Universitat d'Alacant
Universidad de Alicante

I³CE

Memòries del Programa de XARXES-I³CE de qualitat, innovació i investigació en docència universitària. Convocatòria 2018-19

Memorias del Programa de REDES-I³CE de calidad, innovación e investigación en docencia universitaria. Convocatoria 2018-19

Rosabel Roig-Vila (Coord.)
Jordi M. Antolí Martínez, Asunción Lledó Carreres, Neus Pellín Buades (Eds.)

UA | UNIVERSITAT D'ALACANT
ICE Institut de Ciències de l'Educació
Instituto de Ciencias de la Educación

Memòries del Programa de Xarxes-I3CE de qualitat, innovació i investigació en docència universitària. Convocatòria 2018-19

Memorias del Programa de Redes-I3CE de calidad, innovación e investigación en docencia universitaria. Convocatoria 2018-19

Rosabel Roig-Vila (Coord.), Jordi M. Antolí Martínez, Asunción Lledó Carreres, Neus Pellín Buades (Eds.)

UA | UNIVERSITAT D'ALACANT
UNIVERSIDAD DE ALICANTE
ICE Institut de Ciències de l'Educació
Instituto de Ciencias de la Educación

2019

Memòries de les xarxes d'investigació en docència universitària pertanyent al ProgramaXarxes-I3CE d'Investigació en docència universitària del curs 2018-19 / *Memorias de las redes de investigación en docencia universatira que pertence al Programa Redes -I3CE de investigación en docencia universitaria del curso 2018-19*

Organització: Institut de Ciències de l'Educació (Vicerectorat de Qualitat i Innovació Educativa) de la Universitat d'Alacant/ *Organización: Instituto de Ciencias de la Educación (Vicerrectorado de Calidad e Innovación Educativa) de la Universidad de Alicante*

Edició / *Edición*: Rosabel Roig-Vila (Coord.), Jordi M. Antolí Martínez, Asunción Lledó Carreres, Neus Pellín Buades (Eds.)

Comité tècnic / *Comité técnico*: Neus Pellín Buades

Revisió i maquetació: ICE de la Universitat d'Alacant/ Revisión y maquetación: ICE de la Universidad de Alicante

Primera edició: / *Primera edición*: Novembre 2019

ice@ua.es

ISBN: 978-84-09-15746-4

Producció: Institut de Ciències de l'Educació (ICE) de la Universitat d'Alacant / *Producción: Instituto de Ciencias de la Educación (ICE) de la Universidad de Alicante*

ISBN: 978-84-09-15746-4

1. Aplicación de técnicas de inteligencia artificial a la verificación de resultados obtenidos mediante la revisión por pares

Juan Ramón Rico-Juan[*]; Antonio Javier Gallego[*]; Santiago Meliá Beigbeder[*]; Vicente Ferri Coballes[**]; Javier Ortega Bastida[*]

juanramonrico@ua.es (Juan Ramón Rico-Juan), jgallego@dlsi.ua.es (Antonio Javier Gallego), santi@ua.es (Santiago Meliá Beigbeder), chento@ua.es (Vicente Ferri Coballes), jobmnas@gmail.com (Javier Ortega Bastida)

[*] *Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Carretera San Vicente del Raspeig s/n, Alicante, 03690, Spain*

[**] *Rectorado y Servicios Generales, Universidad de Alicante, Carretera San Vicente del Raspeig s/n, Alicante, 03690, Spain*

RESUMEN

El uso de la evaluación por pares para actividades abiertas tiene ventajas tanto para los profesores como para los estudiantes. Los profesores pueden reducir la carga de trabajo del proceso de corrección y los estudiantes logran una mejor comprensión de la materia al evaluar las actividades de sus compañeros. Para facilitar el proceso, es aconsejable proporcionar a los estudiantes una rúbrica sobre la cual realizar la evaluación de sus compañeros; sin embargo, limitarse a proporcionar sólo puntuaciones numéricas es perjudicial, ya que impide proporcionar una retroalimentación valiosa a otros compañeros. Dado que esta evaluación produce dos modalidades de la misma evaluación, a saber, la puntuación numérica y la retroalimentación textual, es posible aplicar técnicas automáticas para detectar inconsistencias en la evaluación, minimizando así la carga de trabajo de los profesores que supervisan todo el proceso. Este trabajo propone un enfoque basado en aprendizaje automático para la detección de tales inconsistencias. Con este fin se han probado diferentes algoritmos de inteligencia artificial para seleccionar el más prometedor. Los experimentos se han realizado con 4 grupos

ISBN: 978-84-09-15746-4

de estudiantes y 2 tipos de actividades muestran que el enfoque propuesto es capaz de producir resultados fiables, lo que representa un enfoque valioso para garantizar un funcionamiento justo del proceso de evaluación por pares.

