

SEP

TecNM

INSTITUTO TECNOLÓGICO DE CULIACÁN



CREACIÓN DE UN CORPUS DE FRASES EN ESPAÑOL  
VALIDADO POR EXPERTOS PARA MINERÍA DE DATOS  
EDUCATIVA

TESIS

PRESENTADA ANTE EL DEPARTAMENTO ACADÉMICO DE ESTUDIOS DE POSGRADO  
DEL INSTITUTO TECNOLÓGICO DE CULIACÁN EN CUMPLIMIENTO PARCIAL DE LOS  
REQUISITOS PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS DE LA COMPUTACIÓN

POR:

OSCAR OMAR SÁNCHEZ ALVARADO  
LICENCIADO EN INGENIERÍA EN INFORMÁTICA

DIRECTOR DE TESIS:

DRA. MARÍA LUCÍA BARRÓN ESTRADA

CULIACÁN, SINALOA

SEPTIEMBRE 2020

# Autorización de impresión



**EDUCACIÓN**  
SECRETARÍA DE EDUCACIÓN PÚBLICA



Instituto Tecnológico de Culiacán

"2020, Año de Leona Vicario, Benemérita Madre de la Patria"

Culiacán, Sin., 14 de Septiembre del 2020.

**DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN**  
OFICIO: DEPI: 333/09/2020


ASUNTO: **Autorización Impresión**

**OSCAR OMAR SÁNCHEZ ALVARADO**  
**ESTUDIANTE DE LA MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN**  
**PRESENTE.**

Por medio de la presente y en virtud de que ha completado los requisitos para el examen de grado de **Maestro en Ciencias de la Computación**, se concede autorización para la impresión de la tesis titulada: **"Creación de un corpus de frases en español validado por expertos para minería de datos educativa"** bajo la dirección del(a) **Dra. María Lucía Barrón Estrada**

Sin otro particular reciba un cordial saludo.

**ATENTAMENTE**  
Excelencia en Educación Tecnológica®

  
**M.C. MARÍA ARACELY MARTÍNEZ AMAYA**  
**JEFE(A) DE LA DIVISIÓN DE ESTUDIOS DE**  
**POSGRADO E INVESTIGACIÓN**

C.c.p. archivo

MAMA/lucy \*

 **EDUCACIÓN**   
SECRETARÍA DE EDUCACIÓN PÚBLICA  
**INSTITUTO TECNOLÓGICO DE CULIACÁN**  
DEPARTAMENTO DE DIVISIÓN DE  
ESTUDIOS DE POSGRADO E INVESTIGACIÓN



Juan de Dios Bátiz 310 Pte. Col. Guadalupe, Culiacán, Sinaloa, C.P. 80220

Tel. 01 (667) 713-3804 y 454-0100

www.tecnm.mx | culiacan.tecnm.mx



# Hoja de firmas



**EDUCACIÓN**  
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO  
NACIONAL DE MÉXICO

Instituto Tecnológico de Culiacán

## “CREACIÓN DE UN CORPUS DE FRASES EN ESPAÑOL VALIDADO POR EXPERTOS PARA MINERÍA DE DATOS EDUCATIVA”

Tesis presentada por:

**OSCAR OMAR SÁNCHEZ ALVARADO**

Aprobada en contenido y estilo por:

Dra. María Lucía Barrón Estrada  
Director de Tesis

Dr. Ramón Zatarain Cabada  
Secretario

M.C. Gloria Ekaterine Peralta Peñuri  
Vocal -1

M.C. Rosalío Zatarain Cabada  
Vocal -2

M.C. María Aracely Martínez Amaya  
Jefe(a) de la División de Estudios de  
Posgrado e Investigación



Juan de Dios Bátiz 310 Pte. Col. Guadalupe, Culiacán, Sinaloa, C.P. 80220

Tel. 01 (667) 713-3804 y 454-0100

[www.tecnm.mx](http://www.tecnm.mx) | [culiacan.tecnm.mx](http://culiacan.tecnm.mx)



# Dedicatoria

Dedico este trabajo de tesis a mis padres Oscar Sánchez S. y Patricia Zulema Alvarado H. por todo su apoyo incondicional en mis estudios y por motivarme a lograr lo que me proponga.

A mis hermanos Cesar Iván Sánchez A. y Estefanía Yamile Sánchez A. que dan alegría a mi vida y siempre están cuando los necesito.

También una dedicación especial para mi novia Karen Paloma Munguía M. que siempre está para apoyarme y ayudarme en cualquier situación. Gracias por motivarme y alentarme a superarme cada día más, cumpliendo cualquier meta.

# Agradecimientos

Agradezco al Instituto Tecnológico de Culiacán por haberme permitido estudiar un posgrado de calidad, en las mejores instalaciones.

A CONACYT por otorgarme el apoyo económico para que pudiera realizar mis estudios.

A todos mis profesores de posgrado que siempre compartieron sus conocimientos y su experiencia profesional, gracias por todo su apoyo incondicional, en especial a la doctora María Lucia Barrón E. que siempre me ayudo en todo lo que necesite, por su paciencia y dedicación como profesora y asesora de tesis.

Al departamento del posgrado que me apoyaron con cualquier duda y estuvieron al pendiente de todo el proceso administrativo.

A todos mis compañeros que diario me motivaban a superar cualquier reto y por todo el apoyo que siempre me brindaron.

A todas las personas que apoyaron con el etiquetado de las frases mediante la aplicación Label, en especial a Ulises Uriel Cabrera R., Flor Jazmín Arcos M., Saraí Armenta M., Karen Paloma Munguía M., Adahedit Ivonne Arias B., Gloria Giselle Álvarez H., Claudia Elena Corrales G., Celina Díaz P., Damián Jasso A. y Aleyda Angélica Salazar H.

# **Declaración de autenticidad**

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Oscar Omar Sánchez Alvarado.

Culiacán, Sinaloa, México, 2020

# Resumen

En este trabajo de tesis se presenta la metodología para crear dos corpus de frases en español para la educación, los cuales fueron etiquetados y validados por expertos en el área de emociones (psicólogos, alumnos de psicología y maestros), utilizando los corpus generados en el Laboratorio de Tecnologías Inteligentes Aplicadas a la Educación (LTIAE) del posgrado en el Instituto Tecnológico de Culiacán, los cuales son: SentiText que contiene 24,543 frases etiquetadas con polaridad (positiva y negativa) y EduSERE el cual contiene 12,083 frases etiquetadas con emociones centradas en el aprendizaje (aburrido, frustrado y comprometido). El proceso de etiquetado y validación se efectuó con el objetivo de incrementar la eficiencia en el reconocimiento automático de emociones en texto que se realizan en ambientes de aprendizaje inteligentes, para que éstos puedan adaptarse mejor a las necesidades cognitivas y afectivas de los estudiantes.

Los corpus originales no presentan algún tipo de validación por expertos en el área de reconocimiento de emociones centradas en el aprendizaje, por este motivo el resultado de la clasificación de emociones en texto no es confiable en su totalidad y los ambientes de aprendizaje inteligentes no pueden predecir correctamente las emociones de los estudiantes. Con el fin de cumplir los objetivos y siguiendo la metodología para crear corpus validados por expertos para mejorar la eficiencia en el reconocimiento de emociones en texto, se desarrolló una aplicación para sistema operativo Android con una arquitectura MVC equilibrada basadas en servicios, que implementa técnicas de gamificación (niveles, trofeos y tablas de puntuación) para el etiquetado y validado de expertos para los corpus SentiText y EduSERE.

Con el uso de *Label* se logró etiquetar y validar por expertos el 20.9% del total de frases con polaridades del corpus SentiText y un 32.8% del total de frases con emociones centradas en el aprendizaje del corpus EduSERE. A partir del etiquetado y validado por expertos se crearon dos nuevos corpus que fueron probados con técnicas de aprendizaje automático supervisado utilizadas para la clasificación de emociones en texto, dando resultados de un incremento del 1% en la precisión del nuevo corpus EduSERE validado con respecto al corpus original y obteniendo el mismo resultado de precisión en el nuevo corpus SentiText validado con respecto al original.

# Palabras clave

Análisis de sentimientos,  
Aplicación móvil,  
Aprendizaje automático,  
Clasificadores de emociones en texto,  
Computación afectiva,  
Corpus de frases en español,  
Emociones centradas en el aprendizaje,  
Entorno inteligente de aprendizaje,  
Gamificación,  
Inteligencia artificial,  
Lógica difusa,  
Minería de opiniones,  
Polaridad en texto,  
Sistema tutor inteligente.



# Índice general

1.	Introducción .....	1
1.1.	Antecedentes .....	2
1.2.	Descripción del problema .....	3
1.3.	Justificación .....	4
1.4.	Objetivos .....	4
1.4.1.	Objetivo general.....	5
1.4.2.	Objetivos específicos .....	5
1.5.	Hipótesis.....	6
1.6.	Organización de la tesis .....	6
2.	Marco teórico .....	7
2.1.	Corpus.....	7
2.1.1.	Tipos de corpus .....	7
2.1.2.	Criterios de diseño de corpus.....	8
2.1.3.	Desarrollo de tecnologías a partir de corpus .....	9
2.2.	Minería de opiniones .....	10
2.2.1.	Aplicaciones de la minería de opiniones .....	11
2.2.2.	Técnicas para la clasificación automática de sentimientos .....	12
2.3.	Computación afectiva .....	14
2.3.1.	Definición de emoción .....	14
2.3.2.	Clasificación de las emociones .....	14
2.3.3.	Emociones centradas en el aprendizaje .....	15
2.3.4.	Técnicas para la detección de emociones.....	15
2.4.	Ambientes educativos .....	16
2.4.1.	Sistema tutor inteligente .....	16
2.4.1.1.	Arquitectura de un sistema tutor inteligente.....	16
2.4.2.	Entorno inteligente de aprendizaje.....	18
3.	Estado del arte.....	19
3.1.	Herramientas para el análisis de sentimientos.....	19
3.1.1.	Tweet sentiment visualization.....	19
3.1.2.	RapidMiner .....	21
3.1.3.	API de análisis de opiniones de MeaningCloud .....	21
3.1.4.	SentiStrength.....	22
3.2.	Corpus de frases en español .....	22
3.2.1.	Corpus Diacrónico del Español (CORDE) .....	22
3.2.2.	Corpus del Español Actual (CEA).....	23
3.2.3.	Corpus del Proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA) .....	24

3.2.4.	Corpus para el estudio del español oral ESLORA.....	24
3.3.	Corpus para el análisis de sentimientos .....	25
3.3.1.	Corpus del Taller de Análisis Semántico en la SPLN (TASS) .....	25
3.3.2.	Corpus del Sistema de Evaluación de Recursos Educativos (SERE) .....	25
3.3.3.	Corpus SentiText .....	26
3.3.4.	Corpus EduSERE .....	26
3.3.5.	Tabla Comparativa .....	27
3.4.	Algoritmo para la clasificación de sentimiento en texto .....	28
3.4.1.	Enfoque evolutivo multilingüe para el análisis de sentimientos (EvoMSA) .....	28
4.	Desarrollo del proyecto .....	30
4.1.	Metodología para la creación de los corpus validados por expertos. ....	30
4.2.	Análisis del sistema de etiquetado de frases .....	31
4.2.1.	Alcance .....	32
4.2.2.	Requerimientos de usuario .....	32
4.2.2.1.	Requisitos funcionales.....	32
4.2.2.2.	Requisitos de calidad.....	33
4.2.2.3.	Restricciones.....	34
4.2.3.	Actores.....	34
4.2.4.	Casos de uso .....	35
4.3.	Diseño del sistema de etiquetado de frases.....	39
4.3.1.	Diagrama de contexto .....	39
4.3.2.	Arquetipos y sus relaciones.....	40
4.3.3.	Vista lógica .....	42
4.3.4.	Vista de desarrollo .....	43
4.3.5.	Vista física .....	44
4.4.	Almacenamiento de los corpus en una base de datos.....	45
4.5.	Desarrollo del sistema de etiquetado de frases.....	46
4.5.1.	Metodología de desarrollo .....	46
4.5.2.	Interfaces de la aplicación Label .....	46
4.6.	Selección de expertos para el etiquetado.....	54
4.7.	Distribución del sistema de etiquetado de frases .....	55
4.8.	Validación de los corpus por expertos.....	55
4.9.	Creación de los nuevos corpus validados por expertos.....	55
5.	Resultados.....	58
5.1.	Tamaño de los nuevos corpus en español validados por expertos .....	58
5.1.1.	Corpus SentiText validado por expertos .....	58
5.1.2.	Corpus EduSERE validado por expertos .....	60
5.1.3.	Nuevo corpus SentiText .....	62
5.1.4.	Nuevo corpus EduSERE.....	63
5.1.5.	Diferencias en el etiquetado de los corpus validados con el original .....	63

5.2.	Clasificadores de polaridad y emociones en texto.....	64
5.2.1.	Comparación de precisiones en los clasificadores .....	65
6.	Conclusiones .....	67
6.1.	Conclusiones .....	67
6.2.	Aportaciones .....	68
6.3.	Trabajo futuro .....	68
	Bibliografía.....	69

# Índice de figuras

Figura 2-1 Proceso de la minería de opiniones (Hemmatian, & Sohrabi, 2019) .....	11
Figura 2-2 Técnicas de clasificación de sentimientos (Medhat, Hassan, & Korashy, 2014) .....	13
Figura 2-3 Arquitectura de cuatro componentes de un STI (Nkambou, Bourdeau, & Mizoguchi, 2010) .....	17
Figura 3-1 Sentimiento estimado de los tweets (Healey, & Ramaswamy, 2019) .....	20
Figura 3-2 Esquema de predicción de EvoMSA (Graff, Miranda Jimenez, Tellez, & Moctezuma, 2020) .....	29
Figura 4-1 Metodología para la creación de los corpus validados por expertos .....	31
Figura 4-2 Diagrama de casos de uso del sistema de etiquetado .....	35
Figura 4-3 Diagrama de contexto del sistema de etiquetado de frases .....	40
Figura 4-4 Arquetipos del sistema de etiquetado de frases .....	41
Figura 4-5 Vista lógica del sistema de etiquetado de frases .....	43
Figura 4-6 Vista de desarrollo del sistema de etiquetado de frases .....	44
Figura 4-7 Vista física del sistema de etiquetado de frases .....	45
Figura 4-8 Interfaz gráfica de inicio de sesión .....	47
Figura 4-9 Interfaz gráfica de registro de usuario .....	48
Figura 4-10 Interfaz gráfica de agregar información personal .....	49
Figura 4-11 Interfaz gráfica de tutorial .....	50
Figura 4-12 Interfaz gráfica que muestra los niveles de la aplicación .....	51
Figura 4-13 Interfaz gráfica de tabla de puntajes .....	52
Figura 4-14 Interfaz gráfica de etiquetado de frases .....	53
Figura 4-15 Interfaz gráfica de trofeos .....	54
Figura 4-16 Algoritmo de puntaje para etiquetar frases .....	57
Figura 5-1 Porcentaje de frases validadas del corpus SentiText .....	59
Figura 5-2 Porcentaje de frases validadas del corpus EduSERE .....	61
Figura 5-3 Diferencias en el etiquetado de los corpus SentiText y EduSERE .....	64

# Índice de tablas

Tabla 3-1 Comparación de corpus en español para el análisis de sentimiento en texto .....	28
Tabla 4-1 Requisitos funcionales del sistema de etiquetado .....	33
Tabla 4-2 Requisitos de calidad del sistema de etiquetado .....	34
Tabla 4-3 Actores en sistema de etiquetado de frases .....	35
Tabla 4-4 CU-01. Registro de usuarios .....	36
Tabla 4-5 CU-02. Selección de nivel .....	37
Tabla 4-6 CU-03. Etiquetar frase .....	38
Tabla 4-7 CU-04. Ver tabla de puntuaciones .....	39
Tabla 4-8 Arquetipos del sistema de etiquetado de frases .....	41
Tabla 5-1 Total de frases por etiqueta del corpus SentiText validado .....	59
Tabla 5-2 Ejemplo de frases etiquetadas por polaridad .....	60
Tabla 5-3 Total de frases por etiqueta del corpus EduSERE validado .....	61
Tabla 5-4 Ejemplo de frases etiquetadas por emoción centrada en el aprendizaje .....	62
Tabla 5-5 Total de frases por etiqueta del nuevo corpus SentiText .....	63
Tabla 5-6 Total de frases por etiqueta del nuevo corpus EduSERE .....	63
Tabla 5-7 Comparación de precisión en los clasificadores de polaridad del corpus SentiText .....	65
Tabla 5-8 Comparación de precisión en los clasificadores de emociones del corpus EduSERE .....	66

# Capítulo 1

## 1. Introducción

En la última década, los estudiantes han adquirido nuevas formas de desarrollar y mejorar su capacidad de aprendizaje con el uso de herramientas tecnológicas que se adaptan a los diversos estados cognitivos y afectivos, tales como los Sistemas Tutores Inteligentes (STI) y los Entornos Inteligentes de aprendizaje (EIA) los cuales influyen positivamente en el aprendizaje del estudiante (D'Mello, Jackson, Craig, Morgan, Chipman, White, Person, Kort, El Kaliouby, Picard, & Graesser, 2008). Estos sistemas y ambientes inteligentes utilizan diversas técnicas de aprendizaje automático, lógica difusa, gamificación, entre otras, para tomar decisiones acertadas sobre el flujo y la dosificación de temas, ejercicios y exámenes que se le administran a cada estudiante (Zatarain Cabada, Barrón Estrada, González Hernández, & Oramas Bustillos, 2015), o el aprendizaje automático para reconocer emociones expresadas en rostro (Zatarain Cabada, Barrón Estrada, Ríos Félix, & Alor Hernández, 2018).

Existen diversas técnicas para reconocer emociones, algunas utilizan dispositivos especiales para captar señales de los usuarios y otras menos invasivas permiten captar imágenes, posturas o texto, mientras el usuario utiliza un sistema de software. Se ha demostrado que las emociones de los estudiantes son una parte esencial en su vida psicológica y pueden influir profundamente en su rendimiento académico (Pekrun, 1992); por este motivo, los sistemas y ambientes inteligentes se desarrollan integrando componentes capaces de empatizar con el estudiante e interpretar su estado emocional para mantenerlo motivado e interesado en los temas que se le imparten.

Para poder desarrollar sistemas y ambientes inteligentes que puedan detectar emociones en texto y personalizar la educación para los estudiantes, es necesario entrenar algoritmos de inteligencia artificial con corpus de frases en español etiquetados con emociones que vayan dirigidas al área de la educación.

Hoy en día es común que los usuarios de diversas plataformas emitan opiniones sobre los bienes o servicios que se ofrecen, estas opiniones están disponibles para otros usuarios los

cuales las usan para tomar decisiones afectando de forma positiva o negativa a una empresa o servicio. Las opiniones generadas por los usuarios en estas plataformas pueden ser usadas para generar corpus de frases y realizar minería de opiniones.

En este trabajo, se detalla el proceso diseñado para permitir que los expertos del área de emociones realicen la validación de corpus de frases en español, las cuales fueron generadas con recursos educativos enfocados al área de programación de computadoras. Una vez que el corpus haya sido validado por los expertos, éste se utilizará para entrenar un sistema clasificador para reconocer emociones en texto, el cual podrá ser utilizado en los ambientes inteligentes de aprendizaje que deseen adaptar dinámicamente el comportamiento del tutor al estado emocional del usuario.

