Universidad de San Andrés

Escuela de Educación

Licenciatura en Ciencias de la Educación

En busca de las *buenas* preguntas

Una aplicación de *machine learning* a
la investigación educativa

Autora: Karina Elizabeth Serfaty

Legajo: 23.240

Mentora: Melina Furman

Victoria, 19 de Abril de 2019

Agradecimientos

Gracias a Melina por su paciencia infinita y por darme la libertad de explorar caminos poco convencionales.

A Romi por acompañarme en todo, por leerme y por ser soldada de mi disciplina cuando yo no lo era.

A mis compañeros/as de clase de quienes aprendí una infinidad durante los años de facultad.

Índice

Resumen

La importancia de la formulación de preguntas para la tarea docente y para el aprendizaje ha sido ampliamente estudiada y validada en la tradición constructivista y en las ciencias de la educación en general. Entender, por ejemplo, qué tipos de preguntas formulan los docentes en clase o cuánto tiempo transcurre hasta que las auto-responden o realizan nuevas puede ayudarnos a comprender mejor cómo pensar la formación docente inicial y continua de modo de fortalecer las prácticas de enseñanza. Por su parte, el análisis de clases apoyándose en herramientas tecnológicas para recolectar datos, tales como sensores que detectan la posición de la mirada de los estudiantes, *softwares* adaptativos que registran y retroalimentan las respuestas de los alumnos a una tarea o herramientas de grabación de clases capaces de identificar diferentes dinámicas en el aula, están comenzando a brindar en los últimos años herramientas novedosas y anteriormente impensables a investigadores y a los propios docentes para el análisis y la gestión de la enseñanza y el aprendizaje y nos presentan potenciales nuevos escenarios para la investigación y la práctica que resulta importante conocer y poner en debate.

En el marco de este campo denominado analíticas de la enseñanza es que nos proponemos llevar adelante el siguiente trabajo que tiene por fin el desarrollo de una herramienta que permita clasificar las preguntas que realizan los docentes en el aula en preguntas abiertas o cerradas, una categoría valiosa para comprender los diálogos que se generan en el marco de una situación de enseñanza-aprendizaje. Para hacerlo utilizaremos inteligencia artificial o más precisamente técnicas de análisis de lenguaje natural. Estas técnicas utilizan *machine learning,* un subcampo de la inteligencia artificial interesada en la aplicación y desarrollo de algoritmos que ayuden a las máquinas a aprender (Segaran, 2007), para crear clasificaciones de objetos como imágenes, palabras u oraciones (o en este caso de preguntas) sin que se necesite de intervención humana para realizar la clasificación. Nos proponemos crear una herramienta que pueda servir como insumo para el análisis de clases y que pueda ir evolucionando de forma autónoma a lo largo del tiempo.

Para llevar a cabo este análisis se optó por un modelo de aprendizaje supervisado que consistió en la clasificación previa de 757 preguntas formuladas por docentes en clases reales y 251 preguntas elaboradas inspiradas en las preguntas reales observadas. Para esta clasificación se tomaron de referencia y simplificaron las tipificaciones de preguntas propuestas por Anijovich y Mora (2010), Burbules (1999) y Martens (1999), categorizándolas en abiertas o cerradas. Con los tipos de preguntas ya clasificados se

entrenó a un modelo sobre el que luego se aplicaron preguntas nuevas para entender con qué porcentaje de acierto era capaz de clasificarlas. Los resultados que obtuvo la herramienta al clasificar preguntas reales formuladas en una clase de Ciencias Naturales de la Provincia de Buenos Aires fueron: 82,91% de seguridad al clasificar preguntas cerradas y 74,82% al clasificar preguntas abiertas. Estos resultados son alentadores para el uso de esta herramienta en la investigación de las prácticas docentes.

Introducción

Preguntas de investigación y objetivos

En los últimos diez años se han incrementado la cantidad de trabajos que aplican inteligencia artificial y técnicas estadísticas al análisis de cómo los niños y niñas aprenden y la tarea docente. Como describe Ferguson (2012) en su notorio trabajo “Learning analytics: drivers, developments and challenges” (Analíticas del aprendizaje: conductores, desarrollos y desafíos) (2012), en donde busca recopilar los avances de este campo en los últimos treinta años, el aumento de este tipo de análisis se propulsó por la incorporación de plataformas tecnológicas para acompañar a la enseñanza utilizando aulas y cursos virtuales, como los *moodles* (Ferguson, 2012), a los que podemos definir como plataformas online para la administración de cursos (About Moodle, n.d.). Según la autora, con el surgimiento de este tipo de herramientas vino aparejado un crecimiento en la cantidad de datos acerca de cómo interactúan los estudiantes, su información personal e información académica (Mazza & Milani, 2004; Romero et al, 2008).

La necesidad de dar sentido a los datos recolectados por estos nuevos sistemas dio lugar al nacimiento del minado de datos educativos cuyo propósito consiste en “desarrollar métodos para explorar los tipos de datos particulares que provienen de entornos educativos, usándolos para entender mejor a los estudiantes, y los entornos en donde aprenden” (International Educational Data Mining Society, n.d.). Esta disciplina se enfoca, entre otras cosas, en el trazado de relaciones entre los datos recolectados y la predicción de resultados (Ferguson, 2012). El propósito del minado de datos tiene, en resumen, un fuerte interés por comprender y mejorar los entornos educativos (Ferguson, 2012). Este interés se observa con claridad en la meta que esta disciplina se propone: “transformar estudiantes en estudiantes mejores y más efectivos” (Zaîane, 2001 como se cita en Ferguson, 2012, p.4).

Con el surgimiento de las plataformas de cursos masivos en línea (MOOCs) y gracias a los avances en la aplicación de técnicas estadísticas al análisis de los datos educativos generados por los *moodles* nace un segundo campo de estudio denominado *analíticas del aprendizaje*. Este campo se encarga de “la medición, recolección, análisis y reporte de data sobre los estudiantes y sus contextos, con el propósito de entender y optimizar el aprendizaje y el entorno en el que ocurre” (Siemens, 2010 como se cita en Ferguson, 2012 p.9). Este campo se caracteriza por la utilización de técnicas de *machine learning* e

inteligencia artificial para la construcción de modelos que sirvan para identificar tendencias y patrones en el comportamiento de los estudiantes (Mor, Ferguson & Wasson, 2015 como se cita en Gunn, Mcdonald, Donald, Milne, & Blumenstein, 2017).

Si bien las análiticas del aprendizaje nacen con el propósito de mejorar las experiencias de aprendizaje de los estudiantes en línea rápidamente se comienzan a volcar también hacia el análisis de las dinámicas que suceden en el aula. Como mencionan Mitri et al (2018) resulta "innegable que es en el aula de clases (...) [en dónde] todavía se llevan a cabo el volumen más grande de actividades para el aprendizaje" (Mitri et al, 2018, p.1). De la necesidad de recolectar data acerca de cómo se desenvuelve una clase en donde no hay dispositivos tecnológicos es que surge un campo denominado *análisis multimodal del aprendizaje*. Este campo se ocupa de la recolección de data en una dinámica de clase en el "mundo físico" utilizando cámaras de video, grabaciones, sensores y otros instrumentos (Prieto & Dillenbourg, 2016).

Buscando comprender mejor cómo guiar a los estudiantes hacia aprendizajes más significativos es que surgen las *analíticas de la enseñanza* cuyo propósito es el de "aplicar técnicas de las analíticas del aprendizaje para entender mejor los procesos de enseñanza y aprendizaje, y que eventualmente se puedan hacer intervenciones para sugerir cómo afianzarlos" (Prieto & Dillenbourg, 2016, p.1). Tal como sostienen Prieto y Dillenbourg (2016), para poder realizar tales intervenciones es necesario primero comprender qué sucede en el aula. Es aquí en donde toman relevancia las técnicas de análisis multimodal que buscan extraer información del aula que pueda ser analizable con grabaciones de clase y otros dispositivos como sensores.

Algunos ejemplos interesantes de trabajos de las analíticas de la enseñanza son, por ejemplo, la experiencia de Prieto y Dillenbourg (2016) en la que se busca identificar, utilizando *machine learning,* los distintos momentos de la clase, es decir, cuándo el docente está desarrollando un tema o realizando preguntas y cuándo los estudiantes contestan o están llevando a cabo actividades. Uno de los argumentos esgrimidos por los autores a la hora de justificar su análisis es que, si bien los trabajos de este tipo, enfocados en grabar u observar clases y tratar de extraer aprendizajes de estas observaciones tienen una larga tradición en las ciencias de la educación, en general las tareas de transcribir e identificar estos momentos son tareas que consumen mucho tiempo para el investigador y que podrían agilizarse con una herramienta que lo hiciera automáticamente y, tal vez, con mayor eficiencia (Prieto & Dillenbourg, 2016).

Es en este contexto en el que surge este trabajo. Si bien existen algunas experiencias recientes realizadas en escuelas en donde se busca identificar cuándo el docente realiza preguntas a lo largo de su clase (Prieto & Dillenbourg, 2016), parece relevante un análisis de qué tipos de preguntas se realizan y especialmente si es posible clasificarlas utilizando las técnicas que utilizan las analíticas de la enseñanza. Nos proponemos, por lo tanto, responder a las siguientes preguntas:

- ¿Es posible desarrollar un modelo o herramienta que sea capaz de identificar preguntas abiertas y cerradas en el aula utilizando *machine learning*?
- ¿Con qué porcentaje de acierto el modelo sería capaz de clasificar preguntas abiertas y cerradas nuevas que nunca haya visto?

Con el objetivo de responder a estas preguntas es que nos proponemos desarrollar una herramienta que permita clasificar las preguntas que los docentes formulan en clase en preguntas abiertas o cerradas utilizando *machine learning*. Definimos *machine learning* como una rama de la inteligencia artificial interesada en el desarrollo de algoritmos que permiten a las computadoras aprender (Segaran, 2007). El interés por utilizar inteligencia artificial para este análisis surge de la posibilidad de poder desarrollar una herramienta que consuma menos tiempo al investigador y que pueda ir evolucionando de manera autónoma a lo largo del tiempo y mejorando su desempeño.

Habiendo presentado el tema de este trabajo en el próximo apartado nos encargaremos de describir las razones por las que resulta relevante un análisis de los tipos de preguntas que los docentes formulan en el aula y las diferencias que tienen para el aprendizaje las preguntas abiertas y cerradas.

Justificación

A menudo se ha sostenido que una buena enseñanza está ligada a las preguntas que los docentes realizan en el aula, especialmente en el marco de la escuela constructivista. Las preguntas son una de las tantas herramientas mediante las que los docentes buscan "despertar el interés de los alumnos, verificar si comprendieron, promover la reflexión, estimular el establecimiento de relaciones entre distintos conceptos" (Anijovich & Mora, 2010, p.37). Es también a través de ellas que buscan establecer un diálogo con sus alumnos que pueda guiarlos hacia descubrimientos nuevos (Burbules, 1999 como se cita en Anijovich & Mora, 2010).

Muchas de las preguntas formuladas tradicionalmente en el aula por los docentes han tenido el foco puesto en que los alumnos puedan recordar o revisitar contenido ya aprendido y eso se ha considerado, a menudo, un éxito en el aprendizaje (Martens, 1999). La escuela constructivista considera, por su parte, que se arriba a aprendizajes más significativos formulándoles a los alumnos preguntas para pensar, que los ayuden a hilar contenido o a sacar nuevas conclusiones. Las "preguntas productivas", como las llama Martens, son un ejemplo de ellas y ayudarían a los estudiantes a "impulsar su pensamiento hacia adelante" (Martens, 1999, p.25).

No podemos dejar de mencionar, sin embargo, que investigaciones recientes han cuestionado la efectividad de las preguntas por sí solas y han puesto el foco en el modo y el contexto en que se formulan (Cotton, 2001). Lo que se ha encontrado es que no solo las preguntas productivas o de orden cognitivo superior son fundamentales a la hora de ayudar a los estudiantes a comprender mejor sino también una serie de factores entre los que se encuentran: si las preguntas habilitan al diálogo o no, el tiempo que los docentes esperan hasta auto-responderlas o formular nuevas preguntas y la cantidad y la forma en que se presentan estas preguntas (en formato escrito u oral) (Cotton, 2001). También se han encontrado diferentes resultados cuando varían los contextos en que se realizan y la edad o qué tan avanzados están los estudiantes en el conocimiento (Cotton, 2001).

Volviendo a los tipos de preguntas que podrían conducir a aprendizajes más significativos, la bibliografía en general las divide en dos grupos: las preguntas de orden cognitivo superior o y las de orden cognitivo inferior. Podemos definir a las preguntas de orden cognitivo inferior como las preguntas en las que se solicita al estudiante "simplemente recordar textualmente o en sus propias palabras el material previamente leído o enseñado por el docente" (Cotton, 2001, p.4). Algunos ejemplos de estas preguntas son las que se utilizan para entender qué sabe un estudiante sobre un tema: ¿cuál es la masa de un cierto elemento?, ¿cómo se define al átomo?, preguntas con una única respuesta correcta posible cuya respuesta requiere recuperar información.

Las preguntas de orden cognitivo superior, por su parte, se definen como las que requieren al estudiante "manipular mentalmente trozos de información aprendidos previamente para crear una respuesta nueva o para apoyar una respuesta con evidencia razonada lógicamente" (Cotton, 2001, p.4). Estas preguntas en general se corresponden con los niveles más altos de la Taxonomía de Bloom (comprensión, aplicación, análisis, síntesis y evaluación) (Cotton, 2001) y son las que vinculamos a mejores aprendizajes. Ejemplos de estas preguntas son las que impulsan a los estudiantes a hacer predicciones sobre un tema: si un papel flota sobre el agua, ¿qué piensan que sucedería si colocamos una moneda en

lugar de un papel?, las que proponen un problema a resolver: ¿cómo podríamos averiguar si una planta es un ser vivo?, las que buscan que los estudiantes razonen sobre un tema: ¿por qué razón piensan que el agua congelada es más “dura” que el agua en estado líquido?, entre otras.

Muy allegadas a las anteriores son las preguntas abiertas, las que se definen como aquellas que admiten más de una respuesta posible y que “contribuyen a que los alumnos aprendan a pensar” (Anijovich & Mora, 2010, p.39). Por el contrario, las preguntas cerradas son aquellas de respuesta única que ayudan a los docentes a “verificar un acuerdo, o bien se las emplea como recursos retórico al devenir de una conversación” (Anijovich & Mora, 2010, p.39). Esta clasificación de preguntas se distingue de la anterior en tanto esta hace referencia a la forma de las preguntas.

Como dijimos, las preguntas también se utilizan como una herramienta para fomentar el diálogo. En este contexto, este se define como “una actividad dirigida al descubrimiento y a una comprensión nueva, que mejora el conocimiento, la inteligencia o la sensibilidad de los que forman parte de ese diálogo” (Burbules, 1999, como se cita en Anijovich & Mora, 2010, p.37). Siguiendo a Benlloch (1992), es a través de la expresión verbal de lo que los niños piensan sobre un fenómeno o situación que son capaces de entender desde qué punto parten en la construcción de sus ideas (Benlloch, 1992 como se cita en Gellon, Rosenvasser Feher, Furman & Golombek, 2005).

En síntesis, la formulación de preguntas es una vía fundamental a través de la cual los docentes buscan establecer un diálogo con sus alumnos y alumnas e intentan fomentar en ellos el pensamiento. Por la relevancia y el lugar central que ocupan en los procesos de enseñanza-aprendizaje es que parece relevante desarrollar una herramienta como la que proponemos que busque simplificar su identificación y clasificación.

Estado del arte

A nivel nacional se encontraron algunas experiencias en la investigación de las analíticas del aprendizaje y el minado de datos educativos en el Instituto de Investigación en Informática (LIDI) de la Universidad de La Plata. Los trabajos encontrados se encuentran abocados fundamentalmente al análisis del aprendizaje y los estudiantes. Algunos ejemplos de estas investigaciones son las que se enfocan en el análisis de grandes conjuntos de datos de los estudiantes de la Universidad de La Plata para identificar patrones de interacción con las plataformas educativas y poder contribuir a su mejoramiento (Díaz et al, 2015) y el estudio de sistemas de recomendación de recursos educativos para estudiantes y docentes (Díaz et al, 2014). No se encontraron experiencias enfocadas específicamente en el reconocimiento de preguntas formuladas por docentes en el aula.

En el contexto internacional, por otro lado, se encontraron algunas experiencias en el análisis de la tarea docente y específicamente en el estudio del reconocimiento de diferentes dinámicas áulicas, lo que incluye la identificación de preguntas. Por fuera del ámbito educativo se encontraron numerosos estudios abocados a la identificación de tipos de preguntas utilizando inteligencia artificial, fundamentalmente para el desarrollo de sistemas inteligentes de preguntas y respuestas. Describiremos brevemente, a continuación, algunos de estos trabajos.

