Revista Internacional de Inteligencia Artificial en Educación (2023) 33:439–466
https://doi.org/10.1007/s40593-022-00290-6

ARTÍCULO

Utilización de un modelo de lenguaje preentrenado (BERT) para clasificar las reflexiones escritas de los futuros profesores de física

Peter Wulff[1] · Lucas Mientus[2] · Anna Nowak[2] · Andreas Borowski[2]

Aceptado: 18 de enero de 2022 / Publicado en línea: 2 de mayo de 2022

Abstracto

El análisis informático de las reflexiones escritas de futuros docentes podría permitir a los investigadores de la educación diseñar medidas de intervención personalizadas y escalables para apoyar la escritura reflexiva. Se ha comprobado que los algoritmos y las tecnologías en el ámbito de la investigación relacionada con la inteligencia artificial son útiles en numerosas tareas relacionadas con el análisis de la escritura reflexiva, como la clasificación de segmentos de texto. Sin embargo, hasta la fecha, se han empleado principalmente algoritmos de aprendizaje superficial. Este estudio explora hasta qué punto los enfoques de aprendizaje profundo pueden mejorar el rendimiento de la clasificación de segmentos de reflexiones escritas. Para ello, se utilizó un modelo de lenguaje preentrenado (BERT) para clasificar segmentos de las reflexiones escritas de futuros docentes de física según los elementos de un modelo que facilita la reflexión. Dado que se ha observado que BERT mejora el rendimiento en numerosas tareas, se planteó la hipótesis de que también mejoraría el rendimiento de la clasificación de las reflexiones escritas. También comparamos el rendimiento de BERT con otras arquitecturas de aprendizaje profundo y examinamos las condiciones para obtener el mejor rendimiento. Descubrimos que BERT superó a las otras arquitecturas de aprendizaje profundo y los rendimientos previamente reportados con algoritmos de aprendizaje superficial para la clasificación de segmentos de escritura reflexiva. BERT empieza a superar a los demás modelos cuando se entrena con entre el 20 % y el 30 % de los datos de entrenamiento. Además, los análisis de atribución de las entradas proporcionaron información sobre características importantes para las decisiones de clasificación de BERT. Nuestro estudio indica que los modelos lingüísticos preentrenados, como BERT, pueden mejorar el rendimiento en tareas relacionadas con el lenguaje en contextos educativos como la clasificación.

Palabras clave Escritura reflexiva · PNL · Aprendizaje profundo · Educación científica

* Peter Wulff
peter.wulff@ph-heidelberg.de

1 Investigación en Educación en Física, Universidad de Educación de Heidelberg, Im Neuenheimer Feld 560-562, 69120 Heidelberg, Alemania

2 Instituto de Física y Astronomía, Universidad de Potsdam, Karl-Liebknecht-Straße 24/25, 14476 Potsdam-Golm, Alemania

Motivación

En su profesión, los docentes deben tomar decisiones complejas en condiciones de incertidumbre (Grossman et al.,2009). Por lo tanto, el crecimiento profesional de los docentes está relacionado con el afrontamiento de la incertidumbre y el aprendizaje de las propias experiencias (Clarke y Hollingsworth,2002; Kolb,1984). Se ha argumentado que la reflexión es un vínculo importante para conectar las experiencias personales con el conocimiento teórico que ayuda a los docentes a comprender y afrontar situaciones inciertas (Korthagen y Kessels, 1999). Se descubrió que la instrucción explícita de reflexión era un ingrediente clave de los programas eficaces de formación docente universitarios (Darling-Hammond et al.,2017). Por lo tanto, un objetivo de los programas de formación docente universitarios es apoyar a los docentes para que se conviertan en profesionales reflexivos: profesionales capaces de aprovechar las experiencias prácticas de enseñanza para crecer profesionalmente (Schön,1987 ;Corthagen,1999; Zeichner,2010).

Para ayudar a los futuros docentes a reflexionar sobre aspectos esenciales de sus experiencias docentes, los instructores a menudo les piden que escriban sobre sus experiencias docentes de una manera estructurada, guiada por modelos que apoyan la reflexión (Lai y Calandra,2010; Poldner y otros,2014; Korthagen y Kessels,1999). Sin embargo, el análisis de la reflexión escrita está poco desarrollado y las expectativas sobre lo que implica una reflexión de calidad varían de un contexto a otro (Buckingham Shum et al.,2017). Además, se encontró que la retroalimentación de los instructores sobre las reflexiones escritas era más holística que analítica (Poldner et al.,2014). Esto se debe a que el análisis de contenido de las reflexiones escritas es un proceso laborioso (Ullmann,2019).

Los métodos computarizados en el ámbito de la investigación de la inteligencia artificial pueden ayudar a analizar reflexiones escritas (Wulff et al.,2020). Por ejemplo, los métodos de aprendizaje automático supervisado pueden categorizar las reflexiones escritas en referencia a modelos que apoyan la reflexión (Ullmann,2019). Investigaciones recientes en el campo de la investigación de inteligencia artificial han descubierto que los avances en las arquitecturas de modelos para redes neuronales profundas, como los transformadores como modelos de lenguaje preentrenados, pueden mejorar aún más el rendimiento de la clasificación para diversas tareas relacionadas con el lenguaje (Devlin et al.,2018). Estas arquitecturas podrían ser aplicables al análisis de la escritura reflexiva y, por lo tanto, proporcionar a los investigadores educativos herramientas novedosas para diseñar aplicaciones que faciliten un análisis preciso de la escritura reflexiva (Buckingham Shum et al., 2017).

El propósito de este estudio es explorar las posibilidades de utilizar modelos de lenguaje preentrenados basados en transformadores para el análisis de reflexiones escritas. Para ello, utilizamos un modelo de lenguaje preentrenado para analizar las reflexiones escritas de futuros profesores de física, basadas en un modelo teórico de apoyo a la reflexión. Las reflexiones escritas se recopilaron durante varios años en prácticas docentes y fueron etiquetadas por evaluadores humanos según el modelo de apoyo a la reflexión. Los textos etiquetados constituyeron los datos de entrenamiento para un modelo de lenguaje preentrenado basado en transformadores, denominado Representaciones de Codificador Bidireccional para Transformadores (BERT) (Devlin et al.,2018). Para evaluar el rendimiento del modelo de lenguaje preentrenado, se ajustaron a los datos varias otras arquitecturas de aprendizaje profundo ampliamente utilizadas para modelar el lenguaje y se compararon con el rendimiento de BERT.

Clasificar segmentos en los datos de prueba retenidos. Para comprender mejor las condiciones óptimas para el entrenamiento de los modelos, comparamos los hiperparámetros y la dependencia del rendimiento de la clasificación con el tamaño de los datos de entrenamiento entre las arquitecturas de aprendizaje profundo. Finalmente, se exploraron características de entrada importantes para las decisiones de clasificación de BERT.

Escritura reflexiva en la formación docente

El pensamiento reflexivo se ha conceptualizado como el antídoto al pensamiento intuitivo y rápido (Dewey,1933; Kahneman,2012). Se ha argumentado que es un proceso de pensamiento que ocurre de manera natural (aunque rara vez) y que se relaciona con la creación de significado a partir de la experiencia y que puede caracterizarse como sistemático, riguroso y disciplinado (Clarà,2015;Dewey, 1933; Rodgers,2002). La reflexión "implica monitorear, evaluar y modificar activamente el propio pensamiento" (Lin et al.,1999, p. 43). En el contexto de la formación docente que apoya la reflexión, Korthagen (2001, p. 58) define la reflexión como "el proceso mental de intentar estructurar o reestructurar una experiencia, un problema o conocimientos o perspectivas existentes". En este sentido, el pensamiento reflexivo se ha caracterizado por incluir (1) un proceso con actividades de pensamiento específicas (p. ej., describir, analizar, evaluar), (2) contenido objetivo (p. ej., experiencia docente, teoría personal, suposiciones) y (3) objetivos y razones específicos para participar en este tipo de pensamiento (p. ej., pensar de manera diferente o más clara, o justificar la propia postura) (Aeppli y Lötscher, 2016).

Dada la complejidad e incertidumbre de la profesión docente, y los desafíos para transferir el conocimiento formal al conocimiento práctico de la enseñanza, los docentes a menudo desarrollan respuestas intuitivas a los eventos del aula que pueden resultar en rutinas ciegas (Fenstermacher,1994 ; Grossman y otros,2009;Corthagen,1999; Camino nuevo, 2007). La exposición sin paliativos a situaciones de enseñanza complejas e inciertas durante un período prolongado de tiempo puede además resultar en el desarrollo de estrategias de control y visiones transmisivas del aprendizaje (Hascher,2005;Corthagen,2005;Loughran y Corrigan,1995). Reflexionar sobre las propias acciones docentes y el desarrollo profesional puede ayudar a integrar el conocimiento formal con conocimientos docentes más prácticos (Carlson et al.,2019). Sin embargo, los docentes en formación que no han tenido oportunidades de reflexionar explícitamente sobre sus acciones de enseñanza y aprendizaje no tienen guiones para el pensamiento reflexivo, por ejemplo, cómo escribir una reflexión (Buckingham Shum et al.,2017;Loughran y Corrigan,1995). Korthagen (1999) observaron que los futuros docentes tienen dificultades para comprender adecuadamente un problema que una determinada experiencia expone. Además, tienden a ser bastante egocéntricos, con una preocupación comparativamente menor por el pensamiento de los estudiantes (Chan et al.,2021; Levin y otros,2009). Los profesores expertos han desarrollado mucha más flexibilidad en su comportamiento en el aula, lo que se relaciona con competencias reflexivas más desarrolladas (Berliner,2001). Los profesores expertos son particularmente capaces de reflexionar mientras enseñan y utilizar sus experiencias para el crecimiento profesional autodirigido (Berliner,2001;Corthagen,1999; Hermoso,1983).

La formación docente reflexiva basada en la universidad puede ayudar a los futuros docentes a exponer sus conocimientos e integrarlos con sus experiencias prácticas de enseñanza (Abels,2011; Darling-Hammond y otros,2017; Lin y otros,1999). Promulgación

Las estructuras de apoyo a la reflexión en la formación docente universitaria requieren que los instructores creen espacios donde los futuros docentes puedan tener experiencias de enseñanza auténticas y estructuradas (Grossman et al.,2009Se argumentó que las prácticas escolares permitían a los futuros docentes actuar en situaciones auténticas de clase que brindaban oportunidades para la reflexión (Grossman et al.,2009;Corthagen,2005; Zeichner, 2010). Las estructuras de apoyo a la reflexión en la formación docente universitaria también implican proporcionar pautas y modelos de apoyo a la reflexión que ayuden a los futuros docentes a participar en el pensamiento reflexivo (Lin et al.,1999; Mena-Marcos y otros,2013).

Las indicaciones y los modelos que apoyan la reflexión en los programas universitarios de formación docente suelen implementarse en el contexto de reflexiones escritas, donde posteriormente se proporciona retroalimentación experta. Entre las formas comunes de estimular el pensamiento reflexivo se incluyen tareas como los cuadernos de registro (Korthagen,1999), diarios reflexivos (Bain et al.,2002), o revistas dialógicas y de respuesta (Roe y Stallman,1994; Paterson,1995), a menudo en el contexto de prácticas escolares. En estos enfoques, los estudiantes inician una conversación escrita en forma de comunicación centrada en el estudiante en la que se exploran experiencias significativas del alumno (Paterson,1995). La escritura ofrece varias ventajas como medio de reflexión para los docentes. Los docentes pueden reflexionar detenidamente sobre lo que escriben, replantear y reelaborar ideas, y establecer una postura clara que pueda discutirse con otros (Poldner et al.,2014).

Sin embargo, a pesar de la implementación generalizada de la escritura reflexiva, se ha recopilado poco conocimiento cuantificable sobre el género de las reflexiones escritas. ¿Cuál es la composición prototípica de una reflexión escrita? ¿Qué aspectos del pensamiento reflexivo predominan en una amplia colección de reflexiones escritas? Argumentamos que responder a estas preguntas requiere un método sistemático y basado en principios para analizar las reflexiones escritas (Buckingham Shum et al.,2017). Además, la retroalimentación de los instructores sobre las reflexiones escritas a menudo se centra en la calidad en lugar de abordar los contenidos específicos de la reflexión (Poldner et al.,2014). Estos problemas pueden atribuirse en parte a la gran cantidad de reflexiones escritas que requieren retroalimentación y a la falta de precisión conceptual sobre el contenido de la reflexión (Aeppli y Lötscher,2016; Rodgers,2002).