A continuación, se presentan brevemente los antecedentes del proyecto, la descripción del problema, la justificación para realizar este trabajo, los objetivos principales y específicos que se acordaron, también se establece la hipótesis, y finaliza con la organización de este trabajo de tesis.

## **1.1. Antecedentes**

Para poder llevar a cabo este trabajo de investigación se requirió usar los corpus SentiText y EduSERE, que son corpus que fueron generados en el Laboratorio de Tecnologías Inteligentes Aplicadas a la Educación (LTIAE) del posgrado en el Instituto Tecnológico de Culiacán. Estos corpus contienen frases en español sobre temas de recursos educativos para el aprendizaje en el área de la programación, cada uno tuvo un proceso diferente de creación. SentiText por su parte contiene 24,543 frases en español con etiquetas positivo y negativo, donde 12,272 son positivas y 12,271 negativas, que fueron recopiladas y extraídas de Twitter de diferentes mensajes emitidos por usuarios sobre temas educativos relacionados a la programación. También se obtuvieron más frases a partir del Sistema de Evaluación de Recursos Educativos (SERE) (Barrón Estrada, Zatarain Cabada, Oramas Bustillos, & Ramírez Ávila, 2017) y de diferentes plataformas

educativas (Ramírez Ávila, 2018), el cual recolecta opiniones de estudiantes sobre recursos educativos para el área de la programación.

Por otra parte, el corpus EduSERE que contiene 12,083 frases con emociones centradas en el aprendizaje (frustrado, aburrido y comprometido), de las cuales 5,600 son comprometido, 3,238 aburrido y 3,245 frustrado. El cual fue creado a partir de la recopilación de opiniones para el área de la programación utilizando diferentes herramientas, plataformas y sistemas, como el sistema SERE, de plataformas educativas en línea (Udemy, Coursera, y Teachlr) y de la extracción de opiniones en Twitter (Ramírez Ávila, 2018).

## **1.2. Descripción del problema**

Actualmente existen diversos STI y EIA en apoyo a la educación, que facilitan el aprendizaje dentro y fuera de las aulas, adaptando los temas educativos a los estados cognitivos y afectivos de los estudiantes. En el LTIAE se han desarrollado STI enfocados a la ayuda y enseñanza del lenguaje de programación Java o la lógica algorítmica, algunos ejemplos son: EasyLogic (Zatarain Cabada, Barrón Estrada, Ríos Félix, & Alor Hernández, 2018), Java Sensei (Zatarain Cabada, Barrón Estrada, González Hernandez, & Alor Hernandez, 2016), Find Error Java (Barrón Estrada, Zatarain Cabada, Valencia Rodríguez, & Peralta Peñuñuri, 2018), ILE Affective Java (Barrón Estrada, Zatarain Cabada, Aispuro Gallegos, Sosa Ochoa, & Lindor Valdez, 2015), entre otros. Estos ambientes y sistemas inteligentes utilizan técnicas de lógica difusa, gamificación y reconocimiento automático de emociones mediante expresiones y gestos del rostro que se capturan mediante una cámara en tiempo real, sin embargo, se sigue investigando diferentes técnicas de reconocimiento automático de emociones para que estos ambientes y sistemas inteligentes funcionen aún mejor al adaptarse a las necesidades cognitivas y afectivas de los estudiantes.

En el LTIAE se continúa trabajando en la creación de corpus de frases en español relacionados a la enseñanza de programación para crear clasificadores automáticos de emociones en texto, con el fin de integrarlos a estos ambientes y sistemas inteligentes para aumentar la eficiencia en el reconocimiento de emociones y así mejorar la enseñanza hacia



los estudiantes. Los corpus existentes que se utilizan para el reconocimiento automático de emociones en texto son SentiText y EduSERE, pero estos fueron etiquetados totalmente por estudiantes universitarios cursando materias de programación de computadoras y por miembros del LTIAE, que no cuentan con suficiente experiencia en la clasificación de polaridades y emociones, y no se garantiza que las etiquetadas sean las correctas para las frases de los corpus, por lo tanto los clasificadores creados a partir de ellos pueden no dar una clasificación de las emociones correctamente, por lo cual en esta investigación se trata de realizar el etiquetado y validado de los corpus por expertos para poder tener corpus confiables que ayuden a aumentar la eficiencia en el reconocimiento automático de emociones en texto.

### **1.3. Justificación**

La calidad de los datos que se usan para el entrenamiento de los clasificadores determina la precisión alcanzada en la clasificación, es por esto que es muy importante que los datos que se van a utilizar tengan etiquetas confiables las cuales hayan sido generadas por personas expertas en el área. La creación de corpus de frases en español con recursos educativos enfocados en la programación de computadoras, etiquetados y validados por expertos en el área de emociones contribuirá a incrementar la eficiencia en el reconocimiento automático de emociones en texto que se utiliza en ambientes de aprendizaje inteligentes, con el fin de que estos puedan adaptarse mejor a las necesidades cognitivas y afectivas de los estudiantes.

### **1.4. Objetivos**

En esta sección se presentan el objetivo general y los objetivos específicos que se definieron para la realización de este trabajo de investigación.

#### **1.4.1. Objetivo general**

Generar dos nuevos corpus de frases en español a partir de los corpus SentiText y EduSERE, que estén etiquetados y validados por expertos en el área de emociones con el fin de incrementar la eficiencia del reconocimiento automático de emociones en texto. Aumentando la precisión de 91% para el clasificador de emociones en texto creado a partir del corpus SentiText y la precisión de 82% para el clasificador creado a partir del corpus EduSERE.

#### **1.4.2. Objetivos específicos**

Para alcanzar el objetivo general, se definieron 9 objetivos específicos que se listan a continuación:

1. Desarrollar un sistema (aplicación móvil) para automatizar el etiquetado de frases de los corpus SentiText y EduSERE con polaridad y emociones centradas en el aprendizaje respectivamente.
2. Convocar a un grupo de expertos en el área de psicología y pedagogía para que realicen el etiquetado de frases.
3. Convocar a un grupo de profesores y estudiantes para que realicen el etiquetado de frases.
4. Definir una metodología para la creación de los corpus etiquetados y validados por expertos.
5. Desarrollar un algoritmo de ponderación para elegir la mejor etiqueta de los expertos para cada frase de los corpus SentiText y EduSERE.
6. Generar un nuevo corpus SentiText con etiquetas de polaridad (positiva o negativa) asignadas por expertos.
7. Generar un nuevo corpus EduSERE con etiquetas de emociones centradas en el aprendizaje (aburrido, frustrado o comprometido) asignadas por expertos.
8. Probar clasificadores de emociones en texto con los dos nuevos corpus generados.
9. Comparar resultados obtenidos de los clasificadores de emociones en texto para los corpus nuevos y los anteriores.

## **1.5. Hipótesis**

La precisión de un clasificador para el reconocimiento automático de emociones centradas en el aprendizaje en texto mejora un 1% cuando se usa un corpus etiquetado y validado por expertos en el área de emociones en comparación con un corpus etiquetado por personas normales.

La precisión de un clasificador para el reconocimiento automático de polaridad en texto mejora un 1% cuando se usa un corpus etiquetado y validado por expertos en el área de emociones en comparación con un corpus etiquetado por personas normales.

## **1.6. Organización de la tesis**

En este documento de tesis, se describe la metodología para crear dos corpus de frases en español con recursos educativos enfocados a la programación de computadoras, validados por expertos, los cuales pueden ser utilizados para el reconocimiento automático de emociones en texto que se implementa en ambientes de aprendizaje inteligente y sistemas tutores inteligentes.

Este documento de tesis está dividido en 5 capítulos principales, marco teórico, estado del arte, desarrollo del proyecto, resultados y conclusiones. A continuación, se ofrece una descripción general de cada uno de los siguientes capítulos.

El capítulo 2 muestra el marco teórico, en el cual se describe el conocimiento que existe y puede guiar al trabajo para la creación de corpus en español. En el capítulo 3 se presenta el estado del arte, donde se refieren y analizan investigaciones y trabajos relevantes relacionados a este tema de tesis. El capítulo 4 presenta detalladamente la metodología que se siguió para la realización de la aplicación de etiquetado de frases, así como la creación de los nuevos corpus validados. En el capítulo 5 se muestran los resultados obtenidos de los nuevos corpus SentiText y EduSERE validados por expertos. Para finalizar en el capítulo 6 se abordan las conclusiones y aportaciones que se realizaron en este trabajo de tesis, así como el trabajo a futuro.

# Capítulo 2

## 2. Marco teórico

En este capítulo se describen los conceptos y fundamentos teóricos que respaldan la realización de este trabajo de tesis. Primero se aborda el concepto de corpus y de minería de opiniones, sus aplicaciones y las técnicas para la clasificación automática de sentimientos, posteriormente se define el concepto de computación afectiva, y por último se habla sobre los ambientes educativos, como los sistemas tutores inteligentes y los ambientes inteligentes de aprendizaje.

### 2.1. Corpus

Un corpus se describe simplemente como un gran cuerpo de evidencia lingüística típicamente compuesta por el uso de un idioma comprobado. Puede ser hablado y escrito. El término corpus solo debe aplicarse correctamente a una colección de datos bien organizada, recopilada dentro de los límites de un marco de muestreo diseñado para permitir la exploración de una determinada característica lingüística (o conjunto de características) a través de los datos recopilados (McEnery, 2012). Todos los datos recopilados en un corpus pueden servir para diferentes propósitos de investigación y por lo general pueden ser leídos fácilmente por una máquina.

#### 2.1.1. Tipos de corpus

Un corpus es una colección de datos lingüísticos recopilados que sirven para diferentes propósitos de investigación y pueden tener variaciones significativas de acuerdo con los datos que manejan:

- **Corpus escritos:** Consisten en textos originalmente escritos en un idioma.

- **Corpus hablados:** Se componen por un conjunto de grabaciones de audio en un idioma, son usados para el reconocimiento del habla.
- **Corpus multimodales:** incluyen voz e imágenes.

También se pueden clasificar por el tipo de documentos que manejan (McEnery, 2012):

- **Corpus monolingües:** Se componen de un solo idioma.
- **Corpus comparables:** son corpus en los que se recopilan una serie de corpus monolingües para una variedad de idiomas, preferiblemente utilizando el mismo marco de muestreo y con un equilibrio y representatividad similares, para permitir el estudio de esos idiomas en contraste.
- **Corpus paralelos:** Estos corpus adoptan un enfoque ligeramente diferente para el estudio de idiomas en contraste, reuniendo un corpus en un idioma y luego las traducciones de esos datos del corpus a uno o más idiomas.

### 2.1.2. Criterios de diseño de corpus

Existen diferentes etapas que se deben llevar a cabo para la creación y planificación de un corpus. Las etapas principales son (Atkins, & Clear, 1992):

- **Especificaciones y diseño:** En esta etapa se deberá establecer qué tipo de corpus se está construyendo, considerando los tamaños de las muestras de texto que se incluirán, la gama de variedades de lengua y el período de tiempo (diacrónico) que se muestreará.
- **Hardware y software:** En esta etapa se requieren estimaciones de las necesidades de hardware y software del proyecto corpus. Estos dependerán en gran medida del tamaño del corpus y de la cantidad de procesamiento y manipulación que se vaya a realizar en él.
- **Captura de datos y marcado:** La captura de datos también requiere mucho tiempo y están determinados en gran medida por la cantidad de texto que se capturará. El material impreso se puede escanear mediante dispositivos OCR con el software adecuado. Alternativamente, el texto se puede introducir en el teclado manualmente, así como la transcripción de grabaciones de audio.

- **Procesamiento de corpus:** En esta etapa es necesario prescindir de herramientas para el procesamiento el corpus. Muchas de estas herramientas ya existen y están en uso, pero están diseñadas para satisfacer necesidades muy específicas por lo que es necesario estandarizar el formato del corpus. Algunas herramientas básicas pueden incluir, frecuencia de palabras, concordancia, búsqueda interactiva, lematización, etiquetado, analizador gramatical, colocación, sentido de desambiguación y enlace a la base de datos léxica.
- **Crecimiento del corpus y retroalimentación:** Para acercarse a un corpus "equilibrado", es práctico adoptar un método de aproximaciones sucesivas. Primero, el constructor del corpus intenta crear un corpus representativo. Luego, este corpus se utiliza y se analiza, se identifica y se reportan las fortalezas y debilidades. A partir de esta experiencia y retroalimentación, el corpus se mejora mediante la adición o eliminación de material y el ciclo se repite continuamente.

### 2.1.3. Desarrollo de tecnologías a partir de corpus

Un corpus bien diseñado y organizado se puede utilizar para desarrollar múltiples tecnologías lingüísticas, como lo son:

- Reconocimiento de emociones en texto
- Conversión de texto en habla
- Reconocimiento automático del habla
- Sistemas de diálogo
- Analizadores morfológicos
- Analizadores sintácticos
- Analizadores semánticos
- Analizadores en el nivel del discurso
- Gramáticas computacionales
- Léxicos computacionales
- Redes léxico-semánticas
- Ontologías

## 2.2. Minería de opiniones

El análisis de sentimientos y minería de opiniones es el campo de estudio que analiza las opiniones, los sentimientos, las evaluaciones, las actitudes y las emociones de las personas a partir del lenguaje escrito (Liu, 2015). La minería de opiniones extrae y analiza la opinión de las personas sobre una entidad, mientras que el análisis de sentimientos identifica el sentimiento expresado en un texto y luego lo analiza (Medhat, Hassan, & Korashy, 2014).

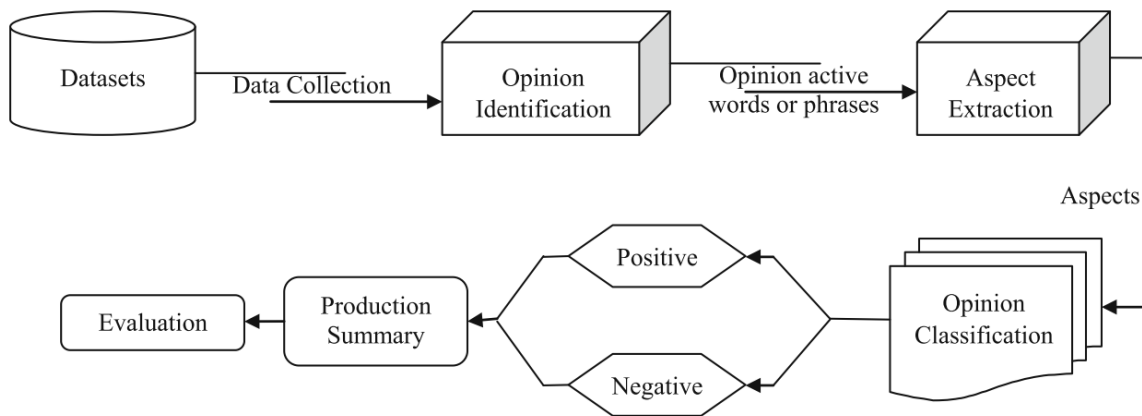
El objetivo principal de la minería de opiniones es automatizar la extracción de los sentimientos expresados por los usuarios a partir de textos no estructurados (Hemmatian, & Sohrabi, 2019), trata de crear herramientas y sistemas capaces de recopilar toda la información que se genera en redes sociales, blogs, discusiones en foros, reseñas, etc, para extraer y clasificar las emociones.

Una de las redes sociales que se ha vuelto popular en la minería de opiniones es Twitter, ya que es un sitio donde los usuarios expresan sus opiniones y sentimientos sobre diferentes temas de interés, además de que Twitter ofrece herramientas para la extracción de datos que se publican en esta red social, lo cual ayuda a recopilar suficiente información para ser analizada y clasificada.

En la figura 2-1 se muestran las 6 etapas del proceso de la minería de opiniones, de acuerdo con Hemmatian, & Sohrabi (2019), que consisten en:

1. **Recopilación de datos:** Tener un conjunto de datos completo y confiable es el primer paso para realizar el proceso de minería de opinión.
2. **Identificación de opiniones:** Todos los comentarios deben estar separados e identificados de los textos presentados en esta fase. Luego, los comentarios extraídos deben procesarse para separar los inapropiados y los falsos.
3. **Extracción de aspecto:** En esta fase se identifican y extraen todos los aspectos existentes según los procedimientos. Seleccionar los aspectos potenciales podría resultar muy eficaz para mejorar la clasificación.
4. **Clasificación de opinión:** Después de la identificación de la opinión y la extracción de los aspectos que se pueden considerar como la fase de preprocesamiento, en este paso se clasifican las opiniones utilizando diferentes técnicas de clasificación.

5. **Resumen de producción:** En este paso se produce un resumen de los resultados de la opinión, que pueden estar en diferentes formatos, como texto, gráficos, etc.
6. **Evaluación:** Aquí se evalúa el desempeño de la clasificación de opiniones, como la precisión y exactitud.



*Figura 2-1 Proceso de la minería de opiniones (Hemmatian, & Sohrabi, 2019)*

### 2.2.1. Aplicaciones de la minería de opiniones

El análisis de sentimientos intenta describir los intereses y sentimientos de los usuarios en internet sobre diferentes temas sociales, políticos, económicos y culturales. Una de las mayores aplicaciones de la minería de opiniones y el análisis de sentimientos es en el área comercial y puede ser analizada desde tres puntos de vista (Hemmatian, & Sohrabi, 2019):

- **El punto de vista de los clientes individuales:** cuando alguien quiere comprar un producto, tener un resumen de las opiniones de los demás puede ser más útil que estudiar la gran cantidad de comentarios de los demás sobre este producto.
- **Las organizaciones empresariales y el punto de vista de los productores:** este tema es importante para que las organizaciones mejoren sus productos. Esta información se utiliza no solo para la comercialización y evaluación de productos, sino también para el diseño y desarrollo de productos. Las empresas de fabricación pueden incluso aumentar, disminuir o cambiar los productos en función de las opiniones de los clientes.



- **El punto de vista de las empresas de publicidad:** las opiniones son importantes para las empresas de publicidad porque pueden obtener ideas de la demanda del mercado. La perspectiva pública de las personas y el tipo de productos que les interesan se puede encontrar entre los elementos que extrae la minería de opinión.

La minería de opiniones también se puede aplicar en la política, donde la recopilación de comentarios puede ayudar a los políticos a beneficiarse de las múltiples opiniones de las personas para tomar decisiones sobre sus campañas.

Otra área importante de aplicación es en la educación, con la minería de datos educativa, que se centra en aplicar varias técnicas de minería de opiniones a varios atributos educativos para descubrir información útil. El objetivo principal es utilizar técnicas para descubrir conocimientos que puedan utilizarse para predecir información importante relacionada con los estudiantes. El desempeño del estudiante juega un papel importante en la decisión de su futuro, por lo que es necesario predecir su desempeño. En caso de que se prediga que los estudiantes fallarán, se pueden hacer esfuerzos adicionales para que el estudiante mejore su resultado. Otra área clave donde se aplica la minería de datos educativa es en la mejora del apoyo pedagógico mediante el uso de software de aprendizaje. Los sistemas educativos modernos proporcionan a los estudiantes una variedad de material de apoyo pedagógico (Dhingra, & Sardana, 2017).

