UNIVERSIDAD DE COSTA RICA
SISTEMA DE ESTUDIOS DE POSGRADO

MINERÍA DE TEXTO: COMPARACIÓN LÉXICO-SEMÁNTICA DEL CORPUS LINGÜÍSTICO DE LAS OBRAS LITERARIAS LATINOAMERICANAS RECOMENDADAS EN EL TERCER CICLO Y EDUCACIÓN DIVERSIFICADA EN EL SISTEMA EDUCATIVO COSTARRICENSE CON EL CORPUS LINGÜÍSTICO DE ÍTEMS DE PRÁCTICA PARA LA PRUEBA DE APTITUD ACADÉMICA (PAA) DE LA UNIVERSIDAD DE COSTA RICA, 2018 -2020

Trabajo final de investigación aplicada sometido a la consideración de la Comisión del Programa de Estudios de Posgrado en Estadística para optar al grado y título de Maestría Profesional en Estadística

MELISSA EDITH VALVERDE HERNÁNDEZ

Ciudad Universitaria Rodrigo Facio, Costa Rica

2025

DEDICATORIA

En primer lugar, lo dedico a mi mamá y a mi papá pues son mis pilares de vida, agradezco su incondicional apoyo y amor. Y por ser la persona y profesional que soy hoy.

Las personas autoras de obras literarias que con su arte abre las puertas a la imaginación humana y nos hace comprender de alguna manera el mundo en el que convivimos.

Por último, dedico este estudio a todas las maestras y maestros que enseñan a leer y a escribir, ya que esto es el primer paso para el cambio y transformación de la vida de las personas en la comprensión del mundo.

AGRADECIMIENTO

En primer lugar, agradezco a la vida y al ser supremo por las oportunidades que se me han brindado a nivel académico y profesional en realizar análisis vinculados a la temática educativa.

En segundo lugar, agradezco al profesor Guaner Rojas por su apoyo en este proceso y brindarme la idea de realizar este análisis para que sea de aporte en la construcción de ítems en el Programa Permanente de la Prueba de Aptitud Académica. Muy agradecida por la retroalimentación recibida.

En tercer lugar, agradezco a las personas lectoras Eugenia Gallardo y Nelson Pérez por contribuir en este documento, agradezco enormemente su tiempo y sugerencias.

También, agradezco a mis amigas y amigos por el constante apoyo y motivación.

Finalmente, me agradezco a mí misma por el tiempo dedicado y la buena actitud en este proceso, pues con esto culmina una meta más en mi formación académica.

Este trabajo final de investigación aplicada fue aceptado por la Comisión del Programa de Estudios de Posgrado en Estadística de la Universidad de Costa Rica, como requisito parcial para optar al grado y título de Maestría Profesional en Estadística.

MARIELOS MURILLO ROJAS (FIRMA)

Firmado digitalmente por MARIELOS MURILLO ROJAS (FIRMA)
Fecha: 2026.01.23 13:37:36 -06'00'

Dra. Marielos Murillo Rojas
Representante del Decano
Sistema de Estudios de Posgrado

Firmado por GUANER DAVID ROJAS ROJAS (FIRMA)
PERSONA FISICA, CPF-05-0328-0120.
Fecha declarada: 26/01/2026 10:54 AM
Razón: Aprobación
Lugar: Montes de Oca Contacto: guaner.rojas@ucr.ac.cr

Dr. Guaner Rojas Rojas
Profesor Guía

EUGENIA GALLARDO ALLEN (FIRMA)

Firmado digitalmente por EUGENIA GALLARDO ALLEN (FIRMA)
Fecha: 2026.01.27 08:23:56 -06'00'

Dra. Eugenia Gallardo Allen
Lectora

NELSON ANTONIO PEREZ ROJAS (FIRMA)
PERSONA FISICA, CPF-02-0606-0347.
Fecha declarada: 27/01/2026 03:12:38 p.m.
Esta es una representación gráfica únicamente, verifique la validez de la firma.

M.L. Nelson Pérez Rojas
Lector

GILBERT BRENES CAMACHO (FIRMA)

Firmado digitalmente por GILBERT BRENES CAMACHO (FIRMA)
Fecha: 2026.01.27 15:59:58 -06'00'

Dr. Gilbert Brenes Camacho
Director
Programa de Posgrado en Estadística

MELISSA EDITH VALVERDE HERNANDEZ (FIRMA)

Firmado digitalmente por MELISSA EDITH VALVERDE HERNANDEZ (FIRMA)
Fecha: 2026.01.23 13:47:43 -06'00'

Melissa Edith Valverde Hernández
Sustentante

TABLA DE CONTENIDO

DEDICATORIA ii

AGRADECIMIENTO iii

HOJA DE APROBACIÓN iv

TABLA DE CONTENIDO v

RESUMEN EN ESPAÑOL viii

ABSTRACT ix

LISTA DE TABLAS x

LISTA DE FIGURAS xi

LISTA DE ABREVIATURAS xii

CAPÍTULO I. INTRODUCCIÓN 1

CAPÍTULO II. ANTECEDENTES 3

CAPÍTULO III. OBJETIVOS 6

3.1 Objetivo general 6

3.2 Objetivos específicos 6

CAPÍTULO IV. MARCO DE REFERENCIA 8

4.1 Aspectos teóricos relevantes 8

4.1.1 Educación como derecho 9

4.1.2 Comprensión lectora 10

4.1.3 Razonamiento verbal 11

4.1.4 Programas de español del Ministerio de Educación Pública 12

4.1.5 Lecturas recomendadas por el Consejo Superior de Educación 14

4.1.6 Pruebas estandarizadas y medición de la habilidad de la comprensión lectora y razonamiento verbal 16

4.1.7 Programa para la Evaluación Internacional de Estudiantes 17
4.1.8 Pruebas Nacionales del Ministerio de Educación Pública 19
4.1.9 Prueba de Aptitud Académica de la Universidad de Costa Rica 22
4.1.10 Comprensión lectora en el ámbito universitario 23
4.1.11 Corpus lingüístico y minería de textos 25
4.2 Análisis de similitud de documentos textuales 27
4.2.1 Escalamiento Multidimensional 27
4.2.2 Análisis Semántico Latente 27
4.3 Medidas de similitud textual 29
4.3.1 Similitud de coseno 29
4.3.2 Similitud de Jaccard o índice de Jaccard 30
4.4 Modelos o clasificadores aplicados al análisis de texto 30
4.4.1 Ingenuo de Bayes o Naive Bayes 30
4.4.2 Máquinas de soporte vectorial o Support vector machine (SVM) 32
4.4.3 Bosques aleatorios o Random Forest 36
4.4.4 Máxima entropía o maximum entropy 38
4.4.5 Asignación latente de Dirichlet o Latent Dirichlet Allocation (LDA) 39
4.5 Medidas de ajuste para validación de la clasificación de los modelos 41
CAPÍTULO V. METODOLOGÍA 43
5.1 Enfoque 43
5.2 Unidades de estudio 43
5.3 Procedimiento y estrategia de análisis 46
CAPÍTULO VI. RESULTADOS 53
6.1 Análisis exploratorio 53
6.1.1 Descriptivos de las lecturas analizadas 53