Palabras clave:
Evaluación por pares, trabajos abiertos, evaluación asistida por ordenador, aprendizaje automático, inteligencia artificial, procesamiento del lenguaje natural.

1. INTRODUCCIÓN

A menudo, los profesores tienen que enfrentarse a aulas saturadas de alumnos (Shin & Teichler, 2014), lo que limita la posibilidad de llevar a cabo ciertas actividades debido a la gran carga de trabajo que suponen. El uso de herramientas informáticas puede aliviar la carga de trabajo del profesor cuando se enfrenta a esta situación. Por ejemplo, las actividades de respuesta cerrada pueden corregirse fácilmente de forma automática, ya que el profesor sólo debe preparar las preguntas y especificar las respuestas esperadas. En cambio los trabajos abiertos permiten estimular la originalidad o practicar la redacción, que no se tiene en cuenta en los de respuesta cerrada. A su vez, estos trabajos abiertos podrían representar una carga de trabajo de corrección inmanejable para el profesor, especialmente en el mencionado escenario de las aulas masificadas.

Una alternativa ampliamente considerada para mitigar la carga de trabajo de corrección es recurrir a la evaluación por pares (EP) entre los estudiantes (Kulkarni et al., 2013): los estudiantes evalúan el trabajo de sus compañeros de clase de los que obtienen una calificación agregada. Este paradigma no sólo es ideal para reducir la carga de trabajo del profesor, sino que también permite a los estudiantes aprender de soluciones alternativas a los mismos problemas propuestos por sus compañeros (Nicol et al., 2014). Es importante enfatizar que el EP por sí mismo no impide que el profesor se involucre en el proceso de corrección, ya que él o ella es eventualmente responsable de que los estudiantes obtengan una calificación justa. Sin embargo, el hecho de obtener varias evaluaciones de un mismo trabajo permite el uso de herramientas estadísticas que ayudan al profesor, como por ejemplo ocuparse sólo de aquellos trabajos en los que no hay consenso entre los evaluadores (Rico et. al., 2018) y evitar la autoevaluación (Falchikov & Goldfinch, 2000).

ISBN: 978-84-09-15746-4

Para facilitar el proceso, el profesor puede proporcionar un conjunto de reglas de evaluación (rúbrica) (Panadero et al., 2013). Los alumnos deben limitarse a evaluar aspectos de los trabajos, y además de verse obligados a proporcionar una puntuación numérica para cada actividad. Sin embargo, es interesante que la evaluación también incluya una revisión textual que pueda servir como retroalimentación para los estudiantes evaluados, además de obligar a los evaluadores a aclarar las razones por las que se determina dicha puntuación numérica (Li et al., 2016). Curiosamente, esto significa que la evaluación de cada actividad produce un doble resultado, a saber, una valoración numérica y otra textual. Los valores numéricos representan la puntuación dada a la sección que se está evaluando - similar a la escala de Likert - y el texto expresa sugerencias de mejora. Esta dualidad representa un escenario interesante para detectar inconsistencias entre ambas puntuaciones propuestas por el evaluador, como por ejemplo asignar una puntuación baja cuando la retroalimentación textual indica que todo es correcto o asignar una puntuación alta cuando la retroalimentación textual incluye varias sugerencias para mejorar. Detectar este tipo de inconsistencias es un paso clave para asegurar una mayor equidad en el proceso, pero hacerlo manualmente representaría una gran carga de trabajo para el profesor. Es por ello que en este trabajo proponemos un sistema para realizar esta detección de forma automática mediante sistemas de aprendizaje automático.

En nuestro trabajo, hemos evaluado varias técnicas que se utilizan comúnmente en el área de procesamiento del lenguaje natural (PNL en adelante, Processing Natural Language) para realizar minería de opiniones o análisis de sentimientos, con el objetivo de estimar qué puntuación numérica corresponde a una retroalimentación textual de specifica. Los recientes avances en las técnicas de PNL sugieren que su aplicación a las respuestas textuales en el proceso de EP es prometedora (Young et al., 2018). Nuestros experimentos, basados en dos actividades de differentes con 1000 revisiones aproximadamente, muestran que el enfoque es prometedor, y que con el uso de modelos apropiados se logran tasas de error muy bajas en las predicciones. Nuestro enfoque se postula como una herramienta interesante para ayudar a los profesores en un proceso de EP con aulas masificadas, haciendo que se tenga que prestar especial atención a ciertas evaluaciones donde nuestro sistema predice un valor muy diferente al propuesto por el evaluador.

