Esta tecnología soporta el estado del arte en el procesamiento de lenguaje natural en la actualidad. Los modelos LLM se apoyan en representaciones en forma de vectores densos (*embeddings*) de palabras y frases. Estos *embeddings* desempeñan un papel esencial al permitir la representación y manipulación de conceptos en un espacio de alta dimensionalidad. Esto facilita la captura de significados y relaciones entre palabras en un espacio de representación operable para los modelos.

Lopez-Ubeda [8] resalta la importancia de adaptar el contexto a campos de conocimiento específicos para habilitar la utilización efectiva de modelos pre-entrenados en la identificación de entidades nombradas, del contexto biomédico, en español, lo que permite abordar tareas como la respuesta de preguntas, la clasificación de textos y la recuperación de información. Este antecedente aporta ideas valiosas que se alinean con esta investigación. Una revisión del documento extendido de López-Ubeda [9] revela que su modelo *emplea embeddings concatenados de modelos estáticos y contextuales* como entrada a una red neuronal Long-Short Term Memory (LSTM), culminando con una capa basada en Conditional Random Fields (CRF). Uno de los puntos destacados de esta propuesta es la adaptación (*fine tuning*) de modelos pre-entrenados al utilizar un corpus específico, lo que resulta en el doble de eficacia en la identificación de entidades en comparación con el modelo base, entrenado con un corpus general.

Sung y colaboradores [10], presentan la estrategia de ajuste fino (*fine-tuning*) de modelos pre-entrenados como una alternativa para su utilización en la evaluación de respuestas

cortas, llegando a la conclusión de que emplear modelos pre-entrenados resulta más apropiado para aplicaciones con datos limitados, como la evaluación de respuestas cortas. Esto se debe a que superan el rendimiento de otros enfoques de deep learning (DL) que requieren conjuntos de datos sustancialmente más extensos. El método presentado [10] consiste en el ajuste fino del modelo *Bidirectional Encoder Representations from Transformers* (BERT) para la clasificación de pares de oraciones, con lo que se propone demostrar la efectividad del modelo BERT para la evaluación de preguntas cortas, contando con datos limitados del dominio específico de conocimiento en el que estas preguntas se inscriben. Estas ideas son particularmente relevantes en el contexto de la tarea específica de evaluación de respuestas a preguntas abiertas que se pretende abordar en esta investigación. No obstante, los autores utilizan esta aproximación para la clasificación de pares de oraciones sin una discusión detallada sobre la elección de este objetivo de clasificación, de lo que se infiere que los pares de oraciones son la repuesta evaluada y la respuesta prototipo. Profundizar en las razones que respalden esta elección podría ser un punto de partida interesante para esta investigación. Además, se podría considerar la evaluación de respuestas basada en otra forma de representación y preprocesamiento de los datos que no requiera el uso de pares de oraciones y que sea aplicable de manera exitosa a respuestas más extensas.

En suma, la revisión de Cetina y colaboradores [6], las contribuciones de Lopez-Ubeda [8] [9], las de Sung y colaboradores [10], soportan la disponibilidad de medios técnicos para emplear tecnologías de IA, especialmente NLP, para generar retroalimentación automática, permitiendo una experiencia de aprendizaje enriquecida, soportada en la evaluación automatizada de respuestas y el análisis de sentimientos que permita estimular la interacción y dinamizar espacios de participación, especialmente virtuales, donde los estudiantes puedan, como consecuencia, enriquecer la construcción de su propio aprendizaje con una retroalimentación alineada con sus necesidades [3] [5].

I-B. Objetivos

I-B1. Objetivo general: Desarrollar una metodología para aplicar modelos de NLP en un contexto pedagógico de enfoque constructivista, soportada en la evaluación de respuestas, el análisis de sentimientos y el uso de *embeddings* personalizados en un campo de conocimiento específico.

I-B2. Objetivos específicos:

1. Seleccionar y ajustar modelos NLP adecuados, incluyendo la incorporación de *embeddings* personalizados si es necesario.
2. Integrar la evaluación de respuestas y el análisis de sentimientos.
3. Diseñar e implementar un sistema de explicabilidad que permita a los usuarios comprender las decisiones tomadas por los modelos en la metodología.
4. Implementar un mecanismo de mejora continua del desempeño de la metodología a través del aprendizaje automático a partir del uso.

5. Dimensionar y realizar un piloto de la infraestructura tecnológica requerida.
6. Integrar el modelo ajustado en una Plataforma LMS.
7. Proponer estrategias pedagógicas alineadas con el enfoque Constructivista.

II. METODOLOGÍA Y PLAN DE TRABAJO

El diseño metodológico se basará en un enfoque de investigación mixto que combina elementos cualitativos y cuantitativos para abordar los objetivos específicos de la investigación. Se llevará a cabo en 7 fases. La fase 0 y la fases 6 se corresponden con las tareas asociadas a la elaboración del plan de investigación y la sistematización final de los resultados. Las fases 1 a 5 están enfocadas en el desarrollo de cada objetivo específico. Estas fases se ejecutarán en tres años, de acuerdo con el cronograma propuesto en la tabla I.

