S. L. Sánchez López, R. P. Díaz Redondo and A. Fernández Vilas, "Can the Type of Content You Share on eLearning Forums Reveal What Kind of Student You Are?" in IEEE-RITA, vol. 16, no. 3, pp. 318-327, AUG. 2021. doi: 10.1109/RITA.2021.3122946

¿El contenido que compartes en los foros eLearning podría revelar qué tipo de estudiante eres?

Sheila Lucero Sánchez López, Rebeca P. Díaz Redondo, Ana Fernández Vilas

***Title*— Can the type of content you share on eLearning forums reveal what kind of student you are?**

***Abstract*— Trending topics on forums or social networks are now being analysed to know our preferences and needs. This line of thought can be applied to the educational context. In this study we want to analyze what information on the profile of students can be inferred by analyzing their performance in educational forums and if it is possible to classify them by applying a taxonomy with a quantitative and qualitative perspective of their social interaction.**

To achieve this, our methodology will be divided into two stages. The first stage consists of preprocessing, classifying and analyzing the messages of the educational forum. In the second stage, the analysis of the messages in the forums is combined with a taxonomy of five different student profiles. These profiles were built according to two types of interactions (social interaction and interaction with content) with the online platform from a quantitative perspective. However, this article proposes to analyze the social interactions of the five proposed profiles from a qualitative perspective.

Our results validate the application of the Extended Taxonomy when qualitative and quantitative approaches are applied.

***Index Terms*— applied computing, learning management systems, e-learning, Learning Analytics, social interaction, learnings forums.**

I. Introducción

La aparición de plataformas de e-learning ha impulsado la innovación en el campo educativo. Entre las diferentes herramientas disponibles en estas plataformas de Learning Management Systems (LMS), destacan los foros de discusión online. Su particularidad radica en la posibilidad que tienen los estudiantes de intercambiar mensajes, opiniones y preguntas libremente de manera sencilla [1]. En la literatura podemos encontrar mucha información sobre el uso de los foros eLearning. En general, muchos investigadores están de acuerdo en que un foro eLearning es un medio de interacción en línea que proporciona un entorno positivo para la discusión y el aprendizaje. Sin embargo, otros investigadores creen que "el grado de contenido de conocimiento involucrado todavía es muy cercano al suelo" [2]. Usualmente, los debates sobre diferentes temas surgen en foros y en este espacio los estudiantes pueden expresar sus opiniones y puntos de vista particulares sobre los conceptos y contenidos relacionados con la asignatura. Estos mensajes a menudo involucran experiencias individuales, preguntas, sugerencias o críticas, ofreciendo una oportunidad para reflexionar sobre los contenidos del curso [3]. Por tanto, no es de extrañar que muchos autores apoyen el uso de estos foros como una herramienta para que los estudiantes adquieran y compartan conocimientos, y a su vez, para que los docentes identifiquen información importante sobre sus alumnos, por ejemplo, como están construyendo su conocimiento [4].

Nuestro enfoque intenta aprovechar el intercambio de información en los foros en línea para conocer el perfil de los estudiantes que comparten contenido en éstos. Además de esto, nuestro objetivo es validar la existencia de los cinco perfiles propuestos por la taxonomía extendida en [5] cuando se aplican enfoques cualitativos y cuantitativos. La taxonomía ampliada se basa en la taxonomía de Bento [6] que clasifica a los estudiantes en cuatro cuadrantes en contextos puramente virtuales. Sin embargo, la taxonomía extendida clasifica a los estudiantes en cinco cuadrantes y se aplica en contextos de aprendizaje mixto bajo un enfoque cuantitativo [7]. No obstante, en este caso, queremos validar esta misma taxonomía en un contexto de aprendizaje mixto bajo un enfoque cualitativo y cuantitativo. Esta taxonomía de estudiantes permite clasificar a los estudiantes en cinco tipos o perfiles diferentes, según la forma en que interactúan con la plataforma de e-learning (considerando sus interacciones sociales y / o sus interacciones de contenido). La metodología que proponemos se divide en dos etapas. Por una parte, la primera se centra en los mensajes intercambiados. Las publicaciones se analizan y clasifican mediante un clasificador bayesiano ingenuo y dos corpus relacionados con la temática del foro eLearning (en este caso, contenido relacionado con los lenguajes de programación y códigos de programación). Al aplicar el clasificador bayesiano obtendremos 3 categorías de mensajes: contenido, código y otros. Para finalizar la primera etapa, se realiza un análisis exploratorio para cada categoría. Por otra parte, la segunda etapa se centra en vincular los diferentes tipos de mensajes con los diferentes tipos de estudiantes. Para ello, se aplica la taxonomía extendida con el objetivo de clasificar a los participantes del foro. Posteriormente, tras la clasificación de los alumnos, se

Manuscrito recibido el 17 de junio de 2021; revisado 24 de junio de 2021; aceptado 30 de julio de 2021. English versión received February, 23th, 20. Revised May, 13th, 21. Accepted May, 13th, 21

Sánchez López Sheila Lucero, Díaz Redondo Rebeca P., Fernández Vilas Ana, ICLAb Universidad de Vigo, España (email sheila.lucero@det.uvigo.es , rebeca@det.uvigo.es, avilas@det.uvigo.es).

estudiará su categorización y se realizará un análisis exploratorio de sus mensajes, también con ayuda del software DepPattern, por cada clasificación de estudiantes, se realizará un análisis minucioso de las principales ideas y temas contenidos en los mensajes que han intercambiado en el foro y finalmente, se estudiará el uso que han hecho los estudiantes de las diferentes funcionalidades disponibles en el foro. Por lo tanto, nuestro objetivo es analizar (y clasificar) los mensajes intercambiados en los foros eLearning por cada perfil de alumno y comprobar si coincide con la descripción del perfil propuesto por la taxonomía.

Como prueba de concepto, nuestro estudio está enfocado en un curso semipresencial del tercer año académico de la Licenciatura en Ingeniería de Telecomunicaciones. Para analizar el contenido de los mensajes y comprobar la aplicación de la taxonomía, se han extraído 580 mensajes del foro de la plataforma oficial e-learning, basada en Moodle. En este curso es de educación mixta, la plataforma Moodle es un complemento a la educación presencial.

Este documento está estructurado de la siguiente forma, en la siguiente sección se proporciona información sobre diferentes trabajos relacionados. Después de eso, se presenta la descripción general y el paso a paso de nuestra metodología; los pasos son el preprocesamiento de datos, la clasificación y análisis de los mensajes, y el análisis de la participación en los foros basado en la taxonomía extendida. La última sección de este documento incluye la discusión de resultados y las conclusiones obtenidas.

II. TRABAJOS RELACIONADOS

A medida que los foros de las plataformas eLearning se han convertido en una enorme colección de perfiles, opiniones y comentarios potencialmente útiles, cada vez más investigadores muestran gran interés en su contenido. En esta sección revisamos el trabajo relacionado y clasificamos las referencias según su enfoque: (i) propuestas que se enfocan en el patrón de conducta del estudiante; (ii) propuestas que establecen distintas taxonomías para clasificar a los estudiantes, y finalmente (iii) trabajos que analizan el contenido de los mensajes.

A. Patrón de comportamiento del estudiante

Uno de los autores que sigue este enfoque es Andresen [8]. Su artículo analiza los componentes de foros exitosos, así como las medidas y limitaciones del aprendizaje en línea. En [9] el comportamiento de deserción se identifica como relevante para las publicaciones del foro de los estudiantes según el análisis de sentimientos de las discusiones de los foros MOOC. Huang et. al. [8] estudia el comportamiento de los usuarios que hacen grandes contribuciones a través de los diferentes foros MOOC ofrecidos por Coursera. De acuerdo con [11], los usuarios que participan constantemente en las discusiones del foro se identifican como usuarios estadísticamente más influyentes, y estos usuarios también tienen un efecto positivo en las discusiones.

Se han presentado muchas propuestas para identificar roles sociales basados únicamente en los patrones estructurales de las conversaciones en foros, como [12], [13] y [14].

Los investigadores en educación han demostrado cómo los estudiantes utilizan diferentes características sociales de los sistemas de gestión del aprendizaje para sus interacciones en grupos pequeños y con toda la clase, y qué implicaciones tiene esto para el diseño del sistema y las estrategias de uso [15]. En [16] se muestran diferentes subgrupos que se podrían utilizar para diseñar sistemas de recomendación que aumentarían la polinización de ideas cruzadas.