### 2.2.2. Técnicas para la clasificación automática de sentimientos

El paso más importante y crítico de la extracción de opiniones es seleccionar una técnica adecuada para clasificar los sentimientos (Hemmatian, & Sohrabi, 2019). A continuación, se muestran tres técnicas de métodos de aprendizaje automático utilizadas para la clasificación de sentimientos:

- **Aprendizaje sin supervisión:** En esta técnica se utilizan algoritmos de agrupamiento (*clustering*) los cuales se dividen en dos, la primera llamada agrupamiento jerárquico (*Hierarchical Clustering*) la cual contiene dos algoritmos: aglomerativo y divisorio; y la segunda llamada agrupamiento particionado (*Partitioning Clustering*) que contiene los algoritmos K-means y fuzzy C-means.

- **Aprendizaje semi-supervisado:** Esta técnica de aprendizaje automático se divide en los siguientes métodos de clasificación: auto-entrenamiento, co-entrenamiento, aprendizaje multi vista, métodos basados en gráficas, modelos generativos.
- **Aprendizaje supervisado:** La técnica de aprendizaje supervisado se clasifica en dos tipos, métodos probabilísticos de clasificación y métodos no probabilísticos de clasificación. Los métodos probabilísticos utilizados son: Naïve Bayes, redes Bayesianas y máxima entropía; los métodos utilizados para la clasificación no probabilística pueden ser: máquina de soporte vectorial, redes neuronales, K-vecinos más cercanos, árboles de decisión y basados en reglas.

En la figura 2-2 se muestra otra propuesta por Medhat, Hassan, & Korashy (2014), para la agrupación de técnicas de clasificación de sentimientos, donde se propone separar en 2 categorías con base en los enfoques: aprendizaje automático y basado en léxico.

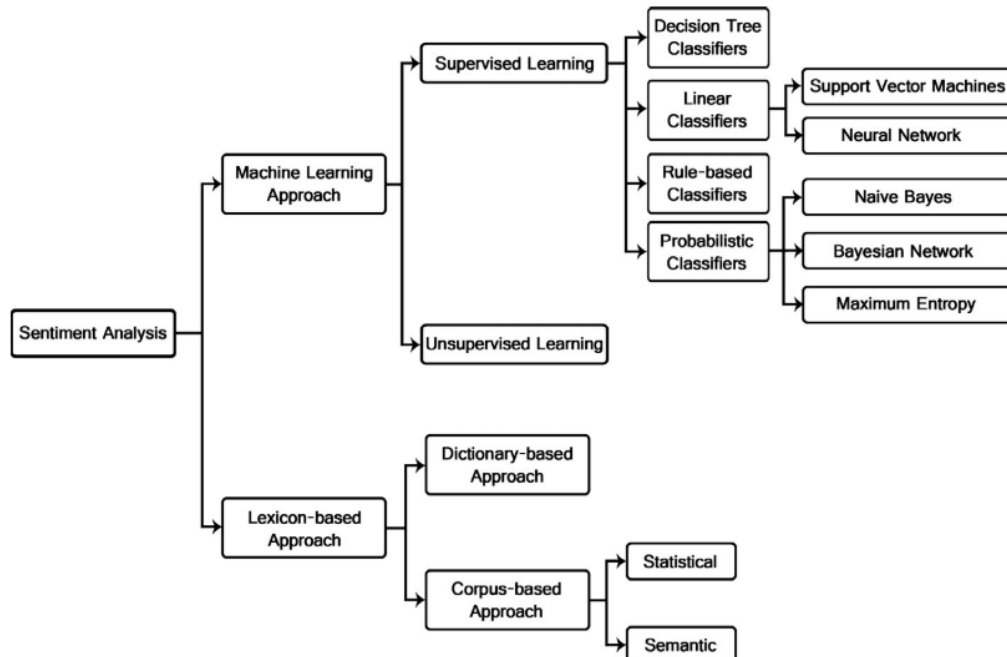


Figura 2-2 Técnicas de clasificación de sentimientos (Medhat, Hassan, & Korashy, 2014)

## **2.3. Computación afectiva**

La computación afectiva es el estudio de técnicas de reconocimiento, expresión y generación de emociones para que una máquina pueda entender y procesar emociones como un ser humano (Picard, 1997). Sus orígenes se basan en el estudio de las emociones para permitir que una máquina puede simular empatía y adaptar su comportamiento de acuerdo con los estados cognitivos y afectivos de los humanos.

### **2.3.1. Definición de emoción**

De acuerdo con Hockenbury, & Hockenbury (2007) el termino emoción es un estado psicológico complejo que involucra tres componentes distintos: una experiencia subjetiva, una respuesta fisiológica y una respuesta conductual o expresiva. Son reacciones orgánicas que se experimentan cuando se recibe algún tipo de estímulo externo, estas reacciones pueden ser fisiológicas, psicológicas o conductuales y tienen una duración más corta y una intensidad mayor que un estado de ánimo.

Las emociones pueden generar comportamientos observables como posturas, movimientos, expresiones faciales, etc.

### **2.3.2. Clasificación de las emociones**

Existen diferentes formas de clasificar las emociones, una de las más conocidas y definitorias, muestra a las emociones clasificadas como primarias y secundarias. Las emociones primarias son reacciones inmediatas en respuesta a un estímulo, suelen ser respuestas corporales y fáciles de identificar. Paul Ekman (1992) definió seis emociones básicas o primarias: miedo, disgusto, ira, sorpresa, felicidad y tristeza. Todas ellas existen en todos los seres humanos y no influyen los factores de cultura o educación.

Por otra parte, las emociones secundarias se aprenden en un ambiente social, mediante interacciones con personas cercanas. Estas emociones son respuestas que se generan después de aparecer una emoción primaria. Las emociones secundarias son difíciles de identificar ya

que a diferencia de las primarias en ellas si intervienen factores como la cultura o educación. Algunas de ellas pueden ser: vergüenza, odio, orgullo, sufrimiento, alivio.

### **2.3.3. Emociones centradas en el aprendizaje**

Las emociones centradas en el aprendizaje son emociones secundarias que se manifiestan en estados afectivos cuando los estudiantes realizan diferentes actividades de aprendizaje. En el proceso de aprendizaje influyen múltiples factores tanto cognitivos como emocionales, por eso el papel de las emociones se ha considerado como uno de los aspectos fundamentales en la formación integral del educando (Ángel, & Retana, 2012). De acuerdo con Pekrun (1992), las emociones son una parte fundamental de la vida psicológica de los estudiantes y pueden influir profundamente en la motivación académica y en las estrategias cognitivas de aprendizaje y logro. Las emociones como comprometido, enfocado y relajado se consideran como positivas en el estado del aprendizaje y ayudan a que los estudiantes puedan aprender mejor, mientras que las emociones como aburrimiento, frustración, ansiedad, decepción muestran un estado negativo en el aprendizaje (Pekrun 1992).

### **2.3.4. Técnicas para la detección de emociones**

Los seres humanos son capaces de expresar emociones en diversas formas, y para que una computadora pueda detectar el estado emocional de una persona es necesario que capte señales del usuario a través de diferentes componentes de hardware como: sensores, cámaras, teclados y micrófonos. También es necesario que mediante software se obtengan estos datos que capta el hardware para que una computadora pueda ser capaz de reconocer y clasificar emociones utilizando técnicas como, el reconocimiento de voz, procesamiento del lenguaje natural, la detección de expresiones faciales y corporales, entre otros. Utilizando los componentes necesarios de hardware junto con técnicas de software, una computadora puede ser capaz de detectar los estados emocionales para facilitar la interacción con los humanos.

## **2.4. Ambientes educativos**

En el ámbito educativo se utiliza la tecnología para proveer ambientes y sistemas que apoyen el proceso de aprendizaje de los estudiantes. En esta sección se presentan las características de dos tipos de ambientes educativos.

### **2.4.1. Sistema tutor inteligente**

Los sistemas tutores inteligentes (STI) son programas informáticos complejos que gestionan varios tipos heterogéneos de conocimiento, que van desde el dominio, hasta el conocimiento pedagógico (Nkambou, Bourdeau, & Psyché, 2010). Construir un STI y representar un proceso de aprendizaje requiere de diferentes disciplinas, incluidas la informática, la psicología y la educación (Woolf, 2010).

Muchos de los métodos y herramientas de la informática, la psicología y la educación son complementarios y, en conjunto, brindan una cobertura casi completa del campo de la inteligencia artificial y la educación. La inteligencia artificial aborda cómo razonar sobre la inteligencia y, por tanto, el aprendizaje. La psicología, en particular su subcampo, ciencia cognitiva, aborda cómo las personas piensan y aprenden, y la educación se centra en cómo mejorar la enseñanza (Woolf, 2010). Estas bases multidisciplinarias de los STI sirven como un primer paso para una teoría cognitivamente correcta del aprendizaje y para adaptarse mejor a las necesidades de los estudiantes.

#### **2.4.1.1. Arquitectura de un sistema tutor inteligente**

Nkambou, Bourdeau, & Mizoguchi, (2010) exponen que la arquitectura de un STI divide básicamente el sistema en cuatro componentes principales que se describen a continuación:

- **Modelo de dominio:** El modelo de dominio (también llamado conocimiento experto) contiene los conceptos, reglas y estrategias de resolución de problemas del dominio a aprender. Puede cumplir varios roles: como fuente de conocimiento experto, para evaluar el desempeño del alumno o para detectar errores.

- **Modelo de estudiante:** Es el componente central de un ITS. Idealmente, debería contener tanto conocimiento como sea posible sobre los estados cognitivos y afectivos del estudiante y su evolución a medida que avanza el proceso de aprendizaje
- **Modelo del Tutor:** El modelo de tutoría recibe información del dominio y del modelo de estudiante, y a partir de ello toma decisiones sobre estrategias y acciones de tutoría. La planificación del contenido y la entrega también forman parte de las funciones del modelo de tutoría.
- **Interfaz:** Las interacciones alumno / tutor suelen producirse a través de la interfaz de aprendizaje, también conocida como componente de comunicación o interfaz. Este componente da acceso a los elementos de conocimiento del dominio a través de múltiples formas de entorno de aprendizaje, que incluyen, simulaciones, hipertexto, micro-mundos, etc.

En la figura 2-3 se muestra los 4 componentes principales de la arquitectura y como se comunican entre ellos.

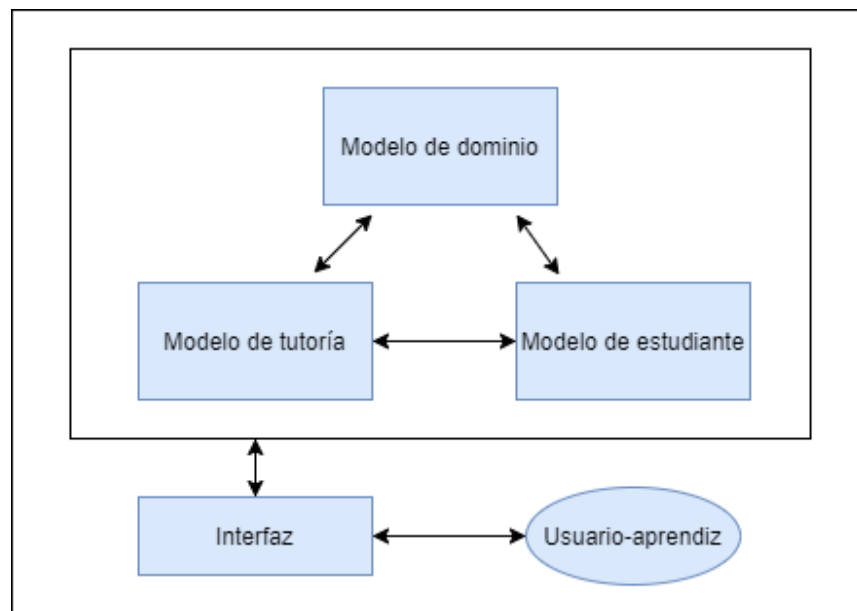


Figura 2-3 Arquitectura de cuatro componentes de un STI (Nkambou, Bourdeau, & Mizoguchi, 2010)

#### **2.4.2. Entorno inteligente de aprendizaje**

Un Entorno Inteligente de Aprendizaje (EIA) es un tipo de sistema educativo inteligente que combina las características de los sistemas tutores inteligentes (STI) tradicionales y los entornos de aprendizaje. Un EIA incluye un componente especial para apoyar el aprendizaje impulsado por el estudiante, el módulo de entorno. El término entorno se usa para referirse a la parte del sistema que especifica o respalda las actividades que realiza el estudiante y los métodos disponibles para que el estudiante realice esas actividades. Algunos EIA incluyen también un componente especial llamado manual, que proporciona acceso a material de instrucción estructurado. Un EIA integrado que incluye el entorno y los componentes manuales, además del componente de tutoría regular, puede respaldar el aprendizaje de conocimientos tanto procedimentales como declarativos y proporcionar estilos de aprendizaje controlados por el sistema y dirigidos por los estudiantes (Brusilovsky, 1994).

# Capítulo 3

## 3. Estado del arte

En este capítulo se presentan las investigaciones relacionados con este trabajo de tesis. A continuación, se describen algunas herramientas utilizadas para la minería de opiniones y el análisis de sentimientos, después se muestran algunos corpus de frases en español existentes para las investigaciones lingüísticas, así como algunos corpus de frases en español para el análisis de sentimientos. Para finalizar se presentan un algoritmo existente para el análisis automático de emociones en texto.

### 3.1. Herramientas para el análisis de sentimientos

A continuación, se muestran algunas herramientas de uso libre y de paga que sirven para la extracción de opiniones y el análisis de sentimientos en texto.

#### 3.1.1. Tweet sentiment visualization

*Tweet sentiment visualization* es una herramienta de uso gratuito para el análisis de sentimientos, que permite al usuario escribir palabras clave en un buscador con el fin de mostrar los Tweets relacionados más recientes. El objetivo es mostrar las propiedades emocionales básicas incorporadas en el texto de los tweets. Para mejorar la visualización de los datos la herramienta presenta diferentes filtros (Healey, & Ramaswamy, 2019):

- **Sentimiento:** Cada *tweet* se muestra como un círculo colocado por sentimiento, una estimación de la emoción contenida en el texto del *tweet*.
- **Temas:** Los *tweets* sobre un tema común se agrupan en grupos de temas.
- **Mapa de calor:** Muestra una cuadrícula de mapeo de *tweets*.
- **Nube de etiquetas:** Se muestran las palabras comunes de las regiones emocionales, Feliz, Relajado e Infeliz.



- **Línea de tiempo:** Los *tweets* se dibujan en un gráfico de barras para mostrar el número de *tweets* publicados en diferentes momentos.
- **Mapa:** Los *tweets* se dibujan en un mapa del mundo en el lugar donde se publicaron.
- **Afinidad:** Los *tweets* frecuentes, las personas, los *hashtags* y las URL se dibujan en un gráfico para mostrar a los actores importantes en el conjunto de *tweets* y cualquier relación o afinidad que tengan entre sí.
- **Narrativa:** Al seleccionar un *tweet* ancla de interés de la lista de *tweets*, se muestra una secuencia de *tweets* ordenados en el tiempo que forman conversaciones o hilos narrativos que pasan por el *tweet* ancla.
- **Tweets:** Los *tweets* se enumeran para mostrar su fecha, autor, sentimiento y texto.

En la figura 3-1 se muestra la segmentación del sentimiento expresado en los tweets relacionados a una palabra clave, también se muestran los diferentes filtros con los que cuenta la herramienta.

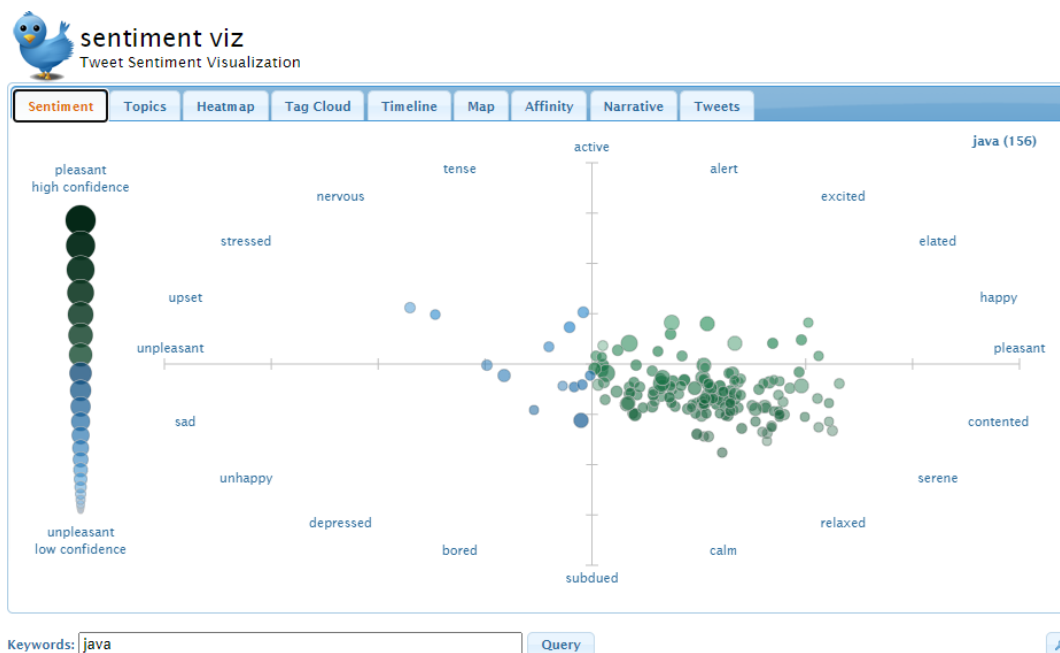


Figura 3-1 Sentimiento estimado de los tweets (Healey, & Ramaswamy, 2019)

### 3.1.2. RapidMiner

*RapidMiner* es una plataforma de software de datos que permite minar texto para ayudar a las marcas a realizar análisis de sentimientos. Es posible analizar las reseñas y posts en redes sociales, al igual que publicaciones oficiales y documentos. Las marcas pueden identificar tendencias de temas que resuenen con consumidores y clientes, recolectar retroalimentación en lanzamiento de productos y encontrar nuevas áreas para expandir su negocio (Carpenter, 2018).

### 3.1.3. API de análisis de opiniones de MeaningCloud

La API de *MeaningCloud* realiza un análisis de sentimiento multilingüe detallado a partir de información proveniente de diversas fuentes. El texto proporcionado se analiza para determinar si expresa un sentimiento positivo, neutro o negativo (o si es imposible detectar algún sentimiento). Además de la polaridad global y a nivel de frase, la API usa técnicas avanzadas de procesamiento del lenguaje natural para detectar la polaridad asociada tanto a las entidades como a los conceptos del texto. También permite al usuario detectar la polaridad de entidades y conceptos que él mismo defina, lo que convierte al servicio en una herramienta aplicable a cualquier tipo de escenario. A continuación, se presentan las características principales de la API de análisis de sentimientos (MeaningCloud, 2020):

- **Sentimiento global:** extrae la opinión general expresada en un *tweet*, publicación o reseña.
- **Sentimiento a nivel de atributo:** detecta un sentimiento específico por un objeto o alguna de sus cualidades, analizando en detalle el sentimiento de cada frase.
- **Identificación de opiniones y hechos:** distingue entre la expresión de un hecho objetivo o una opinión subjetiva.
- **Detección de ironía:** identifica comentarios en los que lo que se expresa es lo contrario de lo que se dice.
- **Polaridad graduada:** distingue opiniones muy positivas y muy negativas, así como la ausencia de sentimiento.
- **Acuerdo y desacuerdo:** identifica opiniones encontradas y mensajes contradictorios o ambiguos.