III. RELEVANCIA

La revisión sistemática de Zawacki-Richter y colaboradores [2] señala la falta generalizada de teorías del aprendizaje explícitas que soporten las investigaciones en el ámbito de la IA aplicada a la educación (IAEd). Esta carencia causa que la mayoría de investigaciones se centren sobre el desempeño de las tecnologías asociadas a la IA, dejando en segundo plano los efectos sobre el aprendizaje y el sentido pedagógico de las aplicaciones. Este vacío constituye una oportunidad de investigación desde una perspectiva interdisciplinar que privilegie el sentido pedagógico de las aplicaciones sin desconocer los aspectos técnicos de IA que las soportan. Asimismo, la literatura del contexto educativo [3] [5] enfatiza el rol central de la retroalimentación en los procesos de aprendizaje, haciendo que los aportes en esta área tengan relevancia tanto tecnológica como pedagógica. Finalmente, los problemas de *question answering* [11] siguen siendo relevantes y vigentes en el contexto del NLP.

IV. AVANCES DEL OBJETIVO ESPECIFICO 1

En su revisión sistemática Qamar y colaboradores [11] listan los principales conjuntos de datos que han sido construidos y curados para su uso en problemas de respuestas a preguntas (QA). En esta revisión se puede evidenciar que uno de los conjuntos de datos más usado es *Stanford Question Answering Dataset (SQuAD)*. Este conjunto, de 150.000 instancias, está compuesto por contextos, preguntas y respuestas en idioma inglés y dada su recurrencia en la literatura, puede considerarse un *estándar de facto* en el contexto. Esta característica del conjunto de datos, mejora la replicabilidad de la investigación y su estructura brinda elementos adicionales a los juegos pregunta-respuesta que resultan de especial interés. Respecto a los métodos para evaluación de las respuesta producidas por los sistemas automáticos de QA se puede evidenciar el uso extendido de la métrica *similitud coseno* respecto a una respuesta prototipo correcta, la valoración de respuestas exactas, el contenido de palabras clave en la respuestas, entre otras.

Tabla I
PLANIFICACIÓN TEMPORAL POR FASES.

Fase	Actividad	Descripción	Año(s)
0	1	Revisión de la literatura (NLP, enfoque constructivista y análisis de sentimientos)	1
0	2	Identificación de modelos de NLP relevantes y herramientas	1
0	3	Definición de un marco teórico preliminar	1
0	4	Diseño del plan de investigación	1
0	5	Escritura y defensa del plan de investigación ante el comité académico	1
1	1	Selección de modelos de NLP apropiados para el contexto educativo	2
1	2	Evaluación y ajuste de los modelos seleccionados	2
1	3	Entrenamiento y evaluación de los modelos con datos/corpora específicos del campo de conocimiento	2
2	1	Identificación de herramientas y métodos para la evaluación de respuestas y análisis de sentimientos	2
2	2	Integración de estas funcionalidades con los modelos de NLP seleccionados	2
2	3	Pruebas exhaustivas para garantizar la efectividad de la integración	3
3	1	Diseño y dimensionamiento de la infraestructura tecnológica necesaria	2
3	2	Adquisición e implementación de hardware y software	2
3	3	Realización de un piloto inicial en un entorno controlado	3
4	1	Adaptación de la metodología y modelos para la integración en la Plataforma LMS	2
4	2	Desarrollo de las interfaces necesarias para la integración	2
4	3	Pruebas y ajustes para garantizar la compatibilidad	3
4	4	Evaluación y elección de algoritmos para explicabilidad y aprendizaje automático	3
4	5	Implementación y ajuste de algoritmos para explicabilidad y aprendizaje automático	3
4	6	Pruebas de funcionalidad e integración completa.	3
5	1	Desarrollo de estrategias pedagógicas basadas en el enfoque constructivista y las capacidades de la metodología	2
5	2	Validación de las estrategias a través de pruebas piloto en entornos educativos reales	2
5	3	Ajustes y mejoras basados en los resultados obtenidos	3
6	1	Recopilación de datos de pruebas piloto y aplicación en entornos educativos reales	3
6	2	Análisis de los resultados y comparación con objetivos y expectativas	3
6	3	Escritura de la tesis doctoral	3
6	4	Preparación y defensa de la tesis ante el comité académico	3

La tendencia del uso de los contextos (textos en los que se presenta un tema) como base para la generación de respuesta en sistemas automáticos de QA y la disponibilidad de estos en el conjunto de datos, permite plantear una hipótesis de trabajo para este componente de la investigación:

La relación entre los contextos, las preguntas y las respuestas permiten proponer métricas adicionales a la similitud coseno con la respuesta correcta para evaluar la bondad de una respuesta.