En el marco del estudio del reconocimiento de dinámicas áulicas, que tiene por objetivo reconocer cuando los docentes realizan preguntas durante las clases sin la intervención de un investigador, podemos mencionar los trabajos de Prieto y Dillenbourg (2016). Esta línea de investigación busca, mediante distintas estrategias, registrar las clases dictadas por los docentes, transcribirlas e identificar los momentos en que docentes y alumnos dialogan. Posteriormente se intenta identificar cuándo los docentes están formulando preguntas. Estas investigaciones se abocan fundamentalmente al desarrollo y mejoramiento de los dispositivos que se utilizan para registrar las clases y realizan una categorización o *taggeo* manual de las diferentes dinámicas áulicas para luego poder reconocer cuando se trata de preguntas.

Se encontraron otras experiencias con objetivos similares en una línea de investigación llevada a cabo en las universidades de Notre Dame, Memphis y Wisconsin-Madison. El objetivo de estos trabajos es identificar cuando los docentes realizan preguntas “no retóricas, no procedimentales y no de manejo del discurso” (Blanchard, Melo & Oley, 2015,

p.283). Para llevarlo a cabo, la metodología que se utilizó fue la de grabar clases y reconocer manualmente los diferentes momentos en que los docentes estaban realizando preguntas. A diferencia de los trabajos llevados a cabo por Prieto y Dillenbourg en este caso no se utilizaron transcripciones de clase sino que el *taggeo* o categorización se hizo sobre las grabaciones de audio. Con esta clasificación se construyó un modelo capaz de reconocer exitosamente los momentos en que un docente realiza preguntas de importancia para el desarrollo de la clase y para el aprendizaje.

Otro ejemplo de inteligencia artificial aplicada a las preguntas docentes son los sistemas de tutelaje inteligentes como Autotutor, capaces de formular preguntas para guiar el aprendizaje de los estudiantes (Graesser et al, 2018). Puntualmente la herramienta Autotutor se propone tres objetivos: (1) evaluar el conocimiento, habilidades, acciones y otros estados psicológicos de los estudiantes de un grupo tomando como base sus acciones y su conversación, (2) generar cambios en el discurso que sean sensibles a los estados psicológicos y los problemas a resolver y (3) ayudarlos a alcanzar la solución a un problema (Graesser et al, 2018). Graesser, Dowell, Hampton, et al mencionan en su trabajo "Building intelligent conversational tutors and mentors for team collaborative problem solving" (Construyendo tutores y mentores inteligentes y conversacionales para la resolución de problemas en equipo colaborativamente) (2018), que tradicionalmente la forma en que estos sistemas trabajan es interactuando con una única persona. En su trabajo los autores exploran la posibilidad de utilizar estos sistemas de tutelaje inteligentes en entornos con muchos estudiantes para guiarlos en el trabajo colaborativo (Graesser et al, 2018).

Otras experiencias en el uso de sistemas de tutelaje inteligente son aquellas que intentan, por medio de la formulación de preguntas, conducir la enseñanza de un tema o habilidad. Podemos observar un ejemplo de esto en la línea de investigación llevada adelante por Shi, Lippert, et al (2018) en la que los investigadores se proponen enseñar y evaluar comprensión lectora en adultos mayores utilizando Autotutor. La forma de llevar esto a cabo fue buscando que los estudiantes conversen con la herramienta. Analizando estas conversaciones luego se estimó la comprensión de los estudiantes y sus habilidades lectoras.

Dentro del campo de las ciencias de la computación, por otra parte, el desarrollo de herramientas capaces de reconocer tipos de preguntas es muy estudiado por sus implicancias para el desarrollo de sistemas inteligentes de preguntas y respuestas. Uno de los trabajos más citados en este sentido corresponde a un estudio llevado a cabo por el equipo de investigación de IBM en el que se desarrolló un sistema capaz de competir con

una persona en el juego de preguntas y respuestas *Jeopardy!* (Ferrucci, et al, 2010). Este tipo de trabajos, entre los que también podemos citar el de Simmons (1970) se enfocan en características gramaticales de las preguntas para clasificarlas y, en general, para poder entrenar una herramienta capaz de contestarlas. Algunas de las clasificaciones tomadas son, por ejemplo, si las preguntas comienzan con 'qué' o 'cuándo'. El objetivo de la herramienta, en estos casos, es comprender qué tipo de respuesta debe dar en cada caso (una definición en el caso de 'qué', por ejemplo, y una fecha en el caso de 'cuándo').

Si bien los trabajos mencionados se enfocan, en muchos casos, en el reconocimiento de preguntas e incluso en su clasificación, no se encontraron trabajos que tomen para el análisis una categorización elaborada dentro de las ciencias de la educación como la que proponemos para este trabajo. Se trata de un campo incipiente en plena expansión sobre el que resulta interesante aportar.

Marco teórico

En este apartado nos proponemos definir algunos de los conceptos centrales de este trabajo de tesis.

Inteligencia artificial

Podemos definir a la inteligencia artificial como a “la teoría y el desarrollo de sistemas informáticos capaces de realizar tareas que normalmente requieren inteligencia humana, como la percepción visual, el reconocimiento de voz, la toma de decisiones y la traducción entre idiomas.” (Artificial intelligence, n.d.). La inteligencia artificial (o AI) es, en resumen, una rama de las ciencias de la computación que permite el desarrollo de programas que funcionen sin intervención humana.

En particular, en relación a las ciencias de la educación la aplicación de la inteligencia artificial tiene una larga trayectoria y suele presentar, para algunos autores, algunos desafíos. Gros y Rodríguez Illera (1991) sostienen en su trabajo “Inteligencia artificial y diseño de programas educativos” que las dos críticas más importantes que han surgido en relación a la implementación de inteligencia artificial a las ciencias de la educación son (1) los relacionados a los supuestos teóricos de quienes desarrollan los sistemas sobre educación y (2) los relacionados a “las dificultades técnicas para implementar adecuadamente algunos de los logros de la AI” (Gros & Rodríguez Illera, 1991, p.40).

En relación a esto último múltiples autores coinciden en que los campos como el minado de datos, las analíticas de enseñanza y aprendizaje y los análisis multimodales tienen el desafío de trazar relaciones con las teorías educativas para dar sentido a los datos recolectados. En palabras de Worsley (2014): “las analíticas de aprendizaje multimodal trabajan para aprovechar los avances en la captura y procesamiento de señales de datos para resolver los desafíos que presenta estudiar una variedad de constructos complejos relevantes para el aprendizaje como los observados en ambientes de aprendizaje.” (Worsley, 2014, p.214).

La inteligencia artificial se divide a su vez en un serie de sub-disciplinas de acuerdo a las herramientas que utiliza, el objetivo que se propone, entre otros. Algunas de estas son la

robótica, el *machine learning* y el procesamiento de lenguaje natural. En este trabajo en especial haremos uso de las últimas dos para abordar el problema propuesto.

Machine learning

Podemos definir a esta disciplina como "a un subcampo de la inteligencia artificial interesada en la aplicación y desarrollo de algoritmos que ayuden a las máquinas a *aprender*" (Segaran, 2007, p.3). En la práctica esto implica que los algoritmos sean capaces de recibir y procesar una serie de datos para luego poder inferir qué características tienen en común. Posteriormente, y a raíz de esas inferencias, estos algoritmos deberían ser capaces de clasificar información nueva, incluso sin haberla procesado antes (Segaran, 2007). Según Segaran (2007), esto es posible debido a que la mayor parte de la información presenta patrones y estos patrones permiten al algoritmo realizar generalizaciones.

A diferencia de cómo se llevaría a cabo un análisis de este tipo sin el uso de *machine learning* estos patrones no deben ser analizados en detalle previamente por una persona, aunque sí necesitan de una clasificación preliminar en algunos casos. Al respecto de esto existen dos maneras de abordar el problema de la clasificación dentro del *machine learning*: un modelo de aprendizaje supervisado y de aprendizaje no supervisado.

Se define a los modelos de **aprendizaje supervisado** como a una serie de "técnicas que usan ejemplos de *inputs* y *outputs* para aprender cómo hacer predicciones (...). Las aplicaciones que usan estos métodos "aprenden" al examinar una serie de *inputs* y de *outputs* esperados. Cuando queremos extraer información usando [este] método, ingresamos una serie de *inputs* y esperamos que la aplicación produzca un *output* basándose en lo que aprendió hasta ese momento." (Segaran, 2007, p.29). En la práctica esto implica tomar una serie de datos y clasificarlos manualmente de acuerdo a alguna característica que creamos que ese conjunto tiene en común. Este proceso es el primer paso en la limpieza de datos y se denomina *tagging* o etiquetado de la información.

Un **aprendizaje no supervisado**, por otro lado, "no [se] entrena con ejemplos de respuestas correctas. Su propósito es encontrar una estructura dentro de una serie de datos (...). El objetivo de [este tipo de algoritmos] es el de tomar la data y encontrar los grupos distintivos que existen dentro de ella." (Segaran, 2007, p.30). La forma en que trabaja este tipo de algoritmos es recibiendo una gran cantidad de datos y realizando una categorización en base a patrones. El trabajo del investigador, en estos casos, suele ser el de dar sentido a

las categorizaciones que realizó la máquina para poder posteriormente afinar la clasificación.

Debido al volumen de datos con que se contaba a la hora de realizar este trabajo y a la complejidad que presentan los modelos no supervisados se optó por un enfoque supervisado en el que el primer paso consistió en categorizar las preguntas con que se contaba. Profundizaremos más sobre la metodología de trabajo en el capítulo siguiente.

Procesamiento de lenguaje natural (NLP)

La expresión *lenguaje natural* se utiliza en las ciencias de la computación para referirse a "lenguajes que se usan para la comunicación diaria entre humanos" (Bird, Klein & Loper, 2009, p.5). Lenguajes como el español y el inglés son ejemplos de lenguaje natural. Se considera que estos lenguajes "al contrario de los lenguajes de programación (...) han evolucionado de generación en generación y es difícil de precisar las reglas explícitas con que funcionan" (Bird, Klein & Loper, 2009, p.5).

Se denomina procesamiento de lenguaje natural, por su parte, a cualquier tipo de manipulación por computadora que se pueda hacer del lenguaje natural. Ejemplos de esto son la contabilización de palabras en un corpus de texto, la comparación de estilos de escritura de distintas personas y la comprensión de oraciones complejas por parte de las computadoras (Bird, Klein & Loper, 2009).

En un modelo de aprendizaje supervisado, como el que utilizaremos en este análisis, existen una serie de pasos que comúnmente se llevan a cabo al procesar lenguaje natural. Este proceso consta de dos etapas: (1) entrenamiento y (2) predicción. Durante la primera etapa del proceso el primer paso consiste en categorizar o *taggear* los inputs de texto para introducirlos al algoritmo de *machine learning*. Durante la segunda parte del proceso este algoritmo construirá un modelo en base a los inputs y sus categorías y luego será capaz de realizar predicciones utilizando este modelo. Es decir que será capaz de categorizar o *taggear* inputs nuevos de texto siguiendo los patrones de las categorías e inputs con que lo entrenamos en la primera parte (Bird, Klein & Loper, 2009).

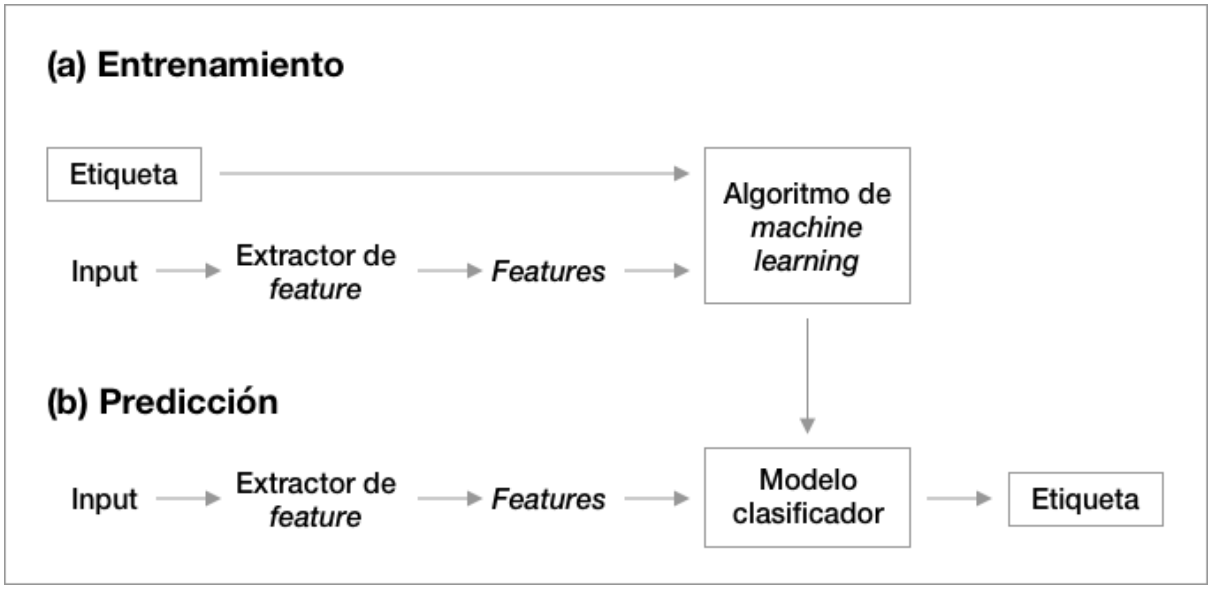

Gráfico en Bird, Klein & Loper (2009)

En este trabajo el proceso de categorizar los inputs será realizado manualmente y luego se aplicarán algoritmos de *machine learning* para construir los modelos.

Preguntas abiertas y cerradas

Hemos definido en apartados previos los tipos de preguntas que intentaremos identificar en este trabajo. En este apartado profundizaremos acerca de la categorización que seleccionamos (preguntas abiertas o cerradas) y presentaremos algunas otras categorizaciones posibles presentes en la literatura especializada partiendo por describir las de Martens (1999), Anijovich y Mora (2010) y Burbules (1999). Describiremos por último las subcategorías de análisis que utilizaremos para identificar a las preguntas.

Martens, en su trabajo “Preguntas productivas: herramientas para soportar un aprendizaje constructivista”, denomina a las preguntas que conducen a aprendizajes más significativos como *preguntas productivas* y las define como aquellas que conducen a la actividad física o mental y que son capaces de llevar a los estudiantes hacia adelante en su pensamiento (es decir, promueven el pensamiento de orden superior) (Martens, 1999). Contrarias a estas preguntas son, para la autora, aquellas que ponen el foco en que los estudiantes recuerden o rememoren contenido que se supone aprendieron previamente o, como se mencionó, que involucran el pensamiento de orden inferior.

La autora divide, a su vez, a las preguntas productivas en los siguientes grupos: de atención, de medición y conteo, de comparación, de acción, de planteo de problemas y de razonamiento. A continuación definiremos a cada una de estas categorías:

- *Preguntas de atención:* aquellas que ayudan a los estudiantes a poner el foco sobre un detalle importante del contenido o la actividad, por ejemplo: ¿qué pudieron observar qué pasa cuando se coloca un papel sobre agua?.

- *Preguntas de medición y conteo:* aquellas que ayudan a los estudiantes a ser más precisos en sus observaciones, por ejemplo: ¿cuántos papeles hay que apilar sobre una superficie de agua para que estos dejen de flotar?

- *Preguntas comparativas:* aquellas que ayudan a los estudiantes a analizar y clasificar, por ejemplo: ¿cuál es la diferencia entre apoyar un papel sobre agua y apoyar una moneda?

- *Preguntas de acción*: aquellas que ayudan a que los estudiantes exploren propiedades de materiales con los que no están familiarizados o que hagan predicciones sobre los fenómenos, por ejemplo: ¿qué pasa si en lugar de apoyar un papel sobre agua coloco una moneda?

- *Preguntas de planteo de problemas:* aquellas que ayudan a los estudiantes a planear e implementar soluciones a problemas, como por ejemplo: ¿qué deberíamos hacer para que una moneda flote sobre una superficie líquida?

- *Preguntas de razonamiento:* aquellas que ayudan a los estudiantes a pensar sobre sus experiencias y construir ideas sobre sus aprendizajes. Un ejemplo de estas últimas es: ¿por qué creen que una moneda no flota en el agua?

La segunda clasificación que mencionaremos es la propuesta por Anijovich y Mora (2010). Las autoras sostienen que las preguntas pueden clasificarse “de acuerdo al nivel de pensamiento que intentan estimular” (Anijovich & Mora, 2010, p.38) y las dividen en: preguntas sencillas, de comprensión, de orden cognitivo superior y metacognitivas.

- *Preguntas sencillas:* aquellas que esperan respuestas breves o una única respuesta correcta, como por ejemplo: ¿en qué año se firmó la declaración de la independencia?

- *Preguntas de compresión:* aquellas que estimulan el procesamiento de información para las que el estudiante necesita “pensar, recolectar datos, clasificar, comparar, etc.” (Anijovich & Mora, 2010, p.38). Un ejemplo de este tipo de preguntas es: ¿qué diferencia hay entre una propaganda y una publicidad?

- *Preguntas de orden cognitivo superior*: aquellas que “demandan respuestas que exigen interpretar, predecir o evaluar críticamente” (Anijovich & Mora, 2010, p.38),

por ejemplo: ¿qué piensan que pasaría si dejamos una naranja guardada en una alacena por muchos días?