6.1.2 Descriptivos de los corpus de las lecturas 60
6.1.3 Descriptivos del corpus de ítems de práctica de la PAA 63
6.1.4 Análisis de similitud textual de los documentos y los corpus 66
6.2 Modelos de estadísticos de clasificación 69
6.2.1 Medidas de ajuste de los modelos 69
CAPÍTULO VII. CONCLUSIONES 72
REFERENCIAS 76
ANEXOS 85
Anexo 1. Lista de textos literarios, y su correspondiente dosificación, para Tercer Ciclo (rama académica y técnica) y Educación Diversificada (académica) recomendada por el Consejo Superior de Educación 85
Anexo 2. Listado de las obras literarias analizadas 94
Anexo 3. Código R del análisis 97

RESUMEN EN ESPAÑOL

La comprensión lectora y el conocimiento del léxico son parte fundamental en el entendimiento de los textos. En este estudio se explora si el corpus lingüístico de diversas obras literarias latinoamericanas, recomendadas a nivel de secundaria en el sistema educativo costarricense, se asocian con el corpus lingüístico de ítems de práctica para la Prueba de Aptitud Académica (PAA) de la Universidad de Costa Rica. En este caso se realiza un análisis bajo el enfoque cuantitativo utilizando métricas y técnicas de minería de textos. En total se analizan 68 obras literarias y se toma en cuenta 150 ítems de práctica de la PAA. Al final, se obtiene un corpus general con más de 73 000 palabras. Según los resultados de asociación de los corpus lingüísticos, se muestra magnitudes de las asociaciones de la similitud de coseno y el índice de Jaccard superiores a 0.9, lo cual indica asociación entre los documentos y los corpus. Además, los modelos de clasificación obtienen precisiones mayores a 0.9 y precisiones equilibradas mayores a 0.5. Los modelos con asignación latente de Direchlet (ALD) presentan las mejores precisiones equilibradas. El modelo de bosques aleatorios clasifica con precisión mayor a 0.93 y precisiones equilibradas mayores a 0.6. Se concluye que existe una notable similitud en el vocabulario utilizado en las obras literarias recomendadas y los ítems de práctica de la PAA. El modelo de bosques aleatorios demostró ser el más eficaz para la clasificación léxica en este contexto. Estos hallazgos sugieren que la familiarización con el léxico de las obras literarias requeridas podría ser beneficiosa para las personas estudiantes al enfrentar la PAA. Sin embargo, se reconoce que la adquisición de vocabulario y la comprensión lectora dependen también de factores individuales (hábitos, actitudes, vivencias) y del enfoque pedagógico empleado en el aula por la persona docente.

Palabras clave: Corpus lingüístico, comprensión lectora, obras literarias, prueba de aptitud, minería de texto.

ABSTRACT

Reading comprehension and lexical knowledge are fundamental for understanding texts. This study aims to determine whether the linguistic corpus of Latin American literary works recommended at the secondary level within the Costa Rican educational system is associated with the linguistic corpus of practice items for the University of Costa Rica's Academic Aptitude Test (PAA). The analysis was conducted using a quantitative approach employing text mining metrics and techniques. A total of 68 literary works and 150 PAA practice items were analyzed, resulting in a general corpus exceeding 73,000 words. According to the linguistic corpora association results, high magnitudes were observed for both cosine similarity and Jaccard index associations (both > 0.9), indicating an association between the documents and the corpora. Furthermore, the classification models achieved accuracy scores greater than 0.9 and balanced accuracy scores greater than 0.5. Models incorporating Latent Dirichlet Allocation (LDA) presented the best balanced accuracy scores. The Random Forest model performed best in classification, achieving an accuracy greater than 0.93 and a balanced accuracy greater than 0.6. It is concluded that there is a notable similarity in the vocabulary used in the recommended literary works and the PAA practice items. The Random Forest model proved to be the most effective for lexical classification in this context. These findings suggest that familiarity with the lexicon of the required literary works could be beneficial for students when facing the PAA. However, it is acknowledged that vocabulary acquisition and reading comprehension also depend on individual factors (such as habits, attitudes, and personal experiences) and the pedagogical approach employed by the teacher in the classroom.

Keywords: Linguistic corpus, reading comprehension, literary works, aptitude test, text mining.

LISTA DE TABLAS

Tabla 1. Estructura matriz de confusión 41
Tabla 2. Obras literarias latinoamericanas consideradas, lecturas recomendadas por el CSE, 2017. 44
Tabla 3. Número y porcentaje de las características de las obras literarias analizadas 53
Tabla 4. Similitud de coseno y Jaccard entre las obras literarias y el corpus de ítems de práctica de la PAA 67
Tabla 5. Medidas de ajuste de los modelos usados en la clasificación.....70
Tabla 6. Medidas de ajuste de los modelos usados en la clasificación considerando ALD .71

LISTA DE FIGURAS

Figura 1. Análisis de escalamiento multidimensional de las obras literarias según nivel educativo 55
Figura 2. Análisis de escalamiento multidimensional de las obras literarias según género literario 56
Figura 3. Análisis de escalamiento multidimensional de las obras literarias según nacionalidad de la persona autora 57
Figura 4. Análisis de escalamiento multidimensional de las obras literarias según evaluación en prueba FARO 59
Figura 5. Análisis de escalamiento multidimensional de las obras literarias en las tres componentes latentes 60
Figura 6. Nube de las palabras más frecuentes en corpus de las obras literarias 61
Figura 7. Biagramas de palabras asociadas del corpus de las obras literarias 63
Figura 8. Nube de las palabras más frecuentes en corpus de los ítems de práctica de la PAA 64
Figura 9. Biagramas de palabras asociadas del corpus de los ítems de práctica de la PAA. 66

LISTA DE ABREVIATURAS

ADL: Asignación Latente de Dirichlet

ASL: Análisis Semántico Latente

CSE: Consejo Superior de Educación

DVS: Descomposición en Valores Singulares

FARO: Fortalecimiento de Aprendizajes para la Renovación de Oportunidades

IA: Inteligencia Artificial

IIP: Instituto de Investigaciones Psicológicas

INEC: Instituto Nacional de Estadística y Censos

KDD: Descubrimiento de Conocimiento en Datos

KDT: Descubrimiento de Conocimiento en Textos

LEE: Lectura y Escritura en Español

LEI: Lectoescritura Inicial

MCJ: Ministerio de Cultura y Juventud

MEP: Ministerio de Educación Pública

ME: Máxima Entropía

ML: Aprendizaje Automático

MSV: Máquinas de Soporte Vectorial

OCDE: Organización para la Cooperación y el Desarrollo Económico

PPPAA: Programa Permanente de la Prueba de Aptitud Académica

PAA: Prueba de Aptitud Académica

PLN: Procesamiento del Lenguaje Natural

PISA: Programa para la Evaluación Internacional de Estudiantes

UCR: Universidad de Costa Rica

CAPÍTULO I. INTRODUCCIÓN

En el presente documento se muestra el análisis realizado de la comparación léxico-semántica del corpus lingüístico de las obras literarias latinoamericanas recomendadas a nivel de secundaria en el sistema educativo costarricense con el corpus lingüístico de ítems de práctica para la prueba de aptitud académica (PAA) (Programa Permanente de la Prueba de Aptitud Académica-IIP, 2025a) de la Universidad de Costa Rica.

La primera parte del documento se centra la descripción de los antecedentes, estos se enfocan en la comprensión lectora, en la aplicación de técnicas de minería de texto y la habilidad de razonamiento verbal en pruebas estandarizadas. Luego, se exponen los objetivos, se describe el objetivo general y los objetivos específicos, los cuales dirigieron el desarrollo del estudio.

