Analítica del aprendizaje: método automático para identificar sentencias que contienen información positiva y negativa utilizando técnicas de minería de texto

Silvana Vanesa Aciar
Instituto de Informática
Universidad Nacional de San Juan
San Juan, Argentina
saciar@unsj-cuim.edu.ar

Carina Soledad Gonzalez
Dep. de Ingeniería, Informática y Sistemas
Universidad de la Laguna
Tenerife, España
cjgonza@ull.edu.es

Gabriela Iris Aciar
FCEFN -Universidad Nacional de San Juan
San Juan, Argentina
gaby_aciar@yahoo.com.ar

***Abstract*— Debido al avance y fácil acceso a la tecnología hoy en día, el uso de los sistemas de enseñanza y aprendizaje online se ha incrementado exponencialmente en los últimos años. Estos sistemas comúnmente llamados Entornos Virtuales de Aprendizaje, proporcionan herramientas para presentar contenido y recursos educativos, facilitar la comunicación e interacción entre los usuarios, herramientas de seguimiento y evaluación de la actividad de los estudiantes y en algunos casos herramientas de autor para crear contenido. Las interacciones de los usuarios con el sistema generan mucha información de gran valor que ayudan a los profesores a tomar decisiones. Los comentarios de los estudiantes en los foros o chats de las plataformas virtuales de aprendizajes son fuentes de información muy valiosas para aplicar analítica del aprendizaje. La información más relevante para obtener ciertas estadísticas de los estudiantes y su contexto está en los comentarios que ellos expresan de forma libre utilizando las herramientas de interacción. En este artículo se presenta una técnica para identificar sentencias que contienen información positiva y negativa relevante e informar al profesor acerca de los aspectos negativos que puedan dar origen a posibles abandonos o problemas en el aprendizaje.**

Keywords—Analítica del aprendizaje; Minería de Texto; Inteligencia Artificial

I. INTRODUCCIÓN

Debido al avance y fácil acceso a la tecnología hoy en día, el uso de los sistemas de enseñanza y aprendizaje online se ha incrementado exponencialmente en los últimos años. Estos sistemas comúnmente llamados Entornos Virtuales de Aprendizaje, proporcionan herramientas para presentar contenido y recursos educativos, facilitar la comunicación e interacción entre los usuarios, herramientas de seguimiento y evaluación de la actividad de los estudiantes y en algunos casos herramientas de autor para crear contenido [1] [2] . Las interacciones de los usuarios con el sistema generan mucha información de gran valor que ayudan a los profesores en tomar decisiones, por ejemplo, evaluar la efectividad del proceso de enseñanza-aprendizaje, mejorar la estructura y contenido del curso, monitorizar el comportamiento de los estudiantes, evaluar el estado de avance de las actividades, diseñar actividades personalizadas, brindar recursos personalizados, obtener información sobre su estado anímico, predecir las situaciones que propician el abandono de los estudiantes, obtener los errores más frecuentes en la ejecución de las actividades, etc.

El análisis de los datos de los estudiantes y el contexto a fin de obtener las mediciones antes mencionadas se denomina Analítica del Aprendizaje (Figura 1).

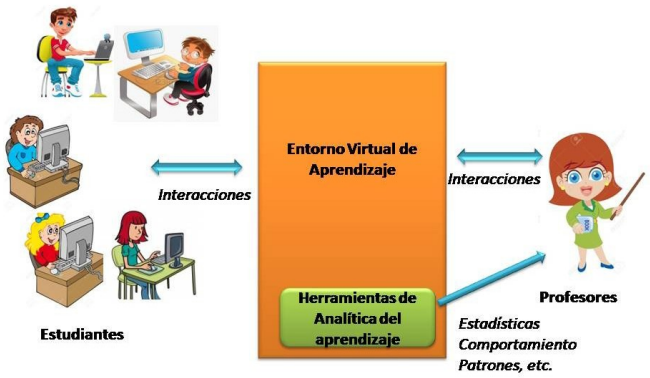

Fig. 1. Analítica del aprendizaje en entornos virtuales de aprendizaje

Las interacciones de los usuarios en los ambientes virtuales de aprendizaje generan mucha información y realizar el análisis en forma manual es imposible, por lo que se necesitan herramientas que sean capaces de procesar, analizar y presentar las estadísticas de forma entendible para los usuarios. Los comentarios de los estudiantes en los foros o chats de las plataformas virtuales de aprendizajes son fuentes de información muy valiosas para aplicar analítica del aprendizaje. La información más relevante para obtener ciertas estadísticas de los estudiantes y su contexto está en los

comentarios que ellos expresan de forma libre utilizando las herramientas de interacción. Los foros y los chat son la forma más popular de intercambiar opiniones, experiencias, problema y sugerencias sobre un curso, una temática, profesores e institución [3].