### **3.1.4. SentiStrength**

*SentiStrength* es una herramienta de análisis de sentimiento para la investigación académica y puede realizar el análisis automático de opiniones de hasta 16,000 textos de redes sociales por segundo. *SentiStrength* estima la fuerza de los sentimientos positivos y negativos en textos breves, incluso para el lenguaje informal. Tiene precisión de nivel humano para textos breves de redes sociales en inglés, excepto textos políticos. *SentiStrength* reporta dos puntos fuertes de sentimiento (*SentiStrength*, s. f.):

- -1 (no negativo) a -5 (extremadamente negativo)
- 1 (no positivo) a 5 (extremadamente positivo)

*SentiStrength* también puede informar resultados binarios (positivos / negativos), trinaros (positivos / negativos / neutrales) y de una sola escala (-4 a +4).

## **3.2. Corpus de frases en español**

De acuerdo con Wynne (2005), un corpus es una colección de fragmentos de texto lingüístico en formato electrónico, seleccionados de acuerdo con criterios externos para representar, en la medida de lo posible, una lengua o variedad lingüística como fuente de datos para la investigación lingüística.

A continuación, se presentarán algunos ejemplos de corpus existentes en idioma español que sirven como base para realizar investigaciones lingüísticas.

### **3.2.1. Corpus Diacrónico del Español (CORDE)**

El Corpus Diacrónico del Español (CORDE) es un corpus textual de todas las épocas y lugares en que se habló español, desde los inicios del idioma hasta el año 1974, en que limita con el Corpus de Referencia del Español Actual (CREA). CORDE está diseñado para extraer información con la cual estudiar las palabras y sus significados, así como la gramática y su uso a través del tiempo (CORDE, s. f.).

CORDE cuenta en la actualidad con 250 millones de registros correspondientes a textos escritos de muy diferente género. Se distribuyen estos en prosa y verso y, dentro de cada modalidad, en textos narrativos, líricos, dramáticos, científico-técnicos, históricos, jurídicos, religiosos, periodísticos, etc. Se pretende recoger todas las variedades geográficas, históricas y genéricas para que el conjunto sea suficientemente representativo (CORDE, s. f.).

Hoy es fuente obligada para cualquier estudio diacrónico relacionado con la lengua española. La Academia utiliza sistemáticamente el CORDE para documentar palabras, para calificarlas de anticuadas o en desuso, para saber el origen de algunos términos, su tradición en la lengua, y las primeras apariciones de las palabras. Sirvió, además, de material básico para la confección del nuevo diccionario histórico del español (CORDE, s. f.).

### **3.2.2. Corpus del Español Actual (CEA)**

El Corpus del Español Actual (CEA) tiene 540 millones de palabras y está lematizado y etiquetado con información morfológica y/o categorial. El CEA está integrado por los siguientes textos (Subits, & Ortega, 2012):

- La parte española del corpus paralelo español-inglés Europarl: European Parliament Proceedings Parallel Corpus v. 6 (1996-2010).
- El módulo en lengua española del Wikicorpus v. 1.0, que contiene una parte importante de la Wikipedia (2006).
- La sección en español del MultiUN: Multilingual UN Parallel Text 2000-2009, un corpus integrado por resoluciones de la Organización de las Naciones Unidas (ONU).

La etiquetación del CEA se ha realizado con una aplicación que utiliza un diccionario electrónico del español de 635,000 formas, generadas automáticamente a partir de un diccionario de 86,000 lemas simples. La desambiguación de la etiquetación se ha llevado a cabo mediante un proceso de intersección de autómatas, que utiliza información léxica y sintáctica (Subits, & Ortega, 2012).

### **3.2.3. Corpus del Proyecto para el Estudio Sociolingüístico del Español de España y América (PRESEEA)**

El Proyecto para el Estudio Sociolingüístico del Español de España y de América (PRESEEA) constituye una iniciativa que se propone la conformación y el estudio de un corpus de lengua española hablada, representativo del mundo hispánico, en su variedad geográfica y social. Esos materiales se reúnen atendiendo a la diversidad sociolingüística de las comunidades de habla hispanohablantes. PRESEEA agrupa a cerca de 40 equipos de investigación sociolingüística. El proyecto es el fruto del trabajo conjunto de investigadores comprometidos con una metodología común, con el propósito de reunir un banco de materiales comparables que posibilite la aplicación de fines educativos y tecnológicos (Moreno Fernández, 1997; PRESEEA, 2014-2020).

### **3.2.4. Corpus para el estudio del español oral ESLORA**

El corpus para el estudio del español oral ESLORA contiene 60 horas de entrevistas semidirigidas y 20 horas de conversaciones de hablantes de Galicia grabadas entre los años 2007 y 2015. Los registros sonoros se transcribieron ortográficamente con alineación texto-voz para facilitar el acceso inmediato al audio desde la transcripción. En el proceso de enriquecimiento del corpus se han desarrollado recursos para la lematización y el etiquetado morfosintáctico de los textos que permiten realizar diversos tipos de búsquedas (ESLORA, 2018).

Hoy en día es posible obtener datos de una parte de los materiales mediante un sistema de consultas simples y combinadas que incluye variables sociales (grupo de edad, nivel de estudios y sexo) junto a categorías lingüísticas (lemas, clases de palabras y categorías morfológicas). La aplicación de consulta tiene además la ventaja de que permite recuperar los fragmentos sonoros correspondientes a las consultas realizadas. En su versión actual, 1.2.2 de noviembre de 2018, el corpus disponible para consulta consta de 56 documentos que incluyen 647,758 palabras ortográficas y 776,260 elementos gramaticales (ESLORA, 2018).

### **3.3. Corpus para el análisis de sentimientos**

El análisis de sentimientos es el campo de estudio que, mediante el procesamiento del lenguaje natural y el análisis de texto, se lleva a cabo la extracción de opiniones y los sentimientos expresados. Actualmente este campo de estudio ha crecido bastante con la globalización del uso de internet, aunado con el auge de las redes sociales, foros, blogs, recomendaciones, reseñas, valoraciones, etc., donde los usuarios de internet generan una gran cantidad de información diaria que es beneficiosa para las empresas e investigadores. Un ejemplo de ello es el análisis de las opiniones de lo que expresan los usuarios sobre un producto, que se hace con el fin de conocer la reputación de un producto para mejorarlo o ver su estado de venta en el mercado.

A continuación, se muestran algunos ejemplos de corpus de frases en español que sirven para el análisis de sentimientos en texto.

#### **3.3.1. Corpus del Taller de Análisis Semántico en la SPLN (TASS)**

TASS es un corpus de textos (principalmente *tweets*) en español, etiquetados para tareas relacionadas con el análisis de sentimientos.

El corpus contiene más de 70,000 tweets, escritos en español por casi 200 personalidades y celebridades del mundo de la política, la economía, la comunicación, los medios de comunicación y la cultura, entre noviembre de 2011 y marzo de 2012. Si bien el contexto de extracción tiene un sesgo centrado en España, la diversa procedencia de los autores, incluyendo personas de España, México, Colombia, Puerto Rico, Estados Unidos y muchos otros países, hace que el corpus alcance una cobertura global en el mundo de habla hispana (Villena Román, 2013).

#### **3.3.2. Corpus del Sistema de Evaluación de Recursos Educativos (SERE)**

Este corpus se realizó a partir de la recopilación de opiniones en español de 53 estudiantes de ingeniería de sistemas computacionales, enfocadas en el área de la programación, que se obtuvieron del Sistema de Evaluación de Recursos Educativos (SERE).

SERE fue diseñado para interactuar con estudiantes a fin de permitirles expresar sus opiniones y comentarios libremente sobre recursos educativos u objetos de aprendizaje de las asignaturas de un curso (Barrón Estrada, Zatarain Cabada, Oramas Bustillos, & Ramírez Ávila, 2017).

Con SERE se logró recopilar 7,386 opiniones etiquetadas con emociones centradas en el aprendizaje, como frustrado (2,082), aburrido (1,361), neutral (945), emocionado (1,697) y comprometido (1,301).

### **3.3.3. Corpus SentiText**

SentiText es un corpus orientado a las polaridades de opiniones, clasificándolas en positivas o negativas. El corpus se construyó a partir de las extracciones de opiniones en *Twitter* con una delimitación geográfica al noroeste de México, con las siguientes palabras claves: programador, desarrollador, profesor, alumno, Java y Python. Estas opiniones fueron clasificadas por polaridad manualmente por un grupo de profesores de posgrado en ciencias de la computación. El tamaño inicial de este corpus fue de 8,412 opiniones, donde 4,164 fueron etiquetadas como positivo y 4,248 como negativo. El corpus generado a partir de *Twitter* era muy pequeño, por lo que se optó realizar una técnica de aumento de datos. Este método consistió en agregar nuevas palabras o frases a una determinada opinión, modificando algunas de ellas con sinónimos de pequeñas variaciones semánticas. Con la técnica de aumento de datos se logró incrementar a 13,010 frases, de las cuales 7,593 corresponden a etiquetas positivas y 5,417 contienen etiquetas negativas (Zatarain Cabada, Barrón Estrada, & Oramas Bustillos, 2018). Más adelante también se integraron las opiniones obtenidas a partir del Sistema de Evaluación de Recursos Educativos (SERE). El corpus SentiText actualmente cuenta con 24,543 frases etiquetadas con polaridad, donde 12,272 son positivas y 12,271 son negativas.

### **3.3.4. Corpus EduSERE**

Este corpus se realizó a partir de la recopilación de opiniones en español enfocadas en el aprendizaje del área de la programación, etiquetadas con emociones centradas en el

aprendizaje (aburrido, frustrado, emocionado y comprometido), que se obtuvieron de diferentes recursos digitales educativos como:

- **Sistema de evaluación de recursos educativos (SERE):** Este sistema realiza una interacción con estudiantes de ingeniería de sistemas computacionales, con el objetivo de permitir que expresen sus opiniones acerca de algunos recursos educativos y así generar un corpus de opiniones en español enfocadas al aprendizaje del tema de programación, etiquetadas con emociones centradas en el aprendizaje (Ramírez Ávila, 2018).
- **Desarrollo de proyectos prácticos:** Los estudiantes que participaron utilizando el sistema SERE, también participaron en la generación de opiniones en formato libre después de desarrollar la solución de un problema a través de un programa Java. A los estudiantes, se les solicitó que expresaran su opinión acerca de los aprendizajes logrados con el desarrollo de la solución del problema planteado. Las opiniones fueron capturadas y etiquetadas para ser incluidas en el corpus de opiniones (Ramírez Ávila, 2018).
- **Plataformas educativas digitales en línea:** Las opiniones expresadas por estudiantes en cursos en línea se recolectaron con un proceso manual visitando diferentes plataformas educativas entre las que se encuentran: Udemy, Coursera, y Teachlr (Ramírez Ávila, 2018).
- **Twitter:** A través de una librería de Python llamada *tweepy*, fue posible recolectar opiniones utilizando palabras clave contenidas en las frases publicadas. Estas frases también fueron etiquetadas con una emoción relacionada con el aprendizaje para incluirlas en el corpus (Ramírez Ávila, 2018).

Actualmente 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 3,245 frustrado.

### 3.3.5. Tabla Comparativa

En la tabla 3-1 se muestra la comparación de algunos de los corpus con frases en español que sirven para el análisis de sentimientos. En ella se puede observar el nombre de los corpus, las etiquetas que contienen y el tamaño total de sus frases.



Tabla 3-1 Comparación de corpus en español para el análisis de sentimiento en texto

Corpus	Etiquetas en frases	Tamaño de frases
<b>TASS</b>	Positivo, negativo, neutral, o ninguno.	70,000 frases
<b>SERE</b>	Emociones centradas en el aprendizaje (frustrado, aburrido, neutral, emocionado o comprometido).	7,386 frases
<b>SentiText</b>	Polaridad (positiva o negativa).	24,543 frases
<b>EduSERE</b>	Emociones centradas en el aprendizaje (frustrado, aburrido o comprometido).	12,083 frases

### 3.4. Algoritmo para la clasificación de sentimiento en texto

A continuación, se muestra un algoritmo que es utilizado para problemas de análisis y clasificación de sentimientos en texto.

#### 3.4.1. Enfoque evolutivo multilingüe para el análisis de sentimientos (EvoMSA)

EvoMSA es un clasificador para análisis de sentimientos, basado en B4MSA y EvoDAG. Es un algoritmo de generalización de pila especializado en problemas de clasificación de texto. Funciona combinando la salida de diferentes modelos de texto para producir la predicción final. EvoMSA es un procedimiento de dos etapas donde la primera etapa está compuesta por varios modelos que transforman un texto en valores de función de decisión; estos valores se combinan, en la segunda etapa, por un clasificador, en particular, EvoDAG que se basa en la programación genética (Graff, Miranda Jimenez, Tellez, & Moctezuma, 2020).

EvoMSA contiene diferentes modelos de texto que se pueden seleccionar usando banderas en el constructor de clases. Los modelos de texto implementados son (Graff, Miranda Jimenez, Tellez, & Moctezuma, 2020):

- `b4msa.textmodel.TextModel`, modelo entrenado con el conjunto de entrenamiento (está configurado por defecto TR)
- Espacio emoji (se evoca usando `EvoMSA.base.EvoMSA (Emo = True, lang = "en")`)
- Modelo basado en el léxico de sentimientos (se evoca utilizando `EvoMSA.base.EvoMSA (TH = True, lang = "en")`)

- Modelo humano anotado (se evoca utilizando `EvoMSA.base.EvoMSA (HA = True, lang = "en")`)
- Modelo de agresividad (se evoca utilizando `EvoMSA.base.EvoMSA (Aggress = True, lang = "en")`)

En la figura 3-2 se muestra el esquema de predicción de EvoMSA, donde un texto es transformado por diferentes modelos componiendo un espacio vectorial, que es la entrada de EvoDAG, para hacer la predicción final (Graff, Miranda Jimenez, Tellez, & Moctezuma, 2020).

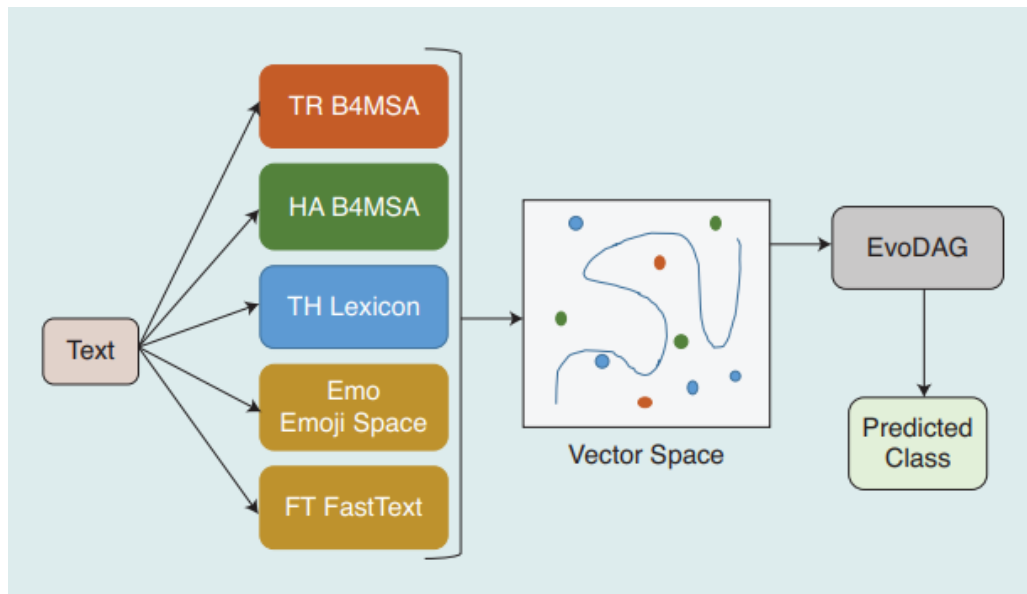


Figura 3-2 Esquema de predicción de EvoMSA (Graff, Miranda Jimenez, Tellez, & Moctezuma, 2020)

# Capítulo 4

## 4. Desarrollo del proyecto

En este capítulo se presenta la metodología diseñada e implementada para la generación de los corpus validados por expertos. Además, se incluye la descripción detallada del desarrollo del sistema de software para el etiquetado de frases, así como el proceso de distribución del sistema entre expertos y el procedimiento de cómo se generaron los nuevos corpus validados.

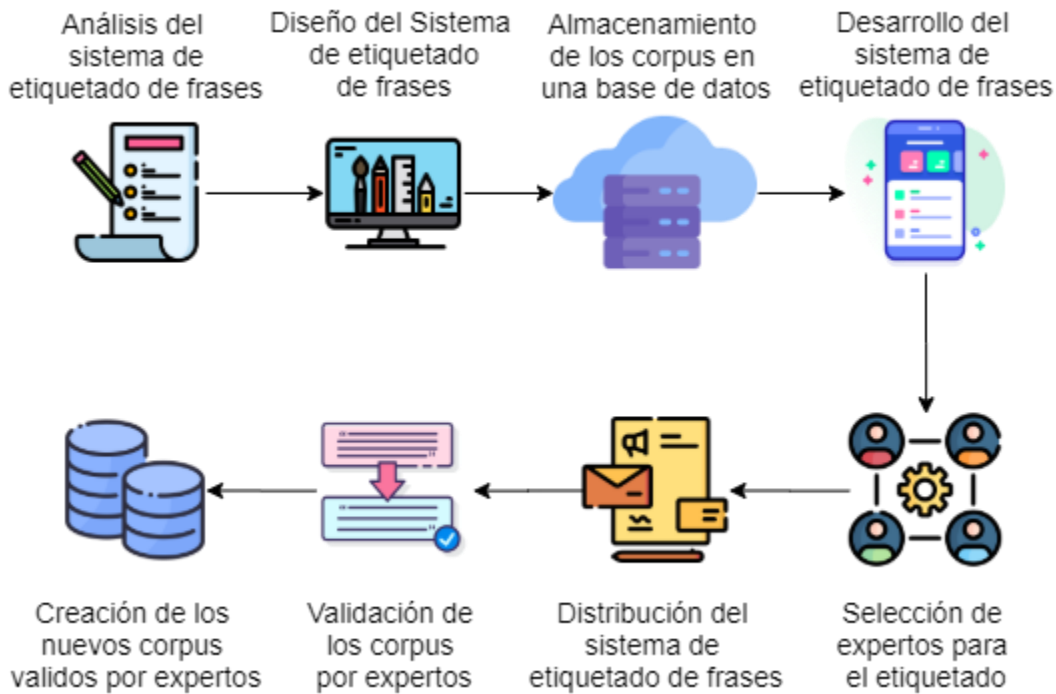
### 4.1. Metodología para la creación de los corpus validados por expertos.