Posteriormente, se muestra el marco de referencia del estudio, en ese apartado se abordan aspectos como la educación como derecho humano, la comprensión lectora, el razonamiento verbal. Adicionalmente, se describen detalles del programa de la materia de español del Ministerio de Educación Pública y además de las lecturas recomendadas por el Consejo Superior de Educación. A su vez, se comenta la importancia de las pruebas estandarizadas en el proceso de aptitudes y habilidades de comprensión lectora. Por último, se comentan algunas aplicaciones de análisis de corpus lingüístico y minería de texto.

En la metodología, se describe el enfoque en el que se enmarca la investigación, las unidades del estudio, la estrategia y procedimiento del análisis, además, el desarrollo técnico aplicado en las métricas de similitud, modelos de clasificación y las medidas de ajuste de los modelos aplicados.

Seguidamente, se exponen los resultados del estudio, primero el análisis descriptivo de las obras literarias analizadas, luego los análisis exploratorios del corpus de las obras

literarias y el corpus de los ítems de práctica de la PAA. Además, se muestran los hallazgos asociados a las métricas de similitud textual y lo resultante de los modelos de clasificación y sus respectivas medidas de ajuste.

Por último, se presenta las principales conclusiones del estudio, se exponen las limitaciones presentadas en el análisis, así como recomendaciones para futuros análisis.

CAPÍTULO II. ANTECEDENTES

En los últimos años, los avances tecnológicos e informáticos han desarrollado herramientas que permiten realizar análisis relacionados al procesamiento en lenguaje natural y el análisis de texto. Estos han promovido la minería de texto (también conocida como análisis de texto) como una tecnología de inteligencia artificial (IA) que utiliza el procesamiento del lenguaje natural (PLN) para transformar el texto libre (no estructurado) en documentos a datos estructurados y normalizados, adecuados para el análisis basado en algoritmos de aprendizaje automático (ML; Linguamatics, 2021). La minería de texto permite capturar temáticas y conceptos clave, identificando asociaciones ocultas y tendencias entre los textos sin necesidad de conocer las palabras exactas que la persona autora del texto ha utilizado para expresar dichos conceptos.

Precisamente, el desarrollo de estas sofisticadas herramientas computacionales encuentra su justificación en la centralidad misma del lenguaje como pilar de la experiencia humana. La humanidad ha utilizado el lenguaje como un medio simbólico para la comunicación y la compresión del mundo. Por lo que, en los sistemas educativos formales, se contemplan un componente de comprensión y habilidad lectora en los procesos de enseñanza y aprendizaje. Esta es una de las competencias básicas que las personas deben desarrollar para acceder de forma directa al currículo de un sistema educativo (APPF, 2020). Sin embargo, la comprensión lectora trasciende en la vida cotidiana y en el desenvolvimiento de la persona en sociedad. Asimismo, esta hace referencia al uso del lenguaje como instrumento para la comunicación escrita, representación, interpretación y comprensión de la realidad, para construir y organizar el conocimiento, así como autorregular el pensamiento, las emociones y la conducta (Perera y Segura, 2004).

No obstante, está reconocida la trascendencia de la comprensión lectora como una competencia cognitiva profunda, su diagnóstico a nivel poblacional presenta un desafío metodológico significativo. A menudo, las mediciones a gran escala deben recurrir a indicadores de hábitos y consumo cultural como un proxy al fenómeno. En este sentido, si

bien los datos sobre la frecuencia de lectura no equivalen a una medición directa de la habilidad interpretativa, sí ofrecen un panorama contextual, sentando las bases para aplicar las herramientas de la minería de texto y así explorar el fenómeno con mayor profundidad.

Para construir un diagnóstico preciso sobre el estado de la comprensión lectora y el razonamiento verbal en el país, es fundamental articular las evidencias provenientes de distintas escalas de medición. Los resultados de encuestas sobre hábitos culturales, evaluaciones de rendimiento internacional y pruebas nacionales no debe interpretarse como una serie de datos inconexos. Por el contrario, cada uno aporta una dimensión al análisis: la Encuesta Nacional de Cultura 2016 (INEC, 2017) perfila el panorama de los hábitos lectores y el contexto sociocultural de partida; la prueba del Programa para la Evaluación Internacional de Alumnos de la OCDE (PISA, siglas en inglés) (OCDE, 2010) sitúa el desempeño estudiantil en un marco comparativo global, permitiendo identificar brechas y fortalezas a escala internacional; y las Pruebas Nacionales FARO (MEP, 2022) evalúan el dominio de habilidades en función de los objetivos específicos del sistema educativo nacional. Por tanto, la síntesis de estas tres perspectivas permite una comprensión integral y multifacética del fenómeno

En el país se realizó la Encuesta Nacional de Cultura 2016 elaborada por el Instituto Nacional de Estadística y Censos (INEC), en esta encuesta se investigaron temas culturales, entre ellos el consumo de publicaciones como la lectura de libros, periódicos y revistas. Según los resultados de la encuesta, el 43,2% de la población de 5 años y más indica leer libros en los últimos 12 meses (INEC, 2017). Este porcentaje desagregado por zona, el 45,6% de las personas de la zona urbana indica leer libros, mientras que en la zona rural representan el 36,4% (INEC, 2017). El dato desglosado por sexo evidencia que el 38,6 % de los hombres leen libros y el 47,8% de las mujeres realizan esta actividad cultural (INEC, 2017). Otro dato que se indica es que el 61,6% de quienes leen libros lo hacen por gusto o entretenimiento y el 20,4% por estudio (INEC, 2017). Finalmente, se obtiene que la población nacional lee en promedio 5,6 libros al año y, en particular, la población entre 12 y 17 años de edad lee 5,3 libros al año, en promedio (INEC, 2017).

A esto se suman los resultados de pruebas estandarizadas como el Programa para la Evaluación Internacional de Alumnos de la OCDE (PISA, siglas en inglés). Según los resultados de la prueba PISA del 2018, Costa Rica obtuvo una tasa de bajo desempeño en lectura con un 42% (Rodríguez, 2019).

Finalmente, a nivel local, el Ministerio de Educación Pública realiza en 2019 las Pruebas Nacionales para el Fortalecimiento de Aprendizajes para la Renovación de Oportunidades (FARO). Las Pruebas Nacionales FARO miden la habilidad lectora con base en textos literarios (MEP, 2021).

Luego de que la persona concluya el nivel de Educación Diversificada, tienen la oportunidad en decidir continuar con estudios superiores universitarios. En los procesos de admisión para el ingreso a carrera de las universidades estatales, se realiza una Prueba de Aptitud Académica (PAA), la cual evalúa habilidades generales de razonamiento en contextos matemáticos y verbales.

En la PAA si la persona aspirante para ingresar a la universidad se encuentra familiarizada con el hábito por la lectura y compresión de textos, podría reconocer ideas centrales y tener la capacidad de utilizar estrategias en el ámbito verbal para brindar una solución a los ítems (Brizuela-Rodríguez et al., 2018). Al ser la PAA una prueba estandarizada y al realizarse con los principios de equidad e igualdad en la medición de razonamiento, se desea que contemple una similitud léxico-semántica con el corpus de las obras literarias recomendadas en el Tercer Ciclo y Educación Diversificada del sistema educativo costarricense. Esto con el fin de que el lenguaje, redacción y léxico de los ítems tenga familiaridad con el vocabulario que la persona aspirante ha leído y comprendido en su proceso educativo.

