Clasificación Automática de Foros de Discusión de Acuerdo al Dominio Cognitivo de la Taxonomía de Bloom Empleando Minería de Texto y un Clasificador Bayesiano

Pincay, J.; Ochoa, X. Ph.D.
Facultad de Ingeniería en Electricidad y Computación
Escuela Superior Politécnica del Litoral (ESPOL)
Campus Gustavo Galindo, Km 30.5 vía Perimetral
Apartado 09-01-5863. Guayaquil-Ecuador
{jpincay, xochoa}@espol.edu.ec

Resumen

Este artículo presenta la implementación de un sistema de clasificación automática para categorizar las respuestas de estudiantes a foros de discusión de acuerdo al dominio cognitivo de la taxonomía de Bloom. Diversos estudios se han llevado a cabo en esta área, en esta ocasión la efectividad y desempeño de un clasificador Bayesiano específicamente el clasificador Naïve Bayes son analizados, empleado conjuntamente con técnicas de minería de texto y un conjunto de datos previamente clasificado por expertos. Otro aspecto que se estudia es el determinar si el sistema de clasificación puede compararse o considerarse como un codificador humano. Varias pruebas fueron llevadas a cabo con el objetivo de obtener métricas cuyos valores permitan evaluar el desempeño del clasificador y calidad de los resultados. Los resultados obtenidos indican que usando la arquitectura propuesta es posible lograr los objetivos planteados, sin embargo la efectividad de la clasificación se ve afectada por la calidad de los datos de entrenamiento provocando que se logre buenos resultados para los niveles de la taxonomía de los que se tienen una cantidad considerable ejemplos y malos resultados para los niveles que poseen pocos ejemplos etiquetados.

Palabras Claves: *Taxonomía de Bloom, Minería de texto, Algoritmos de aprendizaje, Clasificador Bayesiano.*

Abstract

This paper presents the implementation of an automatic classification system to categorize answers of students to discussion forums according to the cognitive domain of Bloom's Taxonomy. Several works had been completed in this area, in this occasion the effectiveness of a Bayesian classifier, specifically the Naïve Bayes classifier solving this problem is analyzed alongside with the usage of text mining techniques and a dataset previously classified by experts; also the possibility of considering the automatic classification system as a human coder is explored. Several tests were conducted with the objective of obtain metrics whose values allowed rate the performance of the classification and the quality of the results. The obtained results point that using the proposed architecture is possible to achieve the objectives; however the effectiveness of the classification is affected by the quality of the training data which cause that good results are obtained for the levels that have a considerable quantity of examples and bad results for the levels that have few examples labeled.

Keywords: *Bloom's taxonomy, Text mining, learning algorithms, Bayesian classifier*.

.

1. Introducción

La taxonomía de Bloom es una taxonomía o clasificación de habilidades cognitivas que ha sido ampliamente empleada para evaluación y medición de objetivos educacionales [8]. Es un concepto esencial que guía a los educadores en la elaboración de objetivos educacionales, preparación de currículums y creación de evaluaciones [10]. Fue propuesta en 1956 por Benjamin Bloom, esta taxonomía divide los objetivos educacionales en tres dominios: cognitivo, psicomotor y afectivo [2].

El dominio cognitivo está relacionado con el conocimiento, habilidades y destrezas intelectuales de cada individuo, las cuales están clasificadas desde un que corresponde a simplemente recordar o reconocer hasta niveles más altos que conllevan la creación de nuevo conocimiento. El dominio psicomotor se enfoca en las habilidades físicas, las cuales incluyen habilidades motoras refinadas, manipulación de

herramientas o cualquier tipo de actividad que requiera de coordinación neuromuscular. El dominio afectivo se enfoca en las emociones, sentimientos, actitudes el grado de aceptación o rechazo. Este dominio incluye la manera en la cual un individuo actúa con situaciones emocionales como sentimientos, valores, apreciaciones, actitudes, y motivaciones [5].

Anderson y Krahwohl modificaron el dominio cognitivo de la taxonomía de Bloom en el año 2001, siendo la modificación más significativa el cambio de la construcción original de naturaleza unidireccional a una construcción bidireccional que incluye al conocimiento con procesos cognitivos; estas modificaciones fueron conocidas como la taxonomía de Bloom revisada [2]. Hay seis categorías o niveles de la versión revisada, los cuales se encuentran listados desde el más simple al más complejo [3] [4]: recordar, entender, aplicar, analizar, evaluar y crear.