Saber que existen estos subgrupos podría permitir a los diseñadores desarrollar herramientas que satisfagan necesidades particulares, por ejemplo, ofreciendo interfaces y servicios personalizados a diferentes grupos de personas, utilizando el historial de otros usuarios del mismo grupo como guía. Un ejemplo de esto es [17], un estudio que muestra las diferentes variaciones que puede tener el mismo foro de discusión para diferentes usuarios (por ejemplo, jóvenes y adultos mayores).

B. Taxonomías: clasificación de estudiantes

La creación de taxonomías tiene como objetivo clasificar a los estudiantes de acuerdo con su interacción. En la literatura, encontramos diferentes propuestas, entre ellas la Taxonomía de Bento [16]. Es una clasificación que consta de cuatro cuadrantes que categoriza a los estudiantes por su tipo de interacción con las plataformas eLearning: (i) interacción interpersonal (también llamada interacción social), definida como “la interacción que ofrece al estudiante la oportunidad de ganar el apoyo motivacional de compañeros e instructores, desarrollar un juicio crítico y participar en la resolución de problemas" [17]. (ii) interacción del estudiante con el contenido, definida como "el proceso de interactuar intelectualmente con el contenido que provoca cambios en la comprensión, en la perspectiva o en las estructuras cognitivas de la mente del alumno"[18].

La figura 1 resume la taxonomía. El cuadrante I incluye a los "*desaparecidos en acción*". A menudo no les importa el contenido de los cursos o aprender de sus compañeros e incluso podrían dejar el curso. El cuadrante II incluye a los "*estudiantes testigos*", que participan activamente en los recursos del curso, inician sesión con frecuencia y realizan todos los ejercicios, pero no contribuyen activamente al foro. Los *"participantes sociales*" del cuadrante III son excelentes conversadores en línea, pero sus habilidades de conversación no se reflejan en la calidad de su contenido y su participación tiene poco orden. El cuadrante IV incluye a aquellos *"aprendices activos*": sus contribuciones son muy buenas para construir su propio aprendizaje y ayudar a la comunidad en línea.

Interpersonal Interaction HIGH	Quadrant III “Social Participants”	Quadrant IV “Active Learners”
Interpersonal Interaction LOW	Quadrant I “Missing in Action”	Quadrant II “Witness Learners”
	Interaction with Content LOW	Interaction with Content HIGH

Fig. 1 Taxonomía de Bento

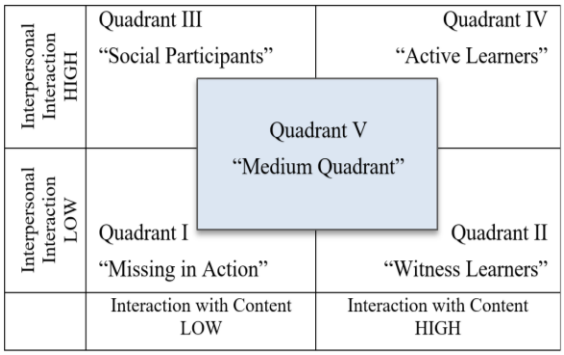


Fig. 2 Taxonomía extendida

En [19] se estudian dos formas de clasificar a los estudiantes. El primero usa técnicas de agrupamiento y el segundo usa lo que ellos llaman taxonomía extendida. Este último obtiene mejores resultados que el primero. La taxonomía extendida [20] propone añadir un cuadrante intermedio a la taxonomía de Bento. Este cuadrante incluye un perfil intermedio con mejores resultados para inferir el desempeño del alumno.

Fig. 2 describe la taxonomía extendida.

Otros enfoques basan su clasificación de los estudiantes en otros tipos de interacciones, como [21], donde se obtienen seis clases de acuerdo con los siguientes tipos de interacción: alumno-profesor, alumno-alumno, profesor-profesor, alumno-contenido, profesor-contenido y contenido-contenido.

La identificación automática de los perfiles de las interacciones de los estudiantes en los foros de discusión se presenta en [22]. Los clasificadores de discurso se desarrollaron para identificar los diferentes roles de los mensajes individuales, como pregunta, respuesta, elaboración y corrección. Estos clasificadores se utilizaron para buscar mensajes que contienen preguntas o respuestas. Para analizar los temas utilizan un conjunto de reglas a fin de encontrar aquellos discursos que pueden tener preguntas sin respuesta y necesitan la atención del profesor.

C. Análisis de contenido del mensaje

Se puede encontrar en la literatura información sobre minería de textos y descubrimiento de conocimiento a partir de datos no estructurados. Un resumen detallado de los principales hallazgos en este campo se encuentra en [23] y [24].

Además, se han realizado muchas investigaciones sobre el procesamiento estadístico del lenguaje natural. Los métodos estadísticos para la minería de texto se describen y analizan en detalle en [25]. El análisis de conversaciones en cadena - que son el patrón de comunicación predominante en la web contemporánea- es un dominio investigado activamente, como lo muestran [26] y [27].

En [2], la idea es medir la presencia de detalles de conocimiento evaluando la profundidad de la información que se está discutiendo y el conocimiento generado a partir de ella por los interlocutores.

En lo que respecta a este trabajo, fusionamos estos enfoques. Primero, se clasifican los mensajes por su contenido. Después, se aplica la taxonomía extendida para analizar el contenido de los mensajes de acuerdo con los cinco perfiles propuestos. Por último, se analiza si cada uno de los perfiles participa de forma diferente en el foro online.

D. Métodos de clasificación

Algunos algoritmos comúnmente utilizados para la clasificación automática de texto son las máquinas de soporte de vectores (VMS), la bolsa de palabras (BoW), el algoritmo de firmas de relevancia, las regresiones logísticas y los clasificadores bayesianos. Los VMS corresponden a máquinas de aprendizaje que toman diferentes características de los elementos a clasificar y los llevan a un espacio vectorial multidimensional. Es en este espacio, donde el algoritmo identifica de forma óptima un hiperplano que separa los vectores de una clase del resto [28]. La representación BoW se utiliza para formar un vector con la frecuencia de los términos o variables predictoras presentes en el documento. Entre sus principales limitaciones está la falta de información sobre las relaciones entre estos términos, especialmente las relaciones entre palabras y las relaciones semánticas que existen a lo largo del documento [29].

El método del algoritmo de firmas de relevancia parte de un conjunto de textos de entrenamiento, se analiza el texto utilizando un software llamado CIRCUS. Los nodos conceptuales que se producen durante el análisis se guardan junto con los elementos léxicos que activaron esos nodos. Aplica diferentes fórmulas estadísticas para cada nodo para decidir correctamente la asociación [30].

La regresión logística es un algoritmo de clasificación que generalmente se usa para averiguar la probabilidad de que una variable esté en una cierta categoría. Tiene varios problemas. Por ejemplo, los clasificadores solo ven los datos de la misma clase como positivos y todos los demás como negativos. Por tanto, existe un desequilibrio a la hora de entrenar el clasificador que utilizará el modelo. En otras palabras, el conjunto de datos de entrenamiento será mucho más negativo que positivo [31].

Los clasificadores bayesianos funcionan utilizando los datos para calcular la probabilidad de que dichos datos de entrada pertenezcan a la clase con la que están etiquetados. Cuando el clasificador se aplica a datos que no están etiquetados, usa esas probabilidades para encontrar la clase más probable.

Después de comparar diferentes métodos, hemos elegido el clasificador bayesiano ingenuo porque es un método clásico, robusto y computacionalmente simple. Este método se usa a menudo cuando muchos atributos de la información deben considerarse simultáneamente para estimar la probabilidad del evento. Además de esto, es un método muy utilizado por sus buenos resultados en la clasificación de textos. En [32] se compara con varios métodos que tienen excelentes resultados incluso en comparación con otros métodos como C4.5.

III. EXPERIMENTO

Para realizar nuestros experimentos, utilizamos datos de un curso de programación en el tercer año académico de la Licenciatura en Ingeniería de Telecomunicaciones. Este es un curso mixto de catorce semanas de septiembre a enero. El conjunto de datos se obtuvo de la plataforma oficial de e-learning basada en Moodle de la Universidad donde se impartió la asignatura.