CAPÍTULO III. OBJETIVOS

El estudio responde el planteamiento mediante el desarrollo de un análisis estadístico para abordar la similitud léxico-semántico de dos corpus lingüísticos aplicando técnicas de análisis de texto o minería de texto. Las preguntas de investigación que se plantean son las siguientes:

- ¿El corpus de las obras literarias recomendadas en el Tercer Ciclo y Educación Diversificada en el sistema educativo costarricense tiene similitud/asociación léxico-semántica con el corpus de los ítems de práctica de la Prueba de Aptitud Académica que aplica la Universidad de Costa Rica?
- ¿Cuál es el índice o grado de similitud/asociación léxico-semántica de ambos corpus lingüísticos?
- ¿Cuál es la técnica de minería de texto o modelo estadístico multivariante que tienen el mejor ajuste y rendimiento para asociar ambos corpus?

3.1 Objetivo general

Comparar la similitud léxico-semántica entre obras literarias latinoamericanas recomendadas en el Tercer Ciclo y Educación Diversificada en el sistema educativo costarricense con el corpus asociado a los ítems de práctica de la Prueba de Aptitud Académica que aplica la Universidad de Costa Rica, mediante técnicas estadísticas multivariantes de minería de texto, 2018 -2020.

3.2 Objetivos específicos

1- Definir un corpus lingüístico basado en las obras literarias latinoamericanas recomendadas en el Tercer Ciclo y Educación Diversificada en el sistema educativo costarricense.

2- Estimar modelos estadísticos para la consistencia y validez estadística y lingüística del corpus de las obras literarias mediante técnicas multivariantes para la clasificación y segmentación de datos.

3- Analizar la estructura léxico-semántica el corpus lingüístico basado en las obras literarias con el corpus asociado a los ítems de práctica de la Prueba de Aptitud Académica que aplica la Universidad de Costa Rica.

CAPÍTULO IV. MARCO DE REFERENCIA

4.1 Aspectos teóricos relevantes

En el presente capítulo se abordarán los conceptos generales y teóricos relacionados con el tema de este trabajo en el ámbito educativo. Inicialmente, el derecho a la educación, consagrado como pilar fundamental del desarrollo social, se materializa a través del dominio de competencias instrumentales por parte de la ciudadanía. Entre estas, la comprensión lectora y el razonamiento verbal emergen como habilidades críticas que posibilitan no solo el aprendizaje continuo, sino también la participación plena en la sociedad.

La comprensión lectora y el razonamiento verbal se encuentran articulados y evaluados por los sistemas educativos formales, tanto a través de los programas curriculares mediante pruebas estandarizadas que marcan hitos clave en la trayectoria del estudiantado. Este continuo evaluativo abarca desde mediciones de alcance internacional como el programa PISA hasta instrumentos de certificación nacional como las Pruebas Nacionales y de ingreso universitario como la Prueba de Aptitud Académica (PAA), todas las cuales presuponen y miden, en distintos grados, estas competencias lingüísticas fundamentales.

La trascendencia de estas habilidades se proyecta como un prerrequisito esencial para el éxito académico en la educación superior; sin embargo, el análisis evaluativo se ha centrado predominantemente en los resultados de rendimiento, dejando en un segundo plano la profundidad de la naturaleza y complejidad de los insumos textuales. Para subsanar esta brecha, esta investigación propone un abordaje innovador que integra los principios de la lingüística de corpus con técnicas de la minería de texto. Mediante la construcción y el análisis computacional de un corpus lingüístico representativo busca identificar patrones léxicos, sintácticos y semánticos objetivos, ofreciendo así una nueva dimensión para comprender y diagnosticar los desafíos en el proceso de enseñanza-aprendizaje y lingüísticos que enfrenta el estudiantado en su recorrido educativo.

En general, se abordan conceptos del ámbito educativo como la comprensión lectora, razonamiento verbal y pruebas estandarizadas nacionales e internacionales. Además, se describen los objetivos de los programas de la asignatura de español establecidos por el Ministerio de Educación Pública (MEP) y hallazgos de investigaciones sobre la comprensión lectora en la población estudiantil universitaria y lo relacionado a las aplicaciones de minería de texto.

4.1.1 Educación como derecho

Primeramente, es fundamental mencionar que para comprender los desafíos en la educación es necesaria la comprensión del origen de la educación como derecho humano. En el año 1948 la Organización de las Naciones Unidas (1948), en el artículo 26 inciso 1 de la Declaración Universal de Derechos Humanos, se establece lo siguiente:

> Toda persona tiene derecho a la educación. La educación debe ser gratuita, al menos en lo concerniente a la instrucción elemental y fundamental. La instrucción elemental será obligatoria. La instrucción técnica y profesional habrá de ser generalizada; el acceso a los estudios superiores será igual para todos, en función de los méritos respectivos.

Asimismo, la educación como derecho se encuentra considerada en múltiples convenciones, pactos, declaraciones y protocolos a saber:

- Declaración de los Derechos del Niño (Naciones Unidas, 1959)
- Convención sobre los Derechos del Niño (Naciones Unidas, 2006)
- Convención sobre la eliminación de todas las formas de discriminación contra la mujer (Naciones Unidas, 1979)
- Declaración de las Naciones Unidas sobre los Derechos de los Pueblos Indígenas (Naciones Unidas, 2007)

Muñoz y Díaz-Soucy (2021) expresan que el derecho de la educación se plantea como una política y una programática de educación basada en el acceso a la educación básica, a la calidad y a la no discriminación. Además, mencionan que esta educación requiere satisfacer la necesidad de oportunidad para el mercado laboral, la participación social y el ejercer una ciudadanía responsable.

El país refleja una realidad en el que el derecho humano a la educación no se ha podido asegurar. Existe un marco jurídico, políticas y programas educativos, sin embargo, se presentan dificultades para hacer cumplir ese derecho. Costa Rica muestra alcances en la cobertura y matrícula, pero se presentan limitaciones en el aprendizaje, el rendimiento escolar y la permanencia en el sistema educativo. Además, se presenta una situación en la que se debe valorar la pertinencia y la calidad de la educación costarricense.

4.1.2 Comprensión lectora

La comprensión lectora es un proceso complejo que va mucho más allá de la simple decodificación de palabras. Por un lado, involucra factores lingüísticos fundamentales y tareas como entender el significado literal de las oraciones. Sin embargo, su verdadera esencia reside en un nivel superior que implica captar el significado complementario del texto. Para lograrlo, la persona lectora debe llevar a cabo un procesamiento dinámico, en el cual integra activamente la nueva información con su bagaje previo de conocimientos, su experiencia personal y el contexto que lo rodea (Santiesteban y Velázquez, 2012).

En general, comprender un texto implica conocer el significado de las palabras que lo componen, entender el significado de las oraciones e interpretar las ideas e intenciones transmitidas por el texto. Según García (1993), los conocimientos previos de la persona lectora, los procesos de lectura (cognitivos, metacognitivos y lingüísticos), las características del texto (tipología, estructura, temática, dificultad) y los objetivos de la persona lectora (motivación, expectativas, propósitos) ante el texto son algunos de los factores asociados a la compresión.

El nivel de conocimiento de palabras y conceptos constituye una medida excelente del desarrollo cognitivo-lingüístico y del potencial de aprendizaje a lo largo de la vida escolar (Gardner, 2004; McKeown y Curtis, 1987). Durante los primeros años de vida del infante, la comprensión del mundo se convierte en la adquisición de palabras (Crais, 1990). En los cursos medios y superiores, el conocimiento de las palabras debe traducirse en el conocimiento del mundo (Puyuelo et al., 2000).