Los beneficios de incluir la utilización de la taxonomía de Bloom en las prácticas educacionales y de enseñanza son innegables, sin embargo el proceso de categorizar preguntas, opiniones y argumentos de estudiantes de acuerdo a esta taxonomía es tedioso y consume una gran cantidad de tiempo, dado que generalmente se lo realiza manualmente [10].

Lo expuesto anteriormente ha desembocado en que se busque automatizar este proceso de clasificación, por ello diversas soluciones han sido propuestas. Estas implementaciones hacen uso de distintas técnicas como la minería de texto [10], procesamiento de lenguaje natural [9], sistemas inteligentes con redes neuronales [11], etc. Lamentablemente la gran mayoría no son suficientemente precisas, no se encuentran disponibles para un uso masivo de usuarios y a fin de cuentas no constituyen una solución real a la problemática planteada [5] [11].

La tarea de clasificar automáticamente de preguntas, opiniones, contribuciones en foros de discusión, etc. puede ser considerada como un problema de clasificación de texto y es la actividad de etiquetar texto en lenguaje natural con categorías temáticas de un conjunto predefinido. La minería de texto se utiliza para denotar las tareas de análisis de grandes cantidades de texto, la detección de patrones útiles y extracción de información que probablemente no es evidente en primera instancia. Según estos criterios, la clasificación de texto es una instancia de la minería de texto [9].

En este trabajo se propone la implementación de un sistema de clasificación automática de las contribuciones en foros de discusión, empleando técnicas de minería de texto y el uso de un clasificador Bayesiano, también se busca medir que tan exacto es este clasificador en la tarea específica de la asignación de una categoría de la taxonomía de Bloom a un texto. El otro propósito de este estudio es determinar si los resultados proporcionados por el sistema de clasificación, son lo suficientemente buenos como para reemplazar a codificadores humanos. La estructura del artículo es la siguiente: en la sección de trabajos relacionados se presentan sistemas similares y algunas investigaciones realizadas en esta área, la sección arquitectura del sistema presenta los componentes utilizados para construir el sistema propuesto de clasificación, los métodos empleados para evaluar la eficacia del sistema se describen en la sección evaluación del sistema, los resultados obtenidos se presentan en la sección de resultados, el documento finaliza con algunas conclusiones.

2. Trabajos relacionados

El disponer de un sistema que clasifique automáticamente texto, opiniones, preguntas, de acuerdo a la taxonomía de Bloom es un tema del cual se han propuesto varias soluciones y enfoques. Uno de ellas es la sugerida por [11], quienes propusieron construir un sistema que clasificara las preguntas empleadas en exámenes de acuerdo al dominio cognitivo de la taxonomía de Bloom. Su solución empleaba una red neuronal como clasificador, la cual fue entrenada empleando un algoritmo de aprendizaje de escala conjugada. Uno de los mayores problemas que experimentaron fue la pobre escalabilidad de la red neuronal, por lo que fue necesaria la aplicación de varios métodos de reducción de características. En este estudió también determinaron que dadas las particularidades de su conjunto de entrenamiento, el método más efectivo de reducción de características fue el DF dado que mantenía la precisión mientras mejoraba la velocidad de convergencia. En cuanto a la efectividad obtenida fue de 65.9%, empleando para el entrenamiento un total de 192 preguntas y obteniendo un total de 605 características.

Una investigación similar llevada a cabo por [10], tenía la misma finalidad que la investigación anteriormente expuesta, pero en este caso en lugar de emplear una red neuronal emplearon máquina de soporte de vectores. Dentro de los resultados que obtuvieron evidenciaron que usando SVM obtuvieron resultados satisfactorios con respecto a la efectividad y precisión de la clasificación, sin embargo el no disponer de un conjunto de entrenamiento grande provocaron que el valor resultante de recuperación sea bajo y con ello no les fue posible obtener resultados más concluyentes. El conjunto de entrenamiento que emplearos estaba compuesto por 190 preguntas las cuales se encontraban uniformemente distribuidas entre los diferentes niveles de la taxonomía, a pesar de no ser un número elevado de ejemplos obtuvieron indicadores de efectividad de alrededor 87% lo cual es superior al obtenido a 65.9% que fue el resultado obtenido empleando una red neuronal como clasificador.

De manera similar [2] proponen un sistema de inferencia para analizar de manera automática la calidad de ítems usados en pruebas basados en la taxonomía de Bloom. En su propuesta, la tarea de la

clasificación era llevada a cabo por un sistema experto basado en reglas, en donde la determinación de las reglas de la clasificación constituyó una de las tareas más complicadas y cruciales. Los mejores resultados que obtuvieron en cuanto a efectividad fueron del 51% y concluyen que si las palabras o verbos empleados en cada nivel no se encuentran en la base de conocimiento, el motor de inferencias no podrá efectuar la tarea con éxito.