El mecanismo de evaluación de este curso se basa en tres tareas obligatorias distribuidas desde la segunda hasta la última semana. El foro es accesible para todos los alumnos y sirve para debatir sobre diferentes aspectos relacionados con

la asignatura (contenidos docentes o cuestiones administrativas), para responder preguntas, resolver dudas, etc. Sin embargo, cabe destacar que no es una actividad obligatoria. Las tareas requeridas se dividen en dos tipos:

1) Laboratorio: estas tareas (3 prácticas) determinan si el alumno ha adquirido todos los conocimientos y habilidades correspondientes a las prácticas de laboratorio.
2) Aplicaciones: estos tests (2 exámenes) determinan si el alumno sabe aplicar los conocimientos del curso para resolver problemas.

La plataforma Moodle almacena en su base de datos no solo toda la información relacionada con los cursos (contenidos docentes, calificaciones, datos personales de estudiantes y profesores, etc.), sino también toda la información sobre la interacción de los estudiantes con la plataforma.

De hecho, Moodle distingue entre diferentes tipos de interacciones, que se clasifican en diez módulos diferentes (Asignación, Blog, Elección, Curso, Foro, Notas, Recurso, Carga, Usuario y Prueba) y eventos.

Para nuestro análisis recopilamos datos (580 mensajes) de tres años académicos (14-15, 15-16, 16-17). Como muestra la Figura 3, los intercambios de mensajes en foros han ido aumentando año a año, ya que la aceptación de la plataforma como complemento educativo es significativa.

IV. DESCRIPCIÓN GENERAL DE LA METODOLOGÍA

En [19] se aplica un enfoque cuantitativo al estudio de la interacción social de los estudiantes con buenos resultados: los autores cuentan el número de interacciones en los módulos "sociales" de la plataforma. A diferencia de este trabajo de 2017, en el presente trabajo queremos comparar y analizar cualitativamente las aportaciones de cada perfil. Nuestro objetivo es validar la existencia de los cinco perfiles propuestos por la taxonomía extendida cuando se aplican enfoques cualitativos y cuantitativos. Para lograrlo, nuestra metodología (Figura 4) se divide en dos etapas. El primero es la clasificación de mensajes. Este paso comienza con el preprocesamiento de los datos, continúa con la clasificación de los mensajes en tres grupos y finaliza con un análisis exploratorio del contenido de los mensajes. La segunda etapa es la aplicación de la taxonomía ampliada [20] y el análisis de posibles diferencias en el contenido de los mensajes de los cinco perfiles propuestos. Estos pasos se detallarán en la siguiente sección V.

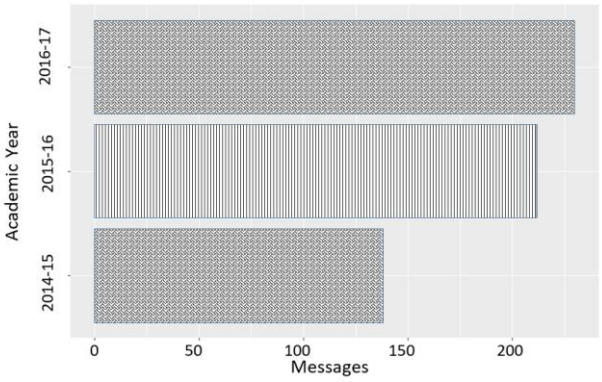


Figura 3 Distribución de mensajes

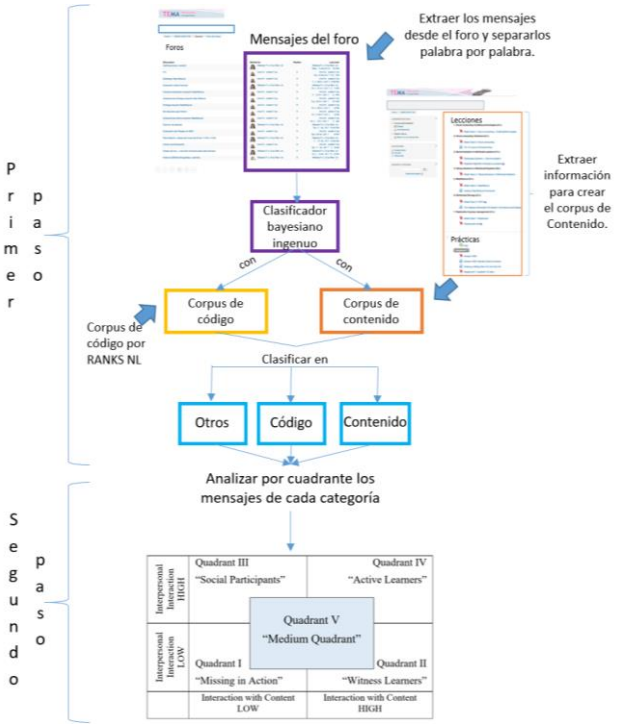


Fig. 4 Metodología

V. CLASIFICACIÓN DE LOS MENSAJES DEL FORO

Como se mencionó anteriormente, esta etapa se divide en tres pasos: el primero corresponde al preprocesamiento de datos; el segundo es una clasificación de los mensajes en tres categorías (contenido, código y otros); y el tercero es el análisis exploratorio del contenido de esos mensajes.

A. Preprocesamiento de datos

Es necesario preparar y transformar la información recopilada para clasificar los mensajes. Inicialmente, se ha creado un corpus de contenido específico para el experimento, extrayendo las palabras principales de 12 archivos pdf: material didáctico (4 archivos pdf), recursos educativos (3 archivos pdf), notas (3 archivos pdf), diapositivas (2 archivos powerpoint), tres ejercicios prácticos (3 archivos) y referencias de las presentaciones en clase (2 archivos pdf). Toda la información está disponible para cualquier alumno matriculado en esta asignatura y con acceso a la plataforma oficial de e-learning de la universidad. De estos documentos hemos obtenido un total de 15.704 palabras. Este conjunto se reduce posteriormente a un corpus de 587 palabras, tras la eliminación de las stopwords, aplicar un proceso de lematización y extraer las palabras más frecuentes de la información. Este corpus se denominará “Corpus de contenido”.

Como se indicó anteriormente, el contenido de la asignatura está relacionado con la informática, especialmente con dos tecnologías de programación: Java y HTML. Como el uso de códigos de programación es muy frecuente, utilizamos un segundo corpus que se denominará “Corpus de Código”. Este corpus, creado por RANKS NL, contiene las palabras principales de los principales lenguajes de programación.

RANKS NL [33] es una herramienta de análisis de palabras clave para URL, sitios web, textos y documentos

para mejorar la optimización de los motores de búsqueda y para otros propósitos. Tiene una colección de listas de stopwords disponibles en más de 40 idiomas y la lista de palabras reservadas de Perl, MySQL, JavaScript, C, C ++ y HTML. Las stopwords planteadas por RANKS NL se eliminarán de todos los mensajes.

En definitiva, los dos corpus que usaremos para clasificar los mensajes son:

1) Corpus de contenido: creado a partir de la extracción de las palabras principales (palabras temáticas) del material didáctico, recursos educativos, notas, diapositivas y prácticas disponibles en la plataforma e-learning. Está compuesto por 587 palabras.

2) Corpus de código: este corpus se utilizará para clasificar mensajes que contengan código de programación y se basa en el corpus de RANKS NL. Contiene 2.500 palabras.

B. Classification

El clasificador bayesiano ingenuo y los dos corpus (Código y Contenido) se utilizarán para clasificar los mensajes intercambiados en los foros. Este clasificador aplica el teorema de Bayes y el supuesto de independencia condicional de las variables predictoras, su estructura no cambia y sigue un criterio generativo o discriminativo. Además, los parámetros se obtienen en base a la estimación de máxima verosimilitud o la estimación máxima a posteriori [34]. Este clasificador ha demostrado buenos resultados en la clasificación de textos [35]. Según el teorema de Bayes, la probabilidad se puede definir como:

(1)

$$p(C|w_1 = y, w_2 = n, w_i = \cdots) = \frac{p(C)p(w_1 = y, w_2 = n, w_i = \cdots)}{p(w_1 = y, w_2 = n, w_i = \cdots)}$$

Donde $p(C)$ es la probabilidad de pertenecer a un corpus específico (Código o Contenido); w_i es el identificador de una palabra;y se refiere al caso cuando pertenece al corpus, y n cuando no es así.