Los métodos en el contexto de la investigación de inteligencia artificial, como el procesamiento del lenguaje natural (PLN) y el aprendizaje automático (AA), pueden proporcionar formas sistemáticas y basadas en principios para analizar las reflexiones escritas de los docentes en formación (Ullmann,2019), porque, entre otras cosas, incentivan a los investigadores a examinar los supuestos y modelos que informan la generación de reflexiones escritas (Breiman,2001; Kovanović y otros,2018; Ullmann,2019).

Análisis automatizado de la escritura reflexiva

Los métodos de PNL y ML se han utilizado para analizar la escritura reflexiva en diversos dominios, como la ingeniería, los negocios, la salud o la formación docente (Buckingham Shum et al.,2017; Luo y Litman,2015; Ullmann,2019; Wulff y otros,2020). Comúnmente, los análisis incluyen un modelo de apoyo a la reflexión utilizado como marco (Ullmann,2019) para definir categorías que se identificarán a través de algoritmos ML.

Como tal, los textos sin procesar se suelen preprocesar con métodos de PNL como: (1) eliminar palabras redundantes y no informativas (por ejemplo, palabras vacías), (2) identificar el papel lingüístico de las palabras (anotación de categorías gramatical) (Gibson et al.,2016; Ullmann y otros,2012), o (3) reducir las palabras a sus formas base (morfológicas) (derivación, lematización) (Ullmann et al.,2012Además de estas técnicas generales de preprocesamiento, también se han ideado métodos de extracción más específicos para la reflexión. Por ejemplo, Ullmann et al. (2012) identificaron autorreferencias como pronombres personales o verbos reflexivos (por ejemplo, “repensar”, “reflexionar”) para anotar segmentos de textos reflexivos basados en diccionarios predefinidos.

Los textos preprocesados se introdujeron en modelos de aprendizaje automático o reglas condicionales para clasificar segmentos (normalmente oraciones) de la escritura reflexiva. Para realizar la clasificación, se aplicaron métodos de aprendizaje automático supervisados (Carpenter et al.,2020; Wulff y otros,2020; Ullmann,2019). En este contexto, Buckingham Shum et al. (2017) utilizaron reglas elaboradas manualmente para distinguir entre oraciones reflexivas e irreflexivas. Utilizaron un total de 30 textos anotados (382 oraciones) para entrenar su sistema. El mejor rendimiento de este sistema fue un kappa de Cohen de 0,43 (calculado en: Ullmann (2019)). Gibson y otros. (2016) utilizaron 6090 reflexiones de estudiantes y las clasificaron en actividad metacognitiva débil o fuerte (relacionada con la reflexión). Utilizaron el etiquetado de categorías gramaticales y métodos basados en diccionarios para representar los textos como características. Según los cálculos de Ullmann (2019), su modelo con mejor rendimiento recibió un valor kappa de Cohen de 0,48. Además, Ullmann (2019) utilizó la distinción entre profundidad reflexiva (texto reflexivo versus texto descriptivo) y amplitud reflexiva (ocho categorías, como la conciencia de un problema y las intenciones futuras). Utilizó técnicas de aprendizaje superficial en aprendizaje automático para clasificar las oraciones según su amplitud reflexiva. Los valores kappa de Cohen para las categorías individuales oscilaron entre 0,53 y 0,85. Finalmente, Cheng (2017) utilizaron un enfoque de análisis semántico latente para clasificar las entradas reflexivas en un sistema de portafolio electrónico mediante el modelo ASER (A: análisis, reformulación y aplicación futura; S: aplicación de la estrategia: analizar la efectividad de la estrategia de aprendizaje de idiomas; E: influencias externas; R: informe de evento o experiencia). Los valores kappa de Cohen para el rendimiento de la clasificación oscilaron entre 0,60 y 0,73.

Carpenter et al. presentaron un avance potencial en esta investigación.2020). Utilizaron incrustaciones de palabras para representar las reflexiones de los estudiantes. Las incrustaciones de palabras son representaciones de alta dimensión de palabras en el espacio vectorial que finalmente encapsulan las relaciones semánticas y sintácticas entre las palabras y resuelven problemas como la sinonimia o la polisemia (Taher Pilehvar y Camacho-Collados,2020). Carpenter y otros. (2020) utilizó el modelo de reflexión de Ullmann (2017) para anotar las respuestas de los estudiantes en un entorno científico basado en un juego, donde los estudiantes representaban a microbiólogos que debían diagnosticar el brote de una enfermedad. En primer lugar, los autores comprobaron que la profundidad reflexiva de las respuestas de los estudiantes predecía sus puntuaciones en la prueba posterior. Además, demostraron que el uso de incrustaciones de palabras (ELMo) con algoritmos de aprendizaje automático (ML) fue más eficaz en comparación con los modelos que solo utilizaban representaciones basadas en conteos de las características de entrada.

La mayoría de las aplicaciones para clasificar segmentos de reflexión escrita según modelos que la apoyan han utilizado modelos de aprendizaje superficiales, como la regresión logística o la clasificación bayesiana ingenua. Sin embargo, el modelado de datos lingüísticos presenta complejidades que no se pueden captar con estos modelos, como el largo alcance.

Dependencias. Se ha descubierto que las arquitecturas de aprendizaje profundo pueden modelar estas complejidades. Además, tienen la ventaja de encontrar automáticamente representaciones eficientes para datos textuales, lo que exime a los investigadores de tareas como la lematización o la eliminación de palabras vacías (Goodfellow et al.,2016).

Modelos de aprendizaje profundo para el modelado del lenguaje

El modelado de datos textuales en general se ha vuelto más eficiente con el advenimiento y la aplicación de arquitecturas de redes neuronales profundas (Goldberg,2017; LeCun y otros, 2015). Las arquitecturas de redes neuronales profundas mitigaron las deficiencias de los modelos lingüísticos de bolsa de palabras empleados anteriormente. Los modelos de bolsa de palabras parten de la suposición simplificadora de que el orden en que aparecen las palabras en un segmento es irrelevante (Jurafsky y Martin,2014). Luego, modelar la dependencia del contexto a través de incrustaciones de palabras y la secuenciación de palabras a través de incrustaciones dinámicas fue un facilitador importante para las mejoras de rendimiento en el modelado del lenguaje (Taher Pilehvar y Camacho-Collados,2020).

En el lado negativo, las arquitecturas de redes neuronales profundas cada vez más sofisticadas también requirieron más datos de entrenamiento, porque la cantidad de parámetros en los modelos aumentó sustancialmente.[1]Para hacer frente a los requisitos excesivos, se desarrollaron métodos de aprendizaje por transferencia donde los investigadores utilizan grandes modelos de aprendizaje profundo previamente entrenados con corpus lingüísticos masivos, como el contenido de internet o Wikipedia. En analogía con la visión artificial, donde la clasificación de imágenes puede mejorarse mediante pesos de modelos preentrenados en lugar de inicializaciones aleatorias de los pesos de los modelos (Pratt y Thrun,1997), los pesos de los modelos preentrenados en los modelos de lenguaje pueden formar una columna vertebral sólida que proporcione estructura para tareas posteriores (Devlin et al.,2018).

Con base en estos hallazgos alentadores en la investigación de ML y PNL, postulamos que la aplicación de modelos lingüísticos preentrenados también puede mejorar el rendimiento en la clasificación de segmentos en las reflexiones escritas de docentes en formación. Sin embargo, no conocemos estudios que hayan utilizado arquitecturas de aprendizaje profundo en la clasificación de la escritura reflexiva. Carpenter et al.2020) observó: “Otra dirección para el trabajo futuro [en análisis de escritura reflexiva] es investigar técnicas alternativas de aprendizaje automático para modelar la profundidad de las reflexiones de los estudiantes, incluidas arquitecturas neuronales profundas (por ejemplo, redes neuronales recurrentes)” (p. 76).

En consonancia con esta sugerencia, empleamos un modelo de lenguaje previamente entrenado llamado representaciones de codificador bidireccional para transformadores (BERT) que ha demostrado ser excelente en muchas tareas relacionadas con el lenguaje no específicas de las reflexiones escritas (Devlin et al.,2018). La siguiente pregunta general de investigación guió el presente estudio: ¿Hasta qué punto se puede utilizar un modelo de lenguaje preentrenado (BERT) en el contexto de las reflexiones escritas de los docentes en formación para clasificar los segmentos según la

[1]Tenga en cuenta que el famoso modelo de lenguaje generativo GPT-3 que fue desarrollado por OpenAI abarca 175 mil millones de parámetros (Brown et al.,2020)

¿Cuáles son los elementos de un modelo que apoya la reflexión? Más específicamente, se responderán las siguientes preguntas de investigación:

1. ¿En qué medida un modelo de lenguaje preentrenado y afinado (BERT) supera a otras arquitecturas de aprendizaje profundo en el desempeño de clasificación de segmentos en las reflexiones escritas de docentes en formación de acuerdo con los elementos de un modelo de apoyo a la reflexión?
2. ¿En qué medida el tamaño de los datos de entrenamiento está relacionado con el rendimiento de clasificación de los modelos de lenguaje de aprendizaje profundo?
3. ¿Qué características pueden explicar mejor las decisiones de los clasificadores?

Método

Reflexiones escritas de profesores de física en prácticas

Con el fin de instruir a los futuros docentes para que escribieran una reflexión sobre sus experiencias docentes, se diseñó un modelo de apoyo a la reflexión basado en un modelo de apoyo a la reflexión existente (Korthagen y Kessels,1999). Los modelos de apoyo a la reflexión como andamiajes para las reflexiones escritas suelen diferenciar entre varias zonas funcionales (Swales,1990) de escritura que deben abordarse para provocar procesos de pensamiento apropiados relacionados con la reflexión (Aeppli y Lötscher, 2016; Bain y otros,1999; Korthagen y Kessels,1999; Poldner y otros,2014; Ullmann,2019). Según muchos modelos, los procesos de pensamiento relacionados con la reflexión comienzan con una recapitulación y descripción de la situación de enseñanza. Esto permite a los futuros docentes establecer evidencia sobre un problema que han observado (Hatton y Smith,1995; van Es y Sherin, 2002). Otra categoría importante es la evaluación y análisis de la situación de enseñanza (Korthagen & Kessels,1999Los futuros docentes juzgan las acciones de los estudiantes y las suyas propias. Finalmente, se les instruye a idear alternativas para sus acciones y a concebir consecuencias para su propio desarrollo profesional. Las alternativas y las consecuencias son aspectos importantes de una reflexión crítica (Hatton y Smith,1995; Korthagen y Kessels,1999), porque –en comparación con el mero análisis– la reflexión tiene como objetivo transformar la experiencia individual del docente de un modo útil para su futuro desarrollo profesional.

En el presente estudio, se utilizó un modelo de apoyo a la reflexión que se desarrolló en estudios anteriores (Nowak et al.,2019En este modelo, los fundamentos del pensamiento reflexivo, descritos anteriormente, se plasman en elementos constitutivos de la escritura reflexiva, basados en la teoría del aprendizaje experiencial. El modelo se basa en el modelo ALACT de Korthagen y Kessels (1999). En pocas palabras, el modelo de Korthagen y Kessels (1999) describe un proceso cíclico de reflexión que implica aplicar la enseñanza, reflexionar sobre aspectos esenciales de la experiencia, tomar conciencia de dichos aspectos, crear métodos alternativos de acción y ponerlos a prueba. Nowak et al. (2019) adaptaron este modelo en el contexto de reflexiones escritas en unas prácticas escolares para la formación docente universitaria (véase también: Wulff et al.,2020). El modelo de Nowak et al. (2019) define una reflexión para incluir los elementos: circunstancias de

la situación de enseñanza, descripción de la situación de enseñanza, evaluación de las acciones de los estudiantes y del profesor, elaboración de alternativas y derivación de consecuencias.