En Costa Rica, existen estudios relacionados a la evaluación del lenguaje y comprensión lectora en los diferentes niveles educativos desde preescolar hasta el nivel de secundaria. Murillo (2009, 2012) relaciona la evaluación del vocabulario en la educación preescolar y menciona que el enriquecimiento de la competencia lingüístico-comunicativa de las niñas y los niños es una responsabilidad del sistema educativo que procure elevar la calidad de vida de las personas, para esto se necesitan investigaciones que describan la producción lingüística del estudiantado en diferentes circunstancias tanto comunicativas como orales y escritas.

4.1.3 Razonamiento verbal

El razonamiento es una aptitud mental básica y forma parte de los componentes de la inteligencia general. Algunos de los instrumentos que miden el razonamiento lo hacen por medio de pruebas o exámenes donde se evalúa la capacidad o aptitud para resolver problemas lógicos, deduciendo ciertos resultados de lo que se plantea. También, con estas evaluaciones se pretende descubrir la capacidad de razonamiento y análisis, ya que existen factores mentales vinculados a la inteligencia general y a las habilidades cognitivas de las personas.

La capacidad de razonamiento verbal es descrita por Bennett et al. (1992) como la posibilidad de entender conceptos formulados en palabras. Más que simplemente conocer muchas palabras o expresarse bien, se refiere a la capacidad de conectar ideas, encontrar patrones y construir nuevos pensamientos a partir de la información. Por otra parte, Moreira-Mora (2021) indica que el razonamiento verbal se define como la capacidad para manejar el lenguaje verbal en el análisis semántico e inferencial cuando se leen diferentes textos.

Parte del razonamiento verbal se encuentra asociado con el hábito de leer y la comprensión lectora. De Mier et al. (2012), Arancibia-Gutiérrez y Leiva (2022) y Arancibia-Gutiérrez et al. (2022) indican que a medida que las niñas y los niños en los primeros años escolares alcancen un reconocimiento de las palabras y desarrollen mayores habilidades de procesamiento, la lectura será mayormente fluida y comprensiva. Por lo que a la hora de leer se tendría un uso apropiado para jerarquizar, organizar información y realizar operaciones inferenciales en los textos.

El razonamiento verbal se encuentra asociado a las características y contextos de la persona. Al respecto, Carreras et al. (2008) hallaron en su estudio con estudiantes universitarios diferencias estadísticamente significativas entre las habilidades de razonamiento verbal y abstracto, según sexo, edad y educación de los padres y madres, este último aspecto favorece a los que tienen mayor formación académica.

4.1.4 Programas de español del Ministerio de Educación Pública

Los programas de las diferentes asignaturas que ofrece el MEP en los diferentes niveles educativos se encuentran enmarcados bajo la Política Educativa y la Política Curricular.

Según el Ministerio de Educación Pública (2017), en el programa de estudio de español "comunicación y comprensión lectora", se establecen tres grandes componentes para ser implementados a través de la mediación pedagógica en el aula, a saber: el eje único "El ser humano se comunica de diversas formas y en contextos distintos como medio de convivencia en la sociedad nacional y global, aprovechando todo tipo de recursos" (MEP, 2017, p.30). El abordaje metodológico se enfoca en los criterios de evaluación transversales y los criterios de evaluación (específicos) para cada nivel. Además, para cada año escolar, se definen la tipología de textos por redactar, así como la tipología de técnicas de comunicación por desarrollar. Esta organización se hace acompañar, también, por la lista sugerida de textos

literarios, aprobada por el Consejo Superior de Educación, y la dosificación para su correspondiente lectura y la de textos no literarios.

El programa de español de Ministerio de Educación (2017) se fundamenta en la competencia comunicativa y pretende desarrollar en las personas estudiantes las competencias específicas como la lingüística, la sociolingüística y sociocultural, la discursiva o textual, la estratégica, la literaria o lectora y la semiológica. Según el Ministerio de Educación Pública (2017) en el programa de español la competencia lectora pretende estar inherente en las personas estudiantes de II y III ciclo y educación diversificada en los siguientes procesos:

1. Dominar diversas estrategias para abordar distintos tipos de texto.
2. Tomar conciencia de la realidad nacional y global con base en textos literarios representativos de la literatura costarricense y foránea.
3. Comprobar hipótesis construidas en diferentes contextos y tipos de textos.
4. Verificar y criticar el contenido de la lectura de textos no literarios y lo compara con sus propias prácticas, y ofrece juicios y opiniones sobre la veracidad, autenticidad, validez, confiabilidad y otras variables.
5. Asumir una actitud crítica frente a las manifestaciones de exclusión, para subsanarlas.
6. Tomar conciencia de su papel activo dentro de una visión local y global.

Ramírez (2018) realiza una investigación sobre la escritura y la lectura en el currículo de sétimo en el país, entre los resultados muestra que la propuesta curricular del programa de estudios de sétimo año carece de conocimientos didácticos que orienten las prácticas de escritura y lectura en secundaria. Ante esto, Ramírez (2018) propone ajustes para que la propuesta curricular permita al personal docente planificar una programación que brinde al estudiantado desarrollar competencias de lectura y escritura.

Asimismo, en el país existen aplicaciones de test para la validación de contenido léxico en textos para la comprensión lectora del test de Lectura y Escritura en Español (LEE). En el año 2013, se realizó una aplicación para adaptar este instrumento en la población

estudiantil de segundo, tercero y cuarto grado. Carpio y Méndez (2016) identificaron pocas palabras en los textos que generaban incertidumbre en la comprensión, esto por un asunto de léxico regional. En este caso, también se consideró el uso de sinónimos como ajuste del test.

En el Noveno Informe Estado de la Educación 2023, Murillo et al. (2023) evidencian que existe una pobreza de los aprendizajes en lectura y escritura en la población estudiantil de I y II ciclo. En ese mismo estudio 82% del personal docente que participó percibe que sus estudiantes presentan conocimientos deficientes en lectura y escritura. Además, recomiendan una atención en cuatro ámbitos curriculares que abarca la planificación de lectura en todos los niveles en el país como: desarrollo de competencias lectoras, promoción de la lectura, creación de hábitos de lectura y análisis de textos literarios y no literarios.

4.1.5 Lecturas recomendadas por el Consejo Superior de Educación

El Consejo Superior de Educación Pública (CSE) se crea en la Ley No 1362, en el artículo Nº 1, el cual indica: “Se crea el Consejo Superior de Educación Pública como órgano de naturaleza constitucional con personalidad jurídica instrumental y presupuesto propio, que tendrá a su cargo la orientación y dirección de la enseñanza oficial”. (Asamblea Legislativa, 1951, párr. 1).

Asimismo, en el artículo 8 de la Ley No 1362 (Asamblea Legislativa, 1951), establece que el CSE deberá aprobar lo relacionado a:

a) Los planes de desarrollo de la educación pública.

b) Los proyectos para la creación, modificación o supresión de modalidades educativas, tipos de escuelas y colegios, y la puesta en marcha de proyectos innovadores experimentales, ya se trate de la educación formal o la no formal.

c) Los reglamentos, planes de estudio y programas a que deban someterse los establecimientos educativos y resolver sobre los problemas de correlación e integración del sistema.

d) Los planes de estudio y los aspectos centrales del currículum y cualquier otro factor que pueda afectar la enseñanza en sus aspectos fundamentales.

El CSE al tener la facultad de aprobar lo relacionado a planes y programas de estudio, en la sesión N°36-2017, conoció y analizó el dictamen presentado por la Comisión de Planes y Programas sobre la propuesta de lista de lecturas recomendadas (no obligatorias) para el I, II y III ciclo de la Educación General Básica y Ciclo Diversificado.