Nuestro interés son las probabilidades relativas de que los mensajes sean un mensaje de código o un mensaje de contenido. En otras palabras, el valor exacto de la probabilidad no es importante porque la clasificación se asignará según el mayor porcentaje de palabras pertenecientes a cualquiera de los corpus.

Por lo tanto, podemos factorizar cualquier término constante, es decir, el denominador de la ecuación anterior es una constante porque depende del número total de mensajes (de ambos corpus). Por esta razón, el numerador de la ecuación (1) se puede escribir como:

(2)

$$p(C)p(|w_1 = y, w_2 = n, w_i = \cdots) = p(C)p(w_1 = y|C)p(w_2 = n|C)p(w_i)$$

La Figura 5 muestra un ejemplo de los mensajes extraídos del foro.

```
Hola. Me han surgido 3 dudas de teoría sobre el Tema 1:
1.- En las transparencias pone que en POP3 la autentificación
del usuario se hace en claro. Pero...¿aún usando ssh se haría
en claro?
2.- En POP3, al hacer un APOP, ¿se está introduciendo un &quot;
identificador&quot; de la cuenta de correo del usuario o es
simplemente la clave como tal? (transparencia 23)
3.- En la transparencia 20 se aprecia un ejemplo del diálogo
cliente-servidor. Veo que al introducir el comando DATA de SMTP
se escribe el mensaje a enviar finalizando en un &quot;.&quot;.
Mi pregunta es: ¿ahí no deberían escribirse además las cabeceras
que vimos en el ejemplo de la transparencia 15?
Gracias.     Un Saludo.
```

Fig. 5 Mensaje extraído del foro.

Cada mensaje seguirá el mismo proceso. El primer paso es dividir cada mensaje en palabras. El segundo es eliminar las stopwords y ejecutar la lematización. El tercero es clasificar los mensajes utilizando el clasificador bayesiano ingenuo, siguiendo el criterio del mayor porcentaje de pertenencia al corpus de código o contenido, siempre y cuando éste sea mayor al 33%. Este porcentaje se recomienda cuando se utiliza este clasificador para detectar spam en correos electrónicos y también se incluye en las recomendaciones de RANKS NL, el creador del corpus de código. Finalmente, obtenemos tres categorías.

1) Mensajes de código: mensajes donde al menos el 33% de su contenido pertenece al corpus de código.

2) Mensajes de contenido: mensajes donde al menos el 33% de su contenido pertenece al corpus de contenido.

3) Otros mensajes: los mensajes que no pertenecen a ninguno de los dos corpus.

El procedimiento considera que un mismo mensaje puede contener palabras que pertenezcan a los dos corpus. Primero, calcula la probabilidad de pertenencia de cada palabra que hace referencia al corpus de código. Luego, calcula la pertenencia al corpus de contenido, palabra por palabra, hasta que se supera el porcentaje de pertenencia del corpus de código o se termine el análisis de todas las palabras del mensaje. Finalmente, el mensaje se clasifica según el mayor porcentaje, siempre que supere el 33%. La Fig. 6 (localizada en la siguiente página) describe el procedimiento para las dos primeras etapas: preprocesamiento y clasificación de datos.

Un profesor que también es experto en programación revisó cada mensaje para clasificarlos manualmente en los tres grupos identificados. Los resultados de la clasificación manual muestran que sólo el 7,2% de los mensajes corresponden a una categoría diferente a la asignada por el clasificador bayesiano ingenuo. Con esta información, hemos calculado otras medidas interesantes, como precisión, recall y medida F, resumidas en la Tabla 1.

Como se mencionó anteriormente, hemos decidido seguir la recomendación de RANKS y usar el 33% como umbral para decidir si un mensaje pertenece a la categoría Contenido. Aplicando este porcentaje se obtuvieron buenos resultados.

TABLA 1
PRUEBA DE EXACTITUD

	Precision	Recall	F-score
Código	88.5%	88.5%	3.54
Contenido	93.7%	94.3%	3.77
Otros	92.7%	91.9%	3.68

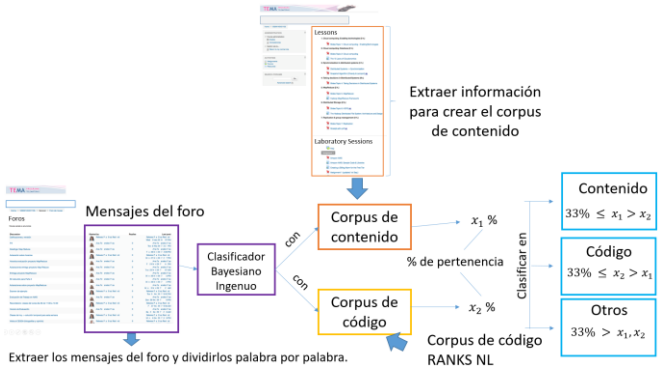


Fig. 6 Preprocesamiento y clasificación

Sin embargo, decidimos realizar algunas pruebas cambiando este umbral. Tras un trabajo exhaustivo, detectamos que si aumentamos el umbral a 52% mejoran los resultados, la tasa de error se redujo al 5,7% y el recall aumentó 1,5%, llegando así al 94,3%. Tras notar estas mejoras en la clasificación de contenido, también se amplió el umbral para la categoría de Código, teniendo que el mejor resultado se obtiene aplican un umbral del 35%. pasando la tasa de errose optimizan los resultados se optimizaron utilizando un umbral superior, concretamente el 52%. Nuestra tasa de error se redujo al 5,7% y el recall aumentó del 92,8% al 94,3%. Esto nos motivó a comprobar qué pasaría si también se modificaba el umbral de la categoría de Código. Después de repetir el análisis, optimizamos nuestros resultados aumentando este umbral al 35%.

Finalmente, la Tabla 2 resume la distribución de mensajes por año académico: los mensajes de contenido son claramente los que se intercambian con mayor frecuencia y los mensajes de código son los más escasos. Dado que los porcentajes son bastante similares en los tres cursos académicos, hemos decidido centrar el análisis en un único conjunto de datos que incluye la información de los tres cursos académicos (2014-15, 2015-16 y 2016-17).

C. *Análisis de mensajes*

Es importante destacar que se analizará el contenido de las tres categorías realizando un análisis exploratorio. Se buscarán las palabras más utilizadas en los mensajes previamente clasificados. Esto permitirá saber cuáles son las palabras principales y si existe una relación entre los diferentes corpus. El siguiente paso será graficar la co-ocurrencia de las palabras, de modo que estas relaciones se muestren claramente, esto se puede observar en la Figura 7, donde n representa la co-ocurrencia de las palabras.

DISTRIBUCIÓN POR CLASIFICACIÓN

	2014-15		2015-16		2016-17		Total	
	No.	%	No.	%	No.	%	No.	%
Código	18	13%	18	8%	25	11%	61	13%
Contenido	68	49%	102	48%	113	49%	283	49%
Otros	52	38%	92	43%	92	38%	236	38%

Se encontró que varias palabras se utilizan indistintamente en la categoría Código y la categoría Contenido, como "entidad", "permisos", "cortafuegos", "navegador" o "ruta". Algunas palabras también aparecen en la categoría Otro y en la categoría Contenido, como "examen", "resultados", "entregar", "archivo adjunto" o "corrección". Finalmente, hay palabras que aparecen en los tres corpus: “php”, “tomcat”, “consulta”, “servidor”, etc. Podemos ver que las palabras de la categoría Otros, suelen referirse a temas relacionados con la administración del curso, por lo tanto, se le asignará este nombre a la categoría. También se considera importante saber conocer cuántos mensajes envían los estudiantes en cada categoría. La Figura 8 muestra la media de mensajes por alumno y por categoría (8,34 mensajes): 2,16 en la categoría Código, 2,47 en la categoría Contenido y 3,71 en la categoría Administración del curso. Para calcular la media de mensajes por categoría, se ha considerado el número total de estudiantes activos en los foros eLearning, independientemente de si el estudiante envió un mensaje o no en esa categoría. En otras palabras, hemos dividido el número de mensajes enviados en cada categoría por el número de estudiantes activos en el foro. Un estudiante se considera activo si ha enviado al menos un mensaje en el foro durante el curso.