El corpus SentiText contiene 24,543 frases etiquetadas con polaridad (positiva y negativa) y el corpus EduSERE contiene 12,083 frases etiquetadas con emociones centradas en el aprendizaje (aburrido, frustrado o comprometido). Para ambos corpus, las frases fueron etiquetadas manualmente por estudiantes de posgrado los cuales no tenían experiencia en docencia ni antecedentes académicos de psicología que permitiera validar la etiqueta asignada a cada frase.

El etiquetado manual es un proceso lento y muy susceptible a errores debido que la tarea de etiquetado se debe realizar leyendo cada una de las frases y asignando la etiqueta que a juicio de la persona sea la que mejor corresponde a la frase.

Debido a esto, en esta investigación se propuso la creación de un sistema de validación que permita a expertos asignar etiquetas a cada frase, a través de una aplicación móvil que administre las frases y almacene dinámicamente las etiquetas en un ambiente progresivo y motivado por la generación de recompensas por la participación. Las frases podrán recibir varias etiquetas de expertos que posiblemente no coincidan y el sistema es capaz de ponderar cada una de las etiquetas asignadas y determinar la etiqueta final para cada frase.

La metodología general de validación de etiquetas de frases en los corpus contiene 8 pasos y se presenta en la figura 4-1.



*Figura 4-1 Metodología para la creación de los corpus validados por expertos*

Los primeros cuatro pasos, corresponden al desarrollo del sistema de etiquetado mientras que los últimos cuatro pasos corresponden al proceso para generar el corpus validado por expertos usando el sistema de etiquetado. En las secciones siguientes, se detalla cada uno de los pasos de la figura 4-1.

## **4.2. Análisis del sistema de etiquetado de frases**

En esta etapa se presenta el análisis detallado del sistema de etiquetado de frases, donde se describe el alcance del proyecto, los requisitos funcionales y de calidad, a partir de ellos se obtuvieron las restricciones del sistema, así como los actores principales involucrados y los procesos principales del sistema que se representan mediante casos de uso.

El sistema de etiquetado de frases se tituló *Label*, debido a que este nombre se asocia la intención del sistema.

#### **4.2.1. Alcance**

Desarrollar un sistema con una interfaz intuitiva que permita que expertos en el área de psicología, pedagogía o computación puedan fácilmente etiquetar las frases de los corpus SentiText y EduSERE con polaridad o emociones centradas en el aprendizaje, y a partir de estas frases etiquetadas y validadas se genere un nuevo corpus.

#### **4.2.2. Requerimientos de usuario**

Con base en la información obtenida del problema y la propuesta de solución, se generaron una serie de requerimientos de usuario que fueron clasificados en funcionales y de calidad, los cuales se describen en las siguientes secciones.

##### **4.2.2.1. Requisitos funcionales**

En esta sección se presentan los requisitos funcionales del sistema de etiquetado, ordenados de acuerdo con la prioridad en el sistema. Se utilizó una escala de tres valores lingüísticos: ALTA, MEDIA, y BAJA, para denotar la importancia de cada requerimiento en el sistema. La tabla 4-1 muestra 13 requisitos funcionales con una descripción y su prioridad.

Tabla 4-1 Requisitos funcionales del sistema de etiquetado

No.	Requisito	Prioridad
<b>RF-01</b>	El sistema registrará usuarios solicitando su información personal como: correo, contraseña, ocupación, edad, nombre completo	<b>ALTA</b>
<b>RF-02</b>	El sistema deberá pedir correo y contraseña para ingresar	<b>ALTA</b>
<b>RF-03</b>	El sistema permitirá acceder mediante una cuenta Gmail de Google	<b>ALTA</b>
<b>RF-04</b>	El sistema mostrará niveles de etiquetado de frases para cada usuario	<b>ALTA</b>
<b>RF-05</b>	El sistema mostrará las frases de los corpus SentiText o EduSERE al seleccionar un nivel	<b>ALTA</b>
<b>RF-06</b>	El sistema permitirá etiquetar frases por polaridad o emociones centradas en el aprendizaje	<b>ALTA</b>
<b>RF-07</b>	El sistema guardará las frases etiquetadas en una base de datos	<b>ALTA</b>
<b>RF-08</b>	El sistema mostrará un tutorial de uso al iniciar sesión	<b>MEDIA</b>
<b>RF-09</b>	El sistema mostrará las frases a etiquetar aleatoriamente	<b>MEDIA</b>
<b>RF-10</b>	El sistema guardará puntajes por cada nivel de frases	<b>MEDIA</b>
<b>RF-11</b>	El sistema tendrá una tabla para los diez primeros usuarios con más puntajes	<b>MEDIA</b>
<b>RF-12</b>	El sistema tendrá una serie de trofeos para mostrar a los usuarios cuando realicen un logro	<b>MEDIA</b>
<b>RF-13</b>	El sistema permitirá que el usuario pueda ver de nuevo el tutorial de uso	<b>BAJA</b>

#### 4.2.2.2. Requisitos de calidad

Los requisitos de calidad fueron clasificados con los atributos de calidad de portabilidad, rendimiento, disponibilidad, seguridad y usabilidad. La tabla 4-2 muestra los requisitos de calidad del sistema de etiquetado los cuales están ordenados por prioridad en escala alta, media y baja.

Tabla 4-2 Requisitos de calidad del sistema de etiquetado

No.	Atributo	Requisito	Prioridad
<b>RC-01</b>	Portabilidad	El sistema debe funcionar en sistemas operativos Android desde la versión 4.4 y versiones posteriores	<b>ALTA</b>
<b>RC-02</b>	Rendimiento	El sistema deberá responder por cada petición en máximo 4 segundos	<b>ALTA</b>
<b>RC-03</b>	Disponibilidad	El sistema debe estar disponible a cualquier hora del día	<b>ALTA</b>
<b>RC-04</b>	Rendimiento	La carga de niveles, frases, tabla de puntuaciones debe tener una espera máxima de 5 segundos	<b>ALTA</b>
<b>RC-05</b>	Seguridad	El sistema debe encriptar la contraseña del usuario	<b>ALTA</b>
<b>RC-06</b>	Usabilidad	El sistema debe tener una interfaz intuitiva y fácil de utilizar	<b>ALTA</b>
<b>RC-07</b>	Usabilidad	El sistema cuenta con una interfaz gráfica amigable	<b>MEDIA</b>
<b>RC-08</b>	Usabilidad	El sistema permitirá etiquetar la frase en menos de 3 clics	<b>BAJA</b>

#### 4.2.2.3. Restricciones

Durante el análisis que se llevó a cabo en los requerimientos se obtuvieron algunas restricciones en el sistema, tales como:

- El sistema funcionara solo en teléfonos móviles con sistema operativo Android.
- El usuario debe contar con una conexión a internet en el teléfono móvil para poder etiquetar las frases.
- El usuario debe registrarse e ingresar al sistema con usuario y contraseña.

#### 4.2.3. Actores

Los actores son todas las entidades externas que tienen alguna interacción con el sistema, pueden ser personas (usuarios), algún otro sistema o hardware que demande alguna funcionalidad. En la tabla 4-3 se muestran los actores principales y una breve descripción de su relación con el sistema.

Tabla 4-3 Actores en sistema de etiquetado de frases

Actor	Descripción
<b>Psicólogo</b>	Actor principal que cuenta con suficiente experiencia en el área de psicología el cual asignará y validará etiquetas de polaridad o emociones centradas en el aprendizaje a las frases almacenadas en los corpus.
<b>Profesor</b>	Actor principal que cuenta con experiencia en pedagogía y el cual asignará y validará etiquetas de polaridad o emociones centradas en el aprendizaje a las frases almacenadas en los corpus.
<b>Estudiante</b>	Actor principal que cuenta con la mínima experiencia requerida en psicología o computación para ayudar a etiquetar las frases con polaridad o emociones centradas en el aprendizaje.

#### 4.2.4. Casos de uso

En esta sección se muestran los principales casos de uso que describen las interacciones de los usuarios con las funcionalidades del sistema para llevar a cabo un proceso, estos que se elaboraron a partir de algunos requisitos y restricciones del sistema. La figura 4-2 muestra las principales interacciones que realiza un usuario con el sistema.

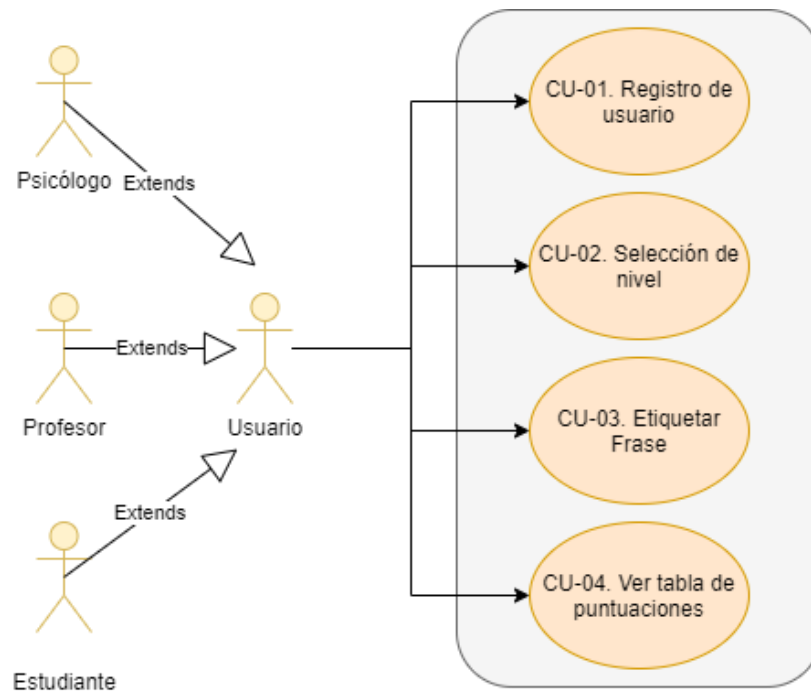


Figura 4-2 Diagrama de casos de uso del sistema de etiquetado



Las tablas 4-4 al 4-7, explican detalladamente los casos de uso principales del sistema de etiquetado. El formato utilizado para documentar los casos de uso contiene, además del título, seis secciones para definir la descripción del caso de uso, los actores involucrados, las precondiciones necesarias para la ejecución, el escenario principal del flujo, las postcondiciones y las referencias utilizadas.

La tabla 4-4 muestra el flujo principal diseñado para el caso de uso 01 que describe el proceso de registro para los usuarios del sistema. Este caso de uso se relaciona con el requisito funcional 01.

Tabla 4-4 CU-01. Registro de usuarios

Caso de uso	
CU-01. Registro de usuario	
<b>Descripción</b>	Proceso en el cual un usuario se registra como psicólogo, profesor o estudiante en el sistema <i>Label</i> .
<b>Actores</b>	Psicólogo, profesor, estudiante.
<b>Precondiciones</b>	<ul style="list-style-type: none"> <li>• Contar con conexión a internet</li> <li>• Contar con un correo electrónico</li> <li>• El correo no pertenece a ningún usuario registrado</li> <li>• La contraseña debe tener mínimo 6 caracteres</li> </ul>
<b>Escenario principal</b>	<ol style="list-style-type: none"> <li>1. El sistema muestra la pantalla de inicio de sesión</li> <li>2. El usuario selecciona la opción para registrar una nueva cuenta</li> <li>3. El usuario ingresa sus datos: correo, contraseña, ocupación, edad y nombre completo</li> <li>4. El usuario presiona el botón de registrar</li> <li>5. El sistema valida campos requeridos</li> <li>6. El sistema valida si el correo que intenta registrar ya existe</li> <li>7. El sistema muestra un mensaje de confirmación al guardar</li> </ol>
<b>Postcondiciones</b>	<ul style="list-style-type: none"> <li>• El sistema registra al nuevo usuario en la base de datos</li> <li>• El sistema redirecciona a la pantalla de inicio de sesión</li> </ul>
<b>Referencias</b>	<ul style="list-style-type: none"> <li>• El sistema registrará usuarios solicitando su información personal como: correo, contraseña, ocupación, edad, nombre completo (<b>RF-01</b>)</li> </ul>

La tabla 4-5 presenta el caso de uso para la selección de nivel, donde el usuario debe seleccionar el nivel actual donde se encuentra para continuar con el etiquetado de frases. Este caso de uso se relaciona con el requisito funcional 04.

Tabla 4-5 CU-02. Selección de nivel

Caso de uso CU-02. Selección de nivel	
<b>Descripción</b>	Proceso en el cual se muestra un listado de niveles para etiquetar frases y el usuario pueda dar clic al nivel en el que se encuentra.
<b>Actores</b>	Psicólogo, profesor, estudiante.
<b>Precondiciones</b>	<ul style="list-style-type: none"> <li>• El usuario debe estar registrado en el sistema</li> <li>• El usuario inicia sesión en el sistema</li> <li>• El usuario ingresa como Psicólogo, profesor o estudiante</li> </ul>
<b>Escenario principal</b>	<ol style="list-style-type: none"> <li>1. El usuario ingresa a la pantalla de niveles</li> <li>2. El sistema muestra en diferentes colores los niveles terminados y en color gris los niveles pendientes</li> <li>3. El usuario hace clic en el nivel que se encuentra</li> </ol>
<b>Postcondiciones</b>	<ul style="list-style-type: none"> <li>• El sistema redirecciona a la pantalla de frases</li> </ul>
<b>Referencias</b>	<ul style="list-style-type: none"> <li>• El sistema mostrará niveles de etiquetado de frases para cada usuario (<b>RF-04</b>)</li> </ul>

La tabla 4-6 muestra el caso de uso para etiquetar frases, el cual permite al usuario realizar el etiquetado de las frases con polaridad o emociones centradas en el aprendizaje. Este caso de uso es el principal y está relacionado con los requisitos funcionales 05, 06, 07, 09, 10 y 12.

Tabla 4-6 CU-03. Etiquetar frase

Caso de uso	
CU-03. Etiquetar frase	
<b>Descripción</b>	Proceso en el cual se muestra un listado de frases con sus respectivas etiquetas de polaridad o emociones centradas en el aprendizaje, donde el usuario puede seleccionar la etiqueta que el crea que es la correcta para la frase.
<b>Actores</b>	Psicólogo, profesor, estudiante.
<b>Precondiciones</b>	<ul style="list-style-type: none"> <li>• El usuario debe estar registrado en el sistema</li> <li>• El usuario inicia sesión en el sistema</li> <li>• El usuario ingresa como Psicólogo, profesor o estudiante</li> <li>• El usuario hizo clic en un nivel</li> </ul>
<b>Escenario principal</b>	<ol style="list-style-type: none"> <li>1. El usuario se encuentra en la pantalla de etiquetado</li> <li>2. Se muestran las frases a etiquetar con sus respectivas etiquetas, dependiendo del nivel con polaridad o con emociones centradas en el aprendizaje</li> <li>3. El usuario selecciona la etiqueta que crea correcta para cada una de las frases del listado</li> <li>4. El usuario hace clic en enviar las frases</li> <li>5. El sistema valida que todas las frases hayan sido etiquetadas</li> <li>6. El sistema muestra un mensaje de confirmación de guardado</li> </ol>
<b>Postcondiciones</b>	<ul style="list-style-type: none"> <li>• El sistema registra en la base de datos un puntaje por el etiquetado de las frases</li> <li>• El sistema registra en la base de datos las frases etiquetadas por el usuario</li> <li>• El sistema muestra un trofeo al realizar un logro</li> <li>• El sistema redirecciona a la pantalla de niveles</li> </ul>
<b>Referencias</b>	<ul style="list-style-type: none"> <li>• El sistema mostrara las frases de los corpus SentiText y EduSERE al seleccionar un nivel (<b>RF-05</b>)</li> <li>• El sistema permitirá etiquetar frases por polaridad o emociones centradas en el aprendizaje (<b>RF-06</b>)</li> <li>• El sistema guardara las frases etiquetadas en una base de datos (<b>RF-07</b>)</li> <li>• El sistema mostrara las frases a etiquetar aleatoriamente (<b>RF-09</b>)</li> <li>• El sistema guardara puntajes por cada nivel de frases (<b>RF-10</b>)</li> <li>• El sistema tendrá una serie de trofeos para mostrar a los usuarios cuando realicen un logro (<b>RF-12</b>)</li> </ul>

La tabla 4-7, presenta el caso de uso 04 que permite visualizar la tabla de puntuaciones de los participantes en el proceso de etiquetado. Este caso de uso se relaciona con el requisito funcional 11.

Tabla 4-7 CU-04. Ver tabla de puntuaciones

Caso de uso	CU-04. Ver tabla de puntuaciones
<b>Descripción</b>	Proceso en el cual un usuario puede ver el listado de las primeras diez personas con más puntajes en la aplicación, así como su puntaje actual.
<b>Actores</b>	Psicólogo, profesor, estudiante.
<b>Precondiciones</b>	<ul style="list-style-type: none"> <li>• El usuario debe estar registrado en el sistema</li> <li>• El usuario inicia sesión en el sistema</li> <li>• El usuario ingresa como Psicólogo, profesor o estudiante</li> </ul>
<b>Escenario principal</b>	<ol style="list-style-type: none"> <li>1. El usuario ingresa a la pantalla de puntajes</li> <li>2. El sistema muestra una tabla de los diez usuarios con mayor puntaje etiquetando</li> <li>3. El sistema muestra el puntaje actual del usuario si es que no se encuentra en la tabla de puntajes</li> </ol>
<b>Postcondiciones</b>	<ul style="list-style-type: none"> <li>• El sistema actualiza automáticamente la tabla de puntajes cada vez que un usuario llega al puntaje de los diez primeros</li> </ul>
<b>Referencias</b>	<ul style="list-style-type: none"> <li>• El sistema tendrá una tabla para los diez primeros usuarios con mayor puntaje (<b>RF-11</b>)</li> </ul>

### 4.3. Diseño del sistema de etiquetado de frases

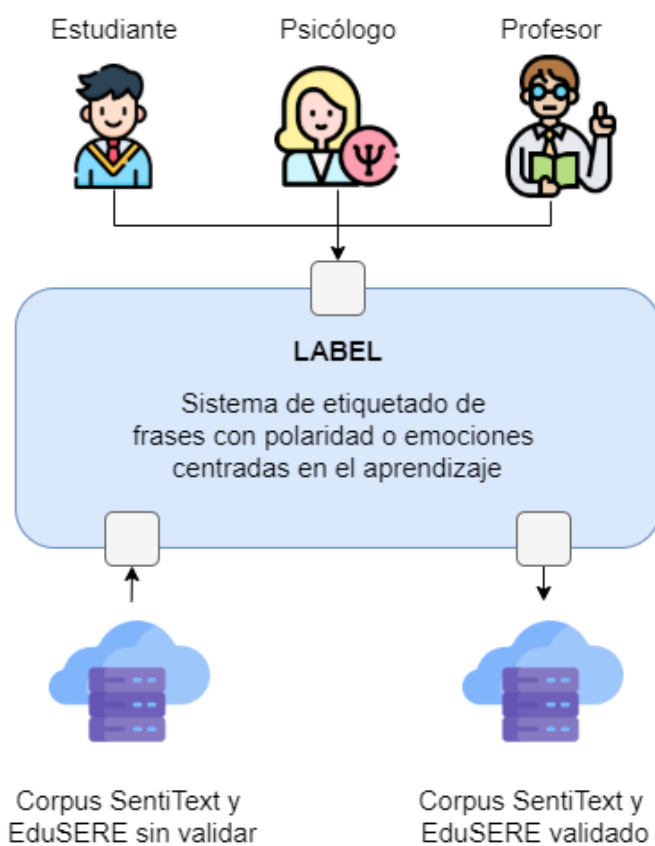
Una vez teniendo el análisis de los procesos principales y los actores involucrados, se continuó con el diseño del sistema de etiquetado de frases donde se describe con detalle cada componente esencial, así como la interacción entre ellos. A partir del diseño se determinó la mejor arquitectura y las tecnologías a utilizar para llevar a cabo el desarrollo del sistema.