ISBN: 978-84-09-15746-4

CONTEXTUALIZACIÓN

En las actividades de respuesta cerrada es habitual presentar sólo una respuesta correcta. Esta característica permite una corrección automatizada de manera relativamente sencilla (Wang et al., 2008). Algunos ejemplos de estas tareas han sido puestos en práctica con éxito en el campo de cursos de programación (Ala-Mutka, 2005), álgebra (Pacheco-Venegas et al., 2015), o en cualquier prueba de evaluación de opción múltiple.

Las actividades de trabajos abiertos no tienen una respuesta en predefineda, y generalmente pueden admitir muchas soluciones válidas. Es por ello que su corrección implica un effort mayor que la corrección de trabajos de respuesta cerrada, y no es fácil recurrir a las tecnologías de corrección automática (Bennett et al., 1997). Además, cuando el profesor quiere proporcionar retroalimentación a los estudiantes sobre sus actividades el esfuerzo general se vuelve inmanejable en clases masificadas (Kulkarni et al., 2013). En este contexto, la EP suele considerarse una opción para reducir la carga de trabajo de corrección. En este caso, las tareas abiertas son evaluadas directamente por otros estudiantes, con algunos beneficios adicionales para ellos mismos, tales como conocer las soluciones diferentes al mismo problema (Panadero & Brown, 2017) o proporciona una serie de comentarios para una retroalimentación oportuna y útil a sus compañeros (Mulder et al., 2014).

El EP ciertamente facilita la corrección de trabajos abiertos. Sin embargo, en el contexto del aula, los evaluadores son otros estudiantes que pueden no tener criterios de evaluación claros. En estas situaciones, es habitual proporcionar una rúbrica como guía para facilitar la evaluación del trabajo en sí y estandarizar los criterios (Anglin et al., 2008). Además, se ha informado de que el uso de rúbricas tiene un efecto positivo en el proceso de aprendizaje de los estudiantes (Panadero & Jonsson, 2013; Brookhart & Chen, 2015). Sin embargo, aunque el uso de las AP realizadas por rúbricas facilita ciertas tareas, la participación del profesor sigue siendo necesaria durante todo el proceso, tanto en la preparación de las rúbricas para la evaluación como en el seguimiento de las entregas para garantizar que no haya fraude o errores de corrección.

Hay pruebas que han explotado este escenario de EP para reducir la carga de trabajo de los profesores. Moodle incluye un módulo para manejar EP: los trabajos se suben a la plataforma, y cada trabajo se asigna automáticamente a determinado número de evaluadores; después de que cada evaluador proporciona una puntuación numérica se calcula la final como la mediana de las anteriores. El trabajo de Rico-Juan et. al. (2018) presenta una metodología basada en

ISBN: 978-84-09-15746-4

estadísticas para evaluar tanto las actividades de los alumnos como la calidad del trabajo realizado por los evaluadores. En el trabajo de Luaces et al. (2018), se utilizan técnicas de factorización matricial para proporcionar tanto calificaciones consistentes como retroalimentación a los estudiantes, y al mismo tiempo reducir la carga de los estudiantes en todo el proceso.

Además, hoy en día es cada vez más común en las publicaciones explorar la posibilidad de aplicar técnicas de aprendizaje automático (ML en adelante, Machine Learning) ―un área de la inteligencia de artificial que estudia cómo los ordenadores pueden aprender de los datos― en el contexto educativo (Barnes et al., 2017). Por ejemplo, para predecir el éxito académico de los estudiantes en los cursos introductorios de programación (Costa et al., 2017), para predecir si los estudiantes completarán con éxito su título universitario (Daud et al., 2017), o para predecir la selección de cursos para un estudiante de educación superior (Kardan et al., 2013). Ha habido intentos de construir métodos ML para la corrección automática de trabajos abiertos con el uso de la PNL (Noorbehbahani & Kardan, 2011; Xiong et al., 2012). Sin embargo, su comportamiento dista mucho de ser robusto y fiable.

Nuestro trabajo presenta una herramienta para ayudar durante el proceso de trabajo abierto en el EP, con el objetivo de detectar automáticamente las inconsistencias entre la puntuación numérica y la retroalimentación textual proporcionada por el evaluador. Nuestra metodología se basa en el uso de la rúbrica, que incluye varias secciones para una actividad. Cada sección se centra en una parte del trabajo de specifico, que debe rellenar con una escala de nivel tipo Likert y un campo abierto con las sugerencias que el evaluador considere (por ejemplo, “todo está correcto”, “esto debería mejorarse”, “la respuesta no es completa”, etc.). Dado un corpus de pares (puntuación numérica, retroalimentación textual), un modelo de PNL basado en ML debería ser capaz de aprender la puntuación numérica que debería corresponder a una retroalimentación textual de specific Este modelo podría ser utilizado para detectar inconsistencias y hacer que el profesor preste atención y corrija si es necesario - sólo esas revisiones.