En este artículo se presenta una técnica para identificar sentencias que contienen información positiva y negativa relevante e informar al profesor acerca de los aspectos negativos que puedan dar origen a posibles abandonos o problemas en el aprendizaje. Como trabajo futuro se estudiará la incorporación y combinación de dicha información en un sistema de predicción de futuros comportamientos de los estudiantes. El resto del trabajo se organiza de la siguiente forma: en la sección II se presenta una breve descripción de analítica del aprendizaje. El análisis de comentarios se presenta en la sección III. La sección IV describe el detalle del proceso de clasificación. Un caso de estudio ejemplificando su implementación se describe en la sección V. Finalmente, la sección VI concluye el documento y proporciona direcciones para la investigación futura a realizar.

II. Analítica del aprendizaje y mejora en el proceso de enseñanza-aprendizaje

La analítica del aprendizaje proporciona información a los profesores sobre el comportamiento de los estudiantes, qué están haciendo, cómo lo están haciendo y cuáles son los obstáculos o problemas que dificultan la consecución de los objetivos del aprendizaje. Esta información le permite al profesor intervenir con medidas correctivas apropiadas para prevenir posibles abandonos, a su vez permite a los profesores preparar actividades personalizas a cada problemática y estudiante. El fracaso o éxito del aprendizaje depende en gran medida de la identificación a tiempo de obstáculos formativos y disminuir el riesgo de abandono. De nada servirá tener muchos datos de los estudiantes si no se tienen las herramientas adecuadas para analizar esos datos que aporte información relevante y permitan tomar decisiones correctas [4] [1] [5].

Hoy en día existen pocos sistemas virtuales de aprendizaje que contienen módulos o herramientas de análisis del aprendizaje automático para medir, recuperar, analizar y presentar información de los estudiantes y su contexto a efectos de optimizar el aprendizaje. Para tal función, la analítica del aprendizaje utiliza herramientas de minería de datos, análisis estadísticos, minería web, recuperación de la información, entre otras. El análisis del aprendizaje no solo brinda beneficios a los profesores, sino que también el estudiante se beneficia, puede recibir recomendaciones de recursos, actividades personalizadas y estímulo en el aprendizaje en función de sus intereses y análisis de las interacciones con el entorno virtual.

Numerosos son los datos y fuentes para obtener información y realizar las mediciones y predicciones: datos personales y demográficos; datos de las interacciones tales como comportamiento de los estudiantes en el sistema, interacciones con otros estudiantes en las redes sociales; datos de contexto tales como ubicación, estaciones climáticas, etc; datos del proceso de aprendizaje, tales como actividades desarrolladas, evaluaciones aprobadas, etc] [Romero, C. R., and S. Ventura, 2010]. El análisis automático de estos datos permitirá al profesor disponer de información sin necesidad de preguntarle al estudiante pudiendo así intervenir ante cualquier situación problemática y motivar al estudiante de forma adecuada.

III. Minería de opinión

Los foros y las redes sociales, se han convertido en los sistemas preferidos de las personas para compartir opiniones y sentimientos [6] [7]. La información obtenida de las redes sociales y foros puede ser muy relevante para mediciones y explotación de la aceptación o rechazo en la sociedad sobre temas y cuestiones concretas [8]. Las investigaciones realizadas y que se están realizando en el análisis de información del texto escrito por usuarios de las redes sociales, tienen dos propósitos: la clasificación de opiniones o comentarios escritos por usuarios en “Positivos” y ”Negativos” lo que se denomina polaridad y la extracción de entidades, los cuales pueden ser conceptos, sujetos, etc. [7] [8] [9] [10] [11] . En este artículo se utilizarán técnicas de minería de texto para detectar sentencias con polaridad negativa en los comentarios de estudiantes en foros [12].