Este modelo se empleó en el presente estudio. Se pidió a los futuros docentes que elaboraran una reflexión escrita sobre una lección autodidacta en su última práctica docente o sobre una lección observada en una viñeta de vídeo. Se les pidió que 1) describieran las circunstancias de su lección, seguidas de 2) una descripción detallada de los eventos sobre los que reflexionarían. Posteriormente, debían 3) evaluar el/los evento(s) descrito(s) y 4) anticipar acciones alternativas. Finalmente, se les pidió que 5) describieran las consecuencias personales para su desarrollo profesional con base en sus evaluaciones. Esta instrucción nos brindó la oportunidad de presentar a los docentes un andamiaje para la reflexión que eventualmente compensara la, a menudo criticada, falta de expectativas sobre lo que realmente implica la reflexión (Buckingham Shum et al.,2017). La simplicidad del modelo lo hace accesible a docentes de diversos grados de experiencia: a todos se les proporciona un marco para estructurar sus escritos.

Todas las reflexiones escritas se recopilaron de profesores de física en prácticas en dos universidades alemanas de tamaño medio. Se consideró útil restringir el estudio a los profesores de física, ya que la experiencia docente se considera específica del dominio (Berliner, 2001). La física puede considerarse un dominio rico en conocimientos (Schoenfeld,2014). La enseñanza actual de física en el aula tiende a estar dominada por una pedagogía centrada en el profesor, en lugar de enfoques de aprendizaje más constructivistas y centrados en el estudiante (Fischer et al.,2010). Se puede esperar que el desarrollo de la atención a las ideas de los estudiantes y la observación de problemas de aprendizaje específicos de la física a través del aprendizaje experiencial facilitado por la reflexión puedan ayudar a los profesores de física a implementar una instrucción más centrada en el estudiante.

Se recopilaron autorreflexiones escritas a lo largo de tres años. En general, N=92 futuros profesores de física escribieron 270 reflexiones durante su última práctica docente en el programa de formación docente universitario. Estos futuros profesores eran nuevos en el modelo de apoyo a la reflexión. Por consiguiente, la instrucción fue explícita sobre lo que se esperaba de los profesores (véase más adelante). La práctica docente duró aproximadamente 15 semanas. Los futuros profesores de física que reflexionaron sobre sus propias lecciones reflexionaron en promedio 8,9, 4,5 y 3,8 veces a lo largo de las tres prácticas docentes consecutivas en las que se recopilaron reflexiones escritas. Al principio del estudio, observamos que 8,9 reflexiones escritas eran demasiadas para los futuros profesores en una práctica docente de tan solo 15 semanas. En consecuencia, se les pidió a los futuros profesores de física que escribieran solo unas seis reflexiones escritas en las siguientes prácticas docentes. El número total de reflexiones que un profesor debía escribir fue, en última instancia, una de las razones por las que algunas reflexiones fueron muy breves. Sin embargo, no esperábamos que la cantidad total de palabras planteara problemas para la tarea de clasificación en cuestión. Si bien las múltiples reflexiones de los mismos futuros docentes plantean problemas de dependencia intraindividual en las reflexiones escritas, consideramos que el mayor tamaño de los datos de capacitación era más relevante dado el propósito de nuestro estudio. Los docentes participaron según dos posibles modalidades. En la primera, podían escribir reflexiones sobre su propia docencia. En la segunda, podían reflexionar sobre la docencia de otros, observada en una viñeta de video. La instrucción para el modelo de apoyo a la reflexión se proporcionó por escrito a los futuros docentes que participaron en la modalidad de viñeta de video. La instrucción para los futuros docentes en la escuela

La colocación se realizó de forma oral y escrita en el seminario correspondiente. En la modalidad de videoviñeta, todos los futuros profesores de física reflexionaron solo una vez sobre el mismo videoclip de 16 minutos. La extensión promedio (en palabras) de todas las reflexiones escritas fue de 667 (*DAKOTA DEL SUR*=433) palabras. La longitud mínima fue de 72 palabras y la máxima de 2808. El vocabulario constaba de 12 518 palabras únicas.

El objetivo de este estudio fue entrenar y evaluar un modelo que clasifica segmentos de las reflexiones escritas de futuros profesores de física según los elementos del modelo de apoyo a la reflexión, a saber: 1) circunstancias, 2) descripción, 3) evaluación, 4) alternativas y 5) consecuencias. Un segmento, como unidad de codificación elemental, se definió como un pasaje de texto que expresa una sola idea (p. ej., la descripción de una acción). En el contexto del análisis de la reflexión escrita, se encontró que las unidades de codificación más pequeñas, como las oraciones, eran más adecuadas para la clasificación de las reflexiones escritas en comparación con unidades más grandes, como textos completos (Ullmann,2019). Las longitudes de los segmentos fueron en promedio de 1,6 (*DE*=1.2) oraciones. La longitud mínima del segmento era de 1 oración y la máxima de 18. Un segmento de texto codificado de ejemplo tenía el siguiente aspecto:

> [El propósito del experimento era observar diferentes fuerzas y determinar sus puntos de contacto.][Código: Circunstancias] ... [La situación experimental comenzó con condiciones desfavorables, ya que debido a un cambio de horario de la clase, los experimentos tuvieron que llevarse a cabo en una sala que no era de Física. Sin embargo, al ser experimentos de mecánica, aún era posible llevarlos a cabo. No obstante, la sala donde se llevó a cabo la clase es mucho más pequeña que la sala que se planeó originalmente. Por lo tanto, la situación experimental era algo estrecha.][Código: Circunstancias] ... [Después de anunciar la tarea, algunos estudiantes permanecieron sentados porque no todos podían experimentar al mismo tiempo.][Código: Descripción] ... [Uno de los experimentos involucraba un carrito conectado a un peso mediante un hilo y una polea. Este experimento fue modificado parcialmente por los estudiantes, en contra de las instrucciones. Sin embargo, como estos experimentos aún se ajustaban al objetivo del experimento, no se realizó ninguna intervención.][Código: Descripción] ... [El experimento tuvo una buena acogida entre los estudiantes.][Código: Descripción] ... [Los estudiantes trabajaron con mucha dedicación; el resultado fue mejor de lo esperado.][Código: Evaluación] ... [Como alternativa, el tiempo de aprendizaje de los estudiantes podría haberse aprovechado mejor realizando los experimentos como experimentos de demostración. Sin embargo, esto habría eliminado uno de los objetivos de aprendizaje de la lección, que se centraba en la adquisición de conocimientos.] [Código: Alternativas] ... [Como profesor, es evidente que la planificación debe ser más detallada. Además, se deben crear reservas en caso de que un experimento tenga que cancelarse por un cambio de aula o algo similar.][Código: Consecuencias] ...

Se capacitó a un evaluador para etiquetar manualmente las reflexiones escritas según los elementos del modelo de apoyo a la reflexión (elementos 1 a 5). El evaluador primero segmentó las reflexiones escritas. La unidad mínima de segmentación se centró principalmente en el nivel de oración, con algunas excepciones, por ejemplo, cuando las oraciones no tenían significado por sí mismas o cuando estaban fuertemente conectadas. En estudios previos, observamos que al relajar el...

La restricción del límite de la oración produjo un acuerdo entre evaluadores notablemente mejor (Wulff et al.,2020). Luego, el evaluador etiquetó los segmentos con uno de los cinco elementos según las definiciones de los elementos descritas anteriormente. La concordancia entre evaluadores se determinó mediante la recodificación de un subconjunto de los textos (generalmente*norte*= 8 reflexiones escritas) realizadas por un segundo evaluador independiente que utilizó los segmentos del primer evaluador. Dado el problema de clasificación comparativamente simple y los prometedores análisis de la concordancia entre evaluadores para este problema de clasificación en estudios previos (Nowak et al., 2018), determinar la concordancia entre evaluadores en un subconjunto tan pequeño se consideró suficiente en este contexto. Los valores kappa de Cohen fueron superiores a 0,73 para todos los elementos (Wulff et al.,2020Dado el acuerdo sustancial, el primer evaluador etiquetó los textos restantes. Con las calificaciones finales de los textos, se calculó que las proporciones de los elementos de los textos eran: circunstancias (26%), descripción (36%), evaluación (23%), alternativas (8%) y consecuencias (7%).

Modelos de aprendizaje profundo

El objetivo de este estudio fue utilizar un modelo de lenguaje preentrenado basado en transformadores (BERT) para clasificar segmentos en las reflexiones escritas de futuros profesores de física y comparar su rendimiento con el de otras arquitecturas de aprendizaje profundo. Antes de describir en detalle el modelo de lenguaje preentrenado, analizaremos las demás arquitecturas de aprendizaje profundo consideradas y describiremos algunas de sus posibles fortalezas y debilidades en relación con las aplicaciones para el análisis del lenguaje. Estas arquitecturas alternativas fueron las redes neuronales de propagación hacia adelante (FFNN) y las redes neuronales de memoria a largo plazo (LSTM), arquitecturas ampliamente empleadas en la investigación del PLN (Goldberg,2017) y tienen ventajas particulares en la clasificación de segmentos de reflexiones escritas.

Las FFNN se encuentran entre las arquitecturas de aprendizaje profundo más sencillas. En su forma más sencilla, constan de una capa de entrada, una capa oculta y una capa de salida, completamente conectadas. La introducción de la capa oculta hace que las FFNN sean más abarcativas que modelos como la regresión logística. De hecho, las FFNN con una sola capa oculta y funciones de activación no lineales pueden aproximar cualquier función concebible (es decir, son aproximadores de funciones generales) (Goodfellow et al.,2016; Jurafsky y Martin, 2014Una deficiencia de las FFNN simples es que la información fluye solo de las capas inferiores a la capa superior, sin que la información fluya de regreso ni lateralmente en una capa. Esperamos que las FFNN sean capaces de clasificar segmentos en las reflexiones escritas, ya que son similares a los algoritmos de aprendizaje automático (ML) más superficiales (Wulff et al.,2020). Sin embargo, no deben considerarse la arquitectura de aprendizaje profundo más eficiente debido a sus fuertes suposiciones sobre el flujo de información. Además, en las FFNN no se tiene en cuenta el orden de entrada. Entre los hiperparámetros importantes de las FFNN se incluye el tamaño de la capa oculta. En este estudio, solo se consideró el ancho, no la profundidad de la capa oculta. Asimismo, las entradas en las FFNN suelen representarse mediante vectores de incrustación. La dimensionalidad de los vectores de incrustación de las entradas puede variar.

Los LSTM superan la suposición de que el orden de entrada es irrelevante. Se basan en redes neuronales recurrentes (RNN). Las RNN se destacan en la captura de patrones en

entradas secuenciales, porque permiten que la salida de una entrada anterior fluya hacia la siguiente predicción (Goldberg,2017). Las RNN pueden proporcionar una representación de tamaño fijo de una secuencia de cualquier longitud, lo que puede ser beneficioso en tareas donde, por ejemplo, el orden de las palabras es importante (Goldberg,2017Las RNN pueden realizar predicciones para la siguiente palabra en función de todas las palabras de entrada anteriores. Por lo tanto, pueden utilizarse para codificar la entrada en una representación significativa. Se identificó una deficiencia de las RNN como la importancia cada vez menor de las relaciones a larga distancia (Jurafsky y Martin,2014). Por lo tanto, las RNN se emplean a menudo en arquitecturas complejas y controladas, como las LSTM, que codifican mejores regularidades estadísticas y relaciones de larga distancia en la secuencia de entrada (Goldberg,2017). En tareas como la clasificación de segmentos, las representaciones resultantes del LSTM se incorporan a una FFNN que asigna la representación a una etiqueta o categoría (Goldberg, 2017; Jurafsky y Martin,2014Esperamos que los LSTM alcancen un mejor rendimiento en la clasificación de segmentos de las reflexiones escritas, ya que conjeturamos que el orden de las palabras en la secuencia de entrada es una característica adicional importante en comparación con la simple ocurrencia de palabras. Los hiperparámetros para ajustar el LSTM incluyen la dimensionalidad oculta y la dimensionalidad de la incrustación de entrada, entre otros. Utilizamos incrustaciones bidireccionales para el modelo LSTM.