2. ENFOQUES DEL PROCESAMIENTO DE LENGUAJE NATURAL

En este trabajo consideramos los algoritmos de ML para la PNL en el contexto de la retroalimentación textual proporcionada durante un proceso de EP. En general, el ML se basa en el uso de ejemplos de la tarea a resolver junto con sus correspondientes predicciones

ISBN: 978-84-09-15746-4

esperadas. Se sabe que el uso de ML es beneficia cuando se generaliza el rendimiento del sistema en contextos y actividades de differentes, en contraposición a la heurística artesanal.

Las técnicas que consideramos son ampliamente utilizadas en otras áreas de la PNL como el análisis de sentimientos o la minería de opinión. En nuestro caso, queremos detectar automáticamente si una retroalimentación textual corresponde a una buena o mala opinión de la actividad evaluada. Estas técnicas usualmente consideran millones de ejemplos para entrenar los sistemas predictivos. Aquí estudiaremos el comportamiento de estos algoritmos en un contexto de unos pocos miles de palabras de differentes, pero aplicadas a un dominio restringido en función de la actividad en el proceso de EP basada en rúbricas. Para construir el conjunto de datos base con el que entrenar los sistemas, utilizaremos la retroalimentación textual proporcionada durante la revisión como entrada y la puntuación numérica propuesta por ese evaluador como salida.

Con el fin de facilitar la tarea de análisis mediante las técnicas mencionadas anteriormente, es habitual representar palabras con un único identificador (entero) para usar el texto como una secuencia de números. Para que identificadores sean realmente útiles para nuestro sistema es aconsejable realizar un preprocesamiento básico del texto original para que el sistema sea más robusto. El preprocesamiento incluye pasos como la conversión a minúsculas, la eliminación de caracteres especiales o signos de puntuación, y la sustitución de las palabras originales por sus lexemas (véase la Fig. 1 arriba).

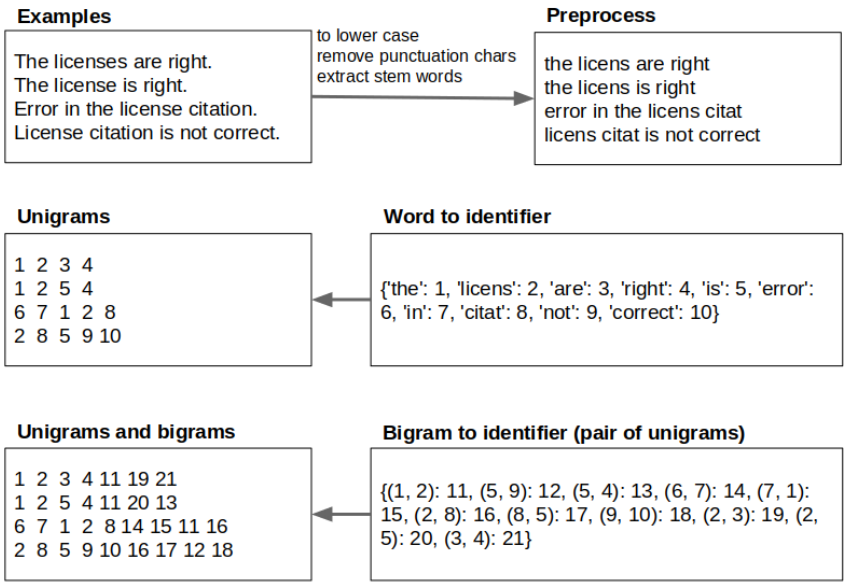

Figura 1: Un ejemplo básico con algunas muestras de texto preprocesadas (arriba); extracción de unigramas (en el medio); extracción de bigrams (abajo).

ISBN: 978-84-09-15746-4

Las sentencias a procesar por el sistema se reducen a una secuencia de longitud variable de identificadores. Los sistemas de predicción requieren que la entrada consista en un vector de longitud fija (Duda et al., 2001), por lo que la situación anterior representa un obstáculo. Para resolver la longitud arbitraria de las frases de entrada se convierten a un tamaño determinado ya sea recortando las palabras sobrantes o rellenando en las oraciones con palabras nulas.
Una vez que el modelo de PNL ha sido adecuadamente entrenado, puede ser usado para predecir la puntuación numérica que debe ser asociada con una retroalimentación textual. Sin embargo, este enfoque debe entenderse como una ayuda para el profesor, y no como un sistema totalmente autónomo. Cuando se trata de ponerlo en práctica, si la puntuación numérica real y la propuesta por el modelo de PNL están de acuerdo, entonces podemos confiar en ella. De lo contrario, el profesor debe ser avisado para verificar manualmente aquellas evaluaciones en las que la inconsistencia supere un determinado umbral.