Otra tentativa de clasificar preguntas de acuerdo a la taxonomía de Bloom fue llevada a cabo por [1], quienes diseñaron un sistema online en el cual los profesores ingresaban las preguntas que deseaban hacer en un test y el sistema respondía en qué nivel de la taxonomía se encontraba dicha pregunta. Era posible también para los estudiantes rendir su examen a través del mismo sistema. La arquitectura del sistema no contempló la inclusión de un clasificador automático, el determinar el nivel en el que se encontraba la pregunta dependía de un conjunto de palabras y verbos asociados a cada nivel los cuales se encontraban almacenados en una base de datos y para llegar a la conclusión era necesario comparar cada palabra de la pregunta de la cual se desea saber el nivel con las palabras que se encontraban en la base de datos y de esa forma el nivel del cual se tenía más palabras era la categoría que se le asignaba a la pregunta. El mejor resultado que obtuvieron fue del 75% de efectividad para el nivel de conocimiento, mientras que para los otros niveles fue muy bajo, pues el nivel de efectividad fue de alrededor 20%.

Algo que tienen en común las propuestas mencionadas anteriormente, es que todas ellas emplean casi los mismos pasos para el pre procesamiento de texto y la reducción de características, mientras que la gran diferencia radica en el método de clasificación que emplean. Comparando estos pocos estudios es rescatable que el uso de algoritmos de aprendizaje para efectuar la clasificación da mejores resultados que simplemente emplear comparaciones contra una base de datos.

Además, solamente tomando en consideración los casos aquí expuestos, es notorio que el emplear máquinas de soportes de vectores provee una mayor efectividad que la obtenida empleando otros tipos de clasificadores como redes neuronales o sistemas expertos. Sin embargo, cabe recalcar que en ninguna de las soluciones propuestas se alcanza una efectividad del 100%, siendo las principales causas la no existencia de un clasificador totalmente efectivo y el no disponer de conjuntos de entrenamiento lo suficientemente grandes y buenos. Esto último resulta ser un factor altamente influyente en el porcentaje de efectividad alcanzado por los sistemas, si no se tienen un número de ejemplos uniformemente distribuido entre los niveles de la taxonomía, los resultados serán buenos en unos casos, regulares en otros y pueden ser también totalmente malos [10].

En este trabajo, el uso de un clasificador bayesiano específicamente *Naïve Bayes* para obtener el modelo de clasificación utilizado para realizar la predicción de la categoría se propone; algunas pruebas se realizan también con objetivos para determinar si el sistema de clasificación automática puede sustituir codificadores humanos. Un aspecto que vale la pena recalcar, es que todos estos sistemas se encuentran implementados para el idioma inglés y los algoritmos usados en pasos previos a la clasificación, como los algoritmos de *stemming*, se encuentran optimizados para este lenguaje por lo que implementar un sistema de este tipo para texto que se encuentra en idioma español constituye un aporte significativo.

3. Arquitectura del Sistema

El motor de inferencia se encuentra formado por dos componentes:

- Componente de aprendizaje
- Componente de clasificación

Para llevar a cabo la predicción de la etiqueta, el motor de inferencia necesita un modelo de clasificación, este modelo es proporcionado por el componente de aprendizaje mediante procesos de minería de texto, un algoritmo de aprendizaje y un conjunto de entrenamiento. La entrada del componente de clasificación es texto cuya etiqueta es desconocida, este texto es pre procesado y las reglas del modelo obtenido anteriormente se aplican, finalmente, la categoría más probable del texto es devuelta como respuesta. Tanto el componente de aprendizaje como el componente de clasificación se implementaron utilizando RapidMiner, que es un software para minería de datos y análisis desarrollado en el lenguaje de programación Java, permite la creación de procesos simples y complejos encadenando una serie de operadores de extracción, procesamiento y visualización de datos a través de un entorno gráfico muy intuitivo [7].

3.1 Componente de Aprendizaje

El componente de aprendizaje está formado por varios subcomponentes que se muestran en la figura 1; la entrada de este componente es un conjunto de entrenamiento, el cual será utilizado para construir el modelo.