IV. Clasificación de las opiniones en positivas o negativas

El proceso de clasificación sigue la aplicación del proceso de minería de texto que se describe en [12] para clasificar las sentencias de las opiniones en el dominio de la cámara digital en buenas, malas y de calidad. En este trabajo aplicamos las herramientas de minería de textos aplicadas en [12] para definir las reglas que nos permiten identificar las sentencias que contienen información positiva y negativa en los foros de estudiantes. Los algoritmos basados en frecuencia de palabras no proporcionan buenos resultados debido al tamaño de las frases que participan en el proceso de clasificación. Así, se emplean técnicas de clasificación basados en reglas. Como se ha descrito antes, dos categorías se han definido para clasificar las frases: "positivas" y "negativas". La categoría "positivas" agrupa aquellas oraciones que contienen información que expresan aspectos satisfactorios para los estudiantes, mientras que la categoría “negativas” agrupa frases que contienen información que expresan aspectos no satisfactorios para los estudiantes.

La herramienta Text- Miner Software Kit (TMSK) y Rule Induction Kit for Text (RIKTEXT) (RIKTEXT) se han utilizado para obtener los conjuntos de reglas de clasificación [13] . El mejor conjunto de reglas se selecciona en base a una combinación de consideraciones de complejidad y de las tasas de error. RIKTEXT encuentra el conjunto de reglas con el índice de error mínimo o razonablemente cerca la tasa de error mínimo.

Se realizó un estudio analizando 5 foros donde participaron estudiantes en el año 2014. Los foros fueron creados en varias temáticas dentro del área de la informática. El objetivo fue conseguir un conjunto de datos para obtener

las reglas de clasificación en las categorías “positivas” y ”negativas” .

Para analizar los foros, fue necesario realizar un pre-procesamiento de los mismos para transformar los datos en el formato propicio para ser procesados por las herramientas de minería de texto. Cada frase es tratada como un documento. Una vez que los datos están en formato XML, están listos para ser procesados por TMSK y generar el diccionario y un conjunto de vectores etiquetados.

Los datos se han dividido en dos grupos una para el entrenamiento y otro para las validaciones de las reglas. Los casos de prueba son seleccionados al azar. Elegimos dos tercios de los casos disponibles para el entrenamiento y el resto para validación. Los resultados se presentan en la Figura 2. Como se puede ver, se muestra una serie de conjuntos de reglas para clasificar frases en la categoría "Positivas". Cada conjunto de reglas está numerado en la columna "RSet". Un "*" delimita el conjunto de reglas con el índice de error mínimo. "**" indica el mejor conjunto de reglas de acuerdo a la tasa de error y la simplicidad. "Rules" es el número de reglas en el conjunto de reglas. La columna "Train Err" da la tasa de error de los conjuntos de reglas en los datos de entrenamiento. "Test Err" es una estimación de tasa de error en los datos de prueba. SD es la desviación estándar de la estimación. "Err / Var" da una indicación de la calidad de la solución. Las reglas elegidas son las que tienen la tasa de error mínimo o están muy cerca al mínimo, pero puede ser más simple que el mínimo (**). Precisión, recall y F-medida obtenida de casos de entrenamiento y la validación se muestran al final de la tabla. La Figura 3 muestra el conjunto de reglas para clasificar sentencias en la categoría "Negativas".

Para cada sentencia de los foros se aplican las reglas y si alguna de las reglas puede ser aplicada se clasifica en esa categoría. Un diccionario con sinónimo de palabras ha sido generad para identificar las palabras relacionadas a las implicadas en las reglas.

V. CASO DE ESTUDIO

Una vez que se han obtenido el conjunto de reglas para clasificar las oraciones en “Positivas” y “Negativas”, se han analizado 10 nuevos comentarios. Cada una de las frases obtenidas de los comentarios fue clasificada utilizando las reglas obtenidas en la sección anterior. 32 frases fueron clasificadas en la categoría positiva, 25 frases en la categoría negativa y 18 frases no pudieron ser clasificadas en ninguna de las dos categorías.

Esta clasificación se ha realizado utilizando el proceso automático descrito en los apartados anteriores. Con el fin de evaluar la precisión de la clasificación automática se realizó manualmente un proceso de clasificación. Se les pidió a los profesores que mediaron en los foros que identificaran manualmente las frases que contenían aspectos negativos y aspectos positivos en los comentarios escritos por los estudiantes en dichos foros.