- *Preguntas metacognitivas:* aquellas que ayudan a los estudiantes a pensar sobre su propio pensamiento y a que comprendan cómo arribaron a determinados resultados o conclusiones. Un ejemplo de estas preguntas es: si tuvieran que aconsejar a un compañero sobre cómo resolver este ejercicio, ¿qué le dirían?

La última clasificación que presentaremos es la de Burbules (1999). El autor divide a las preguntas que se realizan en el aula en dos tipos: las *preguntas de aplicación del pensamiento convergente* y las de aplicación del *pensamiento divergente*.

- *Preguntas de aplicación del pensamiento convergente*: aquellas que se dirigen a una única respuesta, como por ejemplo: ¿qué características tienen los mamíferos?

- *Preguntas de aplicación del pensamiento divergente*: aquellas que buscan una multiplicidad de respuestas, como por ejemplo: ¿por qué piensan que se produjo determinado fenómeno?

Las clasificaciones presentadas comúnmente suelen agruparse en categorías más amplias, relacionadas por lo general con el tipo de pensamiento (de orden superior o inferior) que promueven. Si bien las categorías anteriores resultan relevantes para analizar los tipos de preguntas que los docentes formulan en sus clases, en este trabajo optamos por una categorización más simple que pudiera servir de punto de partida para el desarrollo de la herramienta esperada. Siguiendo a Anijovich y Mora (2010) sostendremos que las preguntas pueden clasificarse de acuerdo a su forma en preguntas abiertas y cerradas y elegiremos dichas categorías para el presente trabajo. Las *preguntas cerradas* serán aquellas que "son útiles para verificar un acuerdo, o bien, se las emplea como recursos retóricos en el devenir de una conversación" (Anijovich y Mora, 2010, p.39). Las *preguntas abiertas*, por otro lado, "son aquellas que contribuyen a que los alumnos aprendan a pensar" (Anijovich & Mora, 2010, p.39).

En este trabajo utilizaremos la clasificación de *preguntas abiertas y cerradas* y, siguiendo la definición de Anijovich y Mora (2010) y una adaptación de las demás clasificaciones previamente presentadas, dividiremos a las preguntas abiertas en preguntas de (1) atención, (2) de medición y conteo, (3) de comparación, (4) de acción, (5) de planteo de problemas, (6) de razonamiento y (7) metacognitivas. Dividiremos a su vez a las preguntas cerradas en preguntas de: (1) verificación o memoria, y (2) procedimentales. Definiremos a las preguntas de "verificación o memoria" como aquellas que admiten una única respuesta

correcta posible y cuyo objetivo es verificar que los alumnos la recuerdan, por ejemplo: ¿en qué año Cristóbal Colón llegó a América? y a las preguntas “procedimentales” como aquellas que repercuten al funcionamiento del aula o de la actividad, por ejemplo: ¿hicieron los deberes?

Metodología

En este apartado nos proponemos describir la metodología de estudio utilizada partiendo por describir (1) el proceso que se utilizó para evaluar, entrenar y testear el modelo, (2) las características del conjunto de datos utilizado, (3) los criterios utilizados para la categorización y la limpieza de los datos, (4) la herramienta de análisis de lenguaje natural y los algoritmos que se emplearon, (5) la categorización de preguntas utilizada y (5) los límites del análisis.

Proceso de construcción del modelo

En el apartado anterior describimos brevemente los pasos que se llevan a cabo en un análisis de tipo supervisado utilizando *machine learning*. En este apartado describiremos cómo se llevaron a cabo esos pasos en este análisis.

El primer paso del análisis consistió en la recolección y transcripción de clases en español, tomando en consideración especialmente la transcripción de las preguntas realizadas por los docentes. En esta transcripción registró también el género de él o la docente y características demográficas de la clase como la nacionalidad, nivel educativo y asignatura.

El segundo paso consistió en la categorización de las preguntas transcriptas. Para esto se utilizó una doble categorización. En primer lugar se clasificaron en preguntas abiertas o cerradas y luego se reclasificaron de acuerdo a subcategorías más específicas: (1) de atención, (2) de medición y conteo, (3) de comparación, (4) de acción, (5) de planteo de problemas, (6) de razonamiento y (7) metacognitivas, (8) verificación o memoria y (9) procedimentales.

El tercer paso del análisis implicó entrenar a un modelo con las preguntas ya categorizadas. Para esto se utilizó una herramienta denominada Natural Language Processing de Google al que se importó el dataset de datos para su entrenamiento. El dataset utilizado es el que comúnmente se conoce como "*traning set*" y tiene por propósito construir el modelo tomando como base los criterios que definimos al hacer la categorización.

El siguiente paso consistió en probar el modelo construido con un nuevo conjunto de preguntas. Este dataset es un conjunto de datos no categorizados que el modelo debía clasificar correctamente. Este dataset comúnmente se denomina "*development set*" y

consistió en un conjunto chico de preguntas de cada una de las categorías a testear. Estas preguntas no fueron extraídas de clases reales sino que fueron elaboradas para el estudio.

El último paso del proceso consistió en probar el modelo final con un conjunto de preguntas denominado "*test set".* Estas preguntas fueron extraídas de la grabación en video de una clase de ciencias naturales de nivel inicial de la Provincia de Buenos Aires de una escuela de gestión privada (Furman et al, 2018). Con este último paso del análisis se elaboraron los porcentajes de acierto y error del modelo.

Características de los conjuntos de datos

Como mencionamos en el apartado anterior, se utilizaron tres conjuntos de datos para entrenar al modelo, el *training set*, el *development set* y el *test set*.

El *training set* consistió en un corpus de 888 preguntas, 638 de las cuales fueron formuladas por docentes de clases reales de aulas de Matemáticas, Ciencias Naturales, Ciencia Sociales, Lengua y Arte de Argentina, Chile, Colombia, Ecuador, España, México, Perú y Venezuela). Estas grabaciones fueron recopiladas de clases encontradas en internet (la lista completa de referencias se encuentra en el anexo 3). De estas 638 preguntas 41 preguntas fueron eliminadas del análisis por estar incompletas o por no cumplir con características que describiremos en el próximo apartado. Las 250 preguntas restantes fueron elaboradas para el análisis.

La distribución de las preguntas extraídas de aulas de clase respecto a la asignatura y país de origen es la siguiente:

País	Total	
Argentina	24	3,76%
Chile	163	25,55%
Colombia	73	11,44%
Ecuador	4	0,63%
España	96	15,05%
México	75	11,76%
Perú	198	31,03%
Venezuela	5	0,78%
Total	**638**	**100,00%**

Asignatura	Total	
Matemáticas	248	38,87%
Ciencias Naturales	205	32,13%
Ciencias Sociales	40	6,27%
Lengua	142	22,26%
Arte	3	0,47%
Total	**638**	**100,00%**

Cuadro 1. Distribución de las preguntas de clase por país y asignatura.

La distribución por asignatura del total de la base de entrenamiento es la siguiente:

Asignatura	Total	
Matemáticas	301	3,90%
Ciencias Naturales	262	2,50%
Ciencias Sociales	61	6,87%
Lengua	151	17,00%
Arte	5	0,56%
N/A	108	12,16%
Total	**888**	**100,00%**

Cuadro 2. Distribución de las preguntas del conjunto de datos de entrenamiento por asignatura.

Por otra parte la distribución de las preguntas por nivel educativo es la siguiente:

Asignatura	Total	
Nivel inicial	41	4.62%
Nivel primario	433	48.76%
Nivel secundario	124	13.96%
Multigrado	39	4.39%
N/A	251	28.27%
Total	**888**	**100.00%**

Cuadro 3. Distribución de las preguntas del conjunto de datos de entrenamiento por nivel educativo.

La cantidad de preguntas de cada dataset respecto a la subcategoría, por su lado, se distribuye de la siguiente manera:

Categoría	Subcategoría	Entrenamiento	Desarrollo	Prueba
Cerradas	Verificación o memoria	438	10	24
	Procedimentales	96	3	8
	Total	**534**	**13**	**32**
Abiertas	Atención	3	10	3
	Medición y conteo	1	10	2
	Comparación	31	10	6
	Acción	25	10	6
	Resolución de problemas	59	10	1
	Razonamiento	127	10	10
	Metacognitivas	73	2	0
	Total	**319**	**62**	**28**

Cuadro 4. Distribución de las preguntas de los tres conjuntos de datos (entrenamiento, desarrollo y prueba) por subcategoría.

Limpieza de datos y consideraciones en el etiquetado

Debido a que todas las preguntas utilizadas para el análisis fueron extraídas de clases reales fue necesario realizar una limpieza de los datos para determinar qué preguntas iban a utilizarse para entrenar al modelo y cuáles no. También fue necesario definir algunos criterios respecto de cómo se iba a realizar el etiquetado o clasificación de las preguntas. Describiremos ambos criterios a continuación.

En primer lugar, fue necesario tomar algunas decisiones respecto de la clasificación de las preguntas. Al analizar las preguntas transcriptas observamos que era posible que algunas de ellas variaran de categoría dependiendo del contexto en que se habían formulado. Un ejemplo es: “¿cuánto es 5 más 5?” Esta pregunta podría considerarse abierta en el contexto de una clase de matemática de primer grado de primaria, ya que, siguiendo la definición de pregunta abierta que propusimos, para responderla los estudiantes deberían elaborar

conocimiento nuevo. En cambio, podríamos considerarla cerrada en un aula de secundaria ya que en este caso los estudiantes recurrirían a la memoria para contestarla.

Un segundo caso son aquellas preguntas que cambian de significado de acuerdo a la entonación con que el o la docente las formula. Observamos un ejemplo en una clase de ciencias de nivel secundario durante una actividad que consistía en pesar diferentes materiales. La docente pregunta "¿para qué va a ser la balanza?", lo que en el contexto de la grabación no suena como una pregunta genuina (porque la docente ocupa un tono burlón al formularla). Por la complejidad que agrega analizar las preguntas respecto a otras variables como el nivel educativo y la entonación, en este análisis nos limitaremos a analizar preguntas aisladas del contexto de su formulación lo que implica no tomar en cuenta la posible intencionalidad del docente al formularla.

En línea con lo anterior, también se tomaron definiciones respecto de qué preguntas no serían consideradas para la elaboración del modelo. Podemos categorizarlas en tres grupos: las que hacen referencia a una pregunta anterior, las incompletas y las que no apuntan al contenido de la clase.

Durante las clases expositivas dialogadas los docentes muchas veces van hilando preguntas para conducir a los estudiantes hacia nuevos saberes. En estos casos, si bien es posible que estas no sean preguntas cerradas, cuando se analizan en solitario carecen de sentido. Observamos un ejemplo de esta dinámica durante una clase de matemática de nivel primario (Ordenes, 2012). El tema de clase son las características de los prismas y la relación entre los tipos de prismas y el número de aristas. Para introducir este tema nuevo el docente plantea un problema al inicio: "vamos a hacer una pirámide ¿con base?" y luego va guiando la dinámica con preguntas formuladas unas tras otras: "¿de? ¿de qué base estamos haciendo la pirámide? ¿base cuadrada? ¿base? ¿qué tiene que hacer? ¿solo una pelotita?". Si bien se observa que durante la dinámica los estudiantes efectivamente arriban a aprendizajes nuevos y en ningún momento parecieran responder desde la memoria, muchas de estas preguntas tienen sentido únicamente en relación a las que vinieron antes o después. En estos casos solo ocuparemos para el análisis las preguntas que por sí solas sean abiertas o cerradas (como por ejemplo, en este caso: "¿de qué base estamos haciendo la pirámide?") y dejaremos por fuera aquellas que son preguntas incompletas o que hacen referencia a alguna otra.

Muy vinculadas a lo anterior son aquellas preguntas incompletas que los docentes utilizan como recurso durante las clases expositivas en donde solo formulan el comienzo de la pregunta para que los estudiantes completen el resto. Algunos ejemplos de estas preguntas

son: “¿que vienen en el...?” “¿y crean nuevas qué?” “¿realizan las...?” “¿podrían ser los...?”. Debido a que estas preguntas podrían ser abiertas o cerradas en contextos diferentes fueron dejadas por fuera del análisis.

Por último, dejamos por fuera todas aquellas preguntas que no apuntan al contenido específico de la clase. Algunos ejemplos de estas son: “¿cómo la pasaron el fin de semana?” “¿se han divertido?”.

En resumen, podemos decir que se tomaron para la construcción del modelo solo aquellas preguntas que estuvieran completas, que fueran pertinentes al contenido de la clase y que pudieran ser consideradas abiertas o cerradas independientemente del contexto de formulación.

Herramienta y algoritmos utilizados

Para este trabajo se utilizó una de las herramientas de procesamiento de lenguaje natural de Google denominada *Google Cloud AutoML Natural Language*. Esta herramienta utiliza dos técnicas para el procesamiento de los datos. La primera se denomina *transfer learning* y es una técnica de análisis de lenguaje natural que consiste en tomar un amplio conjunto de datos para entrenar a un modelo y luego usarlo para entrar un segundo modelo (Elvis, 2018). La segunda técnica algorítmica que utiliza se denomina *Learning to Learn* y consiste en la posibilidad del algoritmo de cambiar por sí solo la forma en que realiza generalizaciones sobre los datos que analiza (Thrun & Pratt, 1998). La utilización de ambas técnicas permite que se puedan realizar análisis con resultados que son pertinentes sobre datasets más pequeños como el que utilizaremos para este trabajo.

Categorización de los datos

Como mencionamos en el apartado anterior utilizaremos dos grandes grupos para categorizar a las preguntas: preguntas abiertas y cerradas. Utilizaremos, además, una subcategorización de estos dos grupos: (1) de atención, (2) de medición y conteo, (3) de comparación, (4) de acción, (5) de planteo de problemas, (6) de razonamiento y (7) metacognitivas, (8) verificación o memoria y (9) procedimentales.

Límites del análisis

Debido a algunas de las decisiones metodológicas que se tomaron y las características de los datasets este trabajo presenta algunas limitaciones que nos gustaría describir a continuación.

La primera limitación con la que contamos es la cantidad de preguntas que fueron usadas para construir el modelo. Como ya mencionamos transcribimos para este trabajo 787 preguntas de las que se utilizaron 638. De estas 638, 193 son preguntas abiertas y 445 son preguntas cerradas. La presencia de más preguntas cerradas para construir al modelo produjo inevitablemente que este sea mucho mejor a la hora de predecir cuando una pregunta es cerrada que cuando es abierta. Es más, durante las primeras dos iteraciones del modelo la probabilidad con era capaz de predecir una categoría por sobre la otra era todavía mayor (84,2% de seguridad al predecir una cerrada vs. 35,7% al predecir una abierta). Para contrarrestar los efectos de contar con más preguntas cerradas se elaboraron preguntas abiertas propias. Después de la incorporación de estas preguntas el total se divide de la siguiente manera: 432 preguntas cerradas y 523 preguntas abiertas.

Una segunda limitación de este análisis podría ser el sesgo con que se categorizaron las preguntas. Si bien el dataset de preguntas fue revisado múltiples veces y los modelos fueron entrenados en más de una ocasión, el hecho de que la categorización haya sido hecha mayormente por una sola persona (si bien hubo triangulación con una segunda persona que validó o propuso modificaciones para las categorizaciones hechas) puede presentar sesgos. Variables como la lengua materna de quien realiza la categorización, su ubicación geográfica, edad y género (Inclusive ML | Google Cloud, n.d.) han sido identificadas como características que pueden sesgar el proceso de etiquetado.

Por último este análisis podría presentar un sesgo respecto de la distribución geográfica de las clases que se transcribieron y las asignaturas a las que pertenecen. Las características de la muestra de clases que se transcribieron podrían influir en la capacidad del modelo de realizar predicciones más erróneas para sectores geográficos menos representados. Para intentar contrarrestar el impacto que esto podría tener se buscó construir una muestra lo más representativa posible de las distintas formas de hablar el español. Algunos lugares geográficos como Venezuela siguen, sin embargo, sin estar lo suficientemente representados.

Análisis de los resultados

En este apartado describiremos los resultados a los que arribamos al probar la herramienta con los diferentes conjuntos de datos. Comenzaremos por describir el grado de acierto y desacierto con que la herramienta pudo realizar predicciones sobre el conjunto de datos de desarrollo y luego describiremos los mismos resultados para el conjunto de datos de prueba.

Resultados del conjunto de datos de desarrollo

Preguntas cerradas

Como describimos en el apartado anterior, el primer mecanismo para probar una herramienta de estas características es analizar un conjunto pequeño de preguntas casi siempre elaboradas para el análisis denominado conjunto de datos de desarrollo. Para este análisis formulamos al menos dos preguntas de cada subcategoría para cada asignatura y las probamos con la herramienta. Recordemos que, si bien las preguntas fueron elaboradas en base a las subcategorías para garantizar que esta pudiera reconocer un bagaje más amplio de preguntas, la herramienta solo será capaz de reconocer preguntas abiertas o cerradas y no a qué subcategoría pertenece cada una.