Recientemente, la redes neuronales profundas (DNN en adelante, Deep Neural Network) ha mejorado el rendimiento en problemas difíciles de ML (Goodfellow et al., 2016). En particular, diferentes arquitecturas de red se han utilizado para problemas de PNL como el análisis de sentimientos o la minería de opinión (Glorot et al., 2011; dos Santos & Gatti, 2014). Cuando se utiliza DNN para tareas de PNL, es común colocar una capa en de datos embebidos al principio. Se espera que esta capa aprenda a mapear cualquier palabra (identificador) en un espacio en el que las palabras relacionadas -para la tarea en cuestión- estén representadas por vectores cercanos. El diseño de la red depende de parámetros como el número de palabras únicas o el vocabulario utilizado, la dimensión de los vectores de destino y la longitud máxima de las secuencias a procesar.
En el caso de nuestros experimentos, no sólo consideraremos secuencias de palabras únicas (unigramas) sino que también representaremos la entrada como una secuencia de bigramas, para lo cual cada par de palabras consecutivas se agrupan para formar un único identificador (ver Fig. 1 posiciones centro y abajo). Esto nos permite comprobar si la agrupación de palabras ayuda a mejorar el poder predictivo.
Independientemente del uso de unigramas o bigramas se probaron diferentes arquitecturas neurales para determinar su idoneidad para el problema en cuestión:

- *Memoria a largo plazo (LSTM)*: Este es un tipo de celda de red neuronal recurrente

ISBN: 978-84-09-15746-4

(Hochreiter & Schmidhuber, 1997), que se utiliza principalmente en el análisis de secuencias (Rumelhart et al., 1986) así como en problemas de modelado del lenguaje (Sundermeyer et al., 2012).

- *LSTM con mecanismos de atención (LSTM+att)*: El mecanismo de atención ayuda a la red neuronal a aprender qué partes de la entrada deben ser ponderadas en cada caso, con la intención tanto de ayudar a la convergencia durante el aprendizaje como de lograr un mejor desempeño en la predicción (Vaswani et al., 2017).
- *CNN+LSTM*: Esta combinación tiene como objetivo extraer las características más relevantes de la secuencia con la capa convolucional al principio, y luego procesar su representación secuencial con capas tipo LSTM.

Como se ha indicado anteriormente, todos estos modelos se complementan con una capa embebida que se coloca antes que el resto de las capas. La descripción detallada utilizada la podemos consultar en la tabla 1.

Red	**Topología**
LSTM	Embedding(embedding_dim=10)
	LSTM(64)
	LSTM(64)
LSTM+att	Embedding(embedding_dim=10)
	AtentionLayer()
	LSTM(64)
	LSTM(64)
CNN+LSTM	Embedding(embedding_dim=10)
	Convolution1D(filters=256, kernel_size=3)
	MaxPooling1D(pool_size=4)
	LSTM(64)
	Dense(10)

Tabla 1: Detalles de la arquitectura usada en las DNN usadas en este trabajo.

3. EXPERIMENTACIÓN Y RESULTADOS

Los experimentos se llevaron a cabo con un conjunto de datos extraídos de dos actividades de diferentes en cursos introductorios a la informática.

ISBN: 978-84-09-15746-4

Los temas considerados para cada una de las actividades fueron:

- **Actividad 1. Licencias Creative Commons:** En esta actividad, los estudiantes deben elegir un tema y buscar en Internet 5 imágenes que cumplan una serie de requisitos en materia de licencias. Por este motivo, la rúbrica también incluye las 5 secciones, cada una de ellas dedicada a recopilar el nivel de resolución de la tarea. Cada sección contiene un campo de texto para que el evaluador rellene sus sugerencias.
- **Actividad 2. Búsqueda web:** En este caso, se debe elegir un tema y dar los pasos correspondientes para crear una Webquest de forma correcta para que los contenidos estén bien estructurados, sean fácilmente navegables, se cite correctamente el material utilizado y se presenten correctamente los créditos. La rúbrica de esta actividad contiene siete secciones con varios campos numéricos por sección y un campo de texto para sugerencias de forma similar a la actividad anterior.

Las rúbricas analíticas de corrección utilizadas en nuestros experimentos se basan en las descritas en Rico-Juan, et. al (2018). Sin embargo, las hemos ampliado para permitir la inclusión de la retroalimentación textual requerida por nuestro enfoque. Recopilamos nuevos datos de cuatro grupos de differentes, compuestos por un total de 354 estudiantes que enviaron 176 trabajos y realizaron 1.925 revisiones. En tabla 2 se ofrece una visión general del caso de prueba.

