



## Creación de un corpus de frases para minería de datos educativa validado por expertos

María Lucia Barrón Estrada<sup>1</sup>, Ramon Zatarain Cabada<sup>1</sup> y Oscar Omar Sánchez Alvarado<sup>2</sup>

1 Tecnológico Nacional de México, Campus Instituto Tecnológico de Culiacán, 2 Tecnológico Nacional de México, Campus Instituto Tecnológico de Culiacán . lbarron@itculiacan.edu.mx

Ante un mundo cada vez más digital e inmerso en la inteligencia artificial se han desarrollado nuevos métodos y tecnologías para apoyar a la educación, los cuales aportan grandes beneficios a los estudiantes en las aulas. Un ejemplo son los ambientes de aprendizaje inteligentes que permiten que los estudiantes reciban una educación más personalizada de acuerdo a sus necesidades cognitivas, elevando así su rendimiento. Los ambientes de aprendizaje inteligentes logran realizar acciones de adaptación utilizando diversas técnicas de inteligencia artificial con el objetivo de tomar decisiones acertadas sobre el flujo y la dosificación de temas, ejercicios y exámenes que se le administran a cada estudiante. La detección de emociones en un ambiente educativo es muy importante porque permite que un sistema se adapte a las necesidades afectivas del usuario. Existen diversas formas de reconocer emociones de forma automática, una de ellas es a través de frases escritas en algún idioma, siendo necesario contar con un corpus de frases etiquetadas. Actualmente existen dos corpus que se crearon a partir de la recopilación de opiniones de alumnos: "SentiText", que contiene frases etiquetadas con polaridades (negativa o positiva), y "EduSere", que contiene frases etiquetadas con emociones centradas en el aprendizaje (aburrido, frustrado o emocionado). Dichos corpus sirven para entrenar algoritmos de detección de emociones y polaridades en texto que se usan en ambientes de aprendizaje inteligentes para clasificar las emociones de los estudiantes que usan el sistema. Los corpus fueron etiquetados por los mismos usuarios y no cuentan con validación por expertos en el área de psicología o pedagogía, por lo cual el resultado de la clasificación no es confiable en su totalidad. Con el fin de generar un corpus validado para mejorar la precisión de la clasificación de frases, se desarrolló una aplicación en Android llamada "Label", la cual tiene una arquitectura basada en capas, utilizando el patrón de diseño MVC. La aplicación implementa técnicas de gamificación (trofeos, tablas de puntuación, etc.) para mejorar el interés de los usuarios y a través de una serie de niveles muestra frases con emociones ya sean polaridades (negativo o positivo) o centradas en el aprendizaje (aburrido, frustrado o emocionado), con el fin de que los expertos del área en emociones (psicólogos), complementando con maestros y alumnos pueden etiquetar dichas frases. A partir de esta aplicación se realizó una recopilación de todas las frases validadas para crear dos nuevos corpus especializados en la educación de los cuales se obtuvo los siguientes resultados, corpus SentiText que consta de 24,543 frases con polaridades, donde 12,272 son positivas y 12,271 negativas, mientras que el corpus EduSere cuenta con 12,083 frases con emociones centradas en el aprendizaje de las cuales 5,600 son comprometido, 3,238 aburrido y 3245 frustrado. La aplicación desarrollada generó dos corpus de frases validadas y será posible desarrollar clasificadores que obtengan un porcentaje de precisión mayor al 90% para reconocer los estados emocionales de los estudiantes y que un ambiente de aprendizaje inteligente adapte su proceso de enseñanza a las necesidades de los estudiantes.