Modelo de lenguaje preentrenado basado en transformadores: BERT

Más recientemente, se han introducido modelos de transformadores para procesar el lenguaje natural y se ha descubierto que superan a los LSTM (Devlin et al.,2018; Vaswani y otros, 2017). En el núcleo de los modelos de transformadores se encuentra el llamado mecanismo de atención, que comprende la consulta, la clave y el valor de cada entrada. Los transformadores tienen la ventaja sobre las arquitecturas RNN de prestar mejor atención a las partes distantes de una oración, sin sesgo de localidad (Taher Pilehvar y Camacho-Collados,2020). Los investigadores demostraron que el lenguaje natural puede caracterizarse por dependencias de largo alcance (Ebeling y Neiman,1995;Zanette,2014), lo que convierte a las arquitecturas de transformadores en modelos prometedores para resolver tareas relacionadas con el lenguaje. Una ventaja técnica es que los cálculos son paralelizables (Taher Pilehvar y Camacho-Collados, 2020Esto significa que, incluso en una computadora personal, la velocidad necesaria para entrenar el modelo puede aumentarse externalizando los cálculos en la unidad de procesamiento gráfico (GPU) diseñada para tareas similares. Una implementación potente de un modelo de lenguaje basado en transformadores se denomina BERT (Devlin et al.,2018). BERT superó a los modelos anteriores en tareas de PNL estandarizadas como respuesta a preguntas, implicación textual, aprendizaje de inferencia de lenguaje natural y clasificación de documentos (Conneau et al., 2019; Devlin y otros,2018).

Además de ser una arquitectura de transformador, BERT también utiliza el aprendizaje por transferencia. La idea del aprendizaje por transferencia es entrenar una representación general del lenguaje a partir de grandes conjuntos de datos lingüísticos. En la fase de preentrenamiento, BERT utiliza datos sin etiquetar, ampliamente disponibles en internet (Devlin et al.,2018). El protocolo de aprendizaje detrás de BERT se basa en enmascarar tokens de entrada elegidos aleatoriamente y tratar de predecirlos, similar a la tarea clozetask (Taylor,1953). Por lo tanto, el modelo cuenta con información de ambos lados de un token de entrada enmascarado para predecirlo. BERT también se entrena con una tarea de predicción de la siguiente oración, que permite al modelo codificar relaciones entre oraciones. Para esta tarea, cada

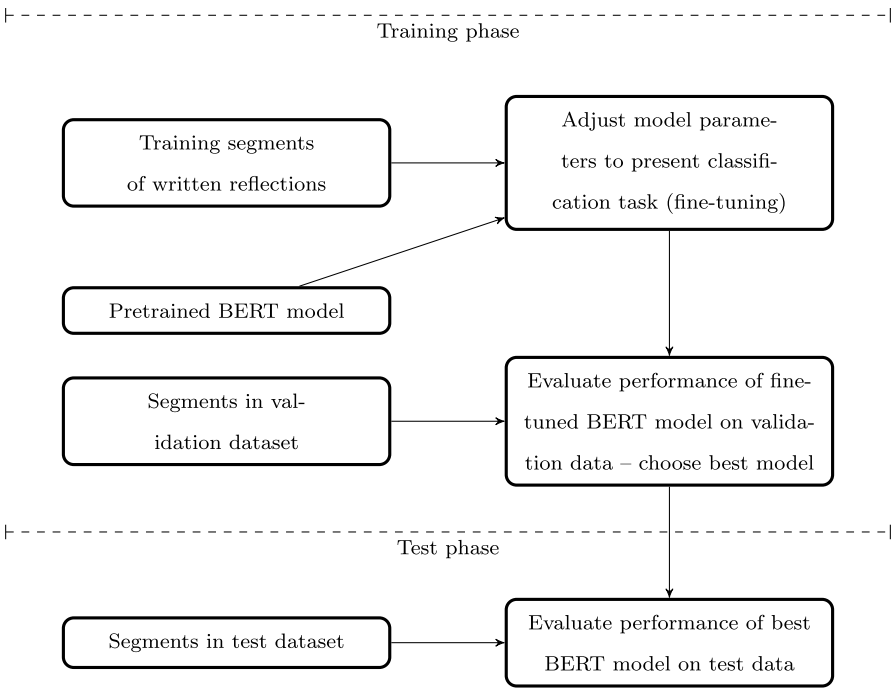

Figura 1 Procedimiento para entrenar, validar y probar el modelo BERT preentrenado. Los demás modelos (FFNN, LSTM) se validaron de forma similar, excepto que no se utilizaron pesos preentrenados.

La secuencia de entrada recibe una codificación adicional en un token especial ("[CLS]") que se añade al vocabulario. La predicción de la siguiente oración es principalmente beneficiosa para tareas como la inferencia del lenguaje natural y la respuesta a preguntas (Taher Pilehvar y Camacho-Collados,2020En general, el entrenamiento de incrustaciones BERT implica minimizar una función de pérdida entre palabras correctas o frases correctas posteriores y predicciones. Con base en los datos de entrenamiento, se entrenan las incrustaciones, las cuales pueden utilizarse para tareas posteriores, como la clasificación de segmentos (Devlin et al.,2018). La arquitectura del modelo base de BERT (en comparación con los grandes) comprende 12 capas de codificador con 768 unidades (Devlin et al.,2018; Vaswani y otros, 2017). Los hiperparámetros son fijos una vez que un investigador decide utilizar un modelo BERT preentrenado en particular.

Después del preentrenamiento, el modelo BERT se puede utilizar en una fase de ajuste fino (ver Fig.1En la fase de ajuste fino, los pesos del modelo preentrenado se ajustan a la tarea dada. Por lo tanto, se requiere menos ajuste cuando se utilizan pesos del modelo preentrenado. Además de los pesos del modelo preentrenado, la incrustación predicha para un segmento por el modelo BERT se introduce en otra capa de clasificación simple que se añade al modelo preentrenado superior (Devlin et al.,2018; Gnehm y Clematide,2020).

En el presente estudio, adoptamos un modelo de idioma alemán previamente entrenado que fue entrenado por una IA profunda.[2]Los pesos de los modelos de IA de Deepset mostraron un mejor rendimiento

[2]Detalles sobre el procedimiento de formación BERT en alemán:https://deepset.ai/german-bert(consultado: 18 dic 2020).

en comparación con los modelos BERT multilingües en tareas de lenguaje compartido (Ostendorff et al.,2019). Este modelo tiene un tamaño total de vocabulario de 30 000 que se utiliza para representar todas las palabras que aparecen en el nuevo contexto de aprendizaje (en ocasiones, los tokens desconocidos se asignan a un token especial, "[UNK]") y produce un buen rendimiento en las tareas, manteniendo al mismo tiempo el consumo de memoria y los recursos computacionales viables para una computadora personal (Ostendorff et al.,2019).

Validación cruzada e implementación técnica

En este estudio, utilizamos dos estrategias de validación cruzada: (1) validación cruzada simple con retención para evaluar el rendimiento predictivo de los modelos entrenados, y (2) validación cruzada iterada de k-fold para tener en cuenta la muestra pequeña y evaluar mejor la generalización. (1) En la validación cruzada simple con retención, el 60% de los datos (seleccionados aleatoriamente) se utilizó como datos de entrenamiento, el 20% se utilizó como datos de validación y otro 20% se utilizó como datos de prueba con retención. Se utilizaron las mismas divisiones para todos los modelos entrenados. (2) Dado el tamaño comparativamente pequeño de nuestros conjuntos de datos para aplicaciones de aprendizaje profundo, se consideró adicionalmente la validación cruzada iterada de k-fold porque permite la evaluación de modelos de aprendizaje profundo con conjuntos de datos comparativamente pequeños con la mayor precisión posible (Chollet, 2018La validación cruzada iterada de k-fold divide aleatoriamente los datos en subconjuntos de entrenamiento y validación. Por lo tanto, primero concatenamos nuestros datos originales de entrenamiento y validación (no los datos de prueba retenidos) y luego realizamos la validación cruzada iterada de k-fold en este nuevo conjunto de datos. En nuestro caso, dividimos el conjunto de datos completo diez veces. Para cada división, se evaluó el rendimiento del modelo. Este procedimiento se repitió diez veces. Los resultados se promediaron para obtener una puntuación final. Realizamos este procedimiento con los modelos de mayor rendimiento, es decir, después de ajustar los hiperparámetros.

Todos los experimentos (especialmente las mediciones de tiempo) se realizaron en una CPU Intel Core i9-10900K de 3,70 GHz y 32 GB de RAM. Las mediciones de la GPU se realizaron en una GeForce RTX 3080 de 10 GB. Además, utilizamos Python 3.8 (Python Software Foundation,2020). Para entrenar y evaluar los modelos de aprendizaje profundo utilizamos la biblioteca Pythonantorcha (Paszke y otros,2019). En particular, se utilizó el optimizador Adam en combinación con la pérdida de entrada cruzada binaria, que se implementan enantorcha.Se accedió al modelo BERT alemán preentrenado a través de cara de abrazo-biblioteca (Wolf et al.,2020).

Comparación de BERT con otras arquitecturas de aprendizaje profundo (RQ1)

Para comparar BERT con otras arquitecturas de aprendizaje profundo, ajustamos los modelos considerados (FFNN, LSTM y BERT) a los datos de entrenamiento y evaluamos su rendimiento con los datos de validación. Además, eliminamos 11 de las 12 capas de codificación de BERT y ajustamos un modelo BERT reducido (denominado: BERT (1)). Esto proporciona indicios de la importancia del número de capas de codificación en la arquitectura BERT para el rendimiento de la clasificación en esta tarea. La tarea consistía en clasificar los segmentos de las reflexiones escritas según los cinco elementos del modelo que las sustenta. Tras encontrar el modelo con mayor rendimiento mediante la optimización de hiperparámetros,

Este modelo de mayor rendimiento se ajustó a los datos de prueba reservados para evaluar la generalización.

La evaluación del desempeño de los modelos para la tarea de clasificación incluye precisión (p), recuperación (r) y la puntuación F1. Estas métricas de rendimiento explican que un clasificador debería detectar correctamente las etiquetas en los datos etiquetados por humanos. Por eso, tanto la precisión como la recuperación se centran en la detección de verdaderos positivos (Jurafsky y Martin, 2014Para un clasificador multidireccional como el del presente caso, se reportarán la precisión, la recuperación y la puntuación F1 para cada elemento, así como los valores promedio. Se reportan promedios para micro (es decir, cálculos globales de p/r/F1), macro (es decir, p/r/F1 promedio para cada elemento, sin ponderar por el apoyo de cada elemento) y ponderado (es decir, p/r/F1 promedio para cada elemento, ponderado por el apoyo de cada elemento). Las métricas de rendimiento se abrevian en adelante como p/r/F1. También calculamos el valor kappa de Cohen para comparar el rendimiento de la clasificación con una métrica ampliamente utilizada en la investigación educativa.

Para encontrar los modelos con mayor rendimiento, realizamos una búsqueda en cuadrícula a través de un espacio de hiperparámetros. Para todos los modelos, variamos las épocas (valores de 3 a 200), los tamaños de lote (valores de 3 a 50) y las tasas de aprendizaje (valores de 10).−3hasta 10−5) y tamaños ocultos (de 100 a 10 000). Se eligieron tamaños de paso adecuados para todos los hiperparámetros. Las épocas indican el número de veces que se realiza el procedimiento de ajuste de los pesos del modelo a los datos de entrenamiento. Generalmente, esto se realiza más de una vez. Sin embargo, en algún momento, el ajuste se restringe demasiado a los datos de entrenamiento, de modo que el modelo los sobreajusta y el rendimiento de la clasificación en los datos de prueba se ve afectado. El tamaño del lote indica el número de muestras de entrenamiento que se introducen simultáneamente en el modelo para generar la salida. Los tamaños más pequeños permiten más actualizaciones y una convergencia más rápida, y los valores más grandes proporcionan mejores estimaciones de los gradientes de todo el corpus (Goldberg,2017La tasa de aprendizaje se relaciona con los algoritmos de optimización y debe ajustarse, ya que valores demasiado grandes o demasiado pequeños provocarán la no convergencia de los pesos del modelo o problemas similares. Finalmente, el tamaño de la capa oculta se relaciona con la capacidad de representación. Sin embargo, valores mayores no son necesariamente mejores. Dado que BERT utiliza un vocabulario muy específico de 30 000 tokens, los FFNN y los LSTM se ajustaron con base en el vocabulario original de BERT (nombres: FFNN y LSTM) y en el vocabulario de entrenamiento original (nombres: FFNN).*y LSTM*). El razonamiento fue que el vocabulario BERT podría ser una representación generalmente ventajosa de los datos lingüísticos.