La administración del curso es el grupo de mensajes que presenta mayor variación. Los resultados del cálculo de la varianza de cada categoría fueron 3,18 para la categoría Código, 5,81 para la categoría Contenido y 11,18 para la

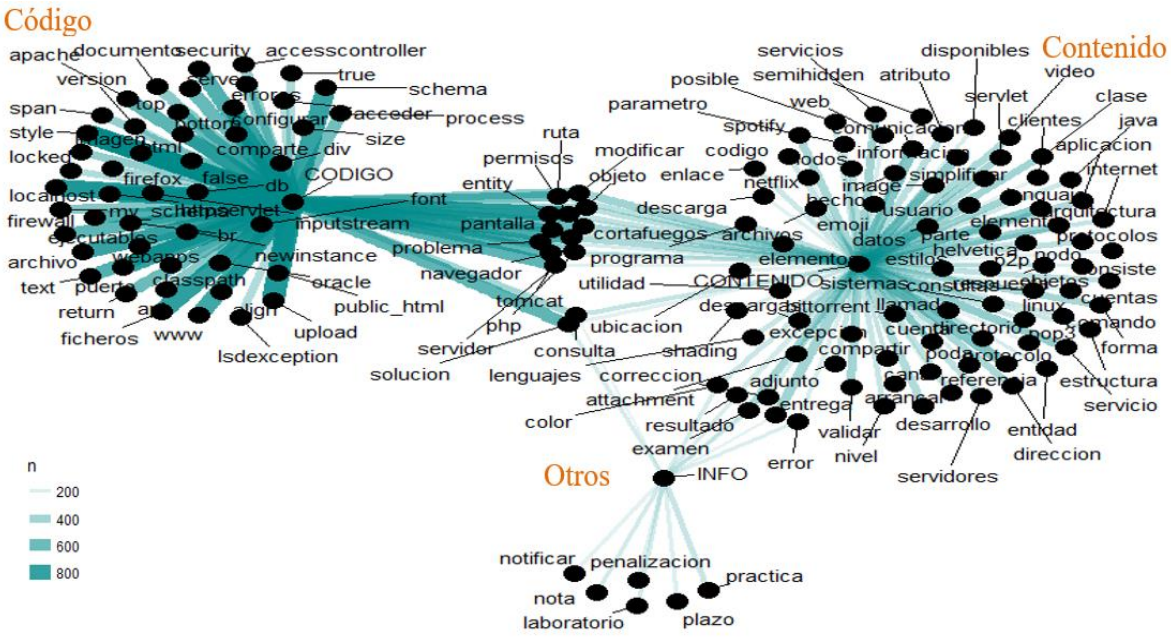


Fig. 7 Co-occurrencia de palabras

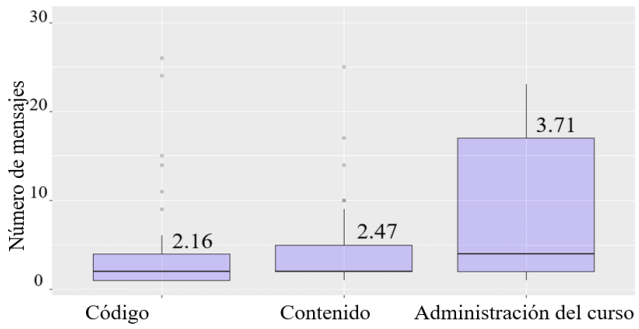


Fig. 8 Media de mensajes por categoría

categoría Administración del curso.

La Figura 9 muestra la distribución de los mensajes para cada categoría a lo largo del período académico. Esto permite visualizar la evolución temporal del intercambio de mensajes. Como se puede observar en la Figura 9, hay 3 picos de mensajes en la categoría Código (círculos azules). El primero corresponde a la entrega del primer ejercicio práctico, el segundo a la revisión del segundo ejercicio práctico y el tercero a la entrega del tercer ejercicio práctico. También hay 2 picos en los mensajes de la categoría Contenido (círculos naranjas), correspondientes a la sesión previa a los exámenes. En cuanto a los mensajes de la categoría Administración del curso, no existe un patrón aparente.

Como parte de nuestro análisis exploratorio, se considera importante saber ¿quién inicia las publicaciones en el foro: el profesor o algún alumno? El resultado de nuestro análisis muestra que el 71% de las conversaciones relacionadas con la categoría código, el 63% de las conversaciones relacionadas con el contenido y el 55% de las conversaciones de información del curso son iniciadas por los estudiantes.

Como la mayoría de los hilos publicados son iniciados por los estudiantes, el siguiente paso sería averiguar ¿quién suele responder estos mensajes con mayor frecuencia: el profesor o los estudiantes? De acuerdo con el análisis, en el 89% de los mensajes la primera respuesta la realiza otro alumno, únicamente el 11% de los mensajes tienen respuesta inicialmente por el profesor. Un dato interesante, en el que se aprecia que los alumnos tienden a alimentar el foro.

VI. PERFILES DE ESTUDIANTES Y PARTICIPACIÓN EN EL FORO

Como se indicó anteriormente, en esta investigación se aplica la taxonomía extendida [20]. Para clasificar a los estudiantes en los cinco perfiles o cuadrantes de esta taxonomía, es necesario definir los umbrales para la

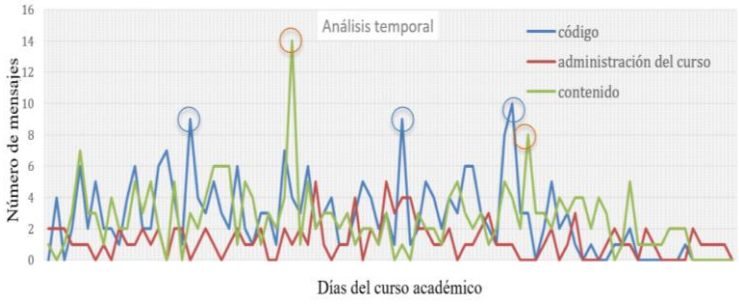


Fig. 9 Análisis Temporal

TABLA 3
UMBRAL PARA APLICAR LA TAXONOMÍA EXTENDIDA

Cuadrante	Interacción interpersonal	Interacción con el contenido
Q1	< 51	< 161.5
Q2	< 51	>218.5
Q3	> 69	< 161.5
Q4	> 69	> 218.5
Q5	51 ≤ ≥69	161.5≤ ≥218.5

interacción social y la interacción con el contenido.

De acuerdo con la metodología de la taxonomía extendida, es necesario establecer la "Interacción Media", para ello se debe calcular el intervalo, considerando el ± 15% del promedio de las interacciones. Una vez calculado esto, se debe aplicar el criterio: Si se excede el intervalo medio, se considera una "Interacción alta", de lo contrario, se considera una "Interacción baja" para cada tipo de interacción. En la Tabla 3 se muestra los parámetros calculados para aplicar la taxonomía ampliada.

Después de calcular los umbrales, se establecen los intervalos para clasificar a los estudiantes según su número de interacciones. Estos parámetros se muestran en la Tabla 4.

Una vez clasificados los alumnos en los cinco cuadrantes aplicando la taxonomía extendida, se obtiene la siguiente distribución: 9 alumnos pertenecen a Q1, 8 alumnos pertenecen a Q2, 17 alumnos a Q3, 21 alumnos a Q4 y 13 alumnos a Q5.

Con el objetivo de relacionar los cuadrantes con las categorías de mensajes, se ha calculado el porcentaje de distribución de cada corpus por cuadrante, estos resultados se muestran en la Tabla 5 (es necesario enfatizar que solo se consideran los datos de los estudiantes que interactúan en el foro).

La segunda columna de la Tabla 5 ("% estudiantes") muestra el porcentaje total de mensajes que pertenecen a cada cuadrante y las siguientes 3 columnas ("código", "curso adm" y "contenido") muestran el porcentaje de pertenencia de cada grupo en ese cuadrante, considerando cada cuadrante como 100%. Por ejemplo, en el cuadrante 1, el 52% de los mensajes son de código, el 20% de administración del curso y el 28% restante, de contenido; sin embargo, estos mensajes solo representan el 4% del número total de mensajes en el foro. Este resultado se representa gráficamente en la Figura 10.