En la sesión se estableció el acuerdo Nº 04-36-2017, el cual considera la promoción de la lectura como un tema que compete a todo el país y que debe comprenderse como un acto de afectividad, recreación, reflexión y de desarrollo de la comprensión y el pensamiento crítico (Consejo Superior de Educación, 2017). En el Anexo 1, se muestran el listado de lecturas recomendadas según los niveles educativos y la dosificación de texto a utilizar en los programas de la asignatura de español.

El análisis de la propuesta de lecturas recomendadas para el Tercer Ciclo, implementada en 2010, revela modificaciones significativas. Al respecto, el estudio de Arias y Vargas (2016) señala que los principales cambios se centraron en tres áreas: la dosificación de textos por nivel, la categorización de los géneros literarios y una mayor flexibilidad; la cual otorga al personal docente la potestad de seleccionar las obras para el estudiantado. No obstante, el mismo estudio concluye que una de las debilidades más notables de esta reforma es que dichas modificaciones carecen de una justificación pedagógica clara.

4.1.6 Pruebas estandarizadas y medición de la habilidad de la comprensión lectora y razonamiento verbal

A nivel internacional y nacional las habilidades y las aptitudes de la comprensión lectora son evaluadas en los sistemas educativos por medio de pruebas estandarizadas. Según The glossary of education reform (2015), se define que las pruebas estandarizadas son un tipo de prueba que exige que todas las personas participantes respondan las mismas preguntas, o una elección de preguntas de un banco común, de la misma forma, y se califica de manera "estándar" o uniforme, lo que facilita la comparación del desempeño relativo de las personas examinadas de manera individual o en grupo.

Las pruebas estandarizadas se vinculan principalmente con exámenes a mayor escala realizados a grandes grupos de personas. Adicionalmente, la estandarización asociada a la medición de las pruebas también se encuentra relacionada a un proceso administrativo uniforme en condiciones y procedimientos (Georgia Department of Education, 2008).

Tristán y Pedraza (2017) destacan algunas características principales de las pruebas estandarizadas entre ellas la especificidad, neutralidad, autonomía, equidad e impersonalidad. Estas son esenciales para interpretación de los resultados, erradicar o disminuir los prejuicios generados por el impacto de estereotipos y preferencias en el diseño de la prueba o en la valoración de las personas evaluadoras y otros elementos que pueden influir en el uso ético de los resultados de las pruebas.

En las pruebas estandarizadas comúnmente se evalúan aspectos de comprensión lectora, Gómez y Silas (2012) examinan el impacto de un programa para potenciar la comprensión lectora en el rendimiento del estudiantado de segundo grado de telesecundaria en el examen estandarizado Complejidad Lingüística Progresiva (CLP). En esta prueba se muestra que las y los estudiantes mejoraron y obtuvieron un avance en el desarrollo de la comprensión lectora en la sección de español.

Adicionalmente, en la Prueba de Aptitud Académica (PAA), para ingreso a la Universidad de Costa Rica, se evalúan habilidades de razonamiento general en contextos verbales y matemáticos. El razonamiento es la habilidad de los individuos para establecer conexiones entre ideas y llegar a conclusiones. En otras palabras, se refiere al proceso de formación de conclusiones que permite informar acerca de los esfuerzos hechos por el individuo en la solución de un problema y en la toma de decisiones que le facilita alcanzar un propósito (Leighton y Sternberg, 2004).

En este contexto de evaluación, las pruebas estandarizadas fungen como mecanismos esenciales para medir de forma objetiva en las personas examinadas su habilidad de la comprensión lectora y su razonamiento verbal para un fin determinado, asimismo contribuyen en la comprensión y mejora de las habilidades y razonamiento.

4.1.7 Programa para la Evaluación Internacional de Estudiantes

El Programa para la Evaluación Internacional de Estudiantes (PISA, siglas en inglés) es un programa que forma parte de la Organización para la Cooperación y el Desarrollo Económicos (OCDE). En los estados miembros se realiza una prueba que tiene como objetivo evaluar los conocimientos y habilidades necesarios que ha adquirido la población estudiantil que se encuentra por finalizar la educación obligatoria, para que pueda ejercer una participación plena en la sociedad del saber (OCDE, 2021).

El PISA mide lectura, matemáticas y ciencia, que van más allá de la memorización. El énfasis está en ver si el estudiantado puede pensar por sí mismo, entender el porqué de las cosas y aplicar su conocimiento para resolver desafíos prácticos en diferentes contextos. Según el Informe de resultados PISA 2009 de la OCDE (2010), se entiende por competencia lectora la capacidad de un individuo para comprender, utilizar y reflexionar sobre textos escritos, con el propósito de alcanzar sus objetivos personales, desarrollar su conocimiento y sus capacidades y participar en la sociedad.

La prueba del PISA agrupa en dos categorías de textos: los textos en prosa continua (narración breve, una nota periodística o una carta) y los textos en prosa discontinua (párrafos separados por imágenes, diagramas y espacios, como manuales de operación de aparatos, textos publicitarios, argumentaciones científicas, entre otros) (OCDE, 2010). En este sentido, la evaluación de la competencia lectora no se centra en una noción del texto literario y se ocupa de una variedad considerable de textos propios de las diferentes circunstancias que puede enfrentar una persona en su vida cotidiana. Las competencias cognitivas que se evalúan a nivel textual son la capacidad para recuperar información, para inferir nueva información a partir de la lectura realizada, para relacionar los contenidos leídos con otros y realizar una reflexión derivada de ellos (OCDE, 2010).

Costa Rica es estado miembro de la OCDE, por lo que forma parte de la evaluación que se le realiza a la población estudiantil. Según BID (2019) los resultados de la prueba PISA del 2018, Costa Rica se ubica en la posición 49 de 77 países pertenecientes a la OCDE. En la competencia lectora se obtuvo una tasa de bajo desempeño en lectura, un 42% (BID, 2019). Costa Rica obtuvo promedios generales en la prueba, sin embargo, descendió un punto en el área de lectura (Rodríguez, 2019).

Ante este panorama, resulta importante destacar los resultados para indagar los factores subyacentes que inciden en el desempeño de la competencia lectora. Estudios académicos han identificado múltiples variables asociadas, destacando la importancia del entorno sociocultural y el acceso a recursos educativos en el hogar. Es precisamente en este ámbito donde el análisis de factores como el capital cultural adquiere una relevancia explicativa fundamental.

Por otro lado, Montero et al. (2014) encuentran asociación mediante un modelo de regresión entre el puntaje de la prueba PISA 2012 y número de libros en el hogar de la persona estudiante. Un número elevado de libros en el hogar forma parte de los factores del perfil de estudiantes con alto rendimiento en la competencia lectora.

4.1.8 Pruebas Nacionales del Ministerio de Educación Pública

En el año 2019 con el decreto N°41686-MEP se le da la facultad al Ministerio de Educación Pública para realizar las Pruebas Nacionales para el Fortalecimiento de Aprendizajes para la Renovación de Oportunidades (FARO), con el fin de proponer una práctica evaluativa continua, que permita la trazabilidad de la información en el desarrollo de los procesos y proyectos adscritos a la implementación de la política educativa y, además, tienen como objetivo determinar el nivel de logro de los aprendizajes y las habilidades esperadas por el estudiantado por concluir el II Ciclo de la Educación General Básica y la Educación Diversificada (La Gaceta Diario Oficial, 2019). Las Pruebas Nacionales FARO miden la habilidad lectora con base en textos literarios, a partir de la lista de lecturas recomendadas en el año 2018 (acuerdo del Consejo Superior de Educación N°04-36-2017) (Ministerio de Educación Pública, 2021).