Dependencia del tamaño de los datos de entrenamiento (RQ2)

Una regla general en la clasificación de imágenes establece que se necesitan aproximadamente 1000 instancias de una clase (por ejemplo, gato o perro) para lograr un rendimiento de clasificación aceptable (Mitchell,2020). En el contexto de la evaluación de la educación científica, Ha et al. (2011), quienes utilizaron un software para calificar las explicaciones de la evolución de los estudiantes (Mayfield y Rose,2010), se encontró que aproximadamente 500 muestras eran suficientes para clasificar las respuestas en ciertas circunstancias. Sin embargo, no conocemos análisis de este tipo que evalúen la dependencia del rendimiento de la clasificación con respecto al tamaño de los datos de entrenamiento en las reflexiones escritas. En particular, no está claro, por ejemplo, si se realizó y cuándo.

El rendimiento de la clasificación para arquitecturas de aprendizaje profundo convergerá dados los conjuntos de datos comparativamente pequeños que se utilizan comúnmente en la investigación educativa disciplinaria. Por lo tanto, realizamos análisis de sensibilidad para el rendimiento de la clasificación de los datos presentes. Para ello, se extrajeron muestras aleatorias de los datos de entrenamiento con una proporción predefinida de 0,05 a 1,00. El análisis se realizó con base en los modelos de mayor rendimiento de la RQ1. El rendimiento de la clasificación se evaluó en el conjunto de datos de prueba.

Interpretación de las decisiones de clasificación de BERT (RQ3)

La interpretabilidad de los modelos de aprendizaje profundo es importante en contextos de investigación para avanzar en nuestra comprensión del funcionamiento de las reflexiones escritas (Rose,2017). Sundararajan y otros. (2017) propusieron gradientes integrados para atribuir importancia a los tokens de entrada en tareas de clasificación en una arquitectura de aprendizaje profundo. Los gradientes integrados permiten que cada característica de entrada reciba una atribución que explica su contribución a la predicción. Los gradientes integrados se calculan con referencia a una entrada de referencia neutral, como una imagen en blanco en el reconocimiento de imágenes o una palabra sin sentido en aplicaciones de PLN. Los gradientes pueden considerarse como productos de los coeficientes del modelo y los valores de las características en una red profunda.

Para los modelos de lenguaje, a los tokens de entrada definidos por el vocabulario (más exactamente, los vectores de incrustación de tokens) se les asignan atribuciones mediante gradientes integrados. Los valores de atribución permiten a los investigadores estimar la importancia de ciertos tokens para una categoría de clasificación. El token de referencia se elige como un token neutral que idealmente tiene una puntuación de predicción neutral. Los hiperparámetros importantes incluyen la longitud del vector de entrada, que sobrecarga la memoria del ordenador. En el presente estudio, el segmento más largo era excesivamente largo, por lo que se eligió un número más razonable (percentil 98 de la longitud de la oración) que incluía casi todos los segmentos, y el consumo de memoria era razonable para un ordenador personal. Otro hiperparámetro que también afecta en gran medida la asignación de memoria del ordenador es el número de pasos que se utilizan para calcular los gradientes integrados. En el presente caso, se eligieron 50 pasos, lo que parecía un número lo suficientemente grande como para evitar cálculos incorrectos, pero lo suficientemente pequeño como para ser razonable para un ordenador personal.captum-Se utilizó una biblioteca en Python para realizar estos cálculos (Sundararajan et al.,2017).

Resultados

Contraste de BERT con otros modelos de lenguaje de aprendizaje profundo para clasificar segmentos en las reflexiones escritas (RQ1)

Nuestra primera pregunta de investigación fue: "¿En qué medida un modelo de lenguaje preentrenado y afinado (BERT) supera a otros modelos de aprendizaje profundo en la clasificación de segmentos en las reflexiones escritas de docentes en formación según los elementos de un modelo que apoya la reflexión?" Para responder a esta pregunta de investigación, se analizó en qué medida el BERT afinado superó a las FFNN y las LSTM. Tabla1representa la clasificación

Tabla 1 Rendimiento del clasificador para los mejores modelos según la evaluación del conjunto de datos de validación

	FFN			FFN*			LSTM			LSTM*			BERT (1)			BERT			BERT (GPU)		
Elemento	pag	o	F1	pag	o	F1	pag	o	F1	pag	o	F1	pag	o	F1	pag	o	F1	pag	o	F1
Circunstancias	0,70	0,69	0,69	0,70	0,73	0,71	0,78	0,74	0,76	0,74	0,74	0,74	0,91	0,73	0,81	0,88	0,82	0,85	0,88	0,82	0,85
Descripción	0,72	0,69	0,71	0,74	0,66	0,70	0,78	0,78	0,78	0,74	0,84	0,78	0,79	0,86	0,82	0,87	0,88	0,87	0,87	0,88	0,87
Evaluación	0,66	0,69	0,60	0,64	0,69	0,71	0,70	0,76	0,82	0,79	0,76	0,82	0,79	0,60	0,29	0,39	0,34	0,61	0,44	0,50	0,72
Alternativas	0,74	0,75	0,76	0,74	0,75																
Consecuencias	0,49	0,36	0,41	0,57	0,33	0,42	0,72	0,47	0,57	0,64	0,51	0,57	0,75	0,56	0,64	0,67	0,60	0,63	0,67	0,60	0,63
Micro	0,63	0,63	0,63	0,63	0,63	0,63	0,72	0,72	0,72	0,72	0,72	0,72	0,77	0,77	0,77	0,82	0,82	0,82	0,82	0,82	0,82
Macro	0,67	0,68	0,74	0,73	0,72	0,79	0,77	0,78	0,79	0,77	0,78	0,64	0,63	0,62	0,65	0,63	0,64	0,73	0,72	0,72	0,72
Ponderado	0,83	0,82	0,82																		

	BERT			BERT (GPU)		
	0,60	0,59	0,58	0,69	0,67	0,67	0,70
	0,72	0,72	0,72	0,78	0,77	0,77	0,83	0,82	0,82
	0,47	0,63	0,54	0,58	0,57	0,57	0,65	0,67
	0,72	0,59	0,68	0,65	0,67	0,55	0,78	0,64	0,76

Tabla 2 Hiperparámetros, número de parámetros y tiempo de entrenamiento para los modelos con mejor rendimiento

	épocas	tamaño del lote	tasa de aprendizaje	tamaño oculto	incrustar dim	# parámetros	tiempo
FFN	10	10	0.001	200	50	14 millones	0h 04min
FFN*	200	5	0.0001	200	300	10 millones	2 h 44 min
LSTM	10	3	0.001	100	300	9 millones	0h 31min
LSTM*	3	50	0.005	100	300	3 millones	0h 08min
BERT (1)	5	50	0.0001	768	768	31 millones	0h 12min
BERT	3	3	1e-05	768	768	109 millones	1 hora y 48 minutos
BERT (GPU)	3	3	1e-05	768	768	109 millones	0h 02min

rendimiento de los mejores modelos que resultaron de la búsqueda en cuadrícula. Se puede ver que los FFNN tuvieron el rendimiento más bajo con promedios F1 ponderados de .62 y .64, respectivamente. Los FFNN fueron seguidos por los LSTM con promedios F1 ponderados de .72 y . 72, respectivamente. Además, BERT (1) tuvo un rendimiento .05 puntos menor en comparación con BERT en el promedio F1 ponderado. En general, BERT fue el modelo con mejor rendimiento para clasificar segmentos en las reflexiones escritas según el modelo de soporte de reflexión, con un promedio F1 ponderado de .82. Los valores kappa de Cohen para el rendimiento de la clasificación fueron: FFNN: 0.48, FFNN*: 0.48, LSTM: 0.56, LSTM*: 0.58, BERT (1): 0.66, BERT: 0.75.

Los modelos con mayor rendimiento se evaluaron mediante una validación cruzada iterada de k-fold para obtener estimaciones más robustas para la generalización. Nótese que los resultados se muestran en la Tabla1, dado que se realizaron con base en una validación cruzada simple de retención, probablemente estén inflados. Los promedios ponderados de F1 (desviaciones estándar) para la validación cruzada iterada de k pliegues fueron los siguientes: FFNN: 0,60 (0,07), FFNN*: 0,57 (0,17), LSTM: 0,71 (0,02), LSTM*: 0,71 (0,02), BERT (1): 0,76 (0,02), BERT: 0,81 (0,02). Para los modelos FFNN, el rendimiento disminuyó cuando se promedió sobre los pliegues. Se puede ver que las desviaciones estándar son comparativamente grandes. Los valores de la validación cruzada iterada de k pliegues para los modelos LSTM se aproximaron al hallazgo para el valor de la validación cruzada de pliegues simple. La desviación estándar también fue menor en comparación con los modelos FFNN. De igual manera, los valores de los modelos BERT también se aproximan a los valores de validación cruzada de pliegues simples con una desviación estándar comparativamente pequeña. Por lo tanto, los modelos LSTM y BERT parecen generalizar mejor que los modelos FFNN. Nuevamente, BERT superó a todos los demás modelos.

Los hiperparámetros finales para los mejores modelos en la fase de entrenamiento, el número total de parámetros para estos modelos y el tiempo de entrenamiento se pueden ver en la Tabla2Se puede observar que todas las arquitecturas tienen millones de parámetros. Los LSTM presentaron el menor número de parámetros generales. Además, los LSTM tuvieron tiempos de entrenamiento razonables. El modelo BERT completo utilizó el mayor número de parámetros. Por consiguiente, guardar el modelo ocupa la mayor cantidad de espacio en el disco duro (> 400 MB). Como era de esperar, el tiempo de entrenamiento de BERT se podría reducir en 54 veces, a un tiempo más manejable de 2 minutos, externalizando el entrenamiento a la GPU.

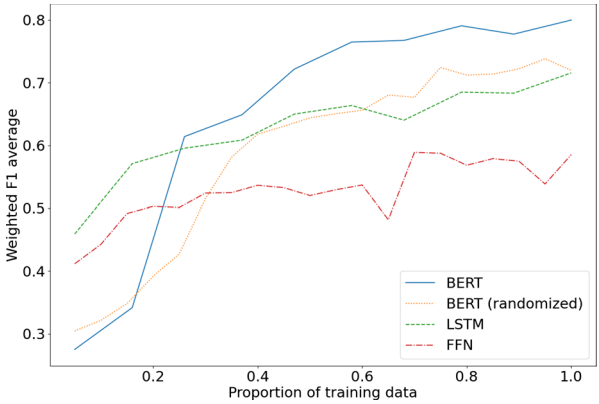

Figura 2 Análisis de la dependencia del rendimiento de la clasificación con respecto al tamaño de los datos de entrenamiento

Finalmente, el mejor modelo general se ajustó a los datos de prueba conservados. Se seleccionó el modelo BERT completamente ajustado, ya que este modelo fue el de clasificación con mayor rendimiento. Al ajustarse a los datos de prueba, BERT alcanzó un promedio ponderado F1 de 0,81. El valor kappa de Cohen fue de 0,74.

Rendimiento de la clasificación en relación con el tamaño de los datos de entrenamiento (RQ2)

Nuestra pregunta de investigación 2 fue "¿En qué medida el tamaño de los datos de entrenamiento se relaciona con el rendimiento de clasificación de los modelos de lenguaje de aprendizaje profundo?". Para analizar esta pregunta de investigación, variamos sistemáticamente el tamaño de los datos de entrenamiento. Para cada proporción, se utilizó una única extracción aleatoria de los datos de entrenamiento. Si bien múltiples extracciones aumentarían la precisión de la estimación del rendimiento, esperábamos que una sola extracción indicara tendencias más globales en el rendimiento. Figura2Muestra el rendimiento de la clasificación (medido mediante el promedio ponderado F1) con respecto al tamaño de los datos de entrenamiento (representado como proporción de los datos de entrenamiento totales). También ajustamos un modelo BERT donde modificamos aleatoriamente el orden de los tokens de entrada. Esto se realizó porque el modelo BERT también codifica la posición de las palabras. Esto nos permitió evaluar la importancia del orden de las palabras para el rendimiento de la clasificación en la tarea de clasificar segmentos según los elementos del modelo de apoyo a la reflexión.