TABLA 4
PARÁMETROS PARA APLICAR LA TAXONOMÍA

Interacción	Media	Baja	Alta
Interpersonal	60	51	69
Contenido	190	161.5	218.5

TABLA 5
DISTRIBUCIÓN DE LA CLASIFICACIÓN

Cuadrante	% estudiantes	Mensaje del foro		
		código	admin del curso	contenido
Q1	4%	52%	20%	28%
Q2	3%	17%	0%	83%
Q3	30%	50%	0%	50%
Q4	47%	42%	22%	36%
Q5	16%	36%	8%	55%

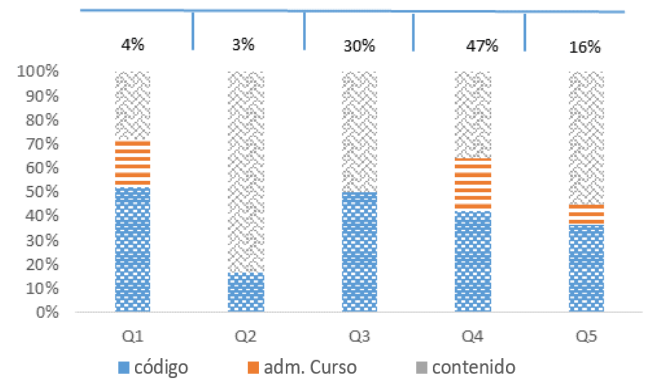


Fig. 10 Representación gráfica de la tabla 5

Adicionalmente, como hemos comentado al principio, queremos comprobar si las características del perfil propuesto por la taxonomía extendida coinciden con el análisis y la clasificación de los mensajes que cada uno de los cinco perfiles propuestos ha intercambiado en el foro. Por lo tanto, se analizan los temas y tópicos de los mensajes intercambiados con un programa llamado DepPattern[36]. Este programa es un paquete lingüístico que proporciona un compilador gramatical, etiquetadores de PoS y analizadores basados en dependencias para varios idiomas, incluidos el español y el gallego. Esta es una característica muy importante porque los mensajes en el foro están escritos en dos idiomas (español y gallego).

La Figura 11 muestra un ejemplo de los resultados obtenidos con el software. Éste se ejecutó 5 veces, una por cuadrante y cada ejecución se realizaron únicamente con los mensajes pertenecientes al cuadrante. Se obtuvo la lista de verbos en infinitivo, signos de puntuación y sustantivos de los mensajes. Al interpretar los resultados, se obtuvieron los siguientes temas:

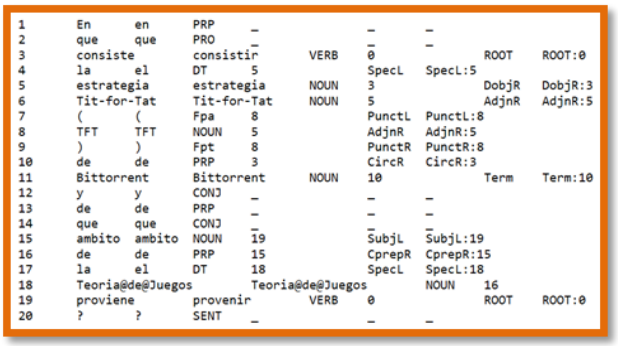
1) Q1. - Preguntas principalmente relacionadas con instrucciones, así como fechas y horarios de entrega de trabajos y exámenes.

2) Q2. - Recomendaciones de contenido adicional y preguntas más específicas, incluyendo preguntas sobre la relevancia de los ejercicios.

3) Q3. - Preguntas sobre fechas de exámenes y fechas de entrega, ciertas suposiciones y preguntas más generales.

4) Q4. – Solicitudes de ejemplos, referencias a diapositivas de clase, enlaces web, comentarios críticos y respuestas a preguntas planteadas por otros alumnos.

5) Q5.- Opciones de respuesta, varias preguntas incluidas en un mismo mensaje, ejemplos y alternativas, recursos extra y mensajes más extensos.



1	En	en	PRP	_		_	_	
2	que	que	PRO	_		_	_	
3	consiste		consistir		VERB	0	ROOT	ROOT:0
4	la	el	DT	5		SpecL	SpecL:5	
5	estrategia		estrategia		NOUN	3	DobjR	DobjR:3
6	Tit-for-Tat		Tit-for-Tat		NOUN	5	AdjnR	AdjnR:5
7	(	(	Fpa	8		PunctL	PunctL:8	
8	TFT	TFT	NOUN	5		AdjnR	AdjnR:5	
9	)	)	Fpt	8		PunctR	PunctR:8	
10	de	de	PRP	3		CircR	CircR:3	
11	Bittorrent		Bittorrent		NOUN	10	Term	Term:10
12	y	y	CONJ	_		_	_	
13	de	de	PRP	_		_	_	
14	que	que	CONJ	_		_	_	
15	ambito	ambito	NOUN	19		SubjL	SubjL:19	
16	de	de	PRP	15		CprepR	CprepR:15	
17	la	el	DT	18		SpecL	SpecL:18	
18	Teoria@de@Juegos		Teoria@de@Juegos			NOUN	16	
19	proviene		provenir		VERB	0	ROOT	ROOT:0
20	?	?	SENT	_		_	_	

Fig. 11 Resultados aportados por DepPattern

Como se puede observar, la distribución de mensajes por cuadrante es muy diferente a excepción de los cuadrantes 1 y 4 (Tabla 5) a pesar de pertenecer a perfiles opuestos. Sin embargo, es necesario tener en cuenta que el tema de los mensajes es diferente: teniendo que ninguno de los mensajes en Q1 es la respuesta a ninguna pregunta propuesta por otro estudiante, la mayoría de los mensajes en Q4 son respuestas a las preguntas de otros estudiantes.

Debemos tener en cuenta que la taxonomía extendida estudia la interacción interpersonal desarrollada en la plataforma Moodle, es decir, se enfoca en los tres módulos donde este tipo de interacción está disponible: blog, encuesta y foro. Sin embargo, en este trabajo se analiza minuciosamente el comportamiento de los estudiantes, considerando únicamente el módulo del foro, lo que nos permite detectar diferentes comportamientos en estudiantes que pertenecen a diferentes cuadrantes. La Figura 12 muestra la media de mensajes por cada categoría de mensajes en cada cuadrante.

Los cuadrantes Q3 y Q4 claramente tienen una tasa de participación más alta en el módulo del foro con 6 y 11,8 mensajes, respectivamente. Estos resultados concuerdan con la teoría de la taxonomía extendida. Q3 y Q4 tienen baja y alta interacción con el contenido respectivamente, pero ambos presentan alta interacción social.

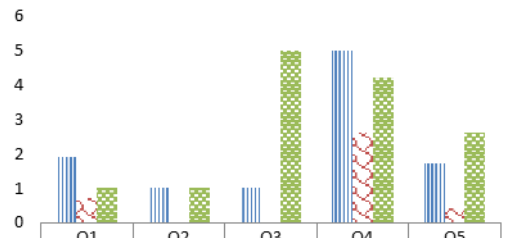
El siguiente paso es analizar toda la interacción de los estudiantes dentro de las opciones que tienen habilitadas el módulo del foro. En otras palabras, el estudio incluye todas las opciones dentro de este módulo, no solamente el número de mensajes escritos en el foro.

El módulo de foro tiene 12 eventos diferentes:

> *ver foro, suscribirse, agregar una discusión, ver discusión, ver foros, informe de usuario, agregar una publicación, actualizar publicación, actualizar, buscar, cancelar suscripción y suscribir todo*.

Sin embargo, no todos los estudiantes usan los mismos eventos, de hecho, cada cuadrante tiene un comportamiento diferente incluso dentro del mismo módulo. No obstante, no todos los eventos se utilizan con la misma frecuencia: la Tabla 6 muestra el porcentaje de uso de cada evento por cuadrante.

Como se puede observar, Q4 usa todos los eventos disponibles, Q3 usa 10, seguidos de Q5 que usa 9 y finalmente, Q1 y Q2, usan 6 y 5 respectivamente.