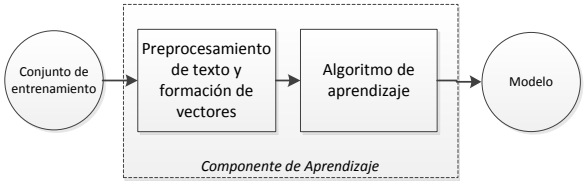

Figura 1. Esquema de la arquitectura del componente de aprendizaje

El conjunto de entrenamiento está constituido por respuestas generadas por estudiantes en foros de discusión y que han sido categorizadas de acuerdo a la taxonomía de Bloom previamente por expertos. Una representación vectorial será la forma que tome el texto, los vectores estarán formados por características extraídas de cada categoría y cada una de esas características tendrá un valor asociado que indicará que tan influyente es una característica para determinada categoría.

Los elementos del conjunto de entrenamiento deben pasar por varias etapas hasta obtener su representación vectorial, es necesario remover cualquier información que no sea relevante ni influyente, es por ello que un pre procesamiento es indispensable. Los procesos ejecutados por el componente de pre procesamiento de texto son *tokenización*, eliminación de palabras vacías y *stemming*; a través de la *tokenización* cada palabra que constituye un texto es aislada, luego palabras vacías como interjecciones, conjunciones y verbos auxiliares que no afectan el sentido del mensaje a ser transmitido son eliminadas, finalmente la raíz de las palabras remanentes son extraídas usando un algoritmo de *stemming*, en nuestro caso el algoritmo *Snowball* es empleado, la forma final de las palabras es llamada stem. En este punto hay muchos stems por categoría, estos stems son los componentes de los vectores de características; dado que el número de características de los vectores puede llegar a ser muy grande y con ello afectar el rendimiento y desempeño del clasificador, un proceso de reducción de características fue aplicado. EL método usado fue el de frecuencia de documento – DF, dado que para este tipo de problemas resulta ser el más efectivo [11]; una vez que las características más representativas han sido seleccionadas, los vectores se encuentran listos para ser la entrada del algoritmo de aprendizaje.

El algoritmo o clasificador es el subcomponente que produce el modelo necesario para ejecutar la predicción de la categoría; a pesar de la gran variedad de clasificadores, *Naïve Bayes* fue escogido debido a su efectividad y simplicidad [6]. Este es un clasificador probabilístico el cual aplica el *Teorema de Bayes* para determinar que un documento representado por un vector $\vec{d}_j = \langle w_{1j}, \ldots, w_{|T|j} \rangle$ de términos pertenece a una categoría c_i.

3.1 Componente de Clasificación

El componente de clasificación es el encargado de determinar cuál es el nivel dentro de la taxonomía al que pertenece algún texto que se desea clasificar. De igual manera que a los elementos del conjunto de entrenamiento, es necesario aplicar pre procesamiento al texto que se está recibiendo puesto que con seguridad contiene palabras poco relevantes para el sentido del mensaje que se desea transmitir, palabras mal escritas y signos de puntuación, de esta forma se extraen las principales características del texto a clasificar y con ello se encuentra preparado para aplicar el modelo de clasificación.

El determinar si un texto pertenece o no a un nivel se lo realiza aplicando el modelo probabilístico aprendido, esto implica que dadas las características del texto se calcula la probabilidad de que pertenezca a una categoría o no. El resultado devuelto es el nivel de la taxonomía al que más probablemente pertenezca el texto, en la figura 2 se muestra el esquema de la estructura del componente de clasificación.

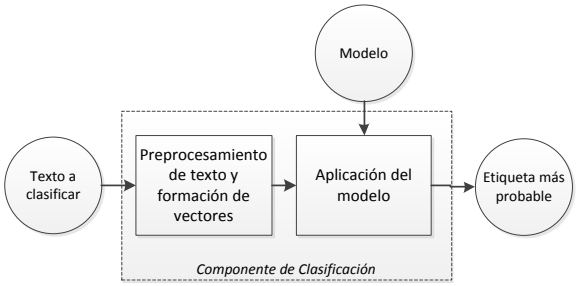

Figura 2. Esquema de la arquitectura del componente de clasificación

4. Evaluación del Sistema

4.1. Definición del Conjunto de Entrenamiento

El conjunto de entrenamiento empleado para obtener el modelo de clasificación está conformado por respuestas emitidas por estudiantes a diferentes temas de discusión planteados de acuerdo a una asignatura en particular, estas respuestas fueron recogidas desde el año 2010 hasta el año 2012 y han sido etiquetados por diferentes expertos.