Cifra2Indica que BERT con orden de palabras no aleatorio tiene un rendimiento inferior para datos de entrenamiento de menor tamaño en comparación con FFNN y LSTM, pero supera a todos los demás modelos en aproximadamente un 20 a 30 % del tamaño total de los datos de entrenamiento. De hecho, BERT con orden de palabras aleatorio tiene un rendimiento de clasificación inferior en comparación con BERT con orden de palabras no aleatorio. BERT con orden de palabras aleatorio se estabiliza aproximadamente en el

Rendimiento del LSTM. Se observa que todas las curvas se aplanan al aumentar el tamaño de los datos de entrenamiento.

Explicación de las decisiones del clasificador (RQ3)

Finalmente, nuestra pregunta de investigación (RQ3) fue "¿Qué características pueden explicar mejor las decisiones de los clasificadores?". Aquí, solo se considera el modelo BERT. Se calcularon gradientes integrados por capas para todas las entradas en los datos de prueba. Estos gradientes se utilizaron para asignar una puntuación de atribución a cada palabra/token en la entrada, según la contribución que esta palabra/token agregó a la predicción del segmento. Dado que BERT tiene múltiples capas de incrustación, se sumaron las puntuaciones de atribución en todas las dimensiones de incrustación para obtener una puntuación de atribución final para cada palabra/token. Ahora, queríamos extraer las palabras más importantes para los elementos en el modelo de apoyo a la reflexión. Por lo tanto, elegimos segmentos clasificados donde el resultado real y el resultado previsto eran iguales. Para cada palabra, promediamos todas las atribuciones.

Mesa3Representa los sustantivos, verbos y adjetivos más importantes para cada elemento. Tenga en cuenta que algunos tokens se dividen mediante almohadillas (#) según el tokenizador BERT (es decir, tokenización de fragmentos de palabra). Esto ocurrió cuando el tokenizador solo conocía una parte de una palabra. Por lo tanto, dividió la palabra en tokens existentes en el vocabulario BERT. Además de estos tokens divididos, la mayoría de las demás palabras son bien interpretables dada su pertenencia al elemento respectivo del modelo de apoyo a la reflexión. Tenga en cuenta, por ejemplo, el modo subjuntivo de los verbos en alternativas. Además, hay palabras de evaluación como "bueno" en evaluación. Asimismo, los sustantivos en circunstancias son algo prototípicos de ese elemento, dado que los futuros maestros debían describir la "clase", establecer sus "objetivos" y caracterizar la "lección". Otras palabras son inesperadas. Por ejemplo, "independiente" no está necesariamente relacionado con circunstancias. Sin embargo, a veces los futuros maestros señalaron que esta era su primera lección "independiente" de sus mentores o similar.

Discusión

El análisis de la escritura reflexiva utiliza aprendizaje automático y procesamiento del lenguaje natural (PLN) para extraer contenido de productos del pensamiento reflexivo, como las reflexiones escritas. El estudio presentado impulsa la investigación en análisis de la escritura reflexiva al aplicar un modelo de lenguaje preentrenado (BERT) para mejorar el rendimiento de la clasificación e identificar elementos de un modelo que apoya la reflexión en los segmentos de las reflexiones escritas.

En este estudio, los segmentos de las reflexiones escritas de los futuros profesores de física fueron identificados por evaluadores humanos de acuerdo con un modelo de apoyo a la reflexión (Nowak et al., 2019). Se entrenó un modelo de lenguaje preentrenado (BERT) para clasificar segmentos, basándose en anotaciones humanas. El entrenamiento se refiere al ajuste de los pesos del modelo en el modelo de lenguaje preentrenado para reproducir las asignaciones de entrada-salida. La calidad de esta asignación se calcula con base en una función de pérdida y los pesos del modelo se ajustan mediante un algoritmo de optimización. El rendimiento de clasificación del modelo de lenguaje preentrenado en los datos de validación se comparó con el de los modelos de uso generalizado.

Tabla 3 Palabras con mayor atribución en cada categoría

Elemento	Sustantivos	Auxiliares	Verbos	Adjetivos
Circunstancias	objetivo, clase, ##objetivo, lección, # # intentar)	tenido, tener, fueron, fue, será	decidido, querido, determinado, previsto, decidido	bueno, te(que), físico(a), independiente, teorético
Descripción	preguntó, se esforzó, minutos, ##i(t), Tuve que	ha sido, han sido, tenido, han, será	dijo, vino, notó, puso, debería	repetido, Posteriormente, Juntos, diferente, siguiente
Evaluación	expectativa, hablar(ed), juzgar, ver, agradable)	era, eran, soy, tenía, es	esperado, perc(eive), encontrado, puede, ayudó	bueno, exitoso, justificable, inseguro, Bien
Alternativas	alternativa, alternativas, habría por ejemplo, necesitar	sería, habría, querría, querría ser, sería	podría, podría, puede, debería, seguro	Alternativamente, Mejor, mejor, posible, más
Consecuencias	consecuencia, sentimiento, ##me(dios), tomar, siguiente	voluntad, había, soy, tengo, es	debería, debería, ver, podría, puede	importante, importante, mejor, personal, bien

Las letras mayúsculas se refieren al comienzo de las oraciones, y el tokenizador BERT utiliza hashes para dividir tokens desconocidos en tokens conocidos.

Algoritmos de aprendizaje profundo (FFNN, LSTM). El modelo BERT completo obtuvo el mejor rendimiento (promedio ponderado de F1 de 0,82). Posteriormente, se ajustó BERT a los datos de prueba retenidos y se obtuvo un promedio ponderado de F1 de 0,81, lo que se relaciona con un valor kappa de Cohen de 0,74. Este hallazgo concuerda con investigaciones recientes en PLN que indican que los modelos de lenguaje basados en transformadores, como BERT, pueden mejorar el rendimiento de la clasificación en diversas tareas relacionadas con el lenguaje y generalizar de forma fiable a datos no vistos (Devlin et al.,2018; Taher Pilehvar y Camacho Collados,2020). Además, demostramos que la ventaja del modelo BERT completo se manifiesta en aproximadamente el 20 al 30 % del tamaño de los datos de entrenamiento. Esto concuerda con la observación de que el uso de modelos lingüísticos preentrenados reduce la necesidad de datos de entrenamiento para lograr un buen rendimiento de clasificación (Devlin et al.,2018Estos hallazgos sugieren que los modelos lingüísticos preentrenados, como BERT, podrían ser aplicables a diversas tareas en el análisis de la escritura reflexiva y más allá, donde solo se dispone de cantidades limitadas de datos. Además, el orden de las palabras en los segmentos parece ser relevante para que el modelo BERT supere a los LSTM en rendimiento de clasificación. Esto podría deberse a que el preentrenamiento en BERT implica enmascarar palabras en las entradas y predecir las palabras enmascaradas a partir del contexto. Asimismo, BERT utiliza incrustaciones posicionales (contextuales) para los tokens de entrada, lo que podría aumentar la importancia del orden de las palabras en la secuencia de entrada (Taher Pilehvar y Camacho-Collados,2020). Finalmente, buscamos interpretar la decisión de clasificación del modelo BERT completo. En consecuencia, se utilizaron gradientes integrados en capas para calcular las atribuciones de las entradas que indicaban en qué medida un token en la entrada contribuía a la clasificación. Las palabras recuperadas con altas atribuciones fueron bien interpretables en términos de los elementos respectivos en el modelo de apoyo a la reflexión. Los gradientes integrados parecen proporcionar una herramienta valiosa para abrir la caja negra de los modelos de aprendizaje profundo y comprender mejor las decisiones del modelo. Por ejemplo, la presencia de palabras mayoritariamente positivas en el elemento de evaluación podría indicar que los futuros docentes tienden a evaluar su propia enseñanza, con el riesgo de consolidar sus propias creencias previas, lo cual se ha reportado en investigaciones previas (Mena-Marcos et al.,2013). Como tal, los gradientes integrados podrían proporcionar a los investigadores herramientas analíticas para las reflexiones escritas.

Limitaciones y mejoras

Varios detalles de implementación del entrenamiento del modelo y del análisis de datos limitan la generalización de nuestros hallazgos. Por ejemplo, elegimos implementaciones específicas de la representación de datos, el procedimiento de optimización y la función de pérdida. Estas decisiones se basaron en parte en investigaciones previas (Devlin et al.,2018; Wulff y otros,2020). Sin embargo, no podemos descartar la posibilidad de que implementaciones alternativas de la representación de datos (p. ej., lematización o eliminación de palabras vacías), procedimientos de optimización (p. ej., descenso de gradiente estocástico) o diferentes funciones de pérdida (p. ej., pérdida cuadrática media) hubieran resultado en un mejor rendimiento de la clasificación. Además, el modelo BERT ajustado se utilizó con un vocabulario fijo de 30 000 palabras. En consecuencia, algunas palabras se dividieron artificialmente en tokens sin significado (p. ej., de «experimento» a «experimento»). Sin embargo, la palabra «experimento» es importante en las reflexiones escritas relacionadas con la física. Nuevos modelos preentrenados que incorporan el vocabulario presentado en el conjunto de datos de entrenamiento.

Podría ser ventajoso, ya que se podría prestar más atención a los límites de palabras y a la estructura de las reflexiones escritas. Esto requeriría que BERT incluyera estas palabras importantes en el vocabulario con el que está entrenado previamente. Los avances en...cara abrazadaLa biblioteca permite en particular el preentrenamiento de modelos de lenguaje basados en transformadores.

Otra limitación se relaciona con la segmentación de las reflexiones escritas. La segmentación se realizó principalmente con base en oraciones, como en estudios previos (Carpenter et al., 2020; Ferreira y otros,2013; Wulff y otros,2020). Una base teórica más sólida, como la teoría del discurso, podría mejorar la segmentación (Stede,2016). Por ejemplo, la teoría de la segmentación del discurso puede ayudar a dividir segmentos basándose en decisiones basadas en principios, en lugar de en la relación o los límites de las oraciones. La segmentación puede tener un gran impacto en el rendimiento de la clasificación (Rosé et al.,2008). Aunque métodos como la segmentación de oraciones arrojaron resultados aceptables en aplicaciones anteriores (Wulff et al.,2020), una base más sólida para la segmentación permitiría a los modelos codificar mejor las dependencias dentro del lenguaje en relación con los elementos del modelo que sustenta la reflexión. Sin embargo, también es importante que la segmentación basada en oraciones pueda automatizarse fácilmente hoy en día con alta precisión, lo que motiva su uso, ya que los datos invisibles pueden segmentarse sin intervención humana. Al mismo tiempo, la segmentación de oraciones asume que toda la información relevante para clasificar un segmento en un elemento se encuentra dentro de la oración. Esta suposición no siempre es cierta, dada la investigación revisada sobre las dependencias de largo alcance que caracterizan el lenguaje (Ebeling y Neiman, 1995). Una forma de resolver potencialmente esta cuestión es a través de múltiples modelos que atiendan diferentes aspectos de las reflexiones escritas.

Finalmente, es importante reconocer los límites del lenguaje y la comunicación humana a través de textos. Las suposiciones y ciertos niveles de conocimiento del mundo rara vez se mencionan explícitamente en los textos (Jurafsky y Martin,2014; McNamara y otros,1996). Más bien, se asume que el lector posee cierta cantidad de conocimiento del mundo común (Jurafsky y Martin,2014; McNamara y otros,1996El conocimiento implícito y no declarado plantea problemas a los algoritmos que toman la información de entrada sin más, ya que solo se utilizan los hechos mencionados explícitamente para la clasificación de segmentos. Además, los evaluadores humanos eventualmente resuelven ambigüedades o correferencias que no se mencionan explícitamente en los textos y que el algoritmo informático no puede utilizar. En consecuencia, la concordancia entre humanos y computadoras probablemente se vea obstaculizada en ciertos contextos, y los métodos de validación que se basan en ella están sesgados.

Trascendencia

Dadas estas limitaciones, nuestros hallazgos tienen implicaciones para la investigación educativa relacionada con las reflexiones escritas y para la implementación de herramientas educativas que utilizan el análisis automatizado de dichas reflexiones. Nuestros hallazgos sugieren que los modelos lingüísticos preentrenados pueden mejorar el rendimiento de la clasificación de segmentos de reflexiones escritas. A partir de esta mejora en el rendimiento de la clasificación, se vislumbran diversas aplicaciones para el análisis automatizado.