	Q1	Q2	Q3	Q4	Q5
Code	1.9	1	1	5	1.7
course adm	0.7	0	0	2.6	0.4
Content	1	1	5	4.2	2.6

Fig. 12 Media de mensajes por cuadrante

TABLA 6
PORCENTAJE DE USO DE CADA EVENTO POR CUADRANTE

Action	Q1	Q2	Q3	Q4	Q5	Total
ver foro	9%	1%	10%	57%	22%	100%
Subscribirse	0%	0%	16%	68%	16%	100%
añadir discusión	4%	0%	5%	76%	14%	100%
ver discusión	9%	1%	11%	55%	24%	100%
Ver foros	9%	2%	12%	52%	26%	100%
Reporte de usuario	8%	0%	21%	50%	21%	100%
Agregar post	3%	1%	6%	82%	7%	100%
actualizar post	0%	0%	5%	85%	10%	100%
Buscar	0%	7%	0%	67%	27%	100%
Dares de baja	0%	0%	50%	50%	0%	100%
Suscribirse a todo	0%	0%	50%	50%	0%	100%

VII. DISCUSIÓN Y CONCLUSIONES

Esta sección resume el experimento. En la fase inicial de nuestra metodología, se utilizaron dos corpus. El primero (corpus de Contenido) fue creado específicamente con el contenido académico del curso y el segundo (corpus de Código) fue creado por RANKS NL. Aplicando el clasificador Bayesiano Ingenuo y estos dos corpus, se han obtenido tres categorías: código, contenido y administración del curso. Los dos primeros, están compuestos por mensajes con un alto porcentaje de palabras relacionadas con el código y el contenido del curso (mínimo 33% de pertenencia). Los mensajes que no se clasifican en estas dos categorías van directamente a la tercera (administración del curso). El nombre de esta última categoría está relacionado con el tema de los mensajes que contiene, en este caso, la administración del curso (preguntas y / o información sobre los exámenes, fechas de revisión, etc.).

Como se afirma ampliamente en la literatura relacionada, la actividad en las plataformas de e-learning permite a los investigadores identificar diferentes comportamientos o perfiles de estudiantes. En este artículo utilizamos los cinco perfiles definidos en la taxonomía ampliada [5]. Después de combinar nuestros resultados en la primera etapa (análisis de mensajes) con estos cinco perfiles, los resultados reforzaron efectivamente la idea inicial de diferentes propuestos por la taxonomía extendida.

Cada categoría de estudiante ya presenta diferencias en los mensajes intercambiados en los foros eLearning: diferente número de mensajes, diferente distribución de los mensajes en las tres categorías (código, contenido y administración del curso), diferentes pautas en el contenido de los mensajes y diferente uso en la sección del mismo foro.

Concluimos que los foros pueden proporcionar información importante. A pesar de que la mayoría de los mensajes y las palabras más recurrentes se centran en los temas revisados en la asignatura, encontramos mensajes que muestran las preferencias de los estudiantes. Sin embargo, cuando analizamos las palabras más recurrentes con las secciones más utilizadas del foro y aplicamos la taxonomía extendida, obtenemos diferencias más notorias entre cada uno de los cinco perfiles. Es decir, cada perfil utiliza cada función del foro de forma diferente y con diferente frecuencia, siguiendo esta misma línea, la clasificación de sus mensajes muestra una diferencia notable en el contenido que comparten en sus mensajes. El análisis exploratorio mostró las palabras más recurrentes en cada una de las 3 categorías detectadas, lo que principalmente nos permite identificar las dudas en cada tema. Al analizar el contenido de los mensajes obtenemos información de la parte teórica del curso y detectamos palabras como *"estructura", "java", "servlet", "atributos", "servidores", "parametros", "cortafuegos"*, etc. Respecto a la parte práctica (mensajes de código) encontramos palabras como *"seguridad", "apache", "AccessController", "firewall", "archivos", "estilo", "bloqueado", etc. y respecto a la administración del curso: "Examen", "plazo", "laboratorio*", etc.

Posteriormente, al aplicar DepPattern pudimos interpretar todos los mensajes de cada cuadrante obteniendo más información, por ejemplo, detección de la estructura sintáctica y el número de interrogantes usados. Detectamos que la mayoría de las palabras recurrentes provienen de preguntas planteadas por los estudiantes, por lo que podemos concluir que las palabras más recurrentes nos permiten identificar las principales dudas.

Finalmente, el análisis muestra una diferencia de actitud entre los estudiantes. Los cinco perfiles son marcadamente diferentes en su desempeño y el contenido que comparten. Concluimos que los estudiantes de Q1 representan aquellos alumnos "perdidos en acción", Q2 representa a los "aprendices testigos", Q3 representa a "participantes sociales", Q4 representa a aquellos "aprendices activos" y Q5 son "estudiantes intermedios".

Profundizando un poco más, de acuerdo con los datos, la taxonomía extendida indica que Q3 y Q4 presentan una alta interacción social. Contrastando esta información, tenemos que estos cuadrantes representan el 77% de todos los mensajes del foro; más específicamente, Q3 representa el 30% y Q4 representa el 47% de los mensajes. En Q3 no hay mensajes clasificados como administración de cursos. Una última diferencia entre ambos cuadrantes es el tipo de eventos: los usuarios de Q3 no realizan búsquedas, lo cual fundamenta la caracterización dada por la taxonomía extendida, la cual menciona que Q3 tiene menos interacción con el contenido que Q4.

Aunado a esto, basado en la taxonomía extendida, Q1 y Q2 se caracterizan por una baja interacción social. Esta idea también se ve reforzada tras el análisis de contenido en los foros. Los mensajes de los estudiantes de Q1 solo representan el 4% de todos los mensajes y los mensajes de Q2 representan el 3%. Ambos cuadrantes muestran el uso más atípico de los eventos disponibles en el módulo del foro: ninguno de ellos se suscribe a las notificaciones del foro ni actualiza las publicaciones. Finalmente, los estudiantes de Q1 no realizan búsquedas y Q2 es el único cuadrante donde los estudiantes no agregan discusiones.

De acuerdo con la taxonomía extendida, Q5 es un perfil intermedio. Nuestro análisis concluye que los estudiantes que pertenecen a este cuadrante emplean todas las funciones habilitadas en el foro con excepción de: suscribirse a todas las publicaciones y darse de baja. Aunado a esto y tomando como referencia el contenido de sus mensajes, está claro que los mensajes muestran proactividad, pero también cierto grado de inseguridad.

Nuestro experimento es una prueba de concepto, ya que ha tenido éxito, se planea expandir a más sujetos.

Estos resultados dejan claro que la taxonomía extendida es válida para clasificar a los estudiantes bajo enfoques cualitativos y cuantitativos. Además de esto podemos decir que sí, el tipo de contenido que compartes en los foros y la forma en la que usas los foros puede revelar el tipo de

estudiante que eres. Esto podría marcar una pauta importante para el proceso educativo, al permitir un análisis del contenido de los mensajes enviados a los foros y una adecuada identificación de los perfiles de los estudiantes. Esto permite analizar datos valiosos sobre el comportamiento de los estudiantes, y tanto los profesores como los estudiantes pueden aplicar modelos de aprendizaje y learning analytics para mejorar la calidad de la educación y la participación de los estudiantes.

Una buena clasificación de estudiantes en cuadrantes en función de su interacción podría ser útil para tomar las medidas más adecuadas para mejorar el proceso de aprendizaje. Tanto es así que actualmente estamos trabajando en enriquecer esta metodología con la predicción de fallos. El objetivo principal será reforzar las interacciones específicas que el alumno necesita mejorar. Además, el análisis de los mensajes puede aportar una retroalimentación a los profesores sobre aquellos temas que se consideran más interesantes o aquellos en los que más suelen surgir dudas.

AGRADECIMIENTOS

Este trabajo está financiado por el Fondo Europeo de Desarrollo Regional (FEDER) y la Xunta de Galicia en el marco del convenio de financiación del Centro Atlántico de Investigación en Tecnologías de la Información y la Comunicación (AtlantTIC); Los autores también agradecen a GRADIANT por su apoyo informático ya la Universidad de Vigo por su servicio de e-Learning, y por su apoyo.

REFERENCIAS

[1] J. Costley, "The effects of instructor control on critical thinking and social presence: Variations within three online asynchronus learning enviroments.," J. Education Online, vol. 13, pp. 109-171, 2016.

[2] M. Zaidi-Abd-Rozan and Y. Mikami, The Presence of Beneficial Knowledge in Web Forum: Analysis by Kipling's Framework, Malaysa: KMICE, 2006.

[3] R. M. Palloff and K. Pratt, "O aluno virtual: um guia para trabalhar com estudantes on-line," in ArtMed, Porto Alegre, 2004.

