XXXVII CONGRESO CHILENO DE EDUCACIÓN EN INGENIERÍA 2025
PROYECCIÓN DE LAS TECNOLOGÍAS DIGITALES EN LA FORMACIÓN EN INGENIERÍA:
LA EDUCACIÓN EN MODALIDAD PRESENCIAL, HÍBRIDA Y VIRTUAL
Concepción, 8 al 10 de octubre 2025

PROTOTIPO DE CORRECTOR INTELIGENTE DE PREGUNTAS DE DESARROLLO BASADO EN PLN Y MACHINE LEARNING

Fredy Troncoso, Universidad del Bío-Bío, ftroncos@ubiobio.cl

RESUMEN

El presente trabajo expone el desarrollo de un prototipo de corrector inteligente de preguntas de desarrollo, basado en técnicas de Machine Learning y Procesamiento del Lenguaje Natural (PLN). El objetivo principal es apoyar la labor docente mediante la evaluación automatizada de respuestas abiertas, reduciendo de manera significativa los tiempos de corrección y aportando mayor objetividad en la asignación de puntajes.

El prototipo fue entrenado y validado con respuestas reales de estudiantes universitarios, logrando reproducir con alta precisión las calificaciones otorgadas por los docentes. En su aplicación práctica se observó que el sistema tiende a ser más riguroso en la corrección, manteniendo una estrecha concordancia con la evaluación humana.

Los resultados alcanzados demuestran que esta herramienta constituye un aporte innovador para la docencia universitaria, ya que optimiza recursos, asegura mayor consistencia en la corrección y complementa el juicio académico del profesor. En este sentido, el prototipo abre el camino hacia una evaluación más eficiente, objetiva y apoyada en Inteligencia Artificial dentro de la educación superior.

PALABRAS CLAVE: Inteligencia Artificial, Machine Learning, Procesamiento del Lenguaje Natural, Evaluación Automatizada

INTRODUCCIÓN

El avance tecnológico ha multiplicado la generación de datos y textos, impactando también en la educación superior, donde la digitalización de procesos académicos se ha vuelto habitual. En este escenario, las preguntas de desarrollo siguen siendo clave para evaluar la comprensión de los estudiantes, aunque su corrección manual implica un alto costo de tiempo y cierta subjetividad.

La pandemia de COVID-19 facilitó la disponibilidad de respuestas digitalizadas y abrió la posibilidad de aplicar Procesamiento del Lenguaje Natural (PLN) y aprendizaje automático para automatizar la evaluación de respuestas abiertas. Hoy, en un contexto de presencialidad, este desafío se retoma con el objetivo de consolidar herramientas innovadoras que apoyen la docencia.

Este trabajo presenta un prototipo de corrección inteligente basado en PLN y Machine Learning, entrenado con respuestas de la asignatura Minería de Datos de la Universidad del Bío-Bío. El sistema compara cada respuesta con una pauta oficial mediante métricas de similitud semántica, generando puntajes automáticos. Más que reemplazar al docente, busca reducir drásticamente los tiempos de corrección, aportar objetividad y abrir nuevas posibilidades de innovación pedagógica en distintos contextos universitarios.

DESARROLLO

Antecedentes

La calificación automática de respuestas cortas ha sido un área activa de investigación en los últimos años, impulsada por los avances en Procesamiento del Lenguaje Natural (PLN). Sus primeras limitaciones estaban asociadas a la cantidad de palabras procesables y a la necesidad de definir patrones o plantillas rígidas para determinar si una respuesta era aceptable. Ejemplos tempranos incluyen C-rater, que analizaba respuestas mediante representaciones canónicas y reconocimiento de sinónimos (Leacock & Chodorow, 2003), y Automark, basado en esquemas de notas predefinidos (Kaur & Sasikumar, 2017).