Actividad	**Evaluadores**	**Trabajos**	**Revisiones**	**Secciones rúbricca**
1. Creative Commons	175	91	956	5
2. Webquest	179	85	969	7

Tabla 1: Estadísticas del proceso de EP en nuestro caso.

Los experimentos de detección automática de inconsistencias se realizaron de dos maneras: la primera considera un modelo único para todas las secciones (anotadas como "All" en las tablas) de la misma actividad, y el segundo considera un modelo separado para cada sección de la rúbrica de cada actividad (anotada como "Sections"). En la tabla 3 podemos observar una serie de descriptores estadísticos sobre el corpus de datos utilizado.

ISBN: 978-84-09-15746-4

n-grama	Actividad	Sección	Palabras únicas	#ejemplos	Número de palabras por frase					
					Avg(std)	Min	Q1	Q2	Q3	Max
1	1	1	614	956	5.9(8.5)	0	1	2	9	78
		2	567		4.9(7.3)	0	2	2	7	46
		3	595		4.9(7.6)	0	2	2	7	52
		4	554		5.2(7.8)	0	2	2	7	54
		5	591		5.3(7.4)	0	2	2	8	46
		All	1103	4780	5.2(7.7)	0	0	2	7	78
	2	1	749	969	7.8(9.5)	0	2	4	11	102
		2	668		7.3(8.3)	0	2	4	10	65
		3	764		9.5(10.3)	0	2	6	12	79
		4	768		10.2(11.1)	0	3	7	14	109
		5	680		7.9(8.9)	0	2	5	10	70
		6	651		6.4(7.3)	0	2	3	9	52
		7	675		9.7(10.4)	0	2	6	13	97
		All	1653	6783	8.4(9.6)	0	2	5	11	109
2	1	1	2985	956	11.1(16.8)	0	2	3	17	155
		2	2572		9.1(14.2)	0	3	3	13	91
		3	2622		9.1(14.9)	0	3	3	13	103
		4	2569		9.8(15.3)	0	3	3	13	107
		5	2772		10.0(14.6)	0	3	3	15	91
		All	6921	4780	9.8(15.2)	0	0	3	13	155
	2	1	4048	969	14.7(19.0)	0	4	7	21	203
		2	3634		13.6(16.5)	0	3	7	19	129
		3	4445		18.0(20.6)	0	3	11	23	157
		4	4738		19.4(22.2)	0	5	13	27	217
		5	3616		14.8(17.9)	0	3	9	19	139
		6	3214		11.9(14.5)	0	3	5	17	103
		7	3789		18.4(20.8)	0	3	11	25	193
		All	14874	6783	15.8(19.1)	0	3	9	21	217

Tabla 3: Estadísticas del corpus textual de nuestros experimentos

Se utilizó un esquema de validación cruzada con 10 bloques (10-CV). Como medida de calidad se utilizó el error absoluto medio (MAE) para medir la precisión de los resultados obtenidos de los algoritmos predictivos. Esta métrica fue elegida por su fácil interpretación en este contexto de enseñanza, ya que representa el valor absoluto entre la predicción automática y el valor real. Por lo tanto, cuanto más bajo sea el MAE, mejor será el modelo.

ISBN: 978-84-09-15746-4

La tabla 4 muestra los resultados de los experimentos para actividad 1 y 2, para cada algoritmo considerado la entrada de datos como 1-gramas o 2-gramas, y mostrando los resultados separado por secciones como de forma conjunta (All).

	Algoritmo	Secciones rúbrica					Avg(std)	All(std)
		1	2	3	4	5		
Actividad 1	LSTM 1-gram	0.29	0.25	0.29	0.36	0.35	0.31(0.15)	**0.22**(0.06)
	LSTM+att 1-gram	0.35	0.31	0.35	0.39	0.41	0.36(0.18)	0.26(0.07)
	CNN+LSTM 1-gram	0.50	0.45	0.47	0.51	0.55	0.50(0.13)	**0.25**(0.09)
	LSTM 2-gram	0.31	0.26	0.28	0.37	0.35	0.31(0.14)	**0.21**(0.05)
	LSTM+att 2-gram	0.33	0.33	0.33	0.41	0.40	0.36(0.17)	**0.23**(0.06)
	CNN+LSTM 2-gram	0.49	0.48	0.52	0.55	0.54	0.52(0.13)	0.28(0.09)
	Min	0.29	0.25	0.28	0.36	0.35	0.31	0.21
	Max	0.50	0.48	0.52	0.57	0.54	0.52	0.50