Por ejemplo, para las preguntas de categoría cerrada elaboramos las siguientes preguntas con contenido de Matemática para las preguntas de la subcategoría “verificación o memoria”: “¿qué es un triángulo?” y “¿cuánto es 10 más 10?”. Para la categoría de preguntas abiertas elaboramos, por otro lado, las siguientes preguntas para la asignatura Ciencias Naturales y subcategoría “de atención”: “¿qué observan que está pasando con el papel cuando lo colocamos en el agua?” y “¿qué pudieron notar que pasa cuando dejamos el compuesto al sol?” (La lista completa de preguntas de este conjunto de datos y los resultados que obtuvo al intentar predecir la categoría de cada una se encuentran en el anexo 1).

Como dijimos, en este apartado intentaremos probar la capacidad del modelo entrenado para predecir exitosamente a qué categoría (abierta o cerrada) pertenece una pregunta. Para hacerlo, el mecanismo que seguimos fue el siguiente: 1) introdujimos una pregunta al

modelo y le pedimos que realizara una predicción de a qué categoría pertenecía, 2) recibimos una respuesta de la herramienta con dos porcentajes correspondientes a la probabilidad de que sea una pregunta abierta o de que sea cerrada. Por ejemplo, al introducir la pregunta "¿qué podríamos hacer para demostrar que este animal vive en el agua?" la herramienta responderá con lo siguiente: "abierta 99,6% / cerrada 0,4%", lo que significa que la herramienta tiene una seguridad del 99,6% de que se trata de una pregunta abierta. Los porcentajes presentados en los cuadros a continuación responden a este porcentaje de seguridad con que la herramienta cree que una pregunta se puede categorizar.

Comenzaremos por analizar la capacidad de la herramienta para reconocer preguntas cerradas. En el Cuadro 4 podemos observar con qué seguridad fue capaz de predecir a qué categoría pertenecen preguntas de las subcategorías 1) de verificación o memoria, y 2) procedimentales para las cinco asignaturas trabajadas.

Respecto de las preguntas de "verificación o memoria" podemos observar que, en términos generales, la herramienta pudo reconocer a este tipo de preguntas con una seguridad del 96,26%. Puntualmente, fue capaz de predecir con exactitud las preguntas de Matemática y Lengua (100% de seguridad en cada caso) y tuvo una seguridad mayor al 99% de acierto para las preguntas de Ciencias Sociales (99,55%). Tuve un éxito levemente menor para reconocer preguntas de Ciencias Naturales (91,55%) y Arte (89,90%).

Para intentar explicar el éxito levemente menor del modelo al reconocer preguntas de Ciencias Naturales comenzaremos por observar qué preguntas se usaron para probar al modelo y su éxito al categorizar cada una. Podemos observar que de las dos preguntas utilizadas: "¿qué es el sistema digestivo?" y "¿cómo definirían el sistema nervioso?" para la primera, el modelo pudo predecir con un 100% de certeza que se trataba de una pregunta cerrada mientras que obtuvo una seguridad menor en la segunda (83,1% cerrada / 16,0% abierta). Esta misma tendencia se observa cuando el modelo intenta clasificar una de las dos preguntas de Arte: "¿cómo definirían el impresionismo?" (79,80% cerrada / 20,20% abierta) (la segunda pregunta con que se testea al modelo para Arte es: "¿este edificio está en perspectiva?" y obtiene una seguridad del 100%). Este error del modelo puede deberse al uso del condicional en las preguntas ("definirían") que en general suele estar asociado a las preguntas abiertas. Una segunda explicación puede ser que las preguntas iniciadas con "cómo" en general hacen referencia a una acción y por lo tanto suelen ser clasificadas como preguntas abiertas. Una pregunta del tipo "¿cómo probarían que se trata de un ser vivo?", por ejemplo, es una pregunta abierta de subcategoría "planteo de problema".

Subcategoría	Tipo	Asignaturas						Total
		N/A	Matemática	Ciencias Naturales	Ciencias Sociales	Arte	Lengua	
Verificación o memoria	Abierta	-	0.00%	8.45%	0.15%	10.10%	0.00%	3.74%
	Cerrada	-	100.00%	91.55%	99.85%	89.90%	100.00%	96.26%
Procedimental	Abierta	20.80%	-	-	-	-	-	20.80%
	Cerrada	79.20%	-	-	-	-	-	79.20%
Promedio acierto cerradas								87.73%

Cuadro 5. Resumen de resultados por subcategoría y asignatura para preguntas cerradas (conjunto de datos de desarrollo)

Subcategoría	Tipo	Asignaturas						Total
		N/A	Matemática	Ciencias Naturales	Ciencias Sociales	Arte	Lengua	
Atención	Abierta	-	25.20%	97.25%	97.35%	58.40%	96.50%	74.94%
	Cerrada	-	74.80%	2.75%	2.65%	41.60%	3.50%	25.06%
Medición y conteo	Abierta	-	7.55%	99.85%	45.60%	99.95%	95.35%	69.66%
	Cerrada	-	92.45%	0.15%	54.40%	0.05%	4.65%	30.34%
Comparación	Abierta	-	68.30%	90.55%	77.75%	70.85%	59.95%	73.48%
	Cerrada	-	31.70%	9.45%	22.25%	29.15%	40.05%	26.52%

Acción	Abierta	-	75.35%	100.00%	100.00%	99.65%	100.00%	95.00%
	Cerrada	-	24.65%	0.00%	0.00%	0.35%	0.00%	5.00%
Resolución de problemas	Abierta	-	100.00%	89.75%	99.40%	99.50%	73.40%	92.41%
	Cerrada	-	0.00%	10.25%	0.60%	0.50%	26.60%	7.59%
Razonamiento	Abierta	-	99.95%	99.95%	99.10%	90.25%	100.00%	97.85%
	Cerrada	-	0.05%	0.05%	0.90%	9.75%	0.00%	2.15%
Metacognitivas	Abierta	94.95%	-	-	-	-	-	94.95%
	Cerrada	5.05%	-	-	-	-	-	5.05%
Total acierto abiertas		94.95%	62.73%	96.23%	86.53%	86.43%	87.53%	

Cuadro 6. Resumen de resultados por subcategoría y asignatura para preguntas abiertas (conjunto de datos de desarrollo)

Para la subcategoría de preguntas “procedimentales”, es decir, todas aquellas que los docentes formulan para en el contexto áulico, como ser “¿quién hizo la actividad?” la seguridad global de la herramienta fue levemente menor a las de tipo memorístico pudiendo predecir con un 79,20% de confianza que se trataba de preguntas cerradas. Para este caso las preguntas que se utilizaron fueron las siguientes: “¿quién quiere pasar al pizarrón?” (95,90% cerrada / 0,41% abierta), “¿quién terminó el ejercicio?” (100,00% cerrada / 0,00% abierta) y “¿cuánto te dió Agustín?” (41,70% cerrada / 58,30% abierta). Es en esta última pregunta que el modelo es incapaz de predecir la categoría “cerrada” correctamente.

Si bien intuitivamente podría parecer que el hecho de que esta última pregunta comience con “cuánto” podría explicar este fracaso, si observamos las preguntas de la base de prueba con las que el modelo fue entrenado de las 40 preguntas que contienen la palabra “cuánto” el 85% son preguntas cerradas. Otra posible explicación podría ser que esta pregunta, a diferencia de las demás, hace referencia a una persona al introducir un nombre propio. Al probar el modelo, sin embargo, con la misma pregunta pero sin una referencia puntual a una persona (“¿cuánto te dió?”) obtenemos un 97,50% de confianza en que se trata de una pregunta abierta, incluso peor que en el caso de la pregunta original. La hipótesis se valida, en cambio, si la pregunta no hace referencia a un individuo sino a un grupo de personas. “¿cuánto les dió?” obtiene un 99,30% de seguridad de que se trata de una pregunta cerrada.

De estos resultados podemos concluir que la herramienta necesita ser entrenada con más ejemplos de preguntas aúlicas que hagan referencia a personas puntuales. Preguntas como “¿quién te ayudó Manuel?” y “¿cómo te fue con el ejercicio Julia?” todas predicen erróneamente que se trata de preguntas abiertas con una confianza del 74,00% y 70,30% respectivamente. También agregar también, que modelos de preguntas como “¿cómo definirían…?” representan un desafío para la herramienta y se la debería entrenar con más ejemplos para mejorar el índice con que puede categorizarlas. Puntualmente, debería hacerse para preguntas de Ciencias Naturales y Arte, ya que preguntas de Lengua, Matemática y Ciencias Sociales son clasificadas con un éxito mayor (“¿cómo definirían a los textos expositivos?” 93,10% cerrada / 7,90% abierta, “¿cómo definirían a los triángulos rectángulos?” 99,70% cerrada / 0,30% abierta y “¿cómo definirían a la democracia?” 97,20% cerrada / 2,80% abierta).

Preguntas abiertas

Respecto a las preguntas de categoría abierta el Cuadro 5 detalla para cada subcategoría y asignatura con qué porcentaje de seguridad el modelo pudo predecir o no que se trataba de preguntas abiertas.

Para las preguntas de subcategoría “atención” podemos observar que el modelo predice con éxito la categoría de las preguntas para las asignaturas Ciencias Naturales, Ciencias Sociales y Lengua, con un 97,25%, 97,35% y 96,50% de confianza en cada caso. Para las preguntas de Arte el porcentaje de acierto es bastante menor (58,40%) y para Matemática, finalmente, el modelo predice erróneamente con un 74,80% de confianza que se trata de preguntas cerradas. A continuación exploraremos cada caso.

Las preguntas utilizadas para probar al modelo de la asignatura Arte fueron las siguientes: “¿notaron qué pasa cuando se mezclan estos dos colores?” y “¿qué observan en la forma en que se construyó esta imagen?”. Para la primera la herramienta predice categoría abierta con un 93,10% de confianza y predice erróneamente que la segunda se trata de una pregunta cerrada (76,30% cerrada / 23,70% abierta). Si probamos la herramienta con preguntas con el mismo significado pero diferentes estructuras gramaticales obtenemos los siguientes resultados:

- “¿qué observan en cómo que se construyó esta imagen?” (51,50% abierta / 48,00% cerrada).
- “¿qué observan en la forma en como que se construyó esta imagen?” (27,80% abierta / 72,20% cerrada).
- “¿qué observan en la forma en que se construyó esta figura?” (11,30% abierta / 88,70% cerrada).
- “¿qué ven en la forma en que se construyó esta imagen?” (0,02% abierta / 97,80% cerrada).

De estas pruebas podríamos extraer la conclusión de que el modismo “la forma” para hacer referencia a cómo o al procedimiento con que se implementó algo podría presentar problemas para el modelo. Esta conclusión se desprende de que la primera pregunta alternativa (“¿qué observan en cómo que se construyó esta imagen?”), en donde reemplazamos “la forma” por “cómo” obtiene un porcentaje de acierto del 51,5%.

Una segunda hipótesis podría ser que la herramienta tiene problemas en particular para categorizar preguntas de Arte lo que podría sustentarse en el bajo volumen de preguntas de esta asignatura en el conjunto de datos de entrenamiento (0,47% del total). Para probar

esto elaboramos preguntas con la misma estructura gramatical aplicadas a otras asignaturas:

- Ciencias Naturales: ¿qué observan en la forma en que la vegetación crece sobre el océano? (98,2% abierta / 1,80% cerrada).
- Lengua: ¿qué observan en la forma en que termina el cuento? (92,1% abierta / 7,9% cerrada).
- Ciencias Sociales: ¿qué observan en la forma en que Irigoyen finaliza su primer discurso presidencial? (76,6% cerrada / 23,4% abierta).
- Ciencias Sociales: ¿qué observan en la forma en que los diputados argumentan su opinión? (99,4% abierta / 0,6% cerrada)

De los resultados obtenidos no podemos extraer una respuesta concluyente a la hipótesis que nos planteamos, fundamentalmente por la clasificación errónea que hace la herramienta de la primera pregunta de Ciencias Sociales. Este error, sin embargo, si nos da lugar a intuir que los temas a los que hacen referencia las preguntas tienen un peso mayor que las estructuras gramaticales, lo que podría soportar la hipótesis de que las preguntas de Arte, de las que encontramos pocos ejemplos en el conjunto de datos de entrenamiento, podrían ser más desafiantes para el modelo.

Respecto a los resultados que obtiene el modelo al clasificar preguntas de Matemáticas, las utilizadas fueron: “¿notaron lo que sucede cuando sumamos -5?” y “¿observaron qué pasa a medida que nos movemos más a la derecha en la recta numérica?” para las que el modelo predijo 0,10% abierta / 99,90% cerrada para la primera y 50,30% abierta / 49,70% cerrada para la segunda.

Una primera explicación de por qué el modelo podría predecir erróneamente la categoría de la primera pregunta y acertar con una confianza relativamente baja en el caso de la segunda podría ser que tan solo el 23,92% de las preguntas de Matemática con que se entrenó al modelo eran abiertas. Esto explicaría el porcentaje de acierto bajo con que el modelo predice preguntas abiertas de Matemáticas para todas las subcategorías (62,73%), en efecto el más bajo de todas las asignaturas. Este resultado no parecería estar influido por la cantidad de preguntas de Matemática que había en el total siendo que, de hecho, el 38.87% de las preguntas del conjunto de datos eran de Matemática (el número más alto de todas las asignaturas).

Puntualmente la pregunta, “¿notaron lo que sucede cuando sumamos -5?” presenta una estructura gramatical nunca antes vista por el modelo. Las preguntas iniciadas con “¿notaron lo que sucede…?” o “¿notaron qué sucede…?” son casi siempre, de hecho,

categorizadas erróneamente como cerradas por el modelo. Un caso donde esto no aplica es con la pregunta utilizada para Arte "¿notaron qué pasa cuando se mezclan estos dos colores?" que es categorizada correctamente como abierta. Si reformulamos esta pregunta para que se parezca más a la pregunta de Matemática ("¿notaron *lo que sucede* se mezclan estos dos colores?"), sin embargo, el modelo vuelve a equivocarse sosteniendo con una seguridad del 51,7% que se trata de una pregunta cerrada.

Una explicación alternativa de por qué el modelo categoriza incorrectamente estas preguntas podría ser que, si bien estas preguntas son consideradas como preguntas de atención pueden ser respondidas con sí/no. Esto no explicaría, sin embargo, por qué un cambio en la reformulación de la pregunta, como el mencionado anteriormente, conduce a que el modelo sea capaz de clasificarla con éxito.

Todo lo mencionado hasta ahora podría llevarnos a concluir que al modelo le resulta desafiante clasificar algunas estructuras gramaticales poco vistas previamente. Observamos esto en el caso de las preguntas iniciadas con "¿notaron qué sucede…?". Una segunda conclusión podría ser que el tema al que hacen referencia las preguntas tiene particular impacto en estos casos. Si los temas de las preguntas fueron analizados en menor medida, como en el caso de las preguntas de Arte, el modelo tiende a cometer más errores. Por último, y muy vinculado a lo anterior, podemos agregar que los ejemplos de preguntas (si estas fueron cerradas o abiertas en el conjunto de datos de entrenamiento) tiene un peso relativamente alto en la capacidad del modelo para predecir preguntas nunca antes vistas. Podemos observar esto en la capacidad de la herramienta para predecir preguntas de Matemáticas en general.

Las preguntas de "medición y conteo" representan el porcentaje más bajo de acierto de todas las subcategorías, prediciendo correctamente que se trata de preguntas abiertas solo en el 69,66% de los casos. De estas, las asignaturas Ciencias Naturales, Arte y Lengua han sido las más exitosas en términos generales (99,85%, 99,95% y 95,35% de seguridad respectivamente). Para las preguntas de Matemática y Ciencias Sociales, sin embargo, la herramienta confunde la categoría de las preguntas sosteniendo que se trata de preguntas cerradas con un 92,45% de confianza en el caso de las de Matemática y de 54,40% en el caso de Ciencias Sociales.

Cabe mencionar que las preguntas de "medición y conteo" tan solo representan el 0,31% de las preguntas con que se entrenó a la herramienta, lo que podría verse reflejado en la falta de certeza del modelo al predecir preguntas de este tipo en términos generales. El hecho de

que, a pesar de esto, sea capaz de predecir de manera global con certeza qué preguntas de esta categoría son abiertas implica un éxito en la versatilidad del modelo para comprender estructuras gramaticales.

Para los casos en que no es capaz de alcanzar al menos un 50% de seguridad al predecir las preguntas abiertas, como en el caso de las de Ciencias Sociales y Matemáticas, podemos realizar algunas conjeturas. Para Ciencias Sociales las preguntas que se utilizaron fueron: "¿siempre podemos encontrar a este tipo de animales en terrenos áridos?" (15,30% abierta / 84,70% cerrada) y "¿cuándo pueden observar que estos animales dejan de migrar?" (75,90% abierta / 24,10% cerrada) lo que implica que el desacierto más notorio se observa para la primera pregunta.

Una primera explicación podría ser la estructura gramatical "¿siempre podemos encontrar qué…?" aunque en la asignatura Lengua la pregunta "¿siempre que colocamos una coma después del sujeto cambia el significado de la oración?" es categorizada como abierta con un 90,70% de confianza. A pesar del éxito del modelo al predecir la categoría de esta pregunta podríamos sostener que la falta de ejemplos anteriores con estructuras similares al entrenar al modelo puede haber tenido un impacto en su éxito a la hora de predecir este tipo de preguntas.