En esta etapa se explica el diseño de los principales componentes, sus interacciones y la arquitectura que conforma el sistema de etiquetado de frases.

#### 4.3.1. Diagrama de contexto

En esta sección se muestra los límites entre el sistema y las entidades que interactúan con él. La figura 4-3 muestra el diagrama de contexto donde se observan los actores principales (usuarios) interactuar con el sistema de etiquetado de frases con polaridades o emociones

centradas en el aprendizaje, con el fin de asignar a cada frase una etiqueta representando polaridad (positiva o negativa) para las frases del corpus SentiText o una emoción centrada en el aprendizaje (aburrido, frustrado o comprometido) para las frases del corpus EduSERE. El sistema obtiene las frases a etiquetar de una base de datos que contiene ambos corpus SentiText y EduSERE; además, almacena en otra base de datos todas las frases etiquetadas por los expertos.



*Figura 4-3 Diagrama de contexto del sistema de etiquetado de frases*

#### 4.3.2. Arquetipos y sus relaciones

Los arquetipos son las entidades abstractas importantes y principales identificadas dentro del sistema de etiquetado las cuales indican cómo se comportará. en la tabla 4-8 se muestran los arquetipos existentes y una descripción.

Tabla 4-8 Arquetipos del sistema de etiquetado de frases

Arquetipos	Descripción
<b>Corpus</b>	Es la representación de los corpus EduSERE y SentiText que se utilizan para el sistema de etiquetado
<b>Frase</b>	Es la representación de las frases que serán mostradas a los usuarios para su etiquetado
<b>Etiqueta</b>	Representa la polaridad (negativo o positivo) y las emociones centradas en el aprendizaje (aburrido, frustrado o comprometido) con la que la frase puede ser etiquetada
<b>Usuario</b>	Es la representación de la persona que se encarga de etiquetar las frases con polaridad o emociones centradas en el aprendizaje
<b>Destino de datos</b>	Es la representación de la base de datos que almacenara las frases etiquetadas y validadas por los usuarios

En la figura 4-4 se muestran las relaciones que tienen los arquetipos dentro del sistema de etiquetado.

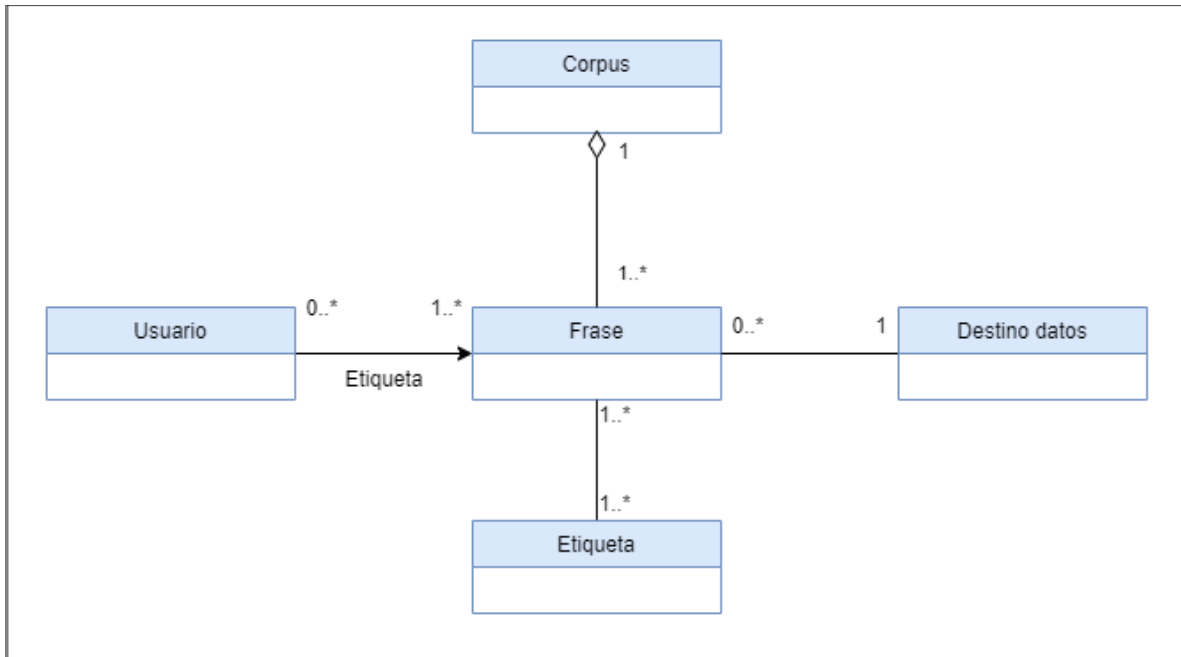


Figura 4-4 Arquetipos del sistema de etiquetado de frases

### 4.3.3. Vista lógica

Para el modelo arquitectónico del sistema se seleccionó la arquitectura MVC equilibrada para desarrollar aplicaciones móviles basadas en servicios (La, & Kim, 2010), adoptando 3 principios arquitectónicos, ser un cliente ligero, realizar el desarrollo en capas con el modelo-vista-controlador (MVC) y equilibrar las interacciones entre el cliente y el servidor. Esta arquitectura permite desarrollar una aplicación ligera y con menos consumo de recursos del dispositivo móvil, ya que se reparten las responsabilidades equilibradamente con el servidor; utilizando MVC es posible realizar una aplicación modular, escalable y con responsabilidades únicas, separando la manipulación de los datos, la lógica de negocios y la interacción del usuario con la interfaz gráfica de la aplicación.

En la figura 4-5 se muestra la vista lógica del sistema de etiquetado de frases la cual contiene 3 capas:

1. **Capa de presentación:** Incluye toda la interfaz gráfica con la que interactúa el usuario. El inicio de sesión, el registro de usuarios, los niveles, el etiquetado de frases, la tabla de puntajes, los trofeos y los tutoriales.
2. **Capa de lógica de negocios:** En esta capa se lleva a cabo el intercambio de información entre la capa de presentación y la capa de datos. Aquí se encuentran los controladores que se encargan de administrar la lógica entre lo que se muestra en la interfaz gráfica, lo que se consulta en los servicios externos y lo que guarda en la base de datos. Estos controladores incluyen la administración de los niveles, el tutorial, el etiquetado de frases, los puntajes, los trofeos y la gestión de usuarios.
3. **Capa de datos:** Incluye los modelos donde se almacena la información del sistema. También se incluye la interacción del sistema con los servicios externos. Para los servicios externos se integró el uso de la plataforma Firebase de Google, de la cual se utilizará Cloud Firestore para el almacenamiento de datos en la nube y Authentication para administrar la seguridad del inicio de sesión del sistema.

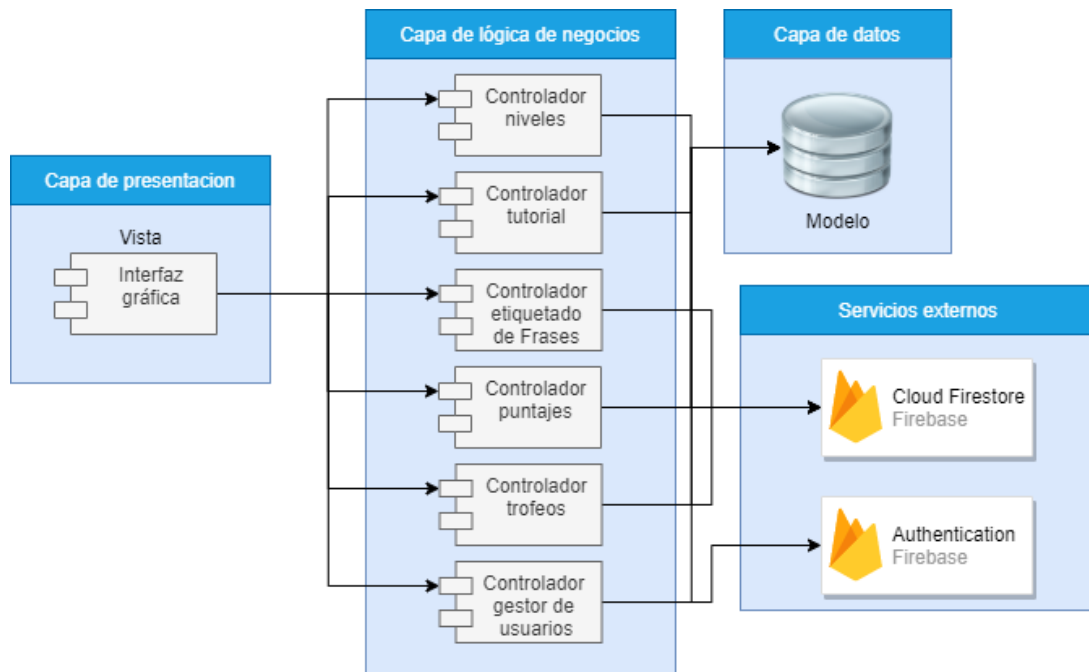
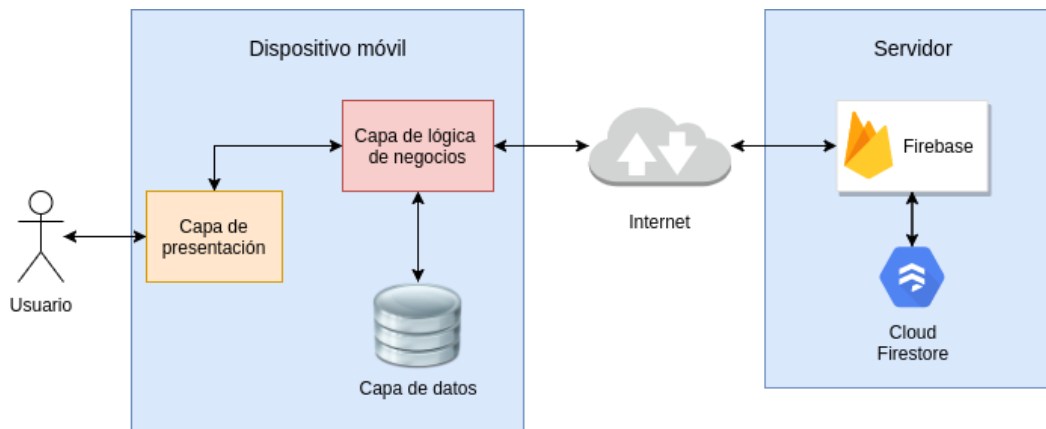


Figura 4-5 Vista lógica del sistema de etiquetado de frases

#### 4.3.4. Vista de desarrollo

La figura 4-6 muestra la arquitectura del implementada para el sistema de etiquetado de frases. La arquitectura se divide en dos secciones importantes, el dispositivo móvil y el servidor, unidas mediante la conexión a internet. La sección del dispositivo móvil se divide en 3 capas, la capa de presentación donde se encuentran todas las interfaces gráficas que se muestran al usuario, la capa de lógica de negocios, que administra el flujo de datos hacia la capa de presentación y las peticiones al servidor, y por último la capa de datos, que guarda en cache la información que se mostrará al usuario y los datos que modifica. En la sección del servidor se muestran los servicios alojados en la nube para la administración de la autenticación de usuarios y el almacenamiento de los corpus no validados y los que serán validados.





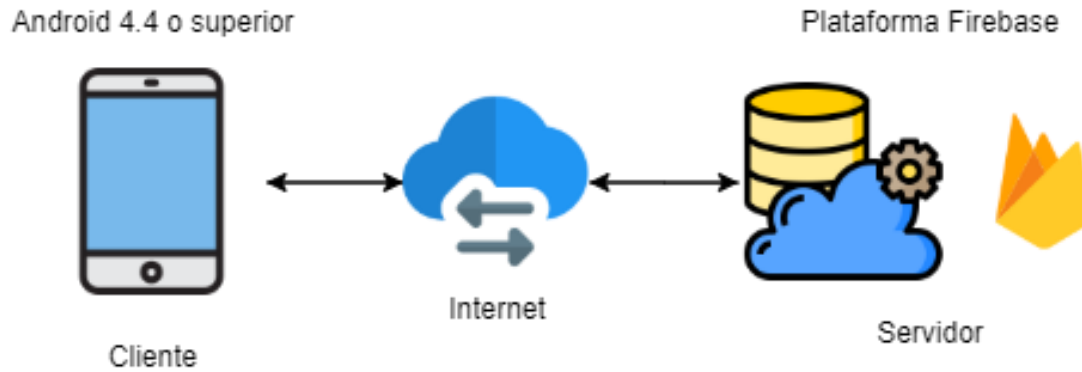
*Figura 4-6 Vista de desarrollo del sistema de etiquetado de frases*

#### 4.3.5. Vista física

A continuación, se explican los diferentes componentes de software de la capa física, así como las interacciones entre ellos. La vista física considera los requisitos de calidad de portabilidad, rendimiento, disponibilidad, seguridad y usabilidad. El sistema de etiquetado se desarrollará para dispositivos móviles con sistema operativo Android que tengan la versión 4.4 o superior, ya que solo menos del 4% de usuarios Android tienen una versión anterior a la 4.4. La aplicación se conectará mediante una conexión de internet a los servidores que administran la plataforma Firebase de Google.

Firebase es una plataforma que ayuda al desarrollo de aplicaciones móviles ofreciendo servicios gratuitos de almacenamiento, autenticación, y distribución entre muchos otros. La aplicación utilizará Cloud Firestore para almacenar los corpus no validados y validados, así como toda la información que generen los usuarios, también se utilizará Firebase Authentication que permite identificar y administrar a todos los usuarios de la aplicación de manera segura.

La figura 4-7 muestra los componentes físicos explicados en esta sección, que se utilizan para llevar a cabo la arquitectura MVC equilibrada para desarrollar aplicaciones móviles basadas en servicios.



*Figura 4-7 Vista física del sistema de etiquetado de frases*

#### **4.4. Almacenamiento de los corpus en una base de datos**

En esta etapa se realizó el guardado de los corpus SentiText y EduSERE en una base de datos alojada en la nube. Estos corpus inicialmente están guardados en un archivo con formato CSV para facilitar el acceso a los datos cuando se entrenan clasificadores de emociones en texto. Para la aplicación se requiere que el corpus se encuentre en un servidor de base de datos en la nube, para conceder acceso a los datos desde cualquier teléfono móvil que tenga la aplicación de etiquetado de frases instalada. Por este motivo, se seleccionó el servicio Cloud Firestore que es una base de datos NoSQL alojada en la nube a la cual pueden acceder las aplicaciones móviles fácilmente; este servicio provee 50,000 lecturas y 20,000 escrituras diarias gratis a la base de datos, por lo cual es un servicio que se adecua a las necesidades del proyecto. En la base de datos se crearon 2 colecciones de datos, una para las frases del corpus SentiText y otra para el corpus EduSERE y para guardar los corpus se creó un script en Java que pudiera leer los corpus del formato CSV e insertara los datos en sus respectivas colecciones de Cloud Firestore.

## **4.5. Desarrollo del sistema de etiquetado de frases**

En esta etapa se describe la metodología de desarrollo utilizada para realizar el sistema de etiquetado de frases, así como la descripción de las interfaces gráficas principales y su funcionamiento.

### **4.5.1. Metodología de desarrollo**

Una vez terminado el análisis y el diseño del sistema de etiquetado se llegó a la conclusión de desarrollar una aplicación móvil llamada *Label*, para sistemas operativos Android con versión 4.4 o superior, la cual se desarrolló con base en los requerimientos funcionales y de calidad, considerando los componentes principales y el diseño de la arquitectura. El desarrollo se llevó a cabo mediante el marco de trabajo Scrum (McKenna, 2016), siguiendo un proceso iterativo de desarrollo y entregas ágil, realizando pequeñas versiones funcionales e incrementales que fueran validadas y probadas antes de lanzar la versión final.

### **4.5.2. Interfaces de la aplicación Label**

En esta sección se describen las interfaces gráficas principales desarrolladas con las que interactúan los usuarios de la aplicación de etiquetado de frases Label.

En la figura 4-8 se muestra el inicio de sesión de la aplicación, donde se puede ingresar mediante correo y contraseña o utilizando el servicio Gmail de Google, también presenta un enlace para poder registrarse como nuevo usuario en caso de no contar con uno.



The image shows a login interface for an application named "LabelApp". At the top, there is a logo consisting of a blue tag icon with a white dot and the text "LabelApp" in a blue sans-serif font. Below the logo, there are two input fields: the first is labeled "Correo electrónico" and the second is labeled "Contraseña". The password field has a small eye icon on the right side to toggle visibility. Below these fields, there is a link that says "¿No tienes una cuenta? Regístrate." in a small, gray font. Underneath the link is a large, light blue button with the text "INGRESAR" in a bold, blue, sans-serif font. Below the button is a horizontal separator line with a small circle in the center. At the bottom, there is a button with the Google logo on the left and the text "Acceder" on the right, both in a gray, sans-serif font.

*Figura 4-8 Interfaz gráfica de inicio de sesión*

En la figura 4-9 se muestra la interfaz gráfica para registrarse como nuevo usuario de la aplicación. Solo se muestran dos campos a llenar, que son el correo electrónico y una contraseña, con el fin de agilizar el proceso de registro.

The image shows a mobile application interface for user registration. At the top, the title 'Nuevo usuario' is centered. Below it, there are two input fields: the first is labeled 'Correo electrónico' and the second is labeled 'Contraseña'. The password field has a small eye icon to its right, indicating a toggle for visibility. Below these fields is a button labeled 'REGISTRARSE' in all caps. The entire form is enclosed in a light gray border.

*Figura 4-9 Interfaz gráfica de registro de usuario*

En la figura 4-10 se muestra la interfaz gráfica de información, donde se captura datos personales del usuario como nombre, edad y ocupación, para llevar un control estadístico de los usuarios que acceden a la aplicación. Esta pantalla solo aparecerá la primera vez que inicien sesión en la aplicación y registren su información.