Según el Ministerio de Educación Pública (s.f), se menciona que el modelo de medición utilizado para analizar los resultados de la prueba FARO es el modelo con referencia a criterios, en este se evalúa el estado de cada persona examinada según con el dominio de las habilidades definidas en los Programas de Estudios de acuerdo con el año escolar. Además, el MEP indica que el modelo de evaluación de la prueba que se utilizó fue la Teoría de Respuesta a los Ítems (TRI), pues permite obtener mediciones más estables y que no varían en función del instrumento empleado.

De acuerdo con el MEP (s.f), en la información técnica de los resultados de las Pruebas FARO, en la evaluación de la prueba de escritura de la materia de español no se lleva a cabo un proceso para determinar puntos de corte, ya que las rúbricas empleadas para puntuar los textos redactados por el estudiantado incluyen la diferenciación y descripción de estos niveles. Asimismo, los indicadores contemplados en la sección de redacción se organizan en tres dimensiones (discursiva, textual y convenciones de legibilidad), cada una de ellas se segmenta en tres niveles de rendimiento. De acuerdo con el rendimiento de cada estudiante en los distintos indicadores, se le otorgará un nivel en cada dimensión (Ministerio de Educación Pública, s.f).

En el año 2022, el MEP dio a conocer los resultados de la prueba FARO, en la asignatura del español aplicaron 66 587 estudiantes. La prueba se dividió en dos partes: lectura y escritura. Para la lectura se establecieron tres niveles de desempeño, sin embargo, en el nivel 2 se establecieron dos subniveles (2A y 2B). Y en la parte de escritura se establecen tres niveles en las dimensiones: textual y convenciones de legibilidad (Ministerio de Educación Pública, 2022).

Específicamente en la parte escrita de la prueba, en la dimensión textual contiene varios elementos de la estructura interna del texto escrito. Estos corresponden a las categorías que aportan congruencia y claridad a los textos, por ejemplo: la concordancia (nominal y verbal), la cohesión (progresión articulada de las oraciones), la coherencia (la unidad temática sostenida) y la estructura de párrafos (delimitación entre la idea fundamental y las complementarias) (Ministerio de Educación Pública, 2022).

Por otro lado, las convenciones de legibilidad agrupan las diversas regulaciones de la comunicación escrita. Se comprende que el lenguaje escrito tiene como objetivo transmitir y, con este propósito, las palabras redactadas de manera incorrecta, la falta o uso incorrecto de tildes o letras mayúsculas, junto con una puntuación inadecuada, modifican el significado o sentido de lo que se busca comunicar (Ministerio de Educación Pública, 2022).

Entre los principales resultados se muestra que en el nivel 3 se encuentra 0,79% del estudiantado, en el nivel 2A el 68,21% y en el nivel 2B el 30,97%. Solo 0,03% del estudiantado se encuentra en el nivel 1. En la parte de escritura de la prueba, en la dimensión textual, se obtuvo que el 55,30% del estudiantado evaluado se encuentra en el nivel 3; 40,67% en el nivel 2 y 4,03% en el nivel 1. En la dimensión de convenciones de legibilidad 29,53% se ubica en el nivel 3; 62,50% en el nivel 2 y 7,97% en el nivel 1 (Ministerio de Educación Pública, 2022).

En general, según los resultados del Ministerio de Educación Pública (2022) los niveles de logro se centran el nivel 2A, lo cual evidencia que la persona estudiante demuestra

un nivel intermedio alto de la comprensión y análisis de textos respecto a los aprendizajes esperados. En el nivel 3 establece que la persona estudiante demuestra un nivel avanzado de la comprensión y análisis de textos respecto a los aprendizajes esperados (Ministerio de Educación Pública, 2022). Por lo que, se podría indicar que en general, el estudiantado ante la prueba FARO tiene un nivel intermedio alto o avanzado en la comprensión y análisis de textos (Ministerio de Educación Pública, 2022).

En la reciente Prueba Nacional Estandarizada Diagnóstica, el MEP indicó que no existen diferencias en los niveles de logros entre los centros educativos públicos y privados en zona urbana y rural (Martínez, 2023). La prueba mostró que 62,4% de estudiantes de primaria se encuentra en nivel intermedio o avanzado de conocimientos y habilidades para resolver problemas y emplear el pensamiento crítico (Martínez, 2023). A nivel de secundaria, el 50% el estudiantado se ubica en los niveles intermedio y avanzado (Martínez, 2023).

En el marco de especificaciones para la Prueba Nacional Estandarizada 2023, en secundaria, el componente "Español" contiene el bloque llamado "Comprensión lectora y análisis literario", el Ministerio de Educación Pública (2024) lo conceptualiza de la siguiente manera:

> Requiere la capacidad de construir los sentidos y significados de lo leído y establecer una reflexiva y crítica interacción entre las ideas del lector y las vertidas por la voz discursiva de los textos literarios y no literarios. Todo esto a la luz de las cuatro fases establecidas en el Programa de Estudio vigente de Español: fase natural, de ubicación, analítica e interpretativa y explicativa (MEP, 2024, p. 4).

El Ministerio de Educación Pública (2024) indica que la meta de la comprensión lectora y análisis literario es conseguir que las personas sean alfabetizadas funcionales; esto es, que mediante la interpretación y estudio de los textos, adquieran destrezas para manejar la vida diaria.

4.1.9 Prueba de Aptitud Académica de la Universidad de Costa Rica

La Prueba de Aptitud Académica (PAA) de la Universidad de Costa Rica (UCR) forma parte de las formas de ingreso del proceso de admisión de la Universidad. La PAA inicia en el año 1957, elaborada por el Comité de Evaluación y esta prueba fue implementada y aplicada por primera vez en el año de 1960, en este año se constituye el Instituto de Investigaciones Psicológicas (IIP) que es la instancia responsable del proceso de la prueba desde su diseño, aplicación y evaluación.

Según el artículo No 1 del Reglamento del Proceso de Admisión Mediante Prueba de Aptitud Académica (La Gaceta Universitaria, 2003, párr. 1), se indica:

> La Universidad de Costa Rica tiene como naturaleza consustancial el logro de sus principios, propósitos y funciones, por lo que el proceso de admisión mediante la Prueba de Aptitud Académica (PAA) se conducirá según los siguientes principios:
>
> - Principio de igualdad de oportunidades a todas las personas interesadas en seguir estudios universitarios.
> - Principio de rectitud, transparencia y justicia en el proceso de admisión.
> - Principio de excelencia académica.
> - Principio de óptima utilización de los recursos con que cuenta la Universidad, en beneficio de la sociedad costarricense.

La PAA de la UCR evalúa habilidades generales de razonamiento en contextos matemáticos y verbales, en este último el desarrollo de hábitos por la lectura es una estrategia de preparación para la prueba (PPPAA-IIP, 2025b). A la persona aspirante a la prueba se le evalúa su capacidad para utilizar material verbal mediante el uso de las estrategias requeridas para resolver los ítems como: suponer, presuponer, parafrasear, oponer, deducir, reducir, generalizar, verificar, indagar y representar (PPPAA-IIP, 2025b).