Una segunda explicación podría ser que para las demás asignaturas las preguntas casi siempre contaban con algún tipo de estructura del tipo "cuánto les parece que…", "cuántas piensan que…", "qué piensan que…", "qué les parece que", lo que podría haber facilitado al modelo su clasificación como preguntas abiertas. Una estructura como la de la pregunta "¿siempre podemos encontrar a este tipo de animales en terrenos áridos?" que pide implícitamente la opinión de los estudiantes podría implicar un desafío mayor para el modelo.

Para las preguntas de Matemáticas las utilizadas para la prueba fueron: "¿siempre que sumamos un número negativo nos da el mismo resultado?" (14,60% abierta / 85,40% cerrada) y "¿cuánto tenemos que movernos en la recta numérica para empezar a ver números positivos?" (0,05% abierta / 99,50% cerrada). La primera pregunta presenta una estructura muy similar a la que analizamos en el apartado anterior por lo que no nos detendremos en ella. La pregunta "¿cuánto tenemos que movernos en la recta numérica para empezar a ver números positivos?" también podría significar un desafío para el modelo en tanto pregunta la opinión de los estudiantes, nuevamente, de manera implícita. Si, en cambio, preguntáramos "¿cuánto les parece que tenemos que movernos en la recta

numérica para empezar a ver números positivos?" el modelo es capaz de predecir que se trata de una pregunta abierta con un 56,9% de confianza.

Podemos extraer como conclusiones del análisis de las preguntas de esta subcategoría que al modelo le resulta especialmente desafiante categorizar preguntas con estructuras gramaticales que nunca vio antes y que tiene problemas también al comprender cuando implícitamente se pide una opinión de los estudiantes (lo que casi siempre implica que se trate de preguntas abiertas). En cuanto a los casos en que predice correctamente las categorías de las preguntas podemos mencionar los casos en que estas piden una opinión ("¿cuánto tiempo les parece que tendrían que dejar el papel sobre el agua para que deje de flotar?" 100% abierta / 0,00% cerrada) incluso cuando implican pregunta acerca de cantidades o se utiliza la estructura gramatical "¿cuánto…?" (muchas veces asociada a preguntas cerradas).

Para la subcategoría de comparación el modelo tiene un éxito general del 73,48% al clasificarlas, siendo Ciencias Naturales, Ciencias Sociales y Arte las predichas con mayor confianza (90,55%, 77,75% y 70,85% respectivamente). Nuevamente observamos que las preguntas de Matemática son categorizadas con una confianza relativamente baja (68,30%) al igual, por primera vez, que las preguntas de Lengua (59,95%).

En el caso de las preguntas de Lengua, las utilizadas fueron: "¿qué tienen en común estos dos personajes?" (99,90% abierta / 0,10% cerrada) y "¿en qué se diferencia este texto de este otro?" (20,00% abierta y 80,00% cerrada). Siendo que solo la segunda pregunta es categorizada erróneamente intentaremos encontrar explicaciones de por qué pudo haber sido.

En primer lugar debe mencionarse que tan solo el 1,32% de las preguntas del conjunto de datos de entrenamiento de Lengua eran de tipo comparativas. La falta de preguntas que comiencen con la estructura "¿en qué se diferencia…?" también podría ayudarnos a explicar la falta de certeza de la herramienta con esta pregunta. No podríamos concluir, por otra parte, que el fracaso del modelo al categorizar la pregunta se deba a que es de asignatura Lengua. Una forma de probar esto es que al intentar predecir la categoría de la pregunta "¿qué tienen de diferente estos dos textos?" la herramienta acierta con un 98,9% de confianza que se trata de una pregunta abierta.

Para las preguntas de Matemáticas las utilizadas fueron: "¿qué creen que tienen en común estas dos figuras?" (100,00% abierta / 0,00% cerrada) y "¿en qué se diferencian los dos

procedimientos que hicimos?" (36,60% abierta / 63,40% cerrada). Los resultados de la herramienta al predecir las categorías de estas preguntas podrían llevarnos a afirmar que puntualmente la estructura gramatical "¿en qué se diferencia…?" representa un desafío para el modelo.

Las preguntas de subcategoría "acción" son las acertadas con mayor confianza por el modelo (95,00% abiertas / 5,00% cerradas en términos generales). Esta tendencia se observa también al nivel de cada asignatura: 100% de confianza en las preguntas de Ciencias Naturales, Ciencias Sociales y Lengua, 99,65% de confianza en las preguntas de Arte y un porcentaje menor de confianza en las preguntas de Matemática (75,35% abiertas / 24,65% cerradas).

Las preguntas que se utilizaron para probar a la herramienta en el caso de Matemática fueron: "¿qué les parece que pasaría si a este número le sumamos -5?" (81,50% abierta / 18,50% cerrada) y "¿qué pasaría si a este figura le sacamos una de sus aristas?" (69,20% abierta / 30,80% cerrada). Tomando en consideración que las preguntas de Matemáticas han sido las que presentan un porcentaje de confianza siempre menor a las de las demás asignaturas partiremos por probar las mismas preguntas para una asignatura diferente reemplazando la oración subordinada condicional ("si a este número le sumamos -5" / "si a esta figura le sacamos una de sus aristas"):

- ¿qué les parece que pasaría si a este compuesto le sacamos el agua? (100% abierta / 0,00% cerrada)
- ¿qué pasaría si a esta mezcla le colocamos arena? (99,40% abierta / 0,60% cerrada)

Se observa a partir de dos pruebas sencillas que las mismas estructuras gramaticales pero aplicadas a Ciencias Naturales tienen un porcentaje de confianza considerablemente mayor (100% para Ciencias Naturales vs. 81,50% para Matemática y 99,4% para Ciencias Naturales vs. 69,20% para Matemática). Esto puede llevarnos a concluir que el modelo es incapaz de abstraer lo suficiente la asignatura a la que se hace referencia y que, la cantidad superior de preguntas cerradas en el conjunto de datos de entrenamiento de Matemática, también puede haber tenido un efecto en la capacidad del modelo para realizar predicciones de esta asignatura.

En cuanto a las preguntas de proposición de problemas, observamos una confianza general del 92,41% al categorizarlas como preguntas abiertas siendo Matemática, Ciencias Sociales y Arte las acertadas con mayor confianza (100%, 99,40% y 99,50% respectivamente). Le siguen Ciencias Naturales con 89,75% de confianza y Lengua con un 73,40%.

Resulta notorio el porcentaje de confianza para categorizar a las preguntas de Matemáticas. Una explicación podría encontrarse en el porcentaje de preguntas para esta subcategoría de esta asignatura. En este sentido se observa que las preguntas de Matemáticas de subcategoría "de proposición de problemas" representan el 27,11% del total, únicamente superadas en número por las de Ciencias Naturales de la misma subcategoría (55,93%). Una segunda explicación puede encontrarse en las estructuras gramaticales de las preguntas utilizadas ("¿qué podríamos hacer para demostrar que la suma de estos números da siempre 5?" y "¿cómo podríamos hacer para estar seguros de que estas dos figuras siempre tienen la misma cantidad de lados?"). La primera estructura ("¿qué podríamos hacer para…?") se encuentra 4 veces en el conjunto de datos de entrenamiento y en todos los casos se trata de preguntas abiertas y la segunda ("¿cómo podríamos hacer para…?") se encuentra 2 veces en el conjunto de datos de entrenamiento también siendo las dos veces preguntas abiertas.

Respecto de los casos en que la herramienta tiene menor confianza al predecir la categoría de las preguntas observamos para Ciencias Naturales las siguientes preguntas: "¿cómo probamos que estos dos compuestos son iguales?" (100,00% abierta / 0,00% cerrada) y "¿cómo probarían ustedes que estos materiales no son los mismos?" (79,50% abierta / 20,50% cerrada). En el segundo caso lo que observamos es que la estructura gramatical "¿cómo probarían ustedes…?" no aparece ni una vez en el conjunto de datos de entrenamiento lo que podría explicar el menor índice de confianza. Esto se confirma si reemplazamos esa estructura por "¿cómo probamos que estos materiales no son los mismos?". Esta pregunta obtiene una confianza del 99,4% al predecir que es una pregunta abierta.

Para el caso de las preguntas de Lengua observamos las siguientes: "¿qué podemos hacer para que el final de este cuento sea abierto?" (74,10% abierta / 25,90% cerrada) y "¿cómo podemos hacer para transformar este texto en informativo?" (72,70% abierta / 27,30% cerrada). Una posible explicación en la confianza menor del modelo al predecir categorías de preguntas de esta asignatura puede ser, como ya se dijo, la cantidad relativamente baja de preguntas de Lengua de subcategoría "proposición de problemas" con que se entrenó al modelo (3,38% del total).

Esto se confirma si reemplazamos ambas preguntas por: “¿cómo podemos hacer para que este compuesto sea más acuoso?” (modificando únicamente la expresión que le sigue a “cómo podemos hacer qué”. Esta pregunta obtiene en cambio una confianza de acierto del 90,3%.

En cuanto a las preguntas de “razonamiento” observamos una confianza general del 97,85% observando una tendencia similar para todas las asignaturas de las preguntas: Matemática (99,95%), Ciencias Naturales (99,95%), Ciencias Sociales (99,10%), Arte (90,25%) y Lengua (100,00%). En efecto, las preguntas de tipo de “razonamiento” son las más numerosas en todo el conjunto de preguntas abiertas con que se entrenó al modelo (39,81% del total) lo que podría ayudarnos a explicar esta tendencia. Esta misma tendencia se observa en la presencia de las preguntas de esta subcategoría por asignatura en el conjunto de datos de entrenamiento: Matemática (18,89%), Ciencias Naturales (32,28%), Ciencias Sociales (16,53%), Lengua (19,68%), Arte (1,57%) y N/A (9,44%).

Por último, las preguntas de tipo metacognitivas fueron categorizadas con éxito por el modelo con una confianza del 94,95%. Para probar al modelo se utilizaron dos preguntas: “¿podría contar cada uno qué aprendió hoy?” (96,60% abierta / 3,40% cerrada) y “¿cómo hiciste para llegar a ese resultado?” (93,30% abierta / 6,70% cerrada).

Resultados

Antes de pasar a trazar algunas conclusiones generales en base a lo que observado en este apartado nos gustaría presentar los resultados en términos generales con que la herramienta pudo realizar predicciones acertadas para las categorías “abierta” y “cerrada” para el conjunto de datos de desarrollo:

		Predicción	
		Abierta	**Cerrada**
Realidad	**Abierta**	85,47%	14,53%
	Cerrada	12,27%	87,73%

Cuadro 7. Predicción de categoría versus categoría real (*conjunto de datos de desarrollo*)

Se observa de este cuadro que la herramienta es capaz de predecir que una pregunta es una pregunta abierta con un 85,46% de confianza (en el conjunto de datos de desarrollo) y que predice con un 87,73% de confianza las preguntas de categoría cerrada.

De los ejemplos analizados en este apartado podemos extraer tres conclusiones. En primer lugar, parecería acertado decir que la herramienta tiene un desafío mayor al clasificar preguntas con estructuras gramaticales nunca vistas. Observamos esto tanto en las preguntas de categoría abierta como cerrada. En segundo lugar, la cantidad de ejemplos con que se entrenó al modelo y la categoría de cada ejemplo para cada asignatura también parecería tener un peso relativamente alto. Podemos observar esto en la tendencia alta del modelo a cometer errores de clasificación en las preguntas de Matemática y Arte en la mayoría de las subcategorías de las preguntas abiertas. Por último, y en relación a los dos puntos anteriores, el modelo parecería tener un desafío especialmente mayor en los casos en donde se dan ambas tendencias: la pregunta presenta una estructura gramatical nueva y el modelo fue entrenado con pocos ejemplos de preguntas para esa asignatura.

Resultados del conjunto de datos de prueba

Analizaremos en este apartado el éxito de la herramienta al realizar predicciones sobre un conjunto de preguntas formuladas por docentes reales en una clase de Ciencias Naturales de nivel inicial la Provincia de Buenos Aires. Para hacerlo se utilizaron preguntas recolectadas de grabaciones de clases que se transcribieron y categorizaron manualmente. A continuación exploraremos con qué índice de éxito o fracaso la herramienta fue capaz de categorizar cada una.

Antes de comenzar haremos algunas salvedades sobre el conjunto de datos seleccionado y las preguntas que se utilizaron para el testeo. A lo largo de la clase analizada la docente formuló 91 preguntas de las cuales se seleccionaron para el análisis 63. Las preguntas que quedaron por fuera fueron todas aquellas que no seguían los principios que propusimos en el apartado metodológico, es decir: preguntas que no tuvieran que ver puntualmente con el tema de la clase (por ejemplo: “¿quieren que venga otro día?”, “¿vieron que yo estoy en la sala de dos?”), todas aquellas que no se pudieran clasificar por fuera del contexto en que se formularon (por ejemplo: “¿de qué es?”, “¿pero por qué?”) y, por último, todas aquellas que se pudieran considerar incompletas (por ejemplo: “¿y la linterna qué nos da?). Las 61 preguntas restantes se clasificaron como abiertas o cerradas y se les asignó, a su vez, una subcategoría. Esta clasificación arroja un total de 29 preguntas abiertas y 34 cerradas.

Preguntas cerradas

Comenzaremos por analizar el éxito de la herramienta al clasificar preguntas cerradas. Estos resultados se observan a continuación:

Subcategoría	Tipo	Asignatura	Total
		Ciencias Naturales	
Verificación o memoria	Abierta	20,68%	20,68%
	Cerrada	79,32%	79,32%
Procedimentales	Abierta	13,49%	13,49%
	Cerrada	86,51%	86,51%
Promedio acierto cerradas			82,91%

Cuadro 8. Resumen de resultados por subcategoría y asignatura para preguntas cerradas (conjunto de datos de prueba)

Podemos extraer del cuadro anterior que el éxito general con que la herramienta fue capaz de predecir preguntas cerradas fue del 82,91%, algunos puntos por debajo del éxito que tuvo al clasificar esta misma categoría en el conjunto de datos de desarrollo (87,73%). Yendo a cada subcategoría podemos observar que fue capaz de clasificar las preguntas de "verificación o memoria" con una seguridad del 79,32% y preguntas "procedimentales" con una confianza del 86,51% en promedio. A continuación analizaremos ambos resultados.

En el caso de las preguntas de "verificación o memoria" las utilizadas para el análisis fueron 24, entre las que observamos que el modelo pudo clasificar 8 de ellas correctamente con una confianza del 100%, 10 con una confianza mayor al 99%, dos con una confianza del 97,30% y del 64,70% respectivamente y clasificó a las 5 restantes erróneamente como preguntas abiertas (en el anexo 2 se encuentra el listado completo de preguntas y los resultados que la herramienta obtuvo al clasificar cada una).

Para el caso de las preguntas clasificadas como cerradas con un 100% de confianza tenemos las siguientes: "¿fuimos exploradores?", "¿los ven ustedes a los animalitos?", "¿a dónde los ven?", "¿ustedes ven acá los animalitos?", "¿estos son crayones?", "¿pasa la luz o no?", "¿porque es de papel?" y "¿pasa la luz?". Para las preguntas clasificadas con una

seguridad mayor al 99% tenemos, por otro lado, las siguientes: “¿ustedes ven algo acá adentro?”, “¿se ven los crayones?”, “¿y se pueden ver los crayones?, “¿y este papel deja ver del otro lado?”, “¿podés verlo?”, “¿pasa la luz a tu mano?”, “ ¿pasaba la luz en todos los materiales?”, “ ¿se ve? ¿se ve la luz allá en el techo?” y “¿y saben qué me preguntó el albañil?”. Por último, con una seguridad del 97,30% y del 64,70% respectivamente tenemos a las siguientes: “¿y este plástico deja pasar la luz?” y “¿pasa la luz adentro de la caja?”.

Nos enfocaremos, a continuación, en analizar las preguntas que la herramienta categorizó incorrectamente. Las preguntas son las siguientes:

- “¿qué utilizan del cuerpo para buscar?” (0,70% cerrada / 99,30% abierta)
- “¿qué material le pongo a la puerta del baño?” (0,70% cerrada / 99,30% abierta)
- “¿y se acuerdan que ayer nos dimos cuenta que además de los ojos necesitamos algo más para poder mirar?” (6,60% cerrada / 93,40% abierta)
- “¿se puede ver bien?” (10,60% cerrada / 89,40% abierta)
- “¿y este plástico deja que pase la luz adentro?” (28,60% cerrada / 71,40% abierta)

Las preguntas de este grupo presentan algunas de las características que mencionamos en el apartado anterior como desafiantes para la herramienta, es decir: 1) son muchas veces preguntas con estructuras gramaticales nunca vistas y 2) hubo pocos ejemplos de preguntas de esa subcategoría o asignatura en el conjunto de datos de entrenamiento.