		1	2	3	4	5	6	7	Avg(std)	All(std)
Actividad 2	LSTM 1-gram	0.16	0.21	0.28	0.37	0.29	0.25	0.44	0.29(0.11)	**0.24**(0.13)
	LSTM+att 1-gram	0.15	0.22	0.28	0.41	0.31	0.25	0.47	0.30(0.12)	**0.25**(0.12)
	CNN+LSTM 1-gram	0.44	0.46	0.40	0.50	0.48	0.39	0.57	0.46(0.10)	0.26(0.12)
	LSTM 2-gram	0.15	0.21	0.28	0.36	0.30	0.25	0.44	0.28(0.11)	**0.23**(0.12)
	LSTM+att 2-gram	0.16	0.23	0.28	0.38	0.33	0.25	0.46	0.30(0.12)	0.26(0.12)
	CNN+LSTM 2-gram	0.41	0.39	0.46	0.49	0.41	0.49	0.52	0.45(0.11)	**0.25**(0.11)
	Min	0.15	0.21	0.28	0.36	0.29	0.25	0.44	0.28	0.23
	Max	0.44	0.46	0.46	0.49	0.48	0.49	0.60	0.46	0.40

Tabla 4: Errores absolutos promedio (MAE) obtenidos a partir del 10-CV de nuestras experimentación. Se destacan en negrita los mejores resultados.

Los mejores resultados son los obtenidos por la red de tipo LSTM con 2-gramas como entrada en el modelo conjunto (All) por actividad. Claramente, los modelos especializados en las secciones obtienen peores resultados debido seguramente a la menor cantidad de datos disponible para su aprendizaje.

4. CONCLUSIONES

En este trabajo asumimos un escenario de EP para aulas masificadas, en el que los alumnos

ISBN: 978-84-09-15746-4

evalúan a sus compañeros en base a una serie de secciones siguiendo una rúbrica. Se pide al evaluador que asigne una puntuación numérica en cada sección, así como una retroalimentación textual que complemente las decisiones tomadas. Para los estudiantes, este procedimiento es beneficioso ya que adquiere un mayor conocimiento de las actividades porque tienen que evaluar las de sus compañeros de clase, a la vez que generan una valiosa retroalimentación. Este escenario nos permite considerar un detector automático de inconsistencias entre la puntuación numérica y la retroalimentación textual proporcionada por los evaluadores, a través de técnicas de PNL. El objetivo real es evitar que los profesores tengan que examinar todas las evaluaciones del EP y que se centren en aquellas en las que se detectan inconsistencias, aliviando así su carga de trabajo.

En este trabajo se han presentado varias aproximaciones basadas en diferente tipos de redes neuronales (DNN) en la cuales se realiza todo el proceso utilizando directamente el texto como entrada (secuencia de palabras en forma de unigrama o bigrama) y la puntuación numérica esperada para la sección como salida.

Este enfoque basado en redes recurrentes como las LSTM obtienen los mejores resultados globales con un MAE de 0,22 para la actividad 1 y 0,23 para la actividad 2.

Este artículo muestra cómo las redes neuronales pueden ser utilizadas con éxito en contextos restringidos (como las EP guiadas por una rúbrica) con un número razonable de muestras de entrenamiento (alrededor de mil por sección) para detectar inconsistencias entre las puntuaciones numéricas y textuales en los resultados de las EP. Este proceso ayuda al profesor en su trabajo de revisar las respuestas y le ayuda a centrar la atención en los casos de inconsistencia para tomar las medidas adecuadas. Además, los resultados obtenidos mediante la aplicación de la metodología propuesta en las diferentes actividades y en varios grupos de estudiantes demuestran que es capaz de obtener resultados precisos revisando sólo un bajo porcentaje de los trabajos. Esta metodología es de hecho similar al caso en el que el profesor revisa todos los trabajos manualmente, pero con un esfuerzo mucho menor.

Podría haber algunas vías prometedoras para un trabajo futuro en lo que respecta a las cuestiones técnicas, con las que mejorar la precisión del sistema. Sin embargo, creemos que la idea más prometedora es cambiar el enfoque hacia el aprendizaje interactivo. Cuando el sistema detecta una inconsistencia, el profesor debe comprobar manualmente lo que ha ocurrido. Si resulta que todo era correcto -porque el sistema ha encontrado una inconsistencia

ISBN: 978-84-09-15746-4

que no lo es- el profesor podría marcarlo, usando esta información para retroalimentar el sistema y aprender de las correcciones humanas.

Otra extensión interesante sería integrar un asistente en línea en el formulario que verifique la puntuación y la retroalimentación escrita por el evaluador para avisar cuando se detecte una inconsistencia. De esta manera, se obtendrían dos ventajas: el evaluador podría verificar la sección antes de presentar su evaluación, y el instructor tendría menos inconsistencias que revisar.

5. TAREAS DESARROLLADAS EN LA RED

A continuación se enumeran los componentes de la red y las tareas que han desarrollado.