Las respuestas a los foros de discusión recolectados fueron etiquetadas cada una por tres diferentes codificadores, quienes las categorizan de acuerdo a su criterio personal, comparan la etiqueta asignada por cada uno de ellos y después de una discusión y exposición de razones del porqué de la etiqueta asignada, llegan a un consenso y proceden asignar el nivel de la taxonomía de Bloom que entre ellos consideran el más apropiado. Cabe recalcar que las respuestas fueron llevadas del idioma español al inglés para que los codificadores procedieran con sus análisis, debido a que no dominaban la lengua española, consideramos que este factor no afecta la validez del estudio debido a que la traducciones fueron hechas de tal forma que no se afectaba el mensaje a transmitir.

Otro aspecto que es necesario recalcar, es que los codificadores fueron diferentes cada año, aunque el número de ellos y el procedimiento para proceder a asignar el nivel de la taxonomía de Bloom a las respuestas fue el mismo.

El subconjunto de entrenamiento uno fue recogido en el año 2010. Se planteó un caso de estudio relacionado a la seguridad informática a un grupo de estudiantes de maestría en seguridad informática de la

ESPOL; los estudiantes fueron separados en cinco grupos de cuatro estudiantes y cada grupo debía proveer una solución a la problemática planteada en el caso de estudio y debían alcanzar conclusiones en un lapso de 48 horas. El total de ejemplos recolectados este año se muestran en la tabla 1.

Tabla 1. Número de respuestas etiquetadas por nivel del subconjunto de entrenamiento del año 2010

Nivel de la taxonomía de Bloom	Número de respuestas
Crear	0
Evaluar	8
Analizar	9
Aplicar	0
Entender	13
Recordar	11
Total	41

El subconjunto de entrenamiento dos fue obtenido en el año 2011 y correspondieron a respuestas de estudiantes de la materia Ingeniería de Software I a un foro de discusión del cual debían entregar como resultado un documento de riesgos correspondientes a los proyectos que se encontraban desarrollando. La actividad fue realizada por treinta y cinco estudiantes y se los dividió en siete grupos de cinco miembros cada uno.

Un particular en la actividad realizada en el 2011, es que los estudiantes fueron capacitados en cuanto como clasificar sus respuestas de acuerdo de la taxonomía de Bloom y debían etiquetar cada una de sus contribuciones de acuerdo a lo que aprendieron. Posteriormente las etiquetas de los estudiantes fueron comparadas con las asignadas por los codificadores expertos con la finalidad de llegar a un consenso y determinar la categoría final. En la tabla 2 se muestra en detalle el número de ejemplos por nivel.

Tabla 2. Número de respuestas etiquetadas por nivel del subconjunto de entrenamiento del año 2011

Nivel de la taxonomía de Bloom	Número de respuestas
Crear	0
Evaluar	2
Analizar	22
Aplicar	7
Entender	46
Recordar	30
Total	107

En el año 2012 las discusiones se llevaron a cabo en la asignatura Aplicaciones Multimedia Interactivas, se efectuaron dos discusiones siendo una de ellas acerca de decidir cuál es el mejor framework web dadas ciertas condiciones y la segunda decidir cuál es el mejor framework para desarrollo de aplicaciones móviles.

Tabla 3. Número de respuestas etiquetadas por nivel del subconjunto de entrenamiento del año 2012

Nivel de la taxonomía de Bloom	Número de respuestas
Crear	0
Evaluar	7
Analizar	32
Aplicar	0
Entender	155
Recordar	78
Total	271

En esta ocasión los estudiantes debían adoptar distintos roles dentro de la discusión. Los estudiantes fueron divididos en cuatro grupos de seis a ocho integrantes y cada equipo de trabajo debía obtener un documento final es donde exponían sus propuestas y explicaban la razón por la cual habían escogido determinada plataforma. El detalle de la cantidad de ejemplos por nivel recolectados este año se muestra en la tabla 3.

El conjunto de entrenamiento final está formado de la unión de los subconjuntos de entrenamiento de los años 2010, 2011 y 2012, quedando establecido finalmente como se muestra en la tabla 4. Un aspecto que es necesario recalcar es que a pesar de que se recogieron datos de tres años distintos, el número de respuestas no se encuentra uniformemente distribuido entre todos los niveles y esto en muchos casos afecta el rendimiento de los clasificadores automáticos.