Table of pruned rule sets

(* = minimum error; ** = within 0-SE of minimum error)

RSet	Rules	Vars	Train Err	Test Err	Test SD	MeanVar	Err/Var
1	45	34	0.1396	0.2171	0.0320	0.0	0.00
2**	12	10	0.1336	0.2171	0.0320	0.0	0.43
3	21	21	0.1645	0.2333	0.0445	0.0	3.00
4	23	23	0.2421	0.2659	0.0444	0.0	2.00
5	18	18	0.2264	0.2496	0.0450	0.0	2.00
6	1	1	0.2413	0.3715	0.0470	0.0	0.89

Random test cases, 88 (33.3%) test cases

****

Selected rule set

1. fantástico>=1 --> pos
2. bueno & resultado --> pos
3. mejor>=1 --> pos
4. lindo --> pos
5. menor --> pos
6. preciso --> pos
7. grandioso --> pos
8. aprobado>=1 --> pos
9. practico>=2 --> pos
10. gusta & me --> pos
11. si -->=1 pos
11. resuelto & si --> pos
12. cuesta & no>=1 --> pos

Additional Statistics (Training Cases):
precision: 76.8040 recall: 87.0317 f-measure: 78.4951

Additional Statistics (Test Cases):
precision: 716.5423 recall: 78.6541 f-measure: 75.4364

Fig. 2. Conjunto de reglas para clasificar frases en la categoría “Positiva”

Table of pruned rule sets

(* = minimum error; ** = within 0-SE of minimum error)

RSet	Rules	Vars	Train Err	Test Err	Test SD	MeanVar	Err/Var
1	32	34	0.2436	0.0141	0.0320	0.0	3.00
2	23	21	0.2345	0.0323	0.0375	0.0	3.00
3**	9	7	0.2126	0.0103	0.0301	0.0	0.65
4	12	12	0.2456	0.0632	0.0454	0.0	2.00
5	10	11	0.2278	0.0494	0.0476	0.0	2.00
6	1	1	0.2653	0.1721	0.0567	0.0	2.00

Random test cases, 65 (33.3%) test cases

****

Selected rule set

1. difícil>=1 --> neg
2. cuesta & si --> neg
3. reprobado>=1 --> neg
4. confuso --> neg
5. peor --> neg
6. complicado --> neg
7. gusta & no --> neg
8. no & resuelto>=1 --> neg
9. no>=2 --> neg

Additional Statistics (Training Cases):
precision: 74.7643 recall: 83.8736 f-measure: 78.7635

Additional Statistics (Test Cases):
precision: 73.4653 recall: 81.8743 f-measure: 77.7362

Fig. 3. Conjunto de reglas para clasificar frases en la categoría “Negativa”

Del análisis realizado por los profesores, 42 frases fueron clasificadas en la categoría positiva, 29 frases en la categoría negativa y 4 frases fueron consideradas irrelevantes. Al analizar las diferencias entre el método automático y el manual, se pudo observar que existían frases que no pudieron aplicarse las reglas obtenidas, fueron clasificadas manualmente en la categoría positiva y negativa por los profesores, esto conduce a tener que revisar las listas de palabras relacionadas en las reglas de cada categoría. También se observó que aquellas sentencias que el sistema automático no pudo clasificar en alguna categoría porque consideraba que esa sentencia pertenecía a ambas categorías, los profesores pudieron clasificarla en solo una categoría, al analizar las cusas se dedujo que ellos tenían información del contexto, que les ayudaba a tomar la decisión de establecer en qué categoría clasificarían las sentencias, esta información no es tomada en cuenta por el sistema.

Los resultados obtenidos en la clasificación manual y automática son presentados en la Tabla 1.