Formulario de información personal con los siguientes campos:

- Título: Información
- Campo de texto: Nombre completo
- Campo de texto: Edad
- Sección Ocupación:
  - ☒ Maestro
  - ☐ Alumno
  - ☐ Psicólogo
  - ☐ Otro
- Botón: GUARDAR

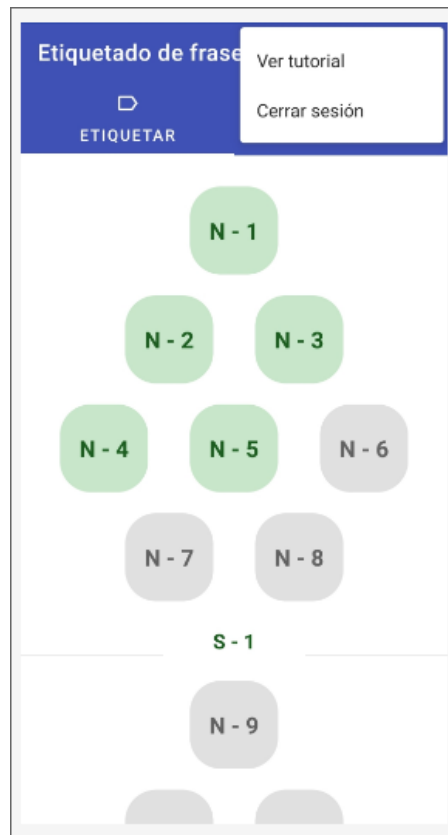
*Figura 4-10 Interfaz gráfica de agregar información personal*

A continuación, en la figura 4-11 se muestra la interfaz gráfica del tutorial de la aplicación. Este tutorial explica con un conjunto de imágenes y texto las secciones de la aplicación y como utilizarlas, con el fin de ayudar al usuario a familiarizarse con los procesos de la aplicación. La navegación es sencilla a través de un botón que despliega la siguiente explicación hasta finalizar.



Figura 4-11 Interfaz gráfica de tutorial

En la figura 4-12 se muestra la sección de niveles, que es una de las pestañas principales de la aplicación y aparece una vez completado el tutorial. Esta sección presenta los niveles de etiquetado que ha realizado el usuario y se implementó usando una técnica de gamificación con la finalidad de que los usuarios se motiven a continuar y finalizar los niveles para que logren un mayor etiquetado de frases. Los niveles fueron definidos agrupando un conjunto de frases y se completan cuando el usuario etiqueta todas las frases contenidas en el nivel. Esto provoca que se cambie el color del nivel y permite que el usuario conozca el número de niveles terminado y los que le faltan aún por terminar.



*Figura 4-12 Interfaz gráfica que muestra los niveles de la aplicación*

En la figura 4-13 se muestra la tabla de puntajes que presenta a los diez usuarios en han logrado obtener mayor puntaje etiquetando frases. Esta tabla de puntajes se actualiza automáticamente cuando los usuarios incrementan sus puntajes al etiquetar frases. Los niveles y la tabla de puntajes se incluyeron como parte de las técnicas de gamificación, cuya finalidad es motivar a los usuarios a obtener el mayor puntaje para que puedan etiquetar un gran número de frases.



Etiquetado de frases		
ETIQUETAR		PUNTAJES
1	Ulises Uriel Cabrera Rubio Alumno	3558
2	Flor Jazmin Arcos Morales Psicólogo	3540
3	Sarahí Armenta Medina Alumno	3520
4	Karen Paloma Munguía Maldona... Maestro	3520
5	Adahedit Ivonne Arias Bracamón... Psicólogo	3520
6	Álvarez Hernandez Gloria Giselle Alumno	3520
7	Claudia Elena Corrales Garcia Psicólogo	3520
8	Celina Díaz Peñuñuri Psicólogo	3520
9	Damian Jasso Antuna	2720
-	Oscar Omar Sánchez Alvarado Alumno	125

Figura 4-13 Interfaz gráfica de tabla de puntajes

En la figura 4-14, se muestran las interfaces gráficas del etiquetado de frases, esta sección se muestra en la aplicación cuando el usuario hace clic en un nivel. Dependiendo del nivel, la interfaz mostrará un listado de frases con diferentes etiquetas a seleccionar, con polaridad o emociones centradas en el aprendizaje. El usuario podrá elegir la etiqueta que crea que es la correcta y el botón finalizar le permite almacenar las etiquetas seleccionadas para cada frase actualizando la información en la base de datos.

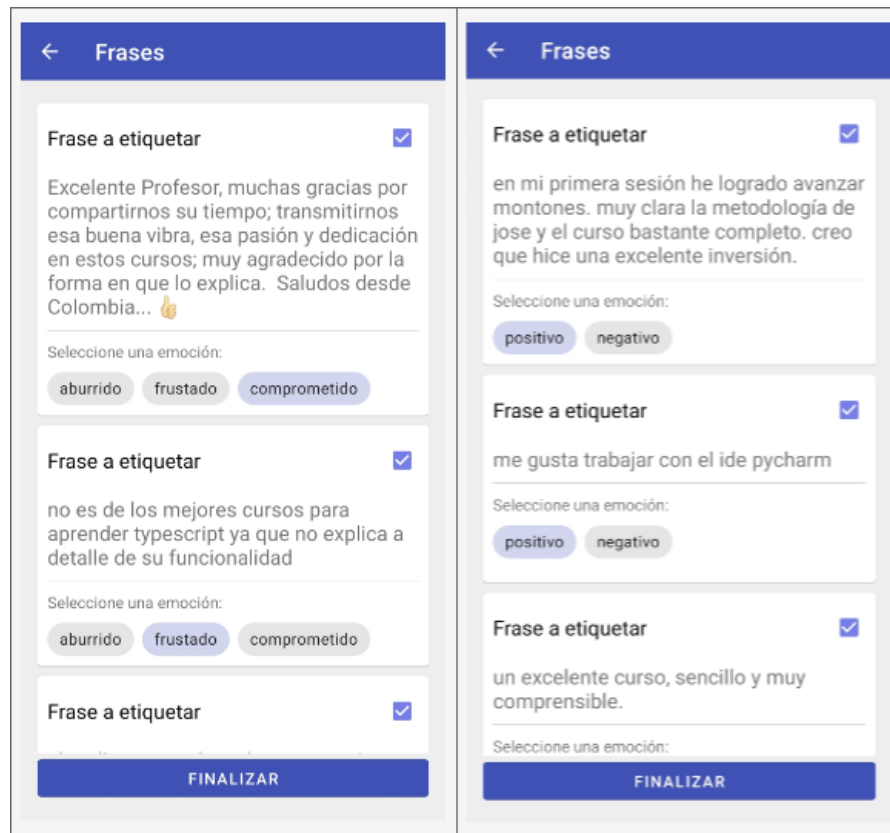
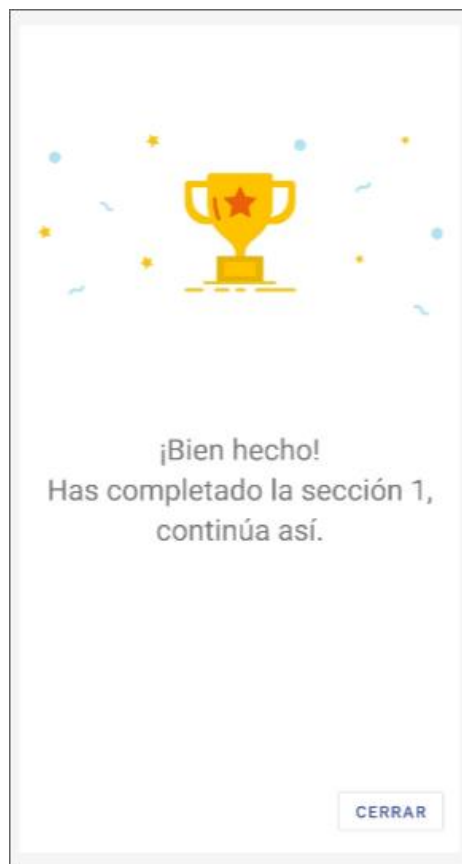


Figura 4-14 Interfaz gráfica de etiquetado de frases

La interfaz gráfica de la figura 4-15 muestra trofeos a los usuarios en diferentes etapas del etiquetado o cuando hayan realizado un logro, esta interfaz también se incluyó como técnica de gamificación para motivar a los usuarios con diferentes mensajes que impulsan a terminar los niveles de etiquetado.



*Figura 4-15 Interfaz gráfica de trofeos*

#### **4.6. Selección de expertos para el etiquetado**

En esta etapa se generó un catálogo de personas con antecedentes curriculares en el área de humanidades y pedagogía e interesadas en el tema de detección de emociones, así como profesores del área de computación. El catálogo se integró con psicólogos, así como también profesores que trabajan directamente con estudiantes y son capaces de detectar algunas emociones centradas en el aprendizaje, y por último estudiantes de nivel licenciatura en psicología que ya cuentan con un acercamiento a la detección de emociones. Todos los miembros del catálogo recibieron una invitación a participar en calidad de experto para etiquetar frases a través de la aplicación y 36 respondieron afirmativamente a la invitación proporcionando sus datos.

#### **4.7. Distribución del sistema de etiquetado de frases**

Una vez terminada la versión final de la aplicación Label, ésta se distribuyó a los participantes mediante el servicio Firebase App Distribution. Este servicio facilita la creación y segmentación de grupos de personas que harán uso de la aplicación. Los participantes fueron agregados mediante su correo electrónico, donde recibieron un enlace para descargar e instalar la aplicación en sus dispositivos móviles. Para realizar la distribución se crearon 3 grupos de usuarios lo cuales se segmentaron de la siguiente forma: 15 Psicólogos, 6 profesores y 15 alumnos.

#### **4.8. Validación de los corpus por expertos**

El proceso de etiquetado de frases de los corpus Sentitext y EduSere es independiente para cada usuario. Al ingresar al sistema el usuario encuentra una interfaz que agrupa por niveles las frases que serán etiquetadas y muestra en un color diferente aquellos niveles que se han completado. Cuando el usuario selecciona el nivel a etiquetar, el sistema presenta una interfaz con las frases y etiquetas que pueden ser usadas para cada una de ellas y que el usuario selecciona con un simple toque en la pantalla. Cada frase etiquetada por un experto es almacenada en una base de datos donde se integran los corpus validados.

#### **4.9. Creación de los nuevos corpus validados por expertos**

Para la creación de los nuevos corpus fue necesario diseñar un algoritmo que eligiera la mejor etiqueta para cada frase, basándose en la experiencia de cada experto. En un inicio se pensó utilizar el coeficiente kappa de Cohen, que sugiere que dos o más jueces categoricen independientemente una muestra de unidades y determinen el grado, la importancia y la estabilidad muestral de su acuerdo (Cohen, 1960). Pero este algoritmo solo se basa en el puntaje que cada juez le da a la muestra.

En nuestro caso, además de puntuar la etiqueta para cada frase, también fue necesario puntuar al experto que ha etiquetado cada frase, ya que el grado de confiabilidad se medirá por el nivel de experiencia de los expertos. El coeficiente kappa no permite realizar esta evaluación, por lo cual se desarrolló un algoritmo para elegir la mejor etiqueta que le corresponda a cada frase evaluando al experto.

El algoritmo asigna un puntaje a cada frase que dependiendo de las etiquetas que seleccionan los expertos para cada frase. Los expertos con mayores conocimientos en el área de emociones y polaridades en frases como son los psicólogos tienen asignado un mayor puntaje en su etiqueta, el segundo puntaje en valor es el de los profesores que tienen experiencia detectando emociones centradas en el aprendizaje en sus alumnos y por último con menor puntaje están los estudiantes que apoyaron en el etiquetado, ya que estos no cuentan con suficiente experiencia etiquetando polaridades y emociones.

Usando el algoritmo de puntuación, se obtiene la sumatoria de todas las etiquetas asignadas por los expertos a cada frase y selecciona la etiqueta con mayor puntuación y se asigna a la frase de forma única y definitiva.

La figura 4-16 muestra un ejemplo del algoritmo de puntaje para la selección de la etiqueta validada por expertos para la frase. Primero el algoritmo obtiene la frase con su respectiva etiqueta sin validar, Después realiza una puntuación a todas las etiquetas de los expertos para la frase, asignando el puntaje dependiendo de la experiencia del experto en el reconocimiento de emociones y polaridades. Posteriormente agrega un punto extra a la etiqueta validada que es igual a la etiqueta original sin validar de la frase, con el fin de evitar obtener empates en el puntaje y así seleccionar solo una etiqueta. Para finalizar se hace una sumatoria de los puntajes de cada etiqueta para así seleccionar la de mayor puntuación.

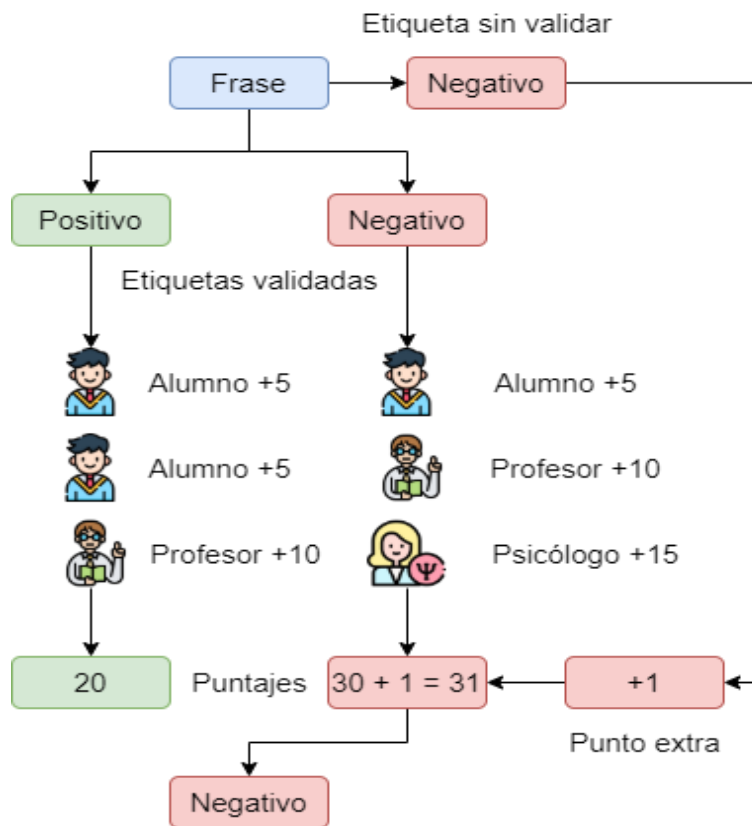


Figura 4-16 Algoritmo de puntaje para etiquetar frases

# Capítulo 5

## 5. Resultados

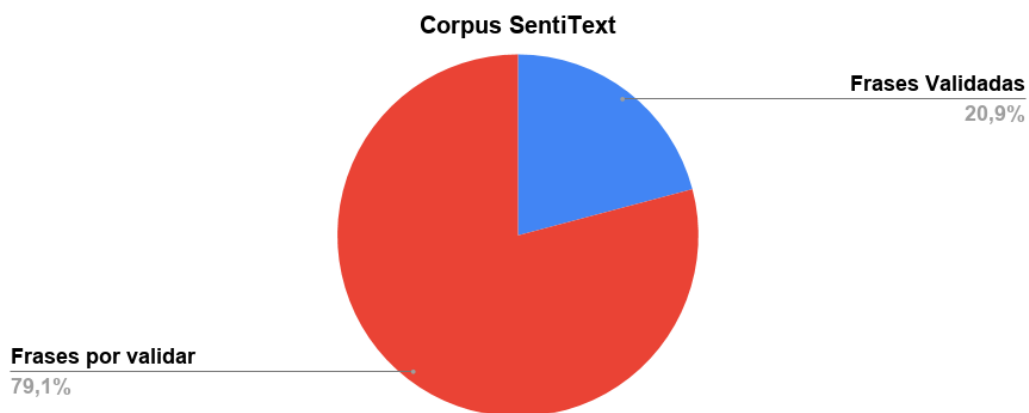
En el siguiente capítulo se muestran los resultados obtenidos de la aplicación Label en el etiquetado de frases por expertos de los corpus SentiText y EduSERE. Se presenta las cifras de frases validadas en los nuevos corpus creados, las diferencias en el etiquetado de los expertos con el etiquetado original, y para finalizar las comparaciones de las precisiones que se obtuvieron en los clasificadores de emociones en texto de los nuevos corpus con el original.

### 5.1. Tamaño de los nuevos corpus en español validados por expertos

En esta sección se muestran las cifras detalladas de las frases etiquetadas y validadas por los expertos en los corpus SentiText y EduSERE, utilizando la aplicación Label, así como el número de frases por cada etiqueta.

#### 5.1.1. Corpus SentiText validado por expertos

A continuación, se muestran las cifras del tamaño del corpus SentiText que se logró validar por expertos. El total de frases que contiene el corpus SentiText inicialmente es de 24,543 frases, de las cuales 5,126 se lograron etiquetar y validar por expertos mediante la aplicación Label. El total de frases etiquetadas por expertos para el corpus SentiText representa un 20.9% del total de frases con polaridad. En la figura 5-1 se muestra el porcentaje del corpus validado por los expertos.



*Figura 5-1 Porcentaje de frases validadas del corpus SentiText*

En la tabla 5-1 se muestra el total de frases etiquetadas de los expertos por etiqueta del corpus SentiText.

*Tabla 5-1 Total de frases por etiqueta del corpus SentiText validado*

Etiqueta	Total de frases etiquetadas por expertos
<b>Positivo</b>	2,649
<b>Negativo</b>	2,477

En la tabla 5-2 se muestra un ejemplo de 6 frases etiquetadas por expertos con polaridad, 3 con polaridad positiva y 3 con polaridad negativa. Las frases se muestran tal como fueron concebidas por su autor.

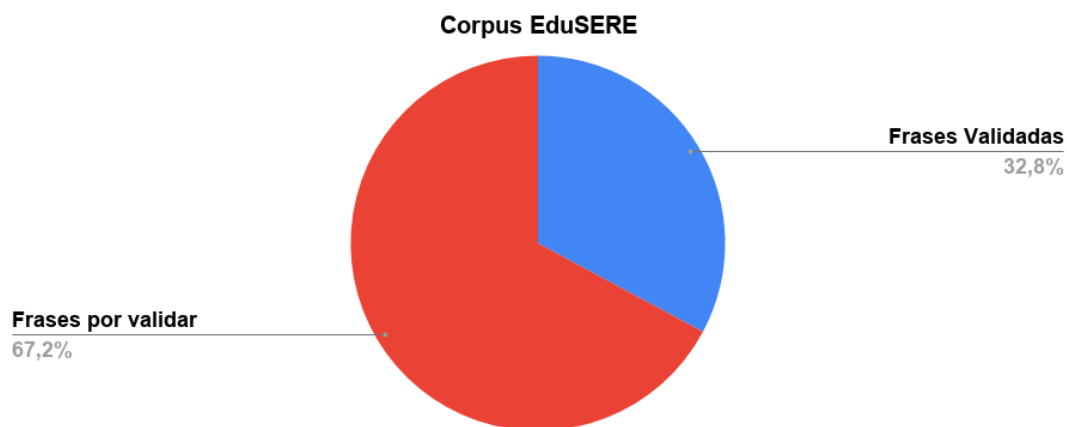


Tabla 5-2 Ejemplo de frases etiquetadas por polaridad

Polaridad	Frase
<b>Positivo</b>	Excelente el curso! he aprendido muchas cosas en poco tiempo.
<b>Positivo</b>	Las explicaciones dadas por la profe me parecieron excelentes sobre en todo en la sección el gran proyecto de java
<b>Positivo</b>	Muchas gracias por tu esfuerzo de crear este nuevo curso que muchos esperábamos.
<b>Negativo</b>	No proporciona material para trabajar en casa
<b>Negativo</b>	Este programa fue sencillo solo sumaba la cantidad de números y les sacaba la media, siento que es algo for dummies
<b>Negativo</b>	El ponente parece apurado, habla muy rápido, sin hacer hincapié en los puntos importantes

### 5.1.2. Corpus EduSERE validado por expertos

A continuación, se muestran las cifras del tamaño del corpus EduSERE que se logró validar por expertos. El total de frases que contiene el corpus EduSERE inicialmente es de 12,083 frases, de las cuales 3,966 se lograron etiquetar y validar por expertos mediante la aplicación Label. El total de frases etiquetadas por expertos para el corpus EduSERE representa un 32.8% del total de frases con emociones centradas en el aprendizaje. En la figura 5-2 se muestra el porcentaje del corpus validado por expertos.