En este sentido y teniendo en cuenta los elementos presentes en el conjunto de datos, las relaciones que pueden explorarse son:

- Relación Contexto - Pregunta (C-P)
- Relación Contexto - Respuesta (C-R)
- Relación Pregunta - Respuesta (P-R)

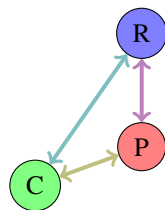


Figura 1. Elementos de texto disponibles en el conjunto de datos y métricas en revisión

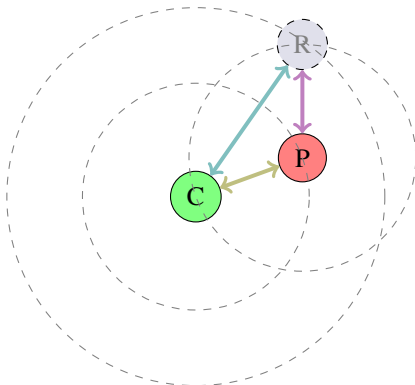


Figura 2. Hipótesis de las relaciones entre distancias coseno

Teniendo en cuenta los antecedentes, una primera exploración calculando la similitud coseno entre estos textos mostró algunas características prometedoras para continuar indagando. En un contexto amplio, la similitud coseno permitió identificar el segmento del contexto en el que se encuentra la información específica solicitada en la pregunta.

La misma métrica calculada entre preguntas y respuesta (sin incluir respuestas dicotómicas) mostró que dado un conjunto de preguntas y respuestas acerca del mismo contexto, la relación pregunta - respuesta aporta información acerca del abordaje o no de la pregunta en la respuesta.

En el mismo contexto, la relación entre pregunta y repuesta no permitiría, en principio, recuperar información respecto a la calidad de la repuesta, toda vez que las mayores similitudes

se encuentran entre preguntas y respuestas no correspondientes. Sin embargo, el uso en conjunto de las métricas debe explorarse con mayor detenimiento para evaluar su idoneidad.

Con lo anterior surge una hipótesis adicional de trabajo:

Es posible evaluar la bondad de una respuesta a una pregunta abierta sin contar con una respuesta prototipo, siempre que se cuente con el contexto y la pregunta.

Esta hipótesis será explorada en las siguientes actividades de la investigación, de cara al abordaje del objetivo específico 1.

REFERENCIAS

- [1] S. Al-Azani and C. Mahender, "Improve hamming character difference based-on derivative lexical similarity and right space padding," *Journal of Engineering Research (Kuwait)*, 2021.
- [2] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?" *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, 2019.
- [3] C. Paterson, N. Paterson, W. Jackson, and F. Work, "What are students' needs and preferences for academic feedback in higher education? a systematic review," *Nurse Education Today*, vol. 85, p. 104236, 2020.
- [4] D. Wood, J. S. Bruner, and G. Ross, "The role of tutoring in problem solving*," *Journal of Child Psychology and Psychiatry*, vol. 17, no. 2, pp. 89–100.
- [5] G. Moretti, A. Giuliani, and A. Morini, "Flexible and dialogic instructional strategies and formative feedback: An observational research on the efficacy of assessment practices in italian high schools," in *ICERI2015: 8TH INTERNATIONAL CONFERENCE OF EDUCATION, RESEARCH AND INNOVATION*, 2015, pp. 8229–8236.
- [6] V. Uc-Cetina, N. Navarro-Guerrero, A. Martin-Gonzalez, C. Weber, and S. Wermter, "Survey on reinforcement learning for language processing," *ARTIFICIAL INTELLIGENCE REVIEW*, vol. 56, no. 2, pp. 1543–1575, 2023.
- [7] V. Krishnamoorthy, "Evolution of reading comprehension and question answering systems," *Procedia Computer Science*, vol. 185, p. 231 – 238, 2021.
- [8] P. Lopez-Ubeda, "Biomedical entities recognition in spanish combining word embeddings," *PROCESAMIENTO DEL LENGUAJE NATURAL*, no. 68, pp. 149–152, MAR 2022.
- [9] P. López-Úbeda, "Biomedical entities recognition in spanish combining word embeddings." [Online]. Available: https://ruja.ujaen.es/bitstream/10953/1094/1/Thesis_plubeda.pdf
- [10] C. Sung, T. I. Dhamecha, and N. Mukhi, "Improving short answer grading using transformer-based pre-training," in *ARTIFICIAL INTELLIGENCE IN EDUCATION (AIED 2019), PT I*, 2019, pp. 469–481.
- [11] F. Qamar, S. Latif, and A. Shah, "Techniques, datasets, evaluation metrics and future directions of a question answering system," *Knowledge and Information Systems*, Dec. 2023.