TABLE I. RESULTADOS DE CLASIFICACIÓN DE SENTENCIAS CON MÉTODOS AUTOMÁTICO PROPUESTO Y DE FORMA MANUAL

	Positivas	**Negativas**	**No Clasificadas**
Método Automático	32	25	18
Manual	42	29	4

VI. CONCLUSIÓN

La analítica del aprendizaje, a través del análisis de los datos, permite mejorar las prácticas pedagógicas y conocer mejor a los estudiantes. Permite obtener información que puede mejorar la toma de decisiones tanto institucionales como pedagógicas, por ejemplo puede servir para una mejor organización y asignación de recursos, o proporcionar indicadores de abandono de los estudiantes. La información más valiosa es la proporcionada por los mismos estudiantes en los foros y los chats, donde ellos pueden expresar en formato libre lo que sienten, piensan, opinan, etc. Extraer información desde texto no es una tarea fácil de automatizar, es necesario la aplicación de técnicas de minería de texto y procesamiento natural del lenguaje para lograrlo. En este artículo se presentó un paso inicial en el proceso de obtención de información relevante para predecir posibles problemas de abandono o de aprendizaje de los estudiantes. Se aplicó un método de minería de texto para obtener un conjunto de reglas que permite identificar aquellas frases que contienen información negativa escritas por los estudiantes. El siguiente paso es la incorporación de dicha información e un sistema de predicción y alertas para el profesor. La clasificación automática se comparó con una clasificación manual realizada por los mismos profesores. De los resultados obtenidos se puede observar que las reglas obtenidas pueden clasificar automáticamente sentencias de comentarios en categorías positivas y negativas.

Como trabajo futuro se propone mejorar el método para obtener resultados más precisos mediante la agregación de más sinónimos en el diccionario de palabras relacionadas, definición de criterios para cortar frases largas las cuales son clasificadas en ambas categorías y la creación de un sistema de predicción y alertas a los profesores que le ayude a realizar acciones tendientes a mejorar el proceso de enseñanza-aprendizaje.

REFERENCES

[1] Berlanga, A. J., García-Peñalvo, F. J., & Sloep, P. B. (2010). Towards eLearning 2.0 University. Interactive Learning Environments, 18(3), 199-201.

[2] Conde, M. Á., García-Peñalvo, F. J., Rodríguez-Conde, M. J., Alier, M., Casany, M. J., & Piguillem, J. (2014). An evolving Learning Management System for new educational environments using 2.0 tools. Interactive Learning Environments, 22(2), 188-204.

[3] Gewerc-Barujel, Adriana and Montero-Mesa, Lourdes and Lama-Penín, Manuel (2014) Collaboration and Social Networking in Higher Education. Colaboración y redes sociales en la enseñanza universitaria. Comunicar, 2014, vol. 21, n. 42, pp. 55-63.

[4] Ángel F. Agudo-Peregrina, Santiago Iglesias-Pradas, Miguel Ángel Conde-González, Ángel Hernández-García (2014), Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning, Computers in Human Behavior, Volume 31, February 2014.

[5] Romero, C. R., and S. Ventura. (2010). "Educational Data Mining: A Review of the State of the Art." IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 40 (6): 601–618.

[6] Sady C. Fuentes Reyes e Marina Ruiz Lobaina (2013), Minería Web: un recurso insoslayable para el profesional de la información. http://scielo.sld.cu/scielo.php?pid=S1024-94352007001000011&script=sci_arttext, Consultado Diciembre 2013.

[7] Eugenio Martínez Cámara, José Ortega, Teresa Martín Valdivia, Alfonso Ureña López (2011). Técnicas de clasificación de opiniones aplicadas a un corpus en español - Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 163-170.

[8] Álvaro José Casado Valverde y Iván Cantador Gutiérrez (2013). Sistema de extracción de entidades y análisis de opiniones en contenidos Web generados por usuarios. Septiembre del 2013.

[9] Krisztian Balog Maarten de Rijke,(2012) Finding Experts and their Details in E-mail Corpora-, University of Amsterdam. 2012.

[10] Gang Liu and Tianyong Hao (2012) .User-based Question Recommendation for Question Answering System- International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012.

[11] Tapia, M., Ruiz, O., Chitinos, (2014) C. Modelo de clasificación de opiniones subjetivas en Redes Sociales– Rev. Ingeniería: Ciencia, Tecnología e Innovación VOL 1/N° 1- 2014.

[12] Silvana Aciar, Debbie Zhang, Simeon Simoff, and John Debenham. (2007). Informed Recommender: Basing Recommendations on Consumer Product Reviews. IEEE Intelligent Systems 22, 3 (May 2007).

[13] Sholom M. Weiss, Nitin Indurkhya, Tong Zhang, and Fred Damerau (2004). Text Mining: Predictive Methods for Analyzing Unstructured Information, Springer-Verlag, New York, 2004.