Tabla 4. Número de respuestas por nivel del conjunto de entrenamiento final

Nivel de la taxonomía de Bloom	Número de respuestas
Crear	0
Evaluar	17
Analizar	63
Aplicar	7
Entender	214
Recordar	119
Total	420

4.2. Definición de Métricas y Conjunto de Pruebas

Una evaluación al sistema fue realizada con el objetivo de determinar la efectividad con la cual el sistema realiza la clasificación. La forma en la que se evalúa la efectividad del sistema es a través de varias medidas que se presentan en una tabla de contingencia, la cual está formada por lo general de los siguientes valores:

- **A:** El número de respuestas que el sistema asigna correctamente a la categoría, es decir, los verdaderos positivos.
- **B:** El número de respuestas que el sistema incorrectamente asigna a la categoría, estos constituyen los falsos positivos.
- **C:** El número de respuestas que pertenecen a la categoría pero que el sistema las asigna a otra, es decir los falsos negativos.
- **D:** El número de respuestas que el sistema correctamente no asigna a la categoría, lo cual constituye los verdaderos negativos.

Las medidas comúnmente empleadas para determinar la efectividad de un clasificador son las siguientes:

- **Exactitud:** Es la probabilidad de que un documento o respuesta d_x es clasificado bajo una categoría c_x esta decisión es la correcta.

$$P = \frac{A}{A+B}$$

- **Recuperación:** Usualmente conocida[1] como *recall*, es la probabilidad que si un texto aleatorio debería ser clasificado bajo una categoría c, esta decisión es tomada.

$$R = \frac{A}{A+C} \quad (2)$$

- **Precisión:** Esta es la medida más comúnmente empleada para medir la efectividad de un clasificador, sin embargo, los valores de precisión son mucho menos renuentes a variaciones en el número de los verdaderos positivos que las medidas de exactitud y reconocimiento.

$$Acc = \frac{A+C}{A+B+C+D} \quad (3)$$

- $\boldsymbol{F_\beta}$: Es la media harmónica del reconocimiento y la exactitud. El valor de β depende de la importancia que se le da a la exactitud y a la recuperación; si la exactitud es considerada más importante que el reconocimiento entonces el valor de β debe ser cero. Si la recuperación es más importante que la exactitud el valor de β debe ser llevado a infinito. Si la exactitud y el reconocimiento son igual de importantes el valor de β debe ser igual a uno. La ecuación de F_β para un valor de β=1 está definida de la siguiente manera:

$$F_\beta = \frac{2RP}{R+P} \quad (4)$$

Para determinar los valores de A, B, C y D es necesario contar con un conjunto de prueba. El conjunto de pruebas o *testing set*, está formado por respuestas de las cuales se conoce el nivel de la taxonomía Bloom al que pertenecen y no formaron parte de conjunto de entrenamiento; la prueba consiste en que el clasificador categorice las respuestas del conjunto de pruebas y la respuesta devuelta por el clasificador es comparada con la etiqueta que realmente posee la respuesta, de esta forma se determinan los verdaderos positivos y negativos y los falsos positivos y negativos. En la tabla 5 se presentan el número de respuestas por nivel del conjunto de pruebas final.

Tabla 5. Número de respuestas por nivel del conjunto de pruebas

Nivel de la taxonomía de Bloom	Número de respuestas
Crear	0
Evaluar	7
Analizar	27
Aplicar	3
Entender	92
Recordar	51
Total	180

Es necesario determinar también si el clasificador automático puede actuar como un codificador humano más, recordemos que el proceso de clasificación por lo general se lo realiza entre varias personas y la decisión que se toma es por consenso. Por ello se aplicará una prueba de *inter-rater reliability* con el estadístico de *alfa de Krippendorff*, el cual indica el grado de concordancia entre los codificadores. El valor de alfa debe encontrarse entre 0.75 y 1 para concluir que los codificadores concuerdan en la mayoría de las etiquetas asignadas, mientras que si se obtiene un valor menor a 0.75 significa que las decisiones emitidas por los codificadores no concuerdan y que la etiqueta resultado del consenso no es confiable.

5. Resultados

El conteo de los valores de verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos, son presentados en la tabla 6 Los resultados son presentados para cada nivel de la taxonomía a excepción del nivel de crear, del cual no se obtuvieron elementos que cayeran en esta categoría.

Es posible definir a los valores de A y D como resultados buenos mientras que B y C son errores. Los valores de A y D indican que la clasificación realizada es la adecuada mientras que B y C son indicadores que las predicciones no son las adecuadas; A representa los verdaderos positivos y C los falsos positivos, esto implica que la suma de estos valores debe ser igual al total de los elementos de un nivel en específico del conjunto de pruebas mientras que la

suma de los valores de A, B, C y D deben ser igual al total de elementos del conjunto de pruebas.