Investigaciones futuras profundizarán en las características de la arquitectura del modelo BERT y el régimen de entrenamiento. Por ejemplo, Ostendorff et al. (2019) emplearon metadatos para tareas de clasificación con BERT. Podría ser útil incluir covariables relacionadas con el autor, como calificaciones de cursos, años de experiencia docente, etc., para mejorar el rendimiento de la clasificación o la evaluación posterior. Además, se podrían considerar características textuales adicionales de los segmentos codificados para la clasificación. Por ejemplo, la posición espacial del segmento en la reflexión escrita y la longitud del texto son dos opciones significativas para covariables adicionales que eventualmente mejoran el rendimiento de la clasificación. Además, se ha demostrado que los modelos de lenguaje generativo como GPT-3 son capaces de aprender nuevas tareas con solo unos pocos ejemplos (Brown et al.,2020Estas capacidades podrían utilizarse en el contexto del análisis de la reflexión escrita. Por ejemplo, los modelos generativos de lenguaje podrían utilizarse para generar consecuencias dada la evaluación de una situación docente. Ciertamente, una consecuencia debería atender la descripción y evaluación de la situación docente. El avance en esta línea de investigación permitiría eventualmente el modelado de procesos complejos para generar reflexiones escritas. La capacidad de modelar procesos complejos se atribuyó a una importante ventaja de los enfoques de datos algorítmicos, como el aprendizaje automático (ML), en comparación con los modelos de datos más tradicionales (Breiman, 2001).

Una advertencia importante en el contexto del análisis de la escritura reflexiva es el vínculo faltante entre la reflexión de calidad y la mejora del desempeño en el aula (Clarà, 2015En cuanto al valor pedagógico de las reflexiones escritas, es necesario investigar la validez externa del modelo que las sustenta. Proponemos que los modelos entrenados en este estudio puedan ayudar a evaluar la eficacia de las reflexiones escritas en los programas de formación docente. Con la ayuda de los modelos entrenados (en particular, BERT), las reflexiones de los futuros profesores de física pueden analizarse de forma escalable en cuanto a la integridad, frecuencia y estructura de los elementos del modelo. A partir de una implementación a gran escala, las relaciones entre la escritura reflexiva y variables como el conocimiento adquirido por los estudiantes o métricas similares pueden explorarse cuantitativamente a mayor escala.

Finalmente, los métodos de aprendizaje profundo presentados pueden ayudar a construir herramientas de retroalimentación fiables y analíticas que proporcionen retroalimentación instantánea para una reflexión escrita. Dados los requisitos para que los futuros docentes aprendan de sus experiencias docentes (Korthagen,1999), así como las limitaciones y el coste de los recursos humanos en la formación docente universitaria (por ejemplo, Nehm y Härtig2012), estos sistemas serían muy necesarios, especialmente dada su eficacia para mejorar los resultados del aprendizaje (Aleven et al.,2016; Chirikov y otros,2020;VanLehn,2011). Los modelos de lenguaje preentrenados pueden desempeñar un papel importante en el desarrollo de dichas herramientas de retroalimentación (Chirikov et al.,2020). BERT podría utilizarse para clasificar automáticamente segmentos en las reflexiones escritas de los futuros profesores de física y reportar estadísticas descriptivas a los futuros profesores e instructores. Las estadísticas descriptivas podrían relacionarse con preguntas como si todos los elementos del modelo que sustenta la reflexión se abordaron en el texto y en qué medida. La autorreflexión también es un elemento importante en sistemas de tutoría inteligentes como AutoTutor o Crystal Island (Carpenter et al.,2020; Graesser y otros, 2005) que utilizan el lenguaje natural como medio para facilitar el aprendizaje (Nye et al.,2014). Carpenter y otros. (2020) utilizaron con éxito representaciones de palabras contextualizadas profundas (ELMo) para representar y predecir la profundidad reflexiva de los estudiantes. El método que apoya la reflexión...

El modelo y el modelo BERT entrenado en nuestro estudio podrían usarse además de estos hallazgos para también conceptualizar y predecir la amplitud reflexiva de la escritura de los estudiantes en sistemas como Crystal Island.

La transferencia de sistemas de tutoría inteligente al contexto de la formación docente en prácticas presentaría un gran potencial para mejorar el aprendizaje experiencial. La profundidad y amplitud reflexiva podrían evaluarse con modelos lingüísticos preentrenados (ELMo, BERT). Se podrían proporcionar pautas y sugerencias a los docentes en prácticas que hayan demostrado ser eficaces en estudios previos (Lai y Calandra,2010). Se relacionarían con las experiencias docentes de los docentes y proporcionarían retroalimentación analítica. Por ejemplo, se sabe en el análisis de la escritura reflexiva que los docentes tienden a ser evaluativos en su descripción de una situación de enseñanza (Mann et al.,2007). Un sistema de tutoría basado en computadora que se basa en el modelo BERT entrenado (o las representaciones ELMo) podría identificar oraciones evaluativas dentro de un párrafo descriptivo y proporcionar sugerencias para reescribirlo.

FondosFinanciación de acceso abierto facilitada y organizada por Projekt DEAL. Este proyecto forma parte de la «Qualitätsoffensive Lehrerbildung», una iniciativa conjunta del Gobierno Federal y los Länder que busca mejorar la calidad de la formación docente. El programa está financiado por el Ministerio Federal de Educación e Investigación. Los autores son responsables del contenido de esta publicación.

Disponibilidad del código Por favor, póngase en contacto con el primer autor.

Declaraciones

Conflicto de intereses Sin conflictos de intereses.

Referencias

Abels, S. (2011). LehrerInnen como 'Practicante reflexivo': Reflexionskompetenz für einen demokratieförderlichen Naturwissenschaftsunterricht, (1. ed. Aufl.). Wiesbaden: VS Verl. für Sozialwiss. Aeppli, J. y Lötscher, HL (2016). EDAMA - Un modelo Rahmen para la reflexión.*Mensajes para profesores Formación de niños y maestros*,*34*(1), 78–97.

Aleven, V., McLaughlin, EA, Glenn, RA y Koedinger, KR (2016). Instrucción basada en adaptativa. Tecnologías de aprendizaje. En RE. Mayer PA Alexander (Eds.)*Manual de investigación sobre aprendizaje e instrucción, Manual de psicología educativa*(págs. 522–560). Taylor y Francis, Florence. Bain, J. D., Ballantyne, R., Packer, J. y Mills, C. (1999). Uso de la escritura de diarios para mejorar la enseñanza de los estudiantes. Reflexividad de los estudiantes durante sus prácticas de campo.*Los profesores y la enseñanza*,*5*(1), 51–73.

Bain, JD, Mills, C., Ballantyne, R. y Packer, J. (2002). Desarrollo de la reflexión sobre la práctica a través del diario. Escritura oral: impactos de las variaciones en el enfoque y el nivel de retroalimentación.*Los profesores y la enseñanza*,*8*(2), 171–196.

Berliner, DC (2001). Aprendiendo sobre y aprendiendo de profesores expertos.*Revista Internacional de Educación-Investigación nacional*,*35*, 463–482.

Breiman, L. (2001). Modelado estadístico: Las dos culturas.*Ciencia estadística*,*16*(3), 199–231. Brown, TB, Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ..., y Amodei, D. (2020). Lan-

Los modelos de lenguaje son aprendices de pocas oportunidades. arXiv.

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Bass, R. y McWilliams, M. (2017). Hacia la reflexión Análisis de la escritura creativa: fundamento, metodología y resultados preliminares.*Revista de análisis del aprendizaje*,*4*(1), 58–84.

Carlson, J., Daehler, K., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., ... y Wilson, CD (2019). El modelo de consenso refinado del conocimiento del contenido pedagógico. En A. Hume, R. Cooper y A. Borowski (Eds.)*Reposicionar el conocimiento del contenido pedagógico en el conocimiento profesional docente*. Singapur: Springer.

Carpenter, D., Geden, M., Rowe, J., Azevedo, R. y Lester, J. (2020). Análisis automatizado de la media Reflexiones escritas de estudiantes durante el aprendizaje basado en juegos. En II Bittencourt, M. Cukurova, K. Muldner, R. Luckin y E. Millán (Eds.)*Inteligencia artificial en la educación*(págs. 67–78). Cham: Springer International Publishing.

Chan, KKH, Xu, L., Cooper, R., Berry, A. y van Driel, JH (2021). La observación docente en la educación científica. ción: ¿Ves lo que yo veo?*Estudios en Educación Científica*,*57*(1), 1–44.

Cheng, G. (2017). Hacia un sistema de clasificación automática para apoyar el desarrollo de habilidades críticas. Habilidades reflexivas en el aprendizaje de L2.*Revista Australasiana de Tecnología Educativa*,*33*(4), 1–21.

Chirikov, I., Semenova, T., Maloshonok, N., Bettinger, E. y Kililcec, RF (2020). educación en línea

Las plataformas amplían la instrucción STEM universitaria con resultados de aprendizaje equivalentes a un menor costo.*Avances científicos, 6*.

Chollet, F. (2018). Aprendizaje profundo con Python. Shelter Island, NY: Manning. Recuperado dehttp://proquest.safaribooksonline.com/9781617294433.

Clarà, M. (2015). ¿Qué es la reflexión? Buscando claridad en una noción ambigua.*Revista del Profesor Educación*,*66*(3), 261–271.

Clarke, D., y Hollingsworth, H. (2002). Elaboración de un modelo de crecimiento profesional docente.*Enseñanza y formación docente*,*18*(8), 947–967.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ..., & Stoyanov, V. (2019). Aprendizaje de representación interlingüística no supervisada a escala. arXiv:1911.02116Darling-Hammond, L., Hammerness, K., Grossman, PL, Rust, F. y Shulman, LS (2017). El diseño

de programas de formación docente. En L. Darling-Hammond y J. Bransford (Eds.)*Preparando a los docentes para un mundo cambiante*. Nueva York: John Wiley & Sons.

Devlin, J., Chang, M.-W., Lee, K. y Toutanova, K. (2018). BERT: Preentrenamiento de la función bidireccional profunda. Transformadores para la comprensión del lenguaje. arXiv:1810.04805.

Dewey, J. (1933). Cómo pensamos: Una reformulación de la relación entre el pensamiento reflexivo y el proceso educativo. proceso ((Nueva ed.) Ed.). Boston usw.: Heath.

Ebeling, W., y Neiman, A. (1995). Correlaciones de largo alcance entre letras y oraciones en textos.*Física-ica A: Mecánica estadística y sus aplicaciones*,*215*(3), 233–241.

Fenstermacher, G. (1994). Capítulo 1: El conocedor y lo conocido: La naturaleza del conocimiento en la investigación Sobre la enseñanza. Revista de Investigación en Educación 20.

Ferreira, R., de Souza Cabral, L., Lins, RD, Pereira e Silva, G., Freitas, F., Cavalcanti, GD, ..., & Favaro, L. (2013). Evaluación de técnicas de puntuación de oraciones para el resumen textual extractivo.*Sistemas expertos con aplicaciones*,*40*(14), 5755–5764.

Fischer, HE, Borowski, A., Kauertz, A. y Neumann, K. (2010). Fachdidaktische Unterrichtsforschung: Unterrichtsmodelle und die Analyse von Physikunterricht.*Zeitschrift für Didaktik der Naturwissenschaften*,*16*, 59–75.

Gibson, A., Kitto, K. y Bruza, P. (2016). Hacia el descubrimiento de la metacognición del aprendiz desde la reflexión. escribiendo.*Revista de análisis del aprendizaje*,*3*(2), 22–36.

Gnehm, A.-S., y Clematide, S. (2020). Zonificación y clasificación de textos para anuncios de empleo en Alemania. Hombre, francés e inglés: Actas del Cuarto Taller sobre Procesamiento del Lenguaje Natural y Ciencias Sociales Computacionales. ACL. Recuperado dehttps://www.aclweb.org/antología/2020.nlpcss-1.10.pdf.

Goldberg, Y. (2017). Métodos de redes neuronales para el procesamiento del lenguaje natural. Morgan y Claypool.

Goodfellow, I., Bengio, Y. y Courville, A. (2016).*aprendizaje profundo*Cambridge, Massachusetts y Londres don, Inglaterra: MIT Press.http://www.deeplearningbook.org/.