Posteriormente, se incorporaron modelos de Machine Learning y minería de textos, aplicando algoritmos como Naive Bayes, Regresión Logística, Árboles de Decisión, Redes Neuronales, SVM, Random Forest y LSTM, con desempeños competitivos (Zhang et al., 2016; Yang et al., 2017). Paralelamente, se han utilizado métricas de similitud de texto como el coseno, la distancia de Levenshtein o el coeficiente de Dice, alcanzando concordancias superiores al 90% con la calificación humana (Olowolayemo et al., 2018; Hazar et al., 2019). Otros enfoques han explorado técnicas de extracción semántica con librerías como NLTK y Pandas, así como representaciones vectoriales tipo Bag of Words (Alrehily et al., 2018; Krithika & Narayanan, 2015). Más recientemente, se han aplicado métodos avanzados como tokenización WordPiece y modelos de transformadores, que han mostrado mejoras en la precisión de clasificación (Rahman & Akter, 2019; Wang et al., 2019).

El campo también se ha diversificado hacia nuevas aplicaciones, incluyendo análisis de perfiles estudiantiles, retroalimentación automática y predicción de rasgos de personalidad (Süzen et al., 2020; Wang et al., 2021). En conjunto, estos antecedentes confirman la viabilidad de combinar enfoques clásicos de similitud textual con modelos modernos de aprendizaje automático. No obstante, la mayoría de los estudios se han desarrollado en inglés, en contextos distintos al latinoamericano o bajo sistemas de licenciamiento restringido. En este escenario, surge la oportunidad de avanzar con aplicaciones en español y en la educación superior chilena, mediante prototipos construidos con datos reales de estudiantes y diseñados para responder a las necesidades locales de evaluación académica.

Diseño Metodológico para la Construcción del Prototipo

El presente estudio se centra en el diseño de un prototipo de corrector inteligente de preguntas de desarrollo, basado en técnicas de Procesamiento del Lenguaje Natural (PLN) y aprendizaje automático. La metodología contempló siete fases: (1) definición de un corpus de respuestas y una pauta de corrección como referencia; (2) preprocesamiento y normalización de los textos; (3) cálculo de métricas de similitud entre respuestas y pauta; (4) establecimiento de rangos de concordancia asociados a la escala de calificación; (5) integración de puntajes parciales en una nota total; (6) entrenamiento de un modelo predictivo capaz de reproducir patrones de calificación; y (7) validación del prototipo con un conjunto independiente de respuestas, evaluando su correspondencia con la corrección manual y el impacto en la reducción de tiempos de revisión.

Contexto y datos de entrenamiento

El modelo de autocorrección se entrenó con datos de la asignatura *Minería de Datos* de la carrera de Ingeniería Civil Industrial de la Universidad del Bío-Bío. Durante dos semestres consecutivos se aplicaron cuatro evaluaciones escritas, cada una con ocho preguntas abiertas. Participaron 80 estudiantes, lo que generó cerca de 150 evaluaciones y un corpus de 1200 respuestas.

La corrección manual fue realizada por cuatro ayudantes con experiencia en la asignatura, siguiendo una pauta establecida por el docente. Cada respuesta se calificó con puntajes entre 0 y 3, según el nivel de cumplimiento. Para reducir la variabilidad entre correctores se evaluaron dos estrategias: usar cada corrección como una entrada independiente, lo que aumentaba la muestra pero introducía sesgos, o calcular un promedio por respuesta, generando un puntaje único y más estable. Se optó por esta última alternativa, lo que permitió obtener una base de entrenamiento más consistente.

Aunque el conjunto de datos provenía de correcciones humanas, el modelo fue diseñado para aprender patrones generales de calificación y no reproducir la subjetividad individual de los correctores. Con ello se construyó un prototipo más fiable y replicable para la corrección automática de respuestas abiertas.