Tabla 6. Resultados del conteo de los valores de A, B, C y D luego de aplicar el modelo de clasificación al conjunto de pruebas

Nivel de la taxonomía de Bloom	A	B	C	D
Crear	7	37	0	135
Evaluar	16	13	11	140
Analizar	3	4	0	173
Aplicar	50	7	42	81
Entender	28	15	23	114
Recordar	7	37	0	135

Los resultados de los valores de las métricas de la aplicación del modelo de clasificación a cada elemento del conjunto de pruebas son mostrados en la tabla 7. Estos valores fueron calculados usando las ecuaciones presentadas previamente y usando los valores de A, B, C y D mostrados en la tabla 6.

Tabla 7. Resultados del conteo de los valores de A, B, C y D luego de aplicar el modelo de clasificación al conjunto de pruebas

Nivel de la taxonomía de Bloom	M1	M2	M3	M4
Crear	0.16	1.00	0.04	0.27
Evaluar	0.55	0.59	0.15	0.57
Analizar	0.43	1.00	0.02	0.60
Aplicar	0.88	0.54	0.51	0.67
Entender	0.65	0.55	0.28	0.60
Recordar	0.16	1.00	0.04	0.27
Promedio	0.53	0.74	0.2	0.54

En la tabla 7, M1 representa los valores de exactitud, M2 a recuperación, M3a precisión y M4 representa a F_β. Los resultados de las cuatro métricas presentan diferencias significativas entre ellas; con respecto a exactitud el valor más alto es de 0.88 para el nivel de entender, mientras que el valor más bajo es de 0.16 para evaluar, esto significa que el 88% de las veces la clasificación para este nivel es hecha completa correctamente mientras que en el nivel de evaluar la predicción es errónea la mayoría de las veces. Estas diferencias entre los valores de las métricas muy seguramente son causadas porque el número de elementos del conjunto de entrenamiento no se encuentran uniformemente distribuidos a través de los niveles de la taxonomía, por ello los valores más altos de exactitud son para los niveles de entender, recordar y analizar, que son los niveles con mayor número de ejemplos.

Los resultados para precisión evidenciaron el mismo patrón que los valores de exactitud; el mejor valor para entender y el peor para evaluar, en general los resultados de precisión son más bajos que los de precisión y no son satisfactorios.

Los valores de recuperación están alrededor del 50%, esto significa que si un texto aleatorio externo al conjunto de pruebas es clasificado correctamente solo la mitad de las veces. Dado que los resultados de recuperación y exactitud son bajos, provocan que los valores de F también lo sean, lo cual en general indica que la efectividad del clasificador es del 67% para el nivel de entender y de 27% para evaluar, una vez más los resultados presentan esta forma debido a la estructura y distribución del conjunto de entrenamiento.

Comparando el mejor resultado obtenido que fue el de exactitud para el nivel de entender con los valores obtenidos por sistemas similares, debe recalcarse que 88% es el valor más alto entre todas las soluciones revisadas, considerando que el resultado obtenido por el sistema de [1] fue de 26% para el nivel equivalente de comprensión y 50% para el clasificador SVM de [10]. Si consideramos la exactitud alcanzada para el nivel de analizar, este es menor que el valor obtenido por el clasificador SVM el cual fue de 75% y más alto que 32% que fue el valor alcanzado por la propuesta de [1]. En general la efectividad lograda por el clasificador Naïve Bayes es inferior a los resultados obtenidos por el clasificador SVM, pero muestra un mejor desempeño que una red neuronal usada como clasificador [11] y que un sistema experto [2] en los niveles que fueron entrenados con un alto número de ejemplos. Aunque algunos resultados puede calificárselos de bueno, el sistema necesita mejoras y emplear un mejor conjunto de datos para el entrenamiento.

Con respecto a la prueba de *inter-rater reliability*, esta fue ejecutada empleando solo los elementos de las respuestas recolectadas en el 2012, puesto que para ejecutar esta prueba es necesario conocer la etiqueta asignada por cada codificador; el objetivo de esta prueba fue determinar si el sistema de inferencia puede actuar como un codificador humano adicional. Es necesario indicar que hasta los codificadores más experimentados tienden a cometer errores, es así como varios test fueron ejecutados y los resultados obtenidos por cada uno se muestran en la tabla 8.

El valor del estadístico alfa de Krippendorff para las etiquetas dadas por los tres codificadores humanos fue de 0.718, lo cual significa que existió un alto grado de acuerdo entre sus respuestas; añadiendo como un cuarto codificador al sistema automático de clasificación, el resultado fue de 0.53. Este valor sugiere que las conclusiones alcanzadas por los codificadores difieren bastante y que el codificador que está introduciendo ruido es el sistema automático, los tres últimos valores presentados en la tabla 8 confirman esto puesto que excluyendo a un codificador humano diferente en cada prueba los valores permanecen bajos.