[4] B. F. Terra-Azevedo, P. A. Behar and E. Berni-Reategui, "Qualitative Analysis of Discussion Forums," International Journal of Computer Information Systems and Industrial Management Applications, vol. 3, pp. pp. 671-678, 2011.

[5] S. L. Sánchez-López, R. P. Díaz-Redondo and A. Fernández-Vilas, "Predicting students' grade based on social & content interactions," International Journal of Engineering Education, vol. 34, no. 3, pp. 940-952, 2018.

[6] M. Andresen, "Asynchronous discussion forums: Success factors, outcomes, assessments, and limitations," Educational Technology & Society , no. 12, p. 249–257 , 2009.

[7] M. Wen, D. Yang and C. Rosè, "Sentiment analysis in MOOC discussion forums: What does it tell us?," in Proceedings of Educational Data Mining, 2014.

[8] J. Huang, A. Dasgupta, A. Ghosh and J. Manning, "Superposter behavior in mooc forums," Proceedings of the First ACM Conference on Learning@ Scale Conference, p. 117–126, 20114.

[9] J.-S. Wong, B. Pursel, A. Divinsky and B. Jansen, "An Analysis of MOOC Discussion Forum Interactions from the Most Active Users," in Social Computing, Behavioral-Cultural Modeling, and Prediction, vol. 9021, Springer International Publishing Switzerland 2015, 2015, pp. 452-457.

[10] H. C. White, S. A. Boorman and R. L. Breiger, "Social-Structure from Multiple Networks: 1. Blockmodels of Roles and Positions," American Journal of Sociology, vol. 81, no. 4, pp. 730-780, 1976.

[11] D. Fisher, M. Smith and H. T. Welser, "You Are Who You Talk To: Detecting Roles in Usenet Newsgroups," in 39th Annual Hawaii International Conference on System Sciences, IEEE Computer Society, 2006.

[12] M. Morzy, "An Analysis of Communities in Different Types of Online Forums," International Conference on Advances in Social Networks Analysis and Mining, pp. 341-345, 2010.

[13] C. Haythornthwaite, "Exploring Multiplexity: Social Network Structures in a Computer-Supported Distance Learning Class," The Information Society, vol. 17, no. 3, p. 211–226, 2001.

[14] S. A. Munson and P. Resnick, "Presenting diverse political opinions: how and how much," in Proceedings of the 28th international conference on Human factors in computing systems, New York, USA: ACM, 2010, pp. 1457-1466.

[15] P. Zaphiris and R. Sarwar, "Trends, similarities, and differences in the usage of teen and senior public online newsgroups," ACM Trans. Comput.-Hum. Interact., vol. 13, no. 3, p. 403–422, 2006.

[16] R. Bento, B. Brownstein, E. Kemery and S. Rawson Zacur, "A Taxonomy Of Participation In Online Courses," Journal of College Teaching & Learning, vol. 2, no. 12, pp. 79-86, 2005.

[17] Z. L. Berge, "Conceptual Frameworks in Distance Training and Education. In Deborah Schreiber and Zane L. Berge, Eds," Distance Training: How innovative organizations are using technology to maximize learning and meet business objectives, pp. 19-36, 1998.

[18] M. G. Moore, "Three Types of Interaction," The American Journal of Distance Education, vol. 3, no. 2, pp. 1-6, 1989.

[19] S. L. Sánchez López, R. P. Díaz Redondo and A. Fernández Vilas, "Discovering knowledge from student interactions: clustering vs classification," in TEEM'17 International Conference Technological Ecosystems for Enhancing Multiculturality, Cádiz, 2017.

[20] S. L. Sánchez López, R. P. Díaz Redondo and A. Fernández Vila, "Is interpersonal participation relevant to pass?," in TEEM'16 International Conference Technological Ecosystems for Enhancing Multiculturality, Salamanca, Spain, 2016.

[21] T. Anderson and D. Garrison, "Learning in a networked world: New roles and responsibilities," in Distance learners in higher education: Institutional responses for quality outcomes, Madison, Atwood Publishing, 1998.

[22] S. Ravi and J. Kim, "Profiling Student Interactions in Threaded Discussions with Speech Act Classifiers," Proceedings of the 13th International Conference on Artificial Intelligence in Education (AI-ED), pp. 357-364, 2007.

[23] R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge: Cambridge University Press, 2006.

[24] S. M. Weiss, N. Indurkhya and T. Zhang, Text Mining: Predictive Methods for Analyzing Unstructured Information, New York: Springer, 2005.

[25] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, Cambridge: MIT Press, 1999.

[26] V. Gómez, A. Kaltenbrunner and V. López, "Statistical analysis of the social network and discussion threads in Slashdot," in 17th International World Wide Web, Beijing, China, 2008.

[27] X. Shi, J. Zhu, R. Cai and L. Zhang, "User Grouping Behavior in Online Forums," in 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, París, France, 2009.

[28] T. Joachims, Learning to classify text using Support Vector Machines., Dordrecht: Kluwer Academic Publishers., 2002.

[29] B. S. G. D. S. &. M. S. Harish, «Representation and classification of text documents: A brief review.,» IJCA Special Issue on Recent Trends in Image Processing and Pattern Recognition, vol. 2, pp. 110-119, 2010.

[30] E. Riloff y W. Lehnert, «Classifying Texts Using Relevancy Signatures,» de DBLP, Amherst, 1992.

[31] J. R. Berrendero, Clasificación y regresión logística, Madrid: Universidad Autónoma de Madrid, 2019.

[32] N. Friedman, D. Geiger y M. Goldszmidt, «Bayesian Network Classifiers,» Machine Learning, vol. 29, nº 2-3, pp. 131-163, 1997.

[33] D. Doyle, "ranks.nl," Ranks, [Online]. Available: https://www.ranks.nl/about. [Accessed 11 2017].

[34] G. Santafé, J. A. Lozano and P. Larrañaga, "Aprendizaje discriminativo de clasificadores Bayesianos," Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial, no. 29, pp. 39-47, 2006.

[35] C. Palazuelos, D. García-Saiz and M. Zorrila, "Social Network Analysis and Data Mining: An Application to the E-learning Context,"

Proceedings of the 5th International Conference on Computational Collective Intelligence, pp. 651-660, 2013.

[36] P. Gamallo-Otero and I. González, "DepPattern: a Multilingual Dependency Parser," in The 10th International Conference on the Computational Processing of Portuguese, Coimbra, Portugal, 2012.

Sheila Lucero Sánchez López es Ingeniera Mecatrónica por la Universidad Tecnológica de México (2012) y Doctora en Tecnologías de la Información y las Comunicaciones por la Universidad de Vigo (2019). Realizó su máster en Investigación en Ingeniería de Procesos y Sistemas en la Universidad de Valladolid (2014). Actualmente, trabaja en el campo de la analítica de datos y la inteligencia empresarial y es profesora de la Universidad Antonio de Nebrija. (email:sheila.lucero@det.uvigo.es)

Rebeca P. Díaz Redondo es Ingeniera de Telecomunicaciones por la Universidad de Vigo (1997), Doctora en Ingeniería de Telecomunicaciones por la misma universidad (2002) y Profesor Titular del Departamento de Ingeniería Telemática de la Universidad de Vigo. Sus intereses de investigación han evolucionado desde la aplicación de técnicas de razonamiento semántico en el campo de las aplicaciones de TV Digital Interactiva a otras técnicas de caracterización de contenidos basadas en el etiquetado colaborativo. Actualmente trabaja en la aplicación de técnicas de minería social y análisis de datos para caracterizar el comportamiento de usuarios y comunidades para diseñar soluciones en áreas de aprendizaje, ciudades inteligentes y negocios. Actualmente participa en las actividades científicas y técnicas de varios proyectos educativos y de investigación nacionales y europeos. (email:rebeca@det.uvigo.es)

Ana Fernández Vilas es profesora asociada del Departamento de Ingeniería Telemática de la Universidad de Vigo e investigadora del Laboratorio de Información y Computación (Centro de Investigación AtlantTIC). Recibió su doctorado en Informática por la Universidad de Vigo en 2002. Su actividad investigadora en el laboratorio de I&C se centra en Inteligencia Semántica-Social y minería de datos. Ella se desarrolla en campos de la planificación urbana y análisis del aprendizaje. Además, está involucrada en varios proyectos de movilidad y cooperación con países del norte de África y los Balcanes Occidentales. (email:avilas@det.uvigo.es)