Sobre el primer caso, es decir aquel en el que las preguntas presentan estructuras gramaticales poco analizadas por la herramienta, podemos mencionar el caso de la pregunta: “¿qué material le pongo a la puerta del baño?”. En esta pregunta la estructura “le pongo” podría funcionar como sinónimo de “se le pone” (“¿qué material se le pone a la puerta del baño?”) por lo que en este caso sería una pregunta cerrada. Podría argumentarse, sin embargo, que la falta de ejemplos de este tipo de estructuras en el conjunto de datos de entrenamiento dificultó su clasificación exitosa. La pregunta “¿qué material se le pone a la puerta del baño?” obtiene, sin embargo, un resultado correcto: 52,8% cerrada / 47,2% abierta.

Observamos un segundo ejemplo de este tipo de preguntas también en: “¿y se acuerdan que ayer nos dimos cuenta que además de los ojos necesitamos algo más para poder mirar?”. En esta pregunta la estructura “y se acuerdan que ayer…” hace alusión a un evento que sucedió en el pasado lo que podría implicar que se deba recurrir a la memoria para contestarla. Si bien la categoría de esta pregunta podría ser evidente para una persona

realizando la clasificación, la falta de ejemplos de preguntas similares durante la fase de entrenamiento vuelve más difícil esta misma tarea para la herramienta.

Otras preguntas difíciles de clasificar para la herramienta son, como mencionamos, aquellas de cuya subcategoría o asignatura hubo pocos ejemplos durante la fase de entrenamiento. Partiremos por mencionar, en este caso, que si bien la asignatura de la clase es Ciencias Naturales el tema (la luz, los materiales y la opacidad) es nuevo para el modelo. Preguntas del tipo: "¿qué material le pongo a la puerta del baño?", "¿se puede ver bien?" y "¿y este plástico deja que pase la luz adentro?" hacen todas referencia a este tema. En todos los casos, si reemplazamos estas preguntas por preguntas iguales pero haciendo referencia a un objeto más conocido por el modelo los resultados son levemente mejores (la pregunta "¿y este material deja que pase la luz adentro?", por ejemplo, obtiene una clasificación correcta con una seguridad del 80,5%). Esto podría implicar que el tema de la clase pudo haber tenido impacto en los resultados de la clasificación.

Respecto a las preguntas "procedimentales", el modelo fue capaz de predecir en términos generales las preguntas de esta subcategoría con una seguridad del 86,51%, obteniendo una seguridad del 100% con dos preguntas, una seguridad mayor al 98% para cinco de ellas, una confianza del 94,80% en una de ellas y una única clasificación errónea.

Para el caso de las preguntas clasificadas correctamente, las acertadas con una confianza del 100% son: "¿quién se acuerda qué hacen los exploradores?" y "¿Mateo ve algo?". Las acertadas con una confianza mayor al 98% son: " ¿tienen ganas de explorar?", "¿ustedes dicen que pasa?", "¿vemos si es poquita o mucha?" y "¿quién dice que sí? ¿vos decís que sí? ¿vos decís que no?". Por último la clasificada con una confianza del 94,80% es: "¿les gustó los materiales que trajeron mis amigas?".

En esta subcategoría la única pregunta clasificada incorrectamente fue una: "¿qué les parece si investigamos?". Al igual que en la subcategoría anterior podríamos adjudicar el error en la clasificación a la falta de preguntas con una estructura similar. La forma "qué les parece si", en efecto, aparece en el conjunto de datos de entrenamiento únicamente dos veces y en ambos casos se trata de preguntas abiertas.

Del análisis de la clasificación de las preguntas cerradas en este conjunto de datos podemos concluir que, en términos generales, parecerían darse las mismas tendencias observadas en el conjunto de datos de desarrollo. Se observa una tendencia global positiva con un porcentaje de acierto tan solo dos puntos por debajo que en el conjunto de datos anterior. Podríamos considerar que este porcentaje es alto si tomamos en cuenta que estas preguntas cuentan con el desafío extra de haber sido formuladas en la oralidad. Respecto a

las preguntas clasificadas incorrectamente también podemos observar las mismas tendencias que en el conjunto de preguntas de desarrollo con algunos errores producto de la falta de ejemplos similares durante el entrenamiento, tanto de estructuras gramaticales como de asignaturas y subcategorías. Estos errores nos permiten sugerir que una próxima iteración de la herramienta debería incluir todavía más ejemplos de preguntas formuladas en la oralidad para poder mejorar el índice con que la herramienta puede clasificar preguntas de este tipo.

Preguntas abiertas

Para el caso de las preguntas abiertas de este conjunto de datos los resultados de la clasificación se encuentran en el cuadro a continuación:

Subcategoría	Tipo	Asignatura	Total
		Ciencias Naturales	
Atención	Abierta	72,33%	72,33%
	Cerrada	27,67%	27,67%
Medición y conteo	Abierta	99,75%	99,75%
	Cerrada	0,25%	0,25%
Comparación	Abierta	61,87%	61,87%
	Cerrada	38,13%	38,13%
Acción	Abierta	52,13%	52,13%
	Cerrada	47,87%	47,87%
Resolución de problemas	Abierta	86,40%	86,40%
	Cerrada	13,60%	13,60%
Razonamiento	Abierta	76,44%	76,44%
	Cerrada	23,56%	23,56%
Metacognitivas	Abierta	-	-
	Cerrada	-	-
Promedio acierto abiertas			74,82%

Cuadro 9. Resumen de resultados por subcategoría y asignatura para preguntas abiertas (conjunto de datos de prueba)

Del cuadro anterior podemos extraer que la herramienta es capaz de clasificar preguntas de esta categoría con una seguridad en promedio del 74,82% unos 10 puntos por debajo del mismo índice para el conjunto de datos de desarrollo. Si bien este índice es considerablemente más bajo que en el conjunto de datos anterior la herramienta es capaz de predecir las preguntas de cada subcategoría con un confianza relativamente alta en cada caso, siendo las más alta las preguntas de "medición y contabilización", con una seguridad del 99,75% y siendo la más baja las preguntas "de acción" con una seguridad del 52,13%.

Si comparamos los resultados de cada subcategoría en relación a los mismos resultados para el conjunto de datos de desarrollo podemos observar que algunas tendencias se mantienen mientras que otras son casi opuestas. En el caso de las preguntas de "medición y conteo", por ejemplo, la herramienta obtiene en este conjunto de datos su porcentaje de acierto más alto mientras que en el conjunto de datos de desarrollo se da la tendencia opuesta. Lo mismo sucede con el caso de las preguntas de la subcategoría de "acción" en donde en el conjunto de datos de desarrollo se observa el porcentaje de acierto más alto mientras que en este conjunto de datos las preguntas se esta subcategoría son las acertadas con menor seguridad. Para hacer sentido a los resultados obtenidos nos enfocaremos a continuación en analizar cada caso puntual.

En el caso de las preguntas de "atención" la herramienta es capaz de acertar la categoría abierta con una confianza del 72,33%. Al nivel de las preguntas, es capaz de acertar dos de tres con una confianza mayor al 94% y predice incorrectamente una de ellas. Esta última es la pregunta: "¿dónde ves un poquito Gabi?".

En el caso de las preguntas de "medición y conteo", el modelo es capaz de clasificar las dos preguntas de esta subcategoría con una confianza mayor al 99%. Las preguntas de este caso son: "¿Pasa la luz? ¿Pero pasa mucha o pasa poquita?" y "¿Se puede ver bien o poquito?".

Respecto de las preguntas de "comparación" los resultados son menos contundentes. De las 6 preguntas analizadas ubica correctamente 4 de ellas con una confianza mayor al 69% y predice erróneamente las otras 2. Las preguntas categorizadas incorrectamente son: "¿Fue más fácil encontrar los crayones que los animales?" (92,40% cerrada / 7,60% abierta) y "¿Es como este?" (98,70% cerrada / 1,30% abierta).

Respecto de las preguntas de "acción" se da una tendencia similar. De las 6 preguntas analizadas ubica correctamente 3 de ellas con una confianza mayor al 91% y predice erróneamente las otras 3. Las preguntas categorizadas incorrectamente son: "¿Y si tengo cerrada la tapa?" (94,10% cerrada / 5.90% abierta), "¿Pasará la luz por este material?"

(84,40% cerrada / 15,60% abierta), y “¿Y esta qué dicen, pasará la luz?” (97,50% cerrada / 2,50% abierta).

En el caso de las preguntas de “resolución de problemas” solamente encontramos una a lo largo de la clase clasificada correctamente como abierta con una confianza del 86,40%.

Por último, en el caso de las preguntas de “razonamiento”, encontramos 10 preguntas de este tipo a lo largo de la clase de las cuales 7 son categorizadas correctamente con una confianza mayor al 92%. Las tres preguntas restantes son: “¿Y por qué no se ve?” (60,70% cerrada / 39,30% abierta), “¿Tomi, por qué no se ve?” (76,30% cerrada / 23,70% abierta) y “¿Por qué se ven?” (87,10% cerrada / 12,90% abierta).

Podemos observar que en términos generales se mantiene la misma tendencia observada en los demás conjuntos de datos así como también en la clasificación de las preguntas cerradas de este grupo. Algunas preguntas con estructuras más complejas, por ejemplo, aquellas que presentan nombres propios, que cuentan con más de una pregunta o que utilizan modismos propios de la oralidad son clasificadas incorrectamente en muchos casos. Lo mismo sucede en el caso de las preguntas que hacen referencia a temas poco trabajados durante el entrenamiento, como ser el tema de la clase (la luz y los materiales). La herramienta, sin embargo, es capaz de obtener un resultado contundentemente bueno en algunas categorías y es capaz de clasificar las preguntas de cualquier subcategoría con una confianza mayor al 52%. En promedio, si tomamos como base este conjunto de datos, la herramienta clasifica preguntas abiertas y cerradas con una confianza del 78,87% y predice, en promedio, correctamente el 75% de las preguntas para las que es consultada.

Resultados

Nos gustaría por último analizar el éxito de la herramienta en términos generales. Los resultados con que es capaz de clasificar las preguntas de cada grupo se encuentran a continuación:

		Predicción	
		Abierta	**Cerrada**
Realidad	**Abierta**	74,82%	25,18%
	Cerrada	17,09%	82,91%

Cuadro 9. Predicción de categoría versus categoría real (*conjunto de datos de prueba*)

Como se observa en el cuadro anterior la herramienta es capaz de predecir con una confianza del 74,85% en promedio preguntas abiertas y es capaz de clasificar con una seguridad levemente mayor, del 82,91%, preguntas cerradas. Podemos extraer como conclusión general del análisis de este conjunto de datos que, si bien la herramienta tiene algunos de los desafíos que mencionamos a lo largo de este apartado, es capaz de clasificar preguntas abiertas y cerradas exitosamente en la mayor parte de los casos. Se observa, como también mencionamos, un éxito menor en la clasificación respecto del que se observó al clasificar preguntas elaboradas, lo que podría responder a los desafíos de clasificar preguntas formuladas en la oralidad y sobre temas no trabajados por la herramienta.

Conclusiones

A lo largo de este trabajo nos abocamos al desarrollo de una herramienta capaz de clasificar preguntas de clase. Para hacerlo transcribimos 757 preguntas y elaboramos 251 de cinco asignaturas: Matemática, Lengua, Ciencias Naturales, Ciencias Sociales y Arte de nivel inicial, primario, secundario y multigrado. Finalmente se utilizaron para el análisis 889 preguntas de este conjunto al que denominamos *conjunto de datos de entrenamiento*. Con esta muestra variada en niveles educativos, países de procedencia y tipos de preguntas entrenamos a un modelo.

Para probar la efectividad del modelo al clasificar preguntas nunca vistas elaboramos un segundo conjunto de datos denominado *conjunto de datos de desarrollo* que consistió en 79 preguntas abiertas y cerradas para las cinco asignaturas que ya mencionamos. Al clasificar las preguntas de este conjunto obtuvimos una certeza del 84,37% al clasificar preguntas abiertas y del 87,73% al clasificar preguntas cerradas.

En segundo lugar, y para obtener el índice definitivo de éxito con que el modelo es capaz de predecir la categoría de preguntas de clase, utilizamos 89 preguntas transcriptas de una clase de Ciencias Naturales de nivel inicial de la Provincia de Buenos Aires. El modelo fue capaz de clasificar exitosamente las preguntas abiertas y cerradas en cada caso con una confianza de 74,82% y del 82,91% respectivamente.

Algunas de las conclusiones generales a las que arribamos al utilizar la herramienta para categorizar preguntas abiertas y cerradas fueron las siguientes: 1) el modelo parecería tener mayores dificultades al clasificar preguntas de una subcategoría de la que hubo pocos ejemplos durante la fase de entrenamiento; 2) parecería también tener dificultades al clasificar preguntas de asignaturas o temas nunca vistos y pudimos observar en el conjunto de datos de prueba que esta variable a veces parecería tener incluso más peso que la anterior; y por último 3) observamos que en los casos en que se dan ambas tendencias (una pregunta corresponde a una subcategoría para la que el modelo fue poco entrenado y esta es de un tema del que hubo pocos ejemplos) es especialmente esperable un error en la clasificación.

Tomando estos resultados en consideración queda por responder la pregunta central de nuestro análisis que es si una herramienta de estas características podría servir a la investigación educativa realizando de manera automática clasificaciones de preguntas de clase. Si nos atenemos exclusivamente a los resultados obtenidos al clasificar preguntas de

clases reales podríamos sostener que esta herramienta será capaz de predecir preguntas correctamente con una confianza promedio del 78,87%. Si el conjunto de preguntas sobre el que se aplicara la herramienta fuera, sin embargo, de preguntas elaboradas la herramienta sería capaz de clasificarlas con una confianza aún mejor del 86,05%.

Este trabajo permite probar también, si se quiere, que incluso con un conjunto de datos relativamente pequeño (menos de 1000 preguntas) y habiendo realizado contadas iteraciones sobre el modelo (menos de 10) se pueden obtener resultados en la clasificación que se acercan considerablemente a la precisión con que una persona no experta podría hacer la misma clasificación. Este resultado, sin embargo, es mejorable y detallaremos a continuación los pasos que podríamos seguir para afinar todavía más sus resultados.

Una primera dimensión a mejorar es la importancia relativa que el modelo le otorga al tema o asignatura de las preguntas. La herramienta que utilizamos para este trabajo (Cloud Natural Language de Google) no admite modificar el peso que se le otorga a esta variable. En una segunda iteración del modelo sería importante poder tener control sobre la importancia que se le asigna a los temas de las preguntas y la relación que esto tiene con la categoría a la que pertenecen (que debería ser cercana a cero).

Un segundo punto a mejorar sería, sin lugar a dudas, la cantidad de preguntas utilizadas para entrenar al modelo. Si bien la cantidad utilizada para este trabajo permitió obtener resultados concluyentes y una seguridad en promedio relativamente alta al interpretar los resultados del conjunto de datos de prueba observamos que las preguntas formuladas en la oralidad son especialmente desafiantes para el modelo. Para afinar su capacidad al clasificar preguntas reales consideramos, entonces, que aportaría más valor sumar transcripciones de clases reales que formular preguntas propias.

Un último punto a mejorar, muy vinculado al anterior, es el tipo de preguntas que deberían sumarse al modelo. Si bien la cantidad de preguntas es de por sí un valor a la hora de entrenar al modelo, intentar en lo posible que estas sean de la mayor cantidad de subcategorías posibles (de atención, de comparación, de acción, etc) puede ayudar, como ya observamos, a mejorar la habilidad de la herramienta al predecir las categorías de este tipo de preguntas en el futuro.

Cabe por último describir qué usos podría tener una herramienta de estas características. Si bien la utilización de dispositivos tecnológicos para el análisis de clases siempre presenta desafíos, una herramienta como esta podría aportar una nueva fuente de información a investigadores, docentes y directivos que busquen comprender las características de las interacciones que se dan en el aula. En especial, podría ayudarnos a comprender grandes

volúmenes de preguntas en tiempo real, algo que de otro modo no sería posible. Un análisis como este, que nos ayude a comprender los tipos de preguntas que se formulan en el aula, podría ayudarnos a repensar algunas dinámicas o a extraer buenas prácticas para ser compartidas con otros docentes, o a diseñar programas de formación y acompañamiento docente que tomen en cuenta las necesidades de cada educador.

Para poder sumar todavía más valor al análisis de clases utilizando una herramienta como esta, sin embargo, deberíamos ser capaces de interpretar las preguntas, ya no de manera aislada, sino en el contexto de su formulación. En segundo lugar, deberíamos poder tomar en consideración las respuestas de los estudiantes y las repreguntas o comentarios del docente que surgen a partir de ellos. Ambas funcionalidades podrían ayudarnos a extraer información todavía más rica acerca de qué preguntas y en qué contextos conducen a aprendizajes más significativos.

En síntesis, este trabajo aporta un desarrollo con potencial promisorio para el campo de la investigación educativa y la analítica de datos, que podría contribuir a la tarea de muchos investigadores y educadores que buscan comprender y mejorar su tarea cotidiana y, por ende, acompañar y favorecer los aprendizajes de sus estudiantes.