Tabla 8. Resultados de las pruebas de inter-rater reliability

Test	Valor de Alfa de Krippendorff
Reliablity de los tres codificadores humanos	0.7189
Reliability de codificadores humanos y sistema automático	0.53
Reliabitlity sin humano uno	0.579
Reliabitlity sin humano dos	0.418
Reliabitlity sin humano tres	0.418

6. Conclusiones

En este trabajo se propone la implementación de un sistema de clasificación automática de contribuciones en foros de discusión de acuerdo a la taxonomía de Bloom, empleando técnicas de minería de texto, el uso de un clasificador Bayesiano y un conjunto de pruebas clasificado previamente. Los resultados obtenidos indicaron que empleando la arquitectura propuesta esto es posible, sin embargo los resultados dependen mucho de la calidad del conjunto de entrenamiento usado para generar el modelo de clasificación. Los valores resultantes para las distintas métricas sugieren que el clasificador Naïve Bayes funciona relativamente bien cuando hay un número considerable de ejemplos para entrenarlo en general, al menor número de ejemplos disponibles, menor será la precisión de la predicción. Este es un problema presentado en la mayoría de los trabajos en este campo y con ello algunas preguntas importantes siguen sin respuesta con respecto a la mejora de la eficacia y los resultados de la clasificación: ¿Cuáles son las características necesarias para ser un buen conjunto de entrenamiento?, ¿cuál es el clasificador óptimo para este tipo de problemas?, ¿cuál es el método más adecuado para la reducción de características?

A pesar de los bajos resultados obtenidos, cabe destacar que el sistema proporciona una precisión bastante buena en lo que respecta a los niveles de recordar, entender y analizar, por lo que los resultados devueltos por el sistema podría ayudar a un profesor a tener una visión rápida y aproximada del nivel alcanzado por sus alumnos y si se están logrando o no un objetivo de aprendizaje. Otro punto importante es que esta aplicación es para texto en español, casi todas las implementaciones de sistemas similares son para documentos en inglés y teniendo en cuenta que el español es uno de los idiomas con más hablantes nativos en el mundo, hace que esta propuesta constituya una contribución significativa.

Los trabajos futuros se centrarán en tratar de obtener mejores resultados al emplear y experimentar con conjuntos de datos más grandes y clasificadores diferentes, también la reducción y transformación de características es un paso en el proceso que debe ser cuidadosamente analizado y seleccionado. Finalmente, si se obtienen mejores resultados los valores del estadístico alfa de Krippendorff serán mayores y esto permitirá llegar a la conclusión de que el sistema puede sustituir eficazmente codificadores humanos.

10. Referencias

[1] Chang, W., and Chung, M., *Automatic Applying Bloom's Taxonomy to Classify the Cognition Level of English Question.* IEEE, 2009.

[2] Chang, Y., and Chen, H., An Automatic Inference System for the Quality Analysis of Test Items Based on The Bloom's Revised Taxonomy. *Proceedings of the Eight International Conference on Machine Learning and Cybernetics*, 2009, pp. 2852-2856.

[3] Churches, A., Bloom's Taxonomy Blooms Digitally. Consultado en Agosto 25, 2012, de techLearning: http://www.techlearning.com/article/blooms-taxonomy-blooms-digitally/44988.

[4] Forehand, M., *Bloom's Taxonomy from Emerging Perspectives on Learning, Teaching and Technology.* The University of Georgia, 2010.

[5] Hui, C., *Feature Reduction for neural Network in Determining the Bloom's Cognitive Level of Question Items.* Universiti Teknologi Malaysia, 2009.

[6] Ikonomakis, M., Kotsiantis, S., and Tampakas, V., Text Classification using Machine Learning Techniques. *WSEAS Transactions on Computers*, 2005, p.p. 966-974.

[7] Jungermann, F., Information Extraction with RapidMiner. *Symposium Sprachtechnologie and eHumanities*, 2009, p.p. 50-61.

[8] Khairuddin, N., and Hashim, K., Application of Bloom's Taxonomy in Software Engineering Assessments. *Proceedings of the 8th International Conference on APPLIED COMPUTER SCIENCE*, 2008, p.p. 66-69.

[9] Sebastiani, F., Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 2009, p.p. 1-47.

[10] Yahya, A., & Osman, A., *Automatic Classification of Questions into Bloom's Cognitive Levels using Support Vector Machines.* Najran: Najran University, 2011.

[11] Yusof, N., & Hui, C. J., *Determination of Bloom's Cognitive Level of Question Items using Artificial Neural Network.* IEEE, 2010.