Aplicación del corrector inteligente

Una vez entrenado y validado, el prototipo puede aplicarse a nuevas evaluaciones. El proceso consiste en: (i) limpiar y normalizar los textos, (ii) calcular métricas de similitud entre cada respuesta y la pauta de corrección, (iii) transformar dichas métricas en puntajes discretos de 0 a 3, (iv) agregar los valores en un vector de características para cada estudiante, y (v) obtener, a través del modelo entrenado, un puntaje total estimado. Finalmente, este puntaje se escala a la nota chilena de 1 a 7 y se entrega como reporte individual por estudiante.

RESULTADOS

El modelo entrenado corresponde a un perceptrón simple, validado mediante un proceso de validación cruzada con k=10. El mejor desempeño se obtuvo con las métricas Synsets, Damerau-Levenshtein y palabras clave, alcanzando un error porcentual absoluto medio (MAPE) de 0.066, es decir, un error inferior al 7%. Al entrenar con el 70% de los datos y validar con el 30% restante, el modelo logró un R^2 de 0.834 y un MAPE de 0.052, lo que confirma su capacidad explicativa y la baja diferencia respecto a las calificaciones manuales.

La elección de un perceptrón simple responde a la necesidad de contar con una técnica eficiente y flexible, capaz de manejar relaciones no lineales básicas con un volumen reducido de datos. A diferencia de modelos más complejos que exigen grandes bases de entrenamiento, este enfoque permitió obtener resultados consistentes en esta primera fase, además de identificar la relevancia relativa de las métricas de similitud empleadas.

En la aplicación práctica con 30 estudiantes y ocho preguntas abiertas, el prototipo asignó un puntaje promedio de 18.43, frente a 22.0 en la corrección tradicional. Dado que el puntaje máximo posible era 24, estas cifras equivalen aproximadamente a notas de 5.31 y 5.92 en la escala chilena de 1 a 7. La Figura 1 muestra las medias y sus intervalos de confianza al 95%, donde se observa que el prototipo tiende a otorgar puntajes más estrictos y con menor dispersión que la corrección manual.

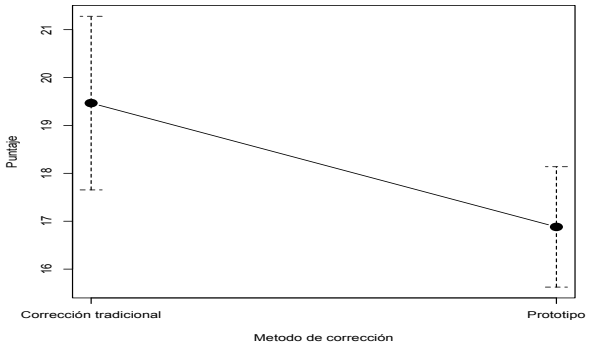

Figura 1: Diferencia de medias por método de corrección con intervalo de confianza
Fuente: Elaboración propia

En cuanto al tiempo de revisión, la mejora es sustancial: la corrección automática redujo el proceso de 1 hora y 27 minutos a solo 8 minutos y 35 segundos, lo que representa un ahorro del 91,19%.

El análisis estadístico respalda estos resultados. La corrección tradicional presentó una varianza de 23.57, mientras que el prototipo alcanzó 11.33. El test de Bartlett indicó diferencias en la homogeneidad de varianzas ($p = 0.053$) y el test de Shapiro-Wilk mostró que las distribuciones no siguen normalidad ($p < 0.01$). Por esta razón, se aplicó la prueba no paramétrica de Kruskal-Wallis, que confirmó diferencias estadísticamente significativas entre ambos métodos ($\chi^2 = 8.16$, $p = 0.004$).

El prototipo asigna puntajes de manera más estricta (diferencia promedio de 0.6 puntos en la nota final), pero con una mayor consistencia en la distribución de resultados respecto a la corrección manual. Esta característica puede considerarse una ventaja, ya que aporta uniformidad y objetividad al proceso evaluativo, reduciendo la variabilidad entre correctores humanos. Sumado al ahorro de tiempo superior al 90%, el sistema se presenta como un apoyo confiable y eficiente para la docencia universitaria, especialmente en asignaturas con alta carga evaluativa.