Anexos

Anexo 1

Conjunto de datos				**Predicción**	
Categoría	Subcategoría	Disciplina	Pregunta	Cerrada	Abierta
Cerrada	Verificación o memoria	Matemática	¿cuánto es 10 más 10?	100.00%	0.00%
			¿qué es un triángulo?	100.00%	0.00%
		Ciencias Naturales	¿qué es el sistema digestivo?	100.00%	0.00%
			¿cómo definirían el sistema nervioso?	83.10%	16.90%
		Ciencias Sociales	¿qué es un ciudadano?	99.80%	0.20%
			¿en qué año se fundó la Argentina?	99.90%	0.10%
		Arte	¿cómo definirían el impresionismo?	79.80%	20.20%
			¿este edificio está en perspectiva?	100.00%	0.00%
		Lengua	¿cuáles son las partes del texto expositivo?	100.00%	0.00%
			¿este es un personaje del cuento?	100.00%	0.00%
		Total		96.26%	3.74%
	Procedural	N/A	¿quién quiere pasar al pizarrón?	95.90%	4.10%

			¿quién terminó el ejercicio?	100.00%	0.00%
			¿cuánto te dió Agustín?	41.70%	58.30%
		Total		79.20%	20.80%
		Total		87.73%	20.80%
Conjunto de datos				**Predicción**	
Categoría	Subcategoría	Disciplina	Pregunta	Cerrada	Abierta
Abierta	Atención	Matemática	¿notaron lo que sucede cuando sumamos -5?	99.90%	0.10%
			¿observaron qué pasa a medida que nos movemos más a la derecha en la recta numérica?	49.70%	50.30%
		Ciencias Naturales	¿qué observan que está pasando con el papel cuando lo colocamos en el agua?	0.50%	99.50%
			¿qué pudieron notar que pasa cuando dejamos el compuesto al sol?	5.00%	95.00%
		Ciencias Sociales	¿qué pueden notar en la forma en que Rosas cambia su argumentación después de la Guerra del Paraguay?	4.80%	95.20%
			¿qué pueden observar que cambia en el discurso de asunción del segundo mandato de Perón?	0.50%	99.50%
		Arte	¿notaron qué pasa cuando se mezclan estos dos colores?	6.90%	93.10%
			¿qué observan en la forma en que se construyó esta imagen?	76.30%	23.70%
		Lengua	¿qué se puede notar que tiene de particular este tipo de texto?	7.00%	93.00%
			¿qué observan que sucede si cambiamos esta coma de lugar?	0.00%	100.00%

		Total		25.06%	74.94%
	Medición y conteo	Matemática	¿siempre que sumamos un número negativo nos da el mismo resultado?	85.40%	14.60%
			¿cuánto tenemos que movernos en la recta numérica para empezar a ver números positivos?	99.50%	0.50%
		Ciencias Naturales	¿cuánto tiempo les parece que tendrían que dejar el papel sobre el agua para que deje de flotar?	0.00%	100.00%
			¿cuántas monedas piensan que deberíamos poner para que el papel deje de flotar?	0.30%	99.70%
		Ciencias Sociales	¿siempre podemos encontrar a este tipo de animales en terrenos áridos?	84.70%	15.30%
			¿cuándo pueden observar que estos animales dejan de migrar?	24.10%	75.90%
		Arte	¿qué cantidad de rojo piensan que tenemos que ponerle a la mezcla para que el color cambie?	0.10%	99.90%
			¿qué longitud les parece que debería tener este edificio para estar en perspectiva?	0.00%	100.00%
		Lengua	¿cuánto tiempo creen que transcurrió desde que arranca el cuento hasta que el personaje se da cuenta de lo que está pasando?	0.00%	100.00%
			¿siempre que colocamos una coma después del sujeto cambia el significado de la oración?	9.30%	90.70%
		Total		30.34%	69.66%
	Comparación	Matemática	¿qué creen que tienen en común estas dos figuras?	0.00%	100.00%
			¿en qué se diferencian los dos procedimientos que hicimos?	63.40%	36.60%
		Ciencias Naturales	¿en qué les parece que se diferencian estas dos sustancias?	7.90%	92.10%
			¿en qué se parece este compuesto con el que hicimos la última vez?	11.00%	89.00%
		Ciencias Sociales	¿en qué notan que se distinguen estos dos terrenos?	5.70%	94.30%
			¿qué tienen de parecido estos dos tipos de terreno?	38.80%	61.20%
		Arte	¿en qué notan que se distinguen estas dos formas de preparación de los colores?	44.10%	55.90%

			¿qué tienen de similar estas dos pinturas?	14.20%	85.80%
		Lengua	¿qué tienen en común estos dos personajes?	0.10%	99.90%
			¿en qué se diferencia este texto de este otro?	80.00%	20.00%
		Total		26.52%	73.48%
	Acción	Matemática	¿qué les parece que pasaría si a este número le sumamos -5?	18.50%	81.50%
			¿qué pasaría si a este figura le sacamos una de sus aristas?	30.80%	69.20%
		Ciencias Naturales	¿qué piensan que pasa si a estos dos compuestos les agregamos agua?	0.00%	100.00%
			¿qué se les ocurrre que pasarían si uno de estos materiales fuera más pesado?	0.00%	100.00%
		Ciencias Sociales	¿qué creen que pasaría si este terreno tuviera un nivel más alto de minerales?	0.00%	100.00%
			¿qué creen que ocurriría si una de estas especies se trasladara a un territorio árido?	0.00%	100.00%
		Arte	¿qué se les ocurre que puede pasar si a este color le ponemos rojo?	0.30%	99.70%
			¿qué sucederá si mezclamos estos dos colores?	0.40%	99.60%
		Lengua	¿qué creen que cambiaría en la historia si este personaje nunca hubiera aparecido?	0.00%	100.00%
			¿qué les parece que hubiera pasado si este hubiera sido un texto argumentativo en vez de informativo?	0.00%	100.00%
		Total		5.00%	95.00%
	Resolución de problemas	Matemática	¿qué podríamos hacer para demostrar que la suma de estos números da siempre 5?	0.00%	100.00%
			¿cómo podríamos hacer para estar seguros de que estas dos figuras siempre tienen la misma cantidad de lados?	0.00%	100.00%
		Ciencias Naturales	¿cómo probamos que estos dos compuestos son iguales?	0.00%	100.00%

			¿cómo probarían ustedes que estos materiales no son los mismos?	20.50%	79.50%
		Ciencias Sociales	¿se les ocurre alguna forma de probar que este terreno tiene más minerales que este otro?	1.20%	98.80%
			¿se les ocurre cómo podemos hacer para ver si estos animales se encuentran siempre en territorios cálidos?	0.00%	100.00%
		Arte	¿cómo podemos ver si estos dos colores mezclados siempre dan violeta?	0.00%	100.00%
			¿qué pasaría si unimos estos dos colores?	1.00%	99.00%
		Lengua	¿que podemos hacer para que el final de este cuento sea abierto?	25.90%	74.10%
			¿cómo podemos hacer para transformar este texto en informativo?	27.30%	72.70%
		Total		7.59%	92.41%
	Razonamiento	Matemática	¿por qué creen que los ángulos interiores de estas dos figuras suman lo mismo?	0.10%	99.90%
			¿qué razón encuentran para que multiplicar dos números negativos de uno positivo?	0.00%	100.00%
		Ciencias Naturales	¿por qué les parece que el papel flotó en el agua en vez de hundirse?	0.00%	100.00%
			¿se les ocurre una regla de por qué siempre que ponemos un papel en el agua flota?	0.10%	99.90%
		Ciencias Sociales	¿por qué les parece que encontramos a estos animales en territorios áridos?	1.30%	98.70%
			¿por qué creen que este tipo de fauna solo se encuentra en la Patagonia?	0.50%	99.50%
		Arte	¿por qué les parece que estos dos colores nos dieron violeta?	0.00%	100.00%
			¿pueden inventar una regla que nos ayude a explicar por qué estos dos colores siempre dan violeta?	19.50%	80.50%
		Lengua	¿por qué un texto argumentativo puede no ser el mejor tipo de texto para un diario?	0.00%	100.00%
			¿por qué piensan que al cambiar el final del cuento cambiamos su moraleja?	0.00%	100.00%
		Total		2.15%	97.85%

	Metacognitiva	N/A	¿podría contar cada uno qué aprendió hoy?	3.40%	96.60%
			¿cómo hiciste para llegar a ese resultado?	6.70%	93.30%
		Total		5.05%	94.95%
		Total		15.63%	84.37%

Anexo 2

Conjunto de datos			Predicción	
Categoría	Subcategoría	Pregunta	Cerrada	Abierta
Cerrada	Verificación o memoria	¿Fuimos exploradores?	100.00%	0.00%
		¿Qué utilizan del cuerpo para buscar?	0.70%	99.30%
		¿Los ven ustedes a los animalitos?	100.00%	0.00%
		¿A dónde los ven?	100.00%	0.00%
		¿ustedes ven acá los animalitos?	100.00%	0.00%
		¿ustedes ven algo acá adentro?	99.30%	0.70%
		¿Pasa la luz adentro de la caja?	64.70%	35.30%
		¿Estos son crayones?	100.00%	0.00%
		¿Se ven los crayones?	99.00%	1.00%
		¿y este plástico deja que pase la luz adentro?	28.60%	71.40%
		¿Y se pueden ver los crayones?	99.70%	0.30%
		¿Y se acuerdan que ayer nos dimos cuenta que además de los ojos necesitamos algo más para poder mirar?	6.60%	93.40%
		¿Y este papel deja ver del otro lado?	99.90%	0.00%
		¿Podés verlo?	99.80%	0.20%
		¿Se puede ver bien?	10.60%	89.40%
		¿Pasa la luz a tu mano?	98.10%	1.90%
		¿pasaba la luz en todos los materiales?	99.00%	1.00%
		¿Pasa la luz o no?	100.00%	0.00%

		¿Porque es de papel?	100.00%	0.00%
		¿Y este plástico deja pasar la luz?	97.30%	2.70%
		¿Pasa la luz?	100.00%	0.00%
		¿Se ve? ¿Se ve la luz allá en el techo?	99.70%	0.30%
		¿Y saben qué me preguntó el albañil?	99.90%	0.10%
		¿Qué material le pongo a la puerta del baño?	0.70%	99.30%
		Total	79.32%	20.68%
	Procedimentales	¿Quién se acuerda qué hacen los exploradores?	100.00%	0.00%
		¿Mateo ve algo?	100.00%	0.00%
		¿Tienen ganas de explorar?	99.70%	0.30%
		¿Qué les parece si investigamos?	2.20%	97.80%
		¿les gustó los materiales que trajeron mis amigas?	94.80%	5.20%
		¿Ustedes dicen que pasa?	98.50%	1.50%
		¿Vemos si es poquita o mucha?	98.60%	1.40%
		¿Quién dice que sí? ¿Vos decís que sí? ¿Vos decís que no?	98.30%	1.70%
		Total	86.51%	13.49%
	Total		82.91%	17.08%
Conjunto de datos			**Predicción**	

Categoría	Subcategoría	Pregunta	Cerrada	Abierta
Abierta	Atención	¿y si yo abro la tapa? ¿Ven algo?	1.40%	98.60%
		¿a dónde? A ver si se ve.… ¿Vos decías que ves poquito?	5.60%	94.40%
		¿Dónde ves un poquito Gabi?	76.00%	24.00%
		Total	27.67%	72.33%
	Medición y conteo	¿Pasa la luz? ¿Pero pasa mucha o pasa poquita?	0.10%	99.90%
		¿Se puede ver bien o poquito?	0.40%	99.60%
		Total	0.25%	99.75%
	Comparación	¿Y si lo cierro?	2.80%	97.20%
		¿Fue más fácil encontrar los crayones que los animales?	92.40%	7.60%
		¿Es como este?	98.70%	1.30%
		¿Son iguales?	2.30%	97.70%
		¿Y dejará pasar la luz mucha como aquel? ¿O poquita?	1.70%	98.30%
		¿Qué dicen, mucha o poquita?	30.90%	69.10%
		Total	38.13%	61.87%
	Action	¿Y si tengo cerrada la tapa?	94.10%	5.90%
		Miren lo que vamos a hacer: ¿vos decís que pasa la luz? ¿quién dice que no pasa la luz?	2.00%	98.00%
		¿Pasará la luz por este material?	84.40%	15.60%
		¿será que pasa la luz por ahí?	8.90%	91.10%
		¿Qué será, dejará pasar la luz?	0.30%	99.70%

		¿Y esta qué dicen, pasará la luz?	97.50%	2.50%
		Total	47.87%	52.13%
	Resolución de problemas	¿La puedo hacer transparente, un material transparente, o le pongo un material opaco?	13.60%	86.40%
		Total	13.60%	86.40%
	Razonamiento	¿y cómo saben que está ahí?	0.30%	99.70%
		¿y cómo sabían que estaban los animalitos acá?	0.90%	99.10%
		¿Y por qué no se ve?	60.70%	39.30%
		¿Tomi, por qué no se ve?	76.30%	23.70%
		¿se puede ver lo que tengo adentro si yo no la destapo?	7.30%	92.70%
		¿Por qué se ven?	87.10%	12.90%
		¿Por qué se puede ver bien con este? Mirá el que tiene Mateo también. ¿Por qué se puede ver bien con este? ¿Alguien sabe por qué se puede ver bien con este?	0.00%	100.00%
		¿Entonces pasa la luz por este material?	2.20%	97.80%
		¿Pero por qué dicen que sí va a pasar la luz?	0.40%	99.60%
		¿Por qué va a pasar la luz?	0.40%	99.60%
		Total	23.56%	76.44%
	Total		25.18%	74.82%

Anexo 3

Nacionalidad	Nivel educativo	Asignatura	Fuente
Chile	Primario	Arte	https://www.youtube.com/watch?v=WdJhRb9iZaw
Méjico	Primario	Ciencias Naturales	https://www.youtube.com/watch?v=N4w56X6ulZo
Colombia	Primario	Ciencias Naturales	https://www.youtube.com/watch?v=pAdYAQhdlno
Chile	Primario	Ciencias Naturales	https://www.youtube.com/watch?v=X4C6q1xPaRE
Argentina	Primario	Ciencias Naturales	https://www.youtube.com/watch?v=oSYLlPC6UyA
Ecuador	Secundario	Ciencias Naturales	https://www.youtube.com/watch?v=qtq6qSznKcE&list=WL
Venezuela	Primario	Ciencias Naturales	https://www.youtube.com/watch?v=Fba_YH9C3IM&list=WL&index=7
Méjico	Primario	Ciencias Naturales	https://www.youtube.com/watch?v=XG0PcTF_mWE
Chile	Secundario	Ciencias Sociales	https://www.youtube.com/watch?v=RTrLA511Xek&list=PL0f0nPlcT_l7KQgz6nl7zd_8_iqiRvKzf&index=2
Chile	Primario	Ciencias Sociales	https://www.youtube.com/watch?v=4mTadTJmil4&list=PL722D300709901077&index=3
Chile	Primario	Lengua	https://www.youtube.com/watch?v=84iV0QFXx9M
Chile	Primario	Lengua	https://www.youtube.com/watch?v=hvs_YtR43eo
Colombia	Inicial	Lengua	https://www.youtube.com/watch?v=aOzrTd2wbe8&list=WL&index=2
Perú	Secundario	Lengua	https://www.youtube.com/watch?v=vEIJM_sMFWw&list=WL&index=8
Perú	Primario	Lengua	https://www.youtube.com/watch?v=EEjiAAlt13A
España	Primario	Lengua	https://www.youtube.com/watch?time_continue=74&v=j_K4vVCbGuI
Chile	Primario	Lengua	https://www.youtube.com/watch?v=31Mcam96QwU&list=PL722D300709901077
Perú	Primario	Lengua	https://www.youtube.com/watch?v=3F04-bDusxc
Méjico	Primario	Lengua	https://www.youtube.com/watch?v=CCbhc3oxbtA
Perú	Secundario	Lengua	https://www.youtube.com/watch?v=HU-jx2MwgpE

Chile	Primario	Matemática	https://www.youtube.com/watch?v=VUPTkKJ8ij8
Chile	Primario	Matemática	https://www.youtube.com/watch?v=fZLHWF4x4Kk
Méjico	Primario	Matemática	https://www.youtube.com/watch?v=7haGk1ro_Jw
Méjico	Inicial	Matemática	https://www.youtube.com/watch?v=-5szvaTitMM
Perú	Primario	Matemática	https://www.youtube.com/watch?v=7X2sSQDeryg
Perú	Primario	Matemática	https://www.youtube.com/watch?v=4es8AsKU3rc
Colombia	Multigrado	Matemática	https://www.youtube.com/watch?v=r0to-FzYv7U

Bibliografía

About Moodle. (n.d.). Retrieved April 15, 2019, from https://docs.moodle.org/36/en/About_Moodle

Aggarwal, P., & Sharma, G. (2017). Questioning Answering Mechanism Using Natural Language Processing. *6*(7), 22020–22026. https://doi.org/10.18535/ijecs/v6i7.20

Aizikovitsh-udi, E., Clarke, D., & Star, J. (2013). Good Questions or Good Questioning: an Essential Issue for Effective Teaching. *Paper Presented at CERME8: 8th Congress of the European Society for Research in Mathematics Education.* Antalya, Turkey.