Graesser, AC, Chipman, P., Haynes, BC y Olney, A. (2005). AutoTutor: Un sistema de tutoría inteligente. sistema con diálogo de iniciativa mixta. *Transacciones IEEE sobre Educación*, *48*(4), 612–618. Grossman,
PL, Compton, C., Igra, D., Ronfeldt, M., Shahan, E. y Williamson, PW (2009). Enseñanza Práctica: una perspectiva interprofesional. *Registro del Teachers College*, *111*(9), 2055–2100.
Ha, M., Nehm, RH, Urban-Lurain, M. y Merrill, JE (2011). Aplicación de modelos de puntuación computarizados. de explicaciones biológicas escritas en los distintos cursos y universidades: perspectivas y limitaciones. *Educación en Ciencias de la Vida del CBE*, *10*(4), 379–393.
Hascher, T. (2005). Die Erfahrungsfalle. *Revista para Lehrerinnen- und Lehrerbildung*, *5*(1), 39–45. Hatton, N., y
Smith, D. (1995). Reflexión en la formación docente: Hacia la definición y la implementación. *Enseñanza y formación docente*, *11*(1), 33–49.
Jurafsky, D., y Martin, JH (2014). Procesamiento del habla y del lenguaje (2.ª ed., Pearson nueva ed. internacional). ed.). Harlow: Pearson Educación.
Kahneman, D. (2012). Schnelles Denken, langsames Denken. Siedler Verlag.
Kolb, D. (1984). *Aprendizaje experiencial: La experiencia como fuente de aprendizaje y desarrollo*. Englewood Acantilados, Nueva Jersey: Prentice Hall.
Korthagen, FA (1999). Vinculando la reflexión y la competencia técnica: El cuaderno de bitácora como instrumento en formación docente. *Revista Europea de Formación del Profesorado*, *22*(2-3), 191–207.
Korthagen, F. A. (2001). *Vinculando la práctica y la teoría: La pedagogía de la formación docente realista*. Mahwah, Nueva Jersey: Erlbaum.http://www.loc.gov/catdir/enhancements/fy0634/00057273-d.html.
Korthagen, FA (2005). Niveles de reflexión: la reflexión central como medio para potenciar el crecimiento profesional. *Los profesores y la enseñanza*, *11*(1), 47–71.
Korthagen, FA, y Kessels, J. (1999). Vinculando la teoría y la práctica: Cambiando la pedagogía de la formación docente. ción. *Investigador Educativo*, *28*(4), 4–17.
Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G. y Dawson, S. (2018). Comprender las autorreflexiones de los estudiantes a través del análisis del aprendizaje: LAK '18, 7 al 9 de marzo de 2018, Sídney, Nueva Gales del Sur, Australia, págs. 389-398.
Lai, G., y Calandra, B. (2010). Examinando los efectos de los andamiajes informáticos en la reflexión de los profesores noveles. Escritura de diario activa. *Investigación y desarrollo de tecnología educativa*, *58*(4), 421–437. LeCun,
Y., Bengio, Y., y Hinton, G. (2015). Aprendizaje profundo. *Naturaleza*, *521*(7553), 436–444.
Levin, DM, Hammer, D., y Coffey, JE (2009). Atención de los profesores noveles al pensamiento de los estudiantes. *Revista de Formación docente*, *60*(2), 142–154.
Lin, X., Hmelo, CE, Kinzer, C. y Secules, T. (1999). Diseño de tecnología para apoyar la reflexión. *Educa- Investigación y desarrollo de tecnología internacional*, *47*(3), 43–62.
Loughran, J., y Corrigan, D. (1995). Portafolios docentes: una estrategia para desarrollar el aprendizaje y la enseñanza en educación previa al servicio. *Maestros y formación docente*, *11*(6), 565–577.
Luo, W. y Litman, D. (2015). Resumen de las respuestas de los estudiantes a las preguntas de reflexión. Actas del Conferencia de 2015 sobre métodos empíricos en el procesamiento del lenguaje natural, 1955-1960.
Mann, K., Gordon, J. y MacLeod, A. (2007). Reflexión y práctica reflexiva en la formación de profesionales de la salud. ción: una revisión sistemática. *Avances en la educación en ciencias de la salud*, *14*(4), 595.
Mayfield, E., y Rose, CP (2010). Una herramienta interactiva para el análisis de errores en la minería de texto: Proceedings de la Asociación Norteamericana de Lingüística Computacional (NAACL) HLT 2010: Sesión de demostración, Los Ángeles, CA, junio de 2010, 25–28.
McNamara, D., Kintsch, E., Butler Songer, N. y Kintsch, W. (1996). ¿Los buenos textos son siempre mejores? interacción ciones de coherencia textual, conocimientos previos y niveles de comprensión en el aprendizaje del texto. *Cognición e instrucción*, *14*(1), 1–43.
Mena-Marcos, J., García-Rodríguez, M.-L., y Tillema, H. (2013). Escritura reflexiva de estudiantes de magisterio: ¿Qué ¿Lo revela? *Revista Europea de Formación del Profesorado*, *36*(2), 147–163. Mitchell, M.
(2020). Inteligencia artificial: Una guía para humanos pensantes. Pelican Books.
Nehm, RH, y Härtig, H. (2012). Diagnóstico humano vs. computacional del conocimiento de selección natural de los estudiantes: Prueba de la eficacia del software de análisis de texto. *Revista de Educación Científica y Tecnología*, *21*(1), 56–73.

Neuweg, GH (2007). ¿Wie grau ist alle Theorie, wie grün des Lebens goldner Baum? LehrerInnenbildung im Spannungsfeld von Theorie und Praxis. bwpat 12.
Nowak, A., Kempin, M., Kulgemeyer, C. y Borowski, A. (2019). Reflexión von Physikunterricht [Reflexión- ción de lecciones de física]. En C. Maurer (Ed.) *Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe: Jahrestagung in Kiel 2018. Regensburg: Gesellschaft für Didaktik der Chemie und Physik* (pág. 838).

Nowak, A., Liepertz, S. y Borowski, A. (2018). Reflexionskompetenz von Praxissemesterstudierenden Soy Fach Physik. En C. Maurer (Ed.)*Qualitätsvoller Chemie- und Physikunterricht- normativa und empirische Dimensionen: Jahrestagung in Regensburg 2017. Universität Regensburg*.

Nye, BD, Graesser, AC y Hu, X. (2014). AutoTutor y su familia: Una revisión de 17 años de lenguaje natural. tutoría.*Revista Internacional de Inteligencia Artificial en Educación*,*24*(4), 427–469.

Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G. y Gipp, B. (2019). enriquecedor BERT con incrustaciones de gráficos de conocimiento para la clasificación de documentos. arXiv:1909.08402v1.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... y Chintala, S. (2019). PyTorch: Una biblioteca de aprendizaje profundo de alto rendimiento y estilo imperativo. En H. Wallach, H. Larochelle, A. Beygelzimer, d'Alché-Buc, E. Fox y R. Garnett (Eds.).*Avances en los sistemas de procesamiento de información neuronal 32*.http://papers.neurips.cc/paper/9015-pytorch-un-estilo-imperativo-alto-rendimiento-aprendizaje-profundo-biblioteca.pdf(págs. 8024–8035). Curran Associates, Inc.

Paterson, BL (1995). Desarrollo y mantenimiento de la reflexión en revistas clínicas.*La educación en enfermería hoy, 15*(3), 211–220.

Poldner, E., van der Schaaf, M., Simons, PR-J., van Tartwijk, J. y Wijngaards, G. (2014). Evaluación de los futuros docentes Escritura reflexiva a través del análisis de contenido cuantitativo.*Revista Europea de Formación del Profesorado*,*37*(3), 348–373. Pratt, L., y Thrun, S. (1997).*Aprendizaje automático, 28*(5).

Fundación del Software Python. (2020). Referencia del lenguaje Python: versión 3.8.http://www.python.orgRodgers, C. (2002). Definiendo la reflexión: Otra mirada a John Dewey y el pensamiento reflexivo.*Profesores Col-Registro de lectura*,*104*(4), 842–866.

Roe, MF y Stallman, AC (1994). Un estudio comparativo de diarios de diálogo y respuesta.*Enseñanza y Formación docente*,*10*(6), 579–588.

Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., y Fischer, F. (2008). Análisis Procesos de aprendizaje colaborativo automáticos: aprovechando los avances de la lingüística computacional en el aprendizaje colaborativo asistido por computadora.*Revista internacional de aprendizaje colaborativo asistido por computadora*,*3*(3), 237–271.

Rose, CP (2017). Un giro social al análisis del lenguaje.*Naturaleza*,*545*, 166–167.

Schoenfeld, AH (2014). ¿Qué contribuye a unas aulas eficaces y cómo podemos apoyar a los docentes en su creación? ¿Ellos? Una historia de investigación y práctica, entrelazadas productivamente.*Investigador Educativo*,*43*(8), 404–412. Schön, DA (1983).*El profesional reflexivo: cómo piensan los profesionales en acción*. Nueva York: Basic Books. http://www.loc.gov/catdir/enhancements/fy0832/82070855-d.html.

Schön, DA (1987).*Educar al profesional reflexivo: hacia un nuevo diseño para la enseñanza y el aprendizaje en las profesiones*, 1.a ed. San Francisco, California: Jossey-Bass.

En M. Stede (Ed.) (2016).*Anotación de texto del manual: Potsdamer Kommentarkorpus 2.0*. Potsdam: Universidad Editorial Potsdam.

Sundararajan, M., Taly, A. y Yan, Q. (2017). Atribución axiomática para redes profundas: Actas de la 34ª Conferencia internacional sobre aprendizaje automático, Sídney, Australia, PMLR 70.

Swales, JM (1990).*Análisis de género: el inglés en el ámbito académico y de investigación*. Cambridge: Cambridge Prensa universitaria.

Taher Pilehvar, M. y Camacho-Collados, J. (2020). Incorporaciones en el procesamiento del lenguaje natural: teoría y Avances en la representación vectorial del significado. Morgan y Claypool.

Taylor, WL (1953). "Procedimiento de cierre": Una nueva herramienta para medir la legibilidad.*Revista trimestral de periodismo*

Ullmann, TD (2017). Análisis de la escritura reflexiva: Palabras clave de la reflexión escrita determinadas empíricamente: Actas de la Séptima Conferencia Internacional sobre Análisis del Aprendizaje y Conocimiento, LAK '17*Serie de actas de conferencias internacionales de la ACM*, 163–167.

Ullmann, TD (2019). Análisis automatizado de la reflexión en la escritura: Validación de enfoques de aprendizaje automático. *Revista Internacional de Inteligencia Artificial en Educación*,*29*(2), 217–257.

Ullmann, TD, Wild, F. y Scott, P. (2012). Comparación de textos reflexivos detectados automáticamente con textos humanos. juicios. En*Actas del 2º taller sobre sensibilización y reflexión en el aprendizaje potenciado por la tecnología (AR-TEL 12), 18 de septiembre de 2013*(págs. 101–116). Saarbrücken, Alemania.

van Es, E., y Sherin, MG (2002). Aprender a observar: andamiaje de las interpretaciones del aula por parte de los nuevos docentes. interacciones.*Revista de tecnología y formación docente*,*10*(4), 571–596.

VanLehn, K. (2011). La eficacia relativa de la tutoría humana, los sistemas de tutoría inteligente y otros sistemas de tutoría, (Vol. 46 págs. 197–221).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN, ... y Polosukhin, I. (2017). Atención es todo lo que necesitas: Conferencia sobre sistemas de procesamiento de información neuronal.*Avances en sistemas de procesamiento de información neuronal*, 6000–6010.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ..., y Rush, A. M. (2020). HuggingFace's Transformers: procesamiento del lenguaje natural de última generación. arXiv.

Wulff, P., Buschhüter, D., Nowak, A., Westphal, A., Becker, L., Robalino, H. y Borowski, A. (2020). Com- Clasificación basada en computadora de reflexiones escritas de futuros profesores de física.*Revista de Educación Científica y Tecnología*.

Zanette, D. (2014). Patrón estadístico en el lenguaje escrito. arXiv:1412.3336.

Zeichner, KM (2010). Repensando las conexiones entre los cursos en el campus y las experiencias prácticas en la universidad. y la formación docente en el ámbito universitario.*Revista de formación docente*,*61*(1-2), 89–99.

Disponibilidad de datos y material Por favor, póngase en contacto con el primer autor.

Nota del editor Springer Nature se mantiene neutral con respecto a los reclamos jurisdiccionales en mapas publicados y afiliaciones institucionales.