CONCLUSIONES

El presente trabajo desarrolló y validó un prototipo de corrección automática para preguntas de desarrollo, fundamentado en técnicas de Procesamiento del Lenguaje Natural y aprendizaje automático. Los resultados muestran que el sistema se aproxima de manera consistente a la evaluación humana, aunque con un criterio más estricto en la asignación de puntajes. Esta diferencia refleja que el prototipo aplica criterios uniformes y objetivos, lo cual puede contribuir a reducir la variabilidad de la corrección manual, además de disminuir de forma significativa el tiempo de revisión, facilitando que los docentes concentren sus esfuerzos en la retroalimentación cualitativa.

El modelo fue implementado con un perceptrón simple, lo que demuestra que incluso con un enfoque básico es posible alcanzar resultados confiables y cercanos a la corrección manual. Sin embargo, esta simplicidad representa también un punto de mejora: con un mayor volumen de datos y la exploración de arquitecturas más complejas, como BERT en español, el prototipo podría incrementar su precisión y evolucionar hacia un sistema capaz de entregar retroalimentación formativa automática, proporcionando a los estudiantes información inmediata sobre sus respuestas.

AGRADECIMIENTOS

El autor agradece el apoyo brindado por el proyecto “INES–Proyectos de Innovación en la Educación Superior” código INES I+D 22-21, que contribuyó al desarrollo de esta investigación.

REFERENCIAS

Alrehily, A. D., Siddiqui, M. A., & Buhari, S. M. (2018). Intelligent electronic assessment for subjective exams. *ACSIT, ICITE, SIPM*, 47–63.

Hazar, M. J., Toman, Z. H., & Toman, S. H. (2019). Automated scoring for essay questions in e-learning. *Journal of Physics: Conference Series, 1294*(4), 042014. https://doi.org/10.1088/1742-6596/1294/4/042014

Kaur, A., & Sasikumar, M. (2017). A comparative analysis of various approaches for automated assessment of descriptive answers. In *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICCIDS.2017.8272650

Krithika, R., & Narayanan, J. (2015). Learning to grade short answers using machine learning techniques. In *Proceedings of the Third International Symposium on Women in Computing and Informatics* (pp. 262–271). Association for Computing Machinery. https://doi.org/10.1145/2791405.2791469

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405. https://doi.org/10.1023/A:1025779619903

Olowolayemo, A., Nawi, S. D., & Mantoro, T. (2018). Short answer scoring in English grammar using text similarity measurement. In *2018 International Conference on Computing, Engineering, and Design (ICCED)* (pp. 131–136). IEEE. https://doi.org/10.1109/ICCED.2018.00034

Rahman, M. M., & Akter, F. (2019). An automated approach for answer script evaluation using natural language processing. *International Journal of Computer Science and Engineering Technology, 9*(39–47), 39–47.

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science, 169*, 726–743. https://doi.org/10.1016/j.procs.2020.02.137

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7–9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology, 30*(2), 269–282. https://doi.org/10.1007/s10956-020-09885-5

Wang, Z., Lan, A. S., Waters, A. E., Grimaldi, P., & Baraniuk, R. G. (2019). A meta-learning augmented bidirectional transformer model for automatic short answer grading. In *Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 574–579). International Educational Data Mining Society.

Yang, X., Zhang, L., & Yu, S. (2017). Can short answers to open response questions be auto-graded without a grading rubric? In *Artificial Intelligence in Education* (pp. 594–597). Springer. https://doi.org/10.1007/978-3-319-61425-0_71

Zhang, Y., Shah, R., & Chi, M. (2016). Deep learning + student modeling + clustering: A recipe for effective automatic short answer grading. *International Educational Data Mining Society*. ERIC. https://eric.ed.gov/?id=ED592653