La PAA se encuentra constituida de ítems de razonamiento en contexto verbal y matemático, sin embargo, se mide un solo constructo, esto forma parte de las evidencias de validez de la PAA en la estructura interna indicadas por Rojas (2014), el cual indica que se evidencia unidimensionalidad del constructo. En la investigación de Montero et al. (2015), en relación con el efecto de los puntajes de la PAA, se concluye que un buen nivel de comprensión lectora potencia o incrementa la capacidad de mejorar en los puntajes.

En un estudio realizado por Calvo et al. (2019), se identificaron elementos irrelevantes para la comprensión de ítems en la PAA. Entre las discusiones que se encuentra el uso del reporte verbal como generador de evidencias de contenido para la validez de las puntuaciones de la prueba de admisión. La retroalimentación de los participantes constituye un recurso de gran valor para el refinamiento de los ítems de una prueba, particularmente en lo que respecta a la detección de dificultades léxicas. En línea con esta idea, el estudio de Calvo et al. (2019) demuestra que el uso de reportes verbales es una técnica sumamente eficaz. Permite, por un lado, comprender los procesos de razonamiento que emplean para resolver un problema y, por otro, identificar con precisión las barreras de vocabulario que enfrentan. Este último punto es crucial, dado que existe una incertidumbre inherente sobre el léxico real que domina el estudiantado en el rango etario de 17 a 19 años.

4.1.10 Comprensión lectora en el ámbito universitario

En Costa Rica, existen algunos estudios relacionados a la comprensión lectora en las personas estudiantes universitarias, lo cual son investigaciones relevantes para conocer los niveles de comprensión de la población estudiantil que ingresa a la educación superior luego de su proceso educativos a nivel de secundaria. La comprensión lectora en la formación de los niveles de educación superior se encuentra inherente en el desempeño y rendimiento universitario, esto desde su ingreso a una carrera universitaria, en el rendimiento académico de los cursos de la malla curricular de carrera hasta su graduación e incorporación al mercado laboral.

En un estudio del Instituto Tecnológico de Costa Rica (ITCR) realizado por Abarca-Petitiean y Romero-Zúñiga (1991), entre los principales resultados se muestra que existe una

alta variabilidad en el nivel de comprensión lectora entre las personas estudiantes del ITCR, particularmente en la capacidad de resolución como en el tiempo para resolver las pruebas. Además, se menciona que la población estudiantil estudiada carece de estrategias efectivas de lectura por lo que su habilidad se ubica en un nivel bajo.

Del mismo modo, Corrella (2018) realiza una investigación relacionada a las habilidades de comprensión lectora y las características de producción de textos escritos de la población estudiantil de primer año de la carrera de Ingeniería en Computación del Instituto Tecnológico de Costa Rica en una Sede Regional, con el fin de brindar una propuesta didáctica. Entre los principales hallazgos se menciona que la población estudiantil presenta un interés por la lectura de noticias y temáticas asociadas a la tecnología. Sin embargo, en la investigación se identificó que el estudiantado no emplea estrategias para la comprensión de textos y, específicamente en los documentos académicos, se caracterizaron por carecer de informatividad, cohesión y coherencia.

En otro estudio realizado por Regueyra y Arguello (2018), sobre los mitos de la comprensión lectora en el estudiantado universitario, mencionan que el proceso de incorporación de la comprensión lectora y la habilidad de escribir en la población universitaria se presenta como una tarea que involucre las instancias universitarias y esta debe ser parte del quehacer académico inclusivo, con el fin de ofrecer una universidad comprometida con la sociedad. Además, indican que la comprensión lectora forma parte integral de la lectoescritura y debe ser analizada desde una perspectiva económico y social que reproduce desigualdad social y cultural.

Por otro lado, otras personas autoras como Brizuela-Rodríguez et al. (2021) concluyen en la investigación que el estudio y el mejoramiento de la comprensión lectora requieren tomar en cuenta mecanismos cognitivos de dominios general y específico, los cuales están en la base de las fortalezas y debilidades que se manifiestan en los instrumentos de evaluación tradicionales de la comprensión lectora. También mencionan que la técnica de seguimiento ocular aplicada al proceso de la lectura es útil para identificar las dificultades al comprender textos escritos.

Actualmente, la Universidad de Costa Rica (UCR) en conjunto con la Universidad del Valle de Guatemala (UVG) crearon la Maestría Académica en Investigación con énfasis en Lectoescritura Inicial (LEI). Este programa de posgrado pretende formar personas profesionales con conocimientos y habilidades para la búsqueda científica relacionada con la lectoescritura inicial. Lo cual es de relevancia para el país y la región centroamericana pues la investigación en comprensión lectora se debe promover y que esta genere conocimiento para la toma de decisión en los procesos de lectoescritura para las autoridades educativas competentes.

4.1.11 Corpus lingüístico y minería de textos

Según Rojo (2014), un corpus es "un conjunto o fragmentos de textos naturales, almacenados en formato electrónico, representativos en su conjunto de una variedad lingüística, en alguno de sus componentes o en su totalidad, y reunidos con el propósito de facilitar su estudio científico".

La semántica léxica es el estudio de los significados de las palabras y los significados de las palabras no pueden observarse directamente (Taylor, 2017). Estos significados se encuentran en la mente de cada persona hablante y, posiblemente, ni siquiera es accesible mediante la introspección, mucho menos susceptibles de validación interpersonal (Taylor, 2017).

La minería de datos se encuentra asociada con el Descubrimiento de Conocimiento en Datos (KDD, siglas en inglés) y es el proceso significativo donde se descubren correlaciones, patrones y tendencias en grandes cantidades de datos utilizando técnicas matemáticas y estadísticas. Sin embargo, en los últimos años la mayor parte de la información se encuentra en formato textual y en particular se define el proceso de Descubrimiento de Conocimiento en Textos (KDT, siglas en inglés).

Según Justicia de la Torre (2017), menciona que el KDD y el KDT son similares en desarrollo, pero existe una diferencia en la falta de estructura en las fases de selección, preprocesamiento y transformación de los datos textuales. Ahora bien, al aplicar el mismo concepto de minería de datos, se podría deducir que cuando se está en presencia de un gran volumen de datos textuales y se aplican técnicas matemáticas y estadística para la identificación de relaciones o patrones se está aplicando la minería de textos.

Existen diversas aplicaciones de la minería de texto, de acuerdo con Justicia de la Torre (2017) se tiene utilidad en las áreas de *web* semántica, redes sociales, filtrado de correos electrónicos, personalización de perfiles *web*, análisis de sentimiento, métodos de síntesis y organización, *marketing* y comercios electrónico, *E-learning* y aplicaciones *Help Desk.*

A nivel internacional se han realizados estudios, en los que se evalúa vocabulario por medio de la lingüística del corpus. Según el estudio de Medellín y Rodríguez (2014), realizan una propuesta metodológica para la evaluación de vocabulario académico a través de una prueba de decisión léxica. Entre los resultados mencionan que el número de palabras que los participantes conocían se incrementó significativamente en relación con las materias que habían cursado y los semestres.

Por otro lado, Vilariño et al. (2014) desarrollan un modelo para resolver el problema de similitud semántica entre textos de diferente longitud, en los resultados encuentran similitud semántica entre párrafo-sentencia y sentencia-frase pero la metodología de expansión propuesta para detectar el grado de similitud semántica entre los pares frase a palabra y palabra a sentido no fue correcta.

Existe múltiples usos y aplicaciones en la minería de textos, no se referencian estudios asociados a la similitud de corpus de textos literarios en programas educativos con ítems de pruebas estandarizadas, por lo que este estudio parece ser innovador en la aplicación de la minería de texto ante estos contextos.

4.2 Análisis de similitud de documentos textuales