*Figura 5-2 Porcentaje de frases validadas del corpus EduSERE*

En la tabla 5-3 se muestra el total de frases etiquetadas por expertos por etiqueta del corpus EduSERE.

*Tabla 5-3 Total de frases por etiqueta del corpus EduSERE validado*

Etiqueta	Total de frases etiquetadas por expertos
<b>Comprometido</b>	1,886
<b>Aburrido</b>	705
<b>Frustrado</b>	1,375

En la tabla 5-4 se muestra un ejemplo de 9 frases etiquetadas por expertos con emociones centradas en el aprendizaje, 3 etiquetas de comprometido, 3 con aburrido y 3 más con frustrado. Las frases se muestran tal como fueron concebidas por su autor.

Tabla 5-4 Ejemplo de frases etiquetadas por emoción centrada en el aprendizaje

Emoción centrada en el aprendizaje	Frase
<b>Comprometido</b>	Me parece muy completo y didáctico, se explica claramente y lo puede realizar una persona que no sabe absolutamente nada de Android.
<b>Comprometido</b>	Es una práctica que cumple con su función, hacerme razonar, pensar y así resolver el problema
<b>Comprometido</b>	El recurso presentado me parece excelente ya que da a conocer la información de forma muy clara
<b>Aburrido</b>	Me enfado cuando no tengo nada que hacer
<b>Aburrido</b>	La clase que imparte la profe de javascript es muy teórica
<b>Aburrido</b>	Mañana tengo examen y que pereza
<b>Frustrado</b>	La información que se muestra no tiene nada que ver con los contenidos
<b>Frustrado</b>	El docente no es perseverante en sus explicaciones y no tiene control del grupo.
<b>Frustrado</b>	La verdad ya entendí la tarea pero el problema es que me equivocó demasiado

### 5.1.3. Nuevo corpus SentiText

A partir del corpus validado se generó una nueva versión del corpus SentiText, juntando las frases no validadas con las validadas por expertos, además de agregar también las frases validadas del corpus EduSERE, convirtiendo las frases con etiquetas de aburrido y frustrado a negativo y las frases con la etiqueta comprometido a positivo, dando un total de 28,675 frases, de las cuales 9,092 están validadas por expertos. En la tabla 5-5 se muestra el total de frases por etiqueta.

Tabla 5-5 Total de frases por etiqueta del nuevo corpus SentiText

Etiqueta	Total de frases
<b>Positivo</b>	14,359
<b>Negativo</b>	14,316

#### 5.1.4. Nuevo corpus EduSERE

Para generar la nueva versión del corpus EduSERE se juntaron las frases validadas y no validas, además se agregaron todas las frases validadas con etiqueta positiva del corpus SentiText, convirtiéndolas de positivo a comprometido. Se logró realizar esta acción ya que comprometido es la única emoción positiva dentro del corpus, para así se aumentar el tamaño de frases. La nueva versión del corpus EduSERE cuenta con 19,385 frases de la cuales 5,604 están validadas por expertos. En la tabla 5-6 se muestra el total de frases por cada etiqueta del corpus.

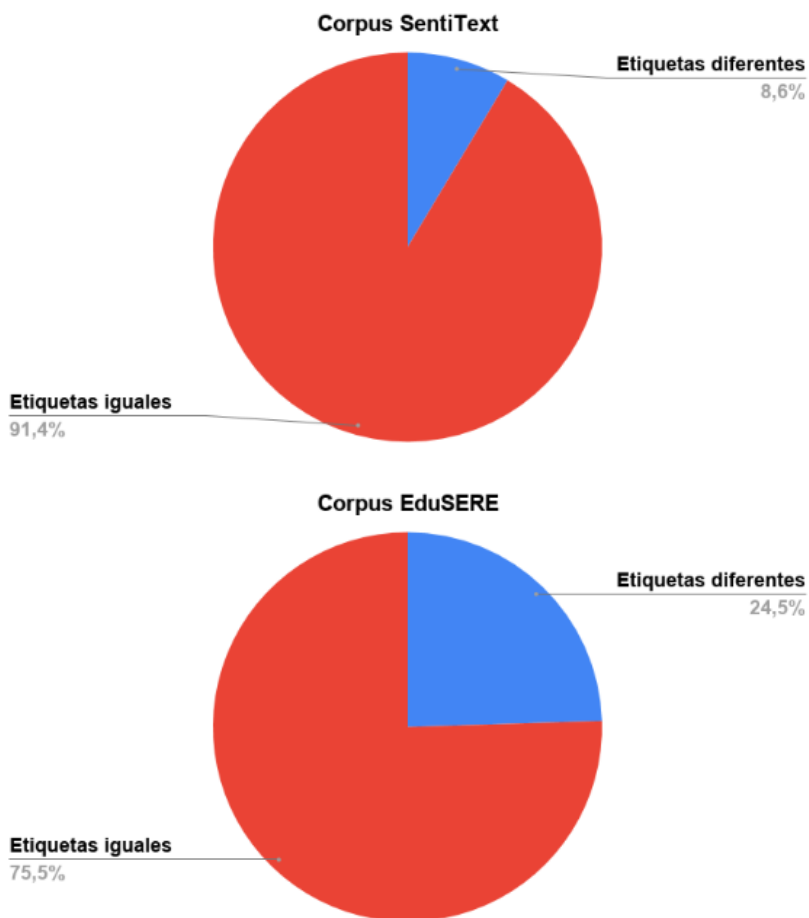
Tabla 5-6 Total de frases por etiqueta del nuevo corpus EduSERE

Etiqueta	Total de frases etiquetadas por expertos
<b>Comprometido</b>	12,936
<b>Aburrido</b>	2,894
<b>Frustrado</b>	3,555

#### 5.1.5. Diferencias en el etiquetado de los corpus validados con el original

Con el nuevo etiquetado de 5,126 frases de SentiText, se obtuvieron 441 diferencias respecto al original, representando aproximadamente un 8.6% de etiquetas diferentes. Las 3,966 etiquetas validadas de EduSERE contienen 972 diferencias, representando un 24.5% de etiquetas diferentes con respecto al etiquetado original. Con estos resultados se puede concluir que los expertos han aportado diferencias significativas en el etiquetado de los corpus originales y por lo tanto se entiende que el corpus original puede no estar etiquetado

correctamente. En la figura 5-3 se muestran los porcentajes de diferencias en los corpus SentiText y EduSERE.



*Figura 5-3 Diferencias en el etiquetado de los corpus SentiText y EduSERE*

## 5.2. Clasificadores de polaridad y emociones en texto

A continuación, se muestran los diferentes clasificadores de polaridad y emociones en texto creados a partir de los nuevos corpus validados por expertos, SentiText y EduSERE, también se presentan las precisiones que alcanzan dichos clasificadores y las comparaciones con los clasificadores generados de los corpus sin validar.

### 5.2.1. Comparación de precisiones en los clasificadores

En esta sección se presentan los diferentes algoritmos que dieron mejor resultado para clasificar emociones y polaridad en texto con los corpus validados SentiText y EduSERE y su comparación con los corpus sin validar. Para el corpus SentiText validado el algoritmo que dio mejor resultado fue EvoMSA con una precisión de 0.9087, quedando en el mismo rango de precisión del corpus original en el mismo algoritmo. Para el corpus EduSERE validado, el algoritmo Stochastic Gradient Descent fue el de mejor resultado dando una precisión de 0.8371, mejorando en un 1% contra el corpus original en el mismo algoritmo.

En la tabla 5-7 se muestra una comparación en la precisión de los clasificadores de polaridad en texto entre el corpus SentiText validado y el original.

*Tabla 5-7 Comparación de precisión en los clasificadores de polaridad del corpus SentiText*

Clasificador	SentiText	SentiText Validado
<b>Multinomial Naive Bayes</b>	0.8743	0.8669
<b>Stochastic Gradient Descent</b>	0.9034	0.8974
<b>Decision Tree</b>	0.8567	0.8552
<b>Linear Support Vector</b>	0.9017	0.8857
<b>EvoMSA</b>	<b>0.9110</b>	<b>0.9087</b>

En la tabla 5-8 se muestra una comparación en la precisión de los clasificadores de emociones centradas en el aprendizaje en texto entre el corpus EduSERE validado y el original.

Tabla 5-8 Comparación de precisión en los clasificadores de emociones del corpus EduSERE

Clasificador	EduSERE	EduSERE validado
<b>Multinomial Naive Bayes</b>	0.7972	0.8043
<b>Stochastic Gradient Descent</b>	<b>0.8287</b>	<b>0.8371</b>
<b>Decision Tree</b>	0.7273	0.7990
<b>Linear Support Vector</b>	0.8130	0.8263
<b>EvoMSA</b>	0.8238	0.7708

# Capítulo 6

## 6. Conclusiones

En este capítulo se presentan las conclusiones alcanzadas en esta investigación, las aportaciones y el trabajo futuro que se planea realizar con los nuevos corpus SentiText y EduSERE.

### 6.1. Conclusiones

Con el desarrollo de la aplicación Label se creó un procedimiento en el que cualquier corpus que se almacene a una base de datos en la nube pueda ser etiquetado por diferentes usuarios de manera colaborativa, con la finalidad de que se puedan crear nuevos corpus que requieran mejorar el etiquetado de sus frases.

A partir del uso de la aplicación Label se logró realizar el etiquetado y validado de los corpus por expertos donde se pudieron etiquetar 5,126 frases para SentiText que equivalen al 20.9% del total del corpus y 3,966 frases para el corpus EduSERE las cuales equivalen al 32.8%. Con estas frases validadas se logró crear una nueva versión de los corpus, para realizar las pruebas con clasificadores de polaridad y emociones en texto, donde se mejoraron los resultados de precisión para el corpus EduSERE y se obtuvieron los mismos resultados para el corpus SentiText.

Con la realización de esta investigación se llega a la conclusión que para la realización de un corpus surgen múltiples factores para que nuestros datos puedan tener cierta calidad, tales como el tamaño, que las etiquetas concuerden con las frases, que las etiquetas estén bien balanceadas en el corpus y no haya mucha diferencia en cantidades entre ellas. Este trabajo de investigación logro que una parte de los corpus fueran validados en su etiquetado y que las frases pudieran concordar mejor con su etiqueta, pero aún queda mucho trabajo por realizar para mejorar aún más los corpus, como balancear las etiquetas del corpus EduSERE y validar todas las frases por expertos en los dos corpus.



## **6.2. Aportaciones**

Las principales aportaciones de esta investigación fueron la creación de una metodología para que expertos puedan etiquetar corpus de frases con polaridades y emociones centradas en el aprendizaje, se desarrolló una aplicación con técnicas de gamificación para que expertos etiqueten corpus de manera fácil e interactiva, también se realizó el registro de 15 Psicólogos, 6 profesores y 15 alumnos que participaron de manera colaborativa en el etiquetado de los corpus SentiText y EduSERE.

A partir de este etiquetado de frases por expertos se obtuvieron dos nuevas versiones de los corpus, para SentiText se obtuvieron 28,675 frases etiquetadas con polaridad, de las cuales 9,092 frases fueron validadas por expertos y para EduSERE se obtuvieron 19,385 frases etiquetadas con emociones centradas en el aprendizaje, de las cuales 5,604 están validadas por expertos. Por último, se logró elevar el porcentaje de precisión de 82.8% a 83.7% en el clasificador automático de emociones en texto para el corpus EduSERE con el algoritmo Stochastic Gradient Descent, el cual fue el mejor resultado.

## **6.3. Trabajo futuro**

Para trabajo futuro se planea agregar más frases en la base de datos en la nube para que se puedan validar mediante la aplicación Label y así formar corpus más grandes, también se planea seguir buscando más expertos para seguir etiquetando y validando el corpus en su totalidad, para poder generar resultados más competitivos con los actuales clasificadores automáticos de polaridad y emociones en texto.

# Bibliografía

- Ángel, J., & Retana, G. (2012). La educación emocional, su importancia en el proceso de aprendizaje The Emotional Intelligence, its importance in the learning process. *Revista Educacion*, 36(1), 1–24.
- Atkins, S., & Clear, J. (1992). *Corpus design criteria*. Oxford University Press, 7(1).
- Barrón Estrada, M. L., Zatarain Cabada, R., Aispuro Gallegos, C. G., Sosa Ochoa, C. de la L., & Lindor Valdez, M. (2015). Affective Environment for Java Programming Using Facial and EEG Recognition. *Research in Computing Science*, 106, 39–47.
- Barrón Estrada, M. L., Zatarain Cabada, R., Oramas Bustillos, R., & Ramírez Ávila, S. L. (2017). Building a Corpus of Phrases Related to Learning for Sentiment Analysis. *Research in Computing Science*, 146, 17–26.
- Barrón Estrada, M. L., Zatarain Cabada, R., Valencia Rodríguez, M., & Peralta Peñuñuri, G. E. (2018). Ambiente inteligente de aprendizaje de depuración de errores en programas. *Research in Computing Science*, 147(8), 277–290.
- Bosch, J. (2000). *Design and Use of Software Architectures: Adopting and Evolving a Product-Line Approach* (1.a ed.). Addison-Wesley Professional.
- Brusilovsky, P. (1994). Student model centered architecture for intelligent learning environments.
- Carpenter, M. (2018). Las mejores herramientas para analizar el sentimiento de marca. Talkwalker. <https://www.talkwalker.com/es/blog/las-mejores-herramientas-analizar-sentimiento>. Consultado: 2020.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. <https://doi.org/10.1177/001316446002000104>.
- CORDE. (s. f.). Corpus diacrónico del español. Real Academia Española: Banco de datos. <http://www.rae.es>. Consultado: 2020.
- Dhingra, K., & Sardana, K. D. (2017). Educational data mining: a review to its future vision. *International Journal of Technology Transfer and Commercialisation*, 15(3), 309. doi:10.1504/ijttc.2017.088730.
- D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., Person, N., Kort, B., El

- Kaliouby, R., Picard, R. W., & Graesser, A. (2008). AutoTutor detects and responds to learners affective and cognitive states. *IEEE Transactions on Education*, 48(4), 612-618.
- Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*. (6), 169-200.
- ESLORA (2018). Corpus para el estudio del español oral. <http://eslora.usc.es>, ISSN: 2444-1430. Consultado: 2020.
- Graff, M., Miranda Jimenez, S., Tellez, E. S., & Moctezuma, D. (2020). EvoMSA: A Multilingual Evolutionary Approach for Sentiment Analysis [Application Notes]. *IEEE Computational Intelligence Magazine*, 15(1), 76–88.
- Healey, C. G., & Ramaswamy. (2019). Twitter Sentiment Visualization. [csc2.ncsu.edu](http://csc2.ncsu.edu). [https://www.csc2.ncsu.edu/faculty/healey/tweet\\_viz/](https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/). Consultado: 2020.
- Hemmatian, F., & Sohrabi, M. K. (2019). A survey on classification techniques for opinion mining and sentiment analysis. *Artificial Intelligence Review*, 52(3), 1495–1545.
- Hockenbury, D. and Hockenbury, S.E. (2007). *Discovering Psychology*. New York: Worth Publishers.
- La, H. J., & Kim, S. D. (2010). Balanced MVC architecture for developing Service-based Mobile Applications. *Proceedings - IEEE International Conference on E-Business Engineering, ICEBE 2010*, 292–299.
- Liu, B., (2015). The Problem of Sentiment Analysis. In *Sentiment Analysis*.
- McEnery, T. (2012). Corpus Linguistics. *The Oxford Handbook of Computational Linguistics*, 9780199276349 (January 2019), 1–18.
- McKenna, D. (2016). The Scrum Framework. *The Art of Scrum*, 27–34.
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113.
- MeaningCloud. (2020). Sentiment analysis and opinion mining API. <https://www.meaningcloud.com/products/sentiment-analysis>. Consultado:2020.
- Moreno Fernández, F. 1997. Metodología del "Proyecto para el Estudio Sociolingüístico del Español de España y de América". *Trabajos de sociolingüística hispánica*. Alcalá de Henares: Universidad de Alcalá, pp. 137-166.

- Nkambou, R., Bourdeau, J. & Mizoguchi, R. (2010). Introduction: What Are Intelligent Tutoring Systems, and Why This Book?. *Studies in Computational Intelligence*. 308. 10.1007/978-3-642-14363-2\_1.
- Nkambou, R., Bourdeau, J., & Psyché, V. (2010). Building Intelligent Tutoring Systems: An Overview. *Studies in Computational Intelligence*. 308, 361-375.
- Pekrun, R. (1992). The Impact of Emotions on Learning and Achievement: Towards a Theory of Cognitive/Motivational Mediators. *Applied Psychology*, 41(4), 359–376.
- Picard, R. W. (1997). *Affective Computing*, MIT Press.
- PRESEEA (2014-2020). Corpus del Proyecto para el estudio sociolingüístico del español de España y de América. Alcalá de Henares: Universidad de Alcalá. <http://preseea.linguas.net>. Consultado: 2020.
- Ramírez Ávila, S. L. (2018). Generación de un corpus de frases relacionadas con el aprendizaje para el análisis de sentimientos. Tesis de Maestría en Ciencias de la Computación. Tecnológico Nacional de México-Instituto Tecnológico de Culiacán.
- SentiStrength. (s. f.). sentiment strength detection in short texts, sentiment analysis and opinion mining. <http://sentistrength.wlv.ac.uk/>. Consultado: 2020.
- Subits, C., & Ortega, M. (2012). Corpus del Español Actual. <http://spanishfn.org/tools/cea/spanish>.
- Villena Román, J. (2013). TASS - Workshop on Sentiment Analysis at SEPLN. *Procesamiento del Lenguaje Natural*. N. 50 (2013). ISSN 1135-5948, pp. 37-44.
- Woolf, B. P. (2010). Building Intelligent Interactive Tutors Student-centered strategies for revolutionizing e-learning. In Morgan Kaufmann.
- Wynne, M. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books.
- Zatarain Cabada, R., Barrón Estrada, M. L., González Hernandez, F., & Alor Hernandez, G., (2016). Evaluando afecto en un entorno de aprendizaje para Java. *Research in Computer Science*, 111, 123–133.
- Zatarain Cabada, R., Barrón Estrada, M. L., González Hernández, F., & Oramas Bustillos, R. (2015). Ambiente inteligente de aprendizaje con manejo afectivo para Java. *Research in Computer Science*, 92, 111-121.

Zatarain Cabada, R., Barrón Estrada, M. L., & Oramas Bustillos, R. (2018). Mining of educational opinions with deep learning. *Journal of Universal Computer Science*, 24(11), 1604–1626.

Zatarain Cabada, R., Barrón Estrada, M. L., Ríos Félix, J. M., & Alor Hernández, G. (2018). A virtual environment for learning computer coding using gamification and emotion recognition. *Interactive Learning Environments*, DOI: 10.1080/10494820.2018.1558256.