PARTICIPANTE DE LA RED	TAREAS QUE DESARROLLA
Juan Ramón Rico Juan	Coordinación. Implementación y aplicación de la metodología. Aportación de ideas base.
Antonio Javier Gallego Sánchez	Implementación y aplicación de la metodología. Aportación de ideas base.
Santiago Meliá Beigbeder	Aportación de ideas a la metodología.
Vicente Ferri Coballes	Aportación de ideas a la metodología.
Javier Ortega Bastida	Aportación de ideas a la metodología.

6. REFERENCIAS BIBLIOGRÁFICAS

Ala-Mutka, K. M. (2005). A survey of automated assessment approaches for programming assignments.*Computer Science Education*, *15*, 83–102.

Anglin, L., Anglin, K., Schumann, P. L., & Kaliski, J. A. (2008). Improving the Efficiency and Effectiveness of Grading Through the Use of Computer-Assisted Grading Rubrics.*Decision Sciences Journal of Innovative Education*, *6*, 51–73.

Barnes, T., Boyer, K., Sharon, I., Hsiao, H., Le, N.-T., & Sosnovsky, S. (2017). Preface for the special issue on ai-supported education in computer science.*International Journal of Artificial Intelligence in Education*, *27*, 1–4.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an Automatically Scorable, Open-Ended Response Type for Measuring Mathematical

Reasoning in Computer-Adaptive Tests.*Journal of Educational Measurement*, *34*, 162–176.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*, 343–368.

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., & Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses.*Computers in Human Behavior*, *73*, 247–256.

Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 415–421). International World Wide Web Conferences Steering Committee.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification*. (2nd ed.). Wiley.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks.*Review of educational research*, *70*, 287–322.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*(pp. 513–520).

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016).*Deep learning* volume 1. MIT press Cambridge.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*, 1735–1780.

Kardan, A. A., Sadeghi, H., Ghidary, S. S., & Sani, M. R. F. (2013). Prediction of student course selection in online higher education institutes using neural network. *Computers & Education*, *65*, 1– 11.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S. R. (2013). Peer and Self Assessment in Massive Online Classes.*ACM Transactions on Computer Human Interaction*, *20*, 1–31.

Li, H., Xiong, Y., Zang, X., L. Kornhaber, M., Lyu, Y., Chung, K. S., & K. Suen, H. (2016).

ISBN: 978-84-09-15746-4

Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education*, *41*, 245–264.

Luaces, O., Díez, J., & Bahamonde, A. (2018). A peer assessment method to provide feedback, consistent grading and reduce students' burden in massive teaching settings. *Computers & Education*, (pp. 283–295).

Mulder, R. A., Pearce, J. M., & Baik, C. (2014). Peer review in higher education: Student perceptions before and after participation.*Active Learning in Higher Education*, *15*, 157–171.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective.*Assessment & Evaluation in Higher Education*, *39*, 102–122.

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education*, *56*, 337–345.

Pacheco-Venegas, N. D., López, G., & Andrade-Aréchiga, M. (2015). Conceptualization, development and implementation of a web based system for automatic evaluation of mathematical expressions.*Computers & Education*, *88*, 15–28.

Panadero, E., & Brown, G. T. (2017). Teachers' reasons for using peer assessment: positive experience predicts use.*European Journal of Psychology of Education*, *32*, 133–156.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review.*Educational Research Review*, *9*, 129–144.

Panadero, E., Jönsson, A., & Alqassab, M. (2018). Providing formative peer feedback: What do we know? In A. A. Lipnevich, & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback*. Oxford: Cambridge University Press.

Panadero, E., Romero, M., & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*, 195–203.

Rico-Juan, J. R., Gallego, A. J., Valero-Mas, J. J., and Calvo-Zaragoza, J. (2018). Statistical semi-supervised system for grading multiple peer-reviewed open-ended works. *Computers & Education*, 126(1):264–282.

ISBN: 978-84-09-15746-4

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning sequential structure in simple recurrent networks.*Parallel distributed processing: Experiments in the microstructure of cognition, 1*.

dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 69–78).

Shin, J. C., & Teichler, U. (2014). The Future of University in the Post-Massification Era: A Conceptual Framework. In *The Future of the Post-Massified University at the Crossroads: Restructuring Systems and Functions*(pp. 1–9). Cham: Springer International Publishing.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*(pp. 5998–6008).

Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2008). Assessing creative problem-solving with automated text grading. *Computers & Education*, *51*, 1450–1466.

Xiong, W., Litman, D., & Schunn, C. (2012). Natural Language Processing techniques for researching and improving peer feedback.*Journal of Writing Research*, *4*, 155–176.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [review article].*IEEE Comp. Int. Mag.*, *13*, 55–75.