Akerson, V. L., Buck, G. A., Donnelly, L. A., Nargund-Joshi, V., & Weiland, I. S. (2011). The Importance of Teaching and Learning Nature of Science in the Early Childhood Years. *Journal of Science Education and Technology*, *20*(5), 537–549. https://doi.org/10.1007/s10956-011-9312-5

Anaya, A. R., & Boticario, J. G. (2011). Application of machine learning techniques to analyse student interactions and improve the collaboration process. *Expert Systems with Applications*, *38*(2), 1171–1181. https://doi.org/10.1016/j.eswa.2010.05.010

Anijovich, R., & Mora, S. (2010). Estrategias de enseñanza: otra mirada al quehacer en el aula. Buenos Aires, Argentina: Aique.

Artificial intelligence [Def. 1]. (n.d.). Artificial intelligence | Definition of artificial intelligence in English by Oxford Dictionaries. Recuperado de https://en.oxforddictionaries.com/definition/artificial_intelligence

Bacca J., Baldiris S., Fabregat R., Guevara J., Calderón D. (2012) A Case-Based Reasoning Approach to Support Teaching of Spanish as a Second Language in Indigenous Communities from Latin America. In: Pavón J., Duque-Méndez N.D., Fuentes-Fernández R. (eds) Advances in Artificial Intelligence – IBERAMIA 2012. IBERAMIA 2012. Lecture Notes in Computer Science, vol 7637. Springer, Berlin, Heidelberg

Baker, R. S. (2016). Stupid Tutoring Systems, Intelligent Humans. *International Journal of Artificial Intelligence in Education*, *26*(2), 600–614. https://doi.org/10.1007/s40593-016-0105-0

Bird, A. S., Klein, E., & Loper, E. (Ed.) (2009). Natural Language Processing in Python (1st Ed.). O'Reilly Media, Inc.

Blanchard, N., Donnelly, P. J., Olney, A. M., Samei, B., Ward, B., Sun, X., … Dame, N. (2016). Semi-Automatic Detection of Teacher Questions from Human-Transcripts of Audio in Live Classrooms, 288–291.

Blanchard, N., Hall, F., Dame, N., Mello, S. D., Hall, F., Dame, N., … Wi, M. (2015). Automatic Classification of Question & Answer Discourse Segments from Teacher's Speech in Classrooms, 282–288.

Breu, F., Guggenbichler, S., & Wollmann, J. (2008). Assesing scientific literacy in PISA 2006 and fostering it in the United States. *Vasa*. http://medcontent.metapress.com/index/A65RM03P4874243N.pdf

Chang, M., Ventura, M., Ahn, J., Foltz, P., Ma, T., Dhamecha, T. I., … Mukhi, N. (2011). Dialogue-based tutoring at scale: Design and Challenges Efficacy of DBTs Scalability of DBTs, (1).

Clarke, J., Srikumar, V., Sammons, M., & Roth, D. (2012). An NLP Curator (or: How I Learned to Stop Worrying and Love NLP Pipelines). *European Language Resources Association (ELRA), Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012),* 3276–3283

Coscia, P. (2013). La importancia del diálogo y las preguntas en el salón de clase. Análisis de estrategias comunicativas en aulas universitarias. Recuperado de https://www.cse.udelar.edu.uy/wp-content/uploads/2013/11/tesis_patrizia_coscia_2017.pdf.

DeBoer, G. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching*, *37*(6), 582–601. Recuperado de http://www.alexiscullerton.com/uploads/2/4/7/2/24729748/scientific_literacy_another_look.pdf

Díaz, F. J., Lanzarini, L. C., Charnelli, M. E., Baldino, G., Schiavoni, M. A., & Amadeo, A. P. (2015, May). Analítica del Aprendizaje y la personalización de la Educación. In *XVII Workshop de Investigadores en Ciencias de la Computación (Salta, 2015)*.

Díaz, F. J., Schiavoni, M. A., Amadeo, A. P., & Charnelli, M. E. (2014). Búsqueda personalizada de recursos educativos basada en el perfil del alumno dentro de un

entorno educativo. In *XVI Workshop de Investigadores en Ciencias de la Computación*.

Dillenbourg, P. (2016). The Evolution of Research on Digital Education. *International Journal of Artificial Intelligence in Education*, *26*(2), 544–560. https://doi.org/10.1007/s40593-016-0106-z

Di Mitri, D., Schneider, J., Specht, M., & Drachsler, H. (2018). The Big Five: Addressing Recurrent Multimodal Learning Data Challenges. In *Companion Proceedings of the 8th International Conference on Learning Analytics and Knowledge: Towards User-Centred Learning Analytics* (pp. 420-424). Syndey, Australia: SoLAR.

Gallou, E., & Abrahams, P. (2018). Creating space for active learning: (Opportunities from) using technology in research-based education. In Tong V., Standen A., & Sotiriou M. (Eds.), *Shaping Higher Education with Students: Ways to Connect Research and Teaching* (pp. 165-175). London: UCL Press. Recuperado de http://www.jstor.org/stable/j.ctt21c4tcm.27

Efklides, A., & Misailidi, P. (2010). Trends and prospects in metacognition research. *Trends and Prospects in Metacognition Research*, (August), 1–479. Recuperado de https://doi.org/10.1007/978-1-4419-6546-2

Elvis. (2018, July 26). A Light Introduction to Transfer Learning for NLP. Recuperado de https://medium.com/dair-ai/a-light-introduction-to-transfer-learning-for-nlp-3e2cb56b48c8

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5/6) pp. 304–317.

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., … Welty, C. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, *31*(3), 59. Recuperado de https://doi.org/10.1609/aimag.v31i3.2303

Forbes, C. T., & Davis, E. A. (2010). Curriculum design for inquiry: Preservice elementary teachers' mobilization and adaptation of science curriculum materials. *Journal of Research in Science Teaching*, *47*(7), 820–839. Recuperado de https://doi.org/10.1002/tea.20379

Forum, I., Technology, E., & Technology, E. (2017). International Forum of Educational Technology & Society Web Intelligence and Artificial Intelligence in Education Author

(s): Vladan Devedžić Published by : International Forum of Educational Technology & Society Linked references are available on JSTOR, *7*(4), 29–39.

Furman, M. (2016). Educar mentes curiosas: la formación del pensamiento científico y tecnológico en la infancia : documento básico, XI Foro Latinoamericano de Educación / Melina Furman. - 1a ed compendiada. - Ciudad Autónoma de Buenos Aires: Santillana.

Furman, M., Luzuriaga, M., Taylor, I., Jarvis, D., Dominguez Prost, E. & Podestá, M. (2018) The use of questions in early years science: a case study in Argentine preschools. *International Journal of Early Years Education*. Recuperado de 10.1080/09669760.2018.1506319

Furman, M., Luzuriaga, M., Taylor, I., Podestá, M. E., & Jarvis, D. (2017). From inception to implementation: an Argentine case study of teachers enacting early years inquiry-based science. *Early Years*, *5146*, 1–18. Recuperado de https://doi.org/10.1080/09575146.2017.1389856

Furman, M. & Zysman, A.. (2011). Ciencias Naturales: Aprender a Investigar en la Escuela. (3era edición). Buenos Aires: Novedades Educativas

Gellon, G., Rosenvasser Feher, E., Furman, M., & Golombek, D. (2005). *La Ciencia en el Aula. Lo que nos Dice la Ciencia de Cómo Enseñarla. (1era ed.). Buenos Aires: Paidós.*

Graesser, A., Dowell, N., Hampton, An., Lippert, A., Li, H., Shaffer, D. (2018), Building Intelligent Conversational Tutors and Mentors for Team Collaborative Problem Solving: Guidance from the 2015 Program for International Student Assessment. En (ed.) *Building Intelligent Tutoring Systems for Teams (Research on Managing Groups and Teams, Volume 19)* Emerald Publishing Limited, pp.173 - 211

Gros, B., & Rodríguez Illera, J. L. (2018). Inteligencia artificial y diseño de programas educativos. *Revista Española de Pedagogía*, *49*(188), 39–57.

Gross, B. (1992). La inteligencia artificial y su aplicación en la enseñanza. *CL & E: Comunicación, lenguaje y educación*, ISSN 0214-7033, Nº 13, 1992, págs. 73-80.

Gunn, C., Mcdonald, J., Donald, C., Milne, J., & Blumenstein, M. (2017). Building an evidence base for teaching and learning design using learning analytics. Wellington: Ako Aotearoa – The National Centre for Tertiary Teaching Excellence.

Heffernan, N. T., Ostrow, K. S., Kelly, K., Selent, D., Van Inwegen, E. G., Xiong, X., & Williams, J. J. (2016). The Future of Adaptive Learning: Does the Crowd Hold the Key? *International Journal of Artificial Intelligence in Education*, *26*(2), 615–644. https://doi.org/10.1007/s40593-016-0094-z

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. … *Journal of Environmental & Science …*, *4*(3), 275–288. Recuperado de http://ijese.com/IJESE_Volume4_Issue3_July_2009.pdf#page=85

Huang, Z., Thint, M., & Qin, Z. (2008). Question Classification using Head Words and their Hypernyms. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)*, (October), 927–936. Recuperado de https://doi.org/10.3115/1613715.1613835.

Inclusive ML | Google Cloud. (n.d.). Retrieved April 15, 2019, from https://cloud.google.com/inclusive-ml/

International Educational Data Mining Society. (n.d.). *educationaldatamining.org*. Recuperado el April 15, 2019, de http://educationaldatamining.org/

Jirout, J., & Zimmerman, C. (2015). *Research in Early Childhood Science Education*. Recuperado de https://doi.org/10.1007/978-94-017-9505-0.

Kinshuk, Chen, N. S., Cheng, I. L., & Chew, S. W. (2016). Evolution Is not enough: Revolutionizing Current Learning Environments to Smart Learning Environments. *International Journal of Artificial Intelligence in Education*, *26*(2), 561–581. Recuperado de https://doi.org/10.1007/s40593-016-0108-x.

Kinzie, M. B., Whittaker, J. V., McGuire, P., Lee, Y., & Kilday, C. (2015). Research on curricular development for pre-kindergarten mathematics and science. *Teachers College Record*, *117*(7), 1–40.

Lemke, J. L. (1990). *Talking Science: Language, Learning, and Values (Language and Educational Processes)*. *Talking Science: Language, Learning, and Values (Language and Educational Processes)*. Recuperado de https://doi.org/citeulike-article-id:748226.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology*, *28*(1), 129–137. https://doi.org/10.1037/h0035519

Lyon, T. D. (2013). Child witnesses and imagination: Lying, hypothetical reasoning, and referential ambiguity. *The Oxford Handbook of the Development of Imagination*, (September), 126–136. Recuperado de https://doi.org/10.1093/oxfordhb/9780195395761.001.0009.

Manning, C. D., Bauer, J., Finkel, J., & Bethard, S. J. (2014). The Stanford CoreNLP Natural Language Processing Toolkit, 55–60.

Mary, B., & Martens, L. (1999). Productive Questions : Tools for Supporting Constructivist Learning. *Science and Children*, *36*(8), 24–27. Recuperado de http://web.missouri.edu/~hanuscind/productivequestions.pdf.

Minner, D. D., Levy, A. J., & Century, J. (2009). Inquiry-Based Science Instruction — What Is It and Does It Matter ? Results from a Research Synthesis Years 1984 to 2002 Center for Elementary Mathematics and Science Education, University of Chicago. Recuperado de https://doi.org/10.1002/tea.20347.

Newton, L. D. (2013). Teachers' Questions: Can they support understanding and higher-level thinking? *Journal of Business Ethics*, *44*(April), 0–12. Recuperado de https://doi.org/10.1063/1.2756072.

Olney, A., Samei, B., Donnelly, P.J., & D'Mello, S. (2017). Assessing the Dialogic Properties of Classroom Discourse: Proportion Models for Imbalanced Classes. *EDM*.

Ordenes, M. O. (2012, January 24). Estudio de Clases - Geometria. Retrieved April 15, 2019, from https://www.youtube.com/watch?v=VUPTkKJ8ij8

Passonneau, R. J., McNamara, D., Muresan, S., & Perin, D. (2017). Preface: Special Issue on Multidisciplinary Approaches to AI and Education for Reading and Writing. *International Journal of Artificial Intelligence in Education*, *27*(4), 665–670. Recuperado de https://doi.org/10.1007/s40593-017-0158-8.

Peña, A. V., & Pérez, D. G. (2001). Una alfabetización científica para el siglo XXI: obstáculos y propuestas de actuación. *Investigación En La Escuela*, (43), 27–37.

Perez, R. S., & Seidel, R. J. (1990). Intelligence in Education : Computer- Based Tools for Instructional Using Artificial Development. *Educational Technology*, *30*(3).

Pinkwart, N. (2016). Another 25 Years of AIED? Challenges and Opportunities for Intelligent Educational Technologies of the Future. *International Journal of Artificial Intelligence*

in Education, *26*(2), 771–783. Recuperado de https://doi.org/10.1007/s40593-016-0099-7.

Prieto, L. P., & Dillenbourg, P. (n.d.). Teaching Analytics: Towards Automatic Extraction of Orchestration Graphs Using Wearable Sensors Categories and Subject Descriptors. *En Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (LAK '16).* ACM, New York, NY, USA, 148-157. Recuperado de https://doi.org/10.1145/2883851.2883927

Resnick, M. (1986). All I Really Need to Know (About Creative Thinking) I Learned (By Studying How Children Learn) in Kindergarten.

Roll, I., & Wylie, R. (2016). Evolution and Revolution in Artificial Intelligence in Education. *International Journal of Artificial Intelligence in Education*, *26*(2), 582–599. Recuperado de https://doi.org/10.1007/s40593-016-0110-3.

Roth, G. L., & McEwing, R. A. (1986). Artificial Intelligence and Vocational Education: An Impending Confluence. *Educational Horizons*, *65*(1), 45–47. Recuperado de http://www.jstor.org/stable/42926856.

Ruzafa Martínez, M., Ruiz García, M. J., & Gómez García, C. I. (2003). Educación ciudadanía y alfabetización científica: mitos y realidades. *Revista de Enfermería (Barcelona, Spain)*, *26*(11), 30–34.

Shi, G., Lippert, A., Shubeck, K., Fang, Y., Chen, S., Pavlik Jr, P., … Graesser, A. (2018). *Exploring an intelligent tutoring system as a conversation-based assessment tool for reading comprehension*. *Behaviormetrika*. https://doi.org/10.1007/s41237-018-0065-9

Siemens, G., & Gasevic, D. (2012). Guest Editorial - Learning and Knowledge Analytics, *15*, 1–2.

Simmons R.F. (1970) Natural Language Question Answering Systems: 1969. In: Banerji R.B., Mesarovic M.D. (eds) Theoretical Approaches to Non-Numerical Problem Solving. Lecture Notes in Operations Research and Mathematical Systems (Economics, Computer Science, Information and Control), vol 28. Springer, Berlin, Heidelberg.

Slotta, J. D., Tissenbaum, M., & Lui, M. (2013). Orchestrating of Complex Inquiry: Three Roles for Learning Analytics in a Smart Classroom Infrastructure. *Proceedings of the Third International Conference on Learning Analytics and Knowledge - LAK '13*, 270–274. Recuperado de https://doi.org/10.1145/2460296.2460352

Tayyar Madabushi, H., & Lee, M. (2016). High Accuracy Rule-based Question Classification using Question Syntax and Semantics. *Proceedings of the 26th International Conference on Computational Linguistics (COLING-16)*, (2002), 1220–1230. Recuperado de https://www.aclweb.org/anthology/C/C16/C16-1116.pdf

Tennyson, R. D., Ferrara, J., Tennyson, R. D., & Ferrara, J. (2018). Introduction to Artificial Intelligence in Education Special Issue :, *27*(5), 7–8.

Thrun S., Pratt L. (1998) Learning to Learn: Introduction and Overview. In: Thrun S., Pratt L. (eds) Learning to Learn. Springer, Boston, MA

Timms, M. J. (2016). Letting Artificial Intelligence in Education out of the Box: Educational Cobots and Smart Classrooms. *International Journal of Artificial Intelligence in Education*, *26*(2), 701–712. Recuperado de https://doi.org/10.1007/s40593-016-0095-y

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Computers & Education Automatic classi fi cation of activities in classroom discourse. *78*. Recuperado de https://doi.org/10.1016/j.compedu.2014.05.010

Weston, J., & Karlen, M. (2011). Natural Language Processing (Almost) from Scratch, *12*, 2493–2537.

Worsley, M. (2014). Multimodal learning analytics as a tool for bridging learning theory and complex learning behaviors. *3rd Multimodal Learning Analytics Workshop and Grand Challenges*, MLA 2014, 1–4. http://doi.org/10.1145/2666633.2666634

Yazdani, M. & Lawler, R. (1986). Artificial Intelligence and Education: An Overview. Instructional Science, Vol. 14 , No. 3/4 , 1985 Conference on Artificial Intelligence and Education (Mayo 1986), pp . 197-206 Published b, *14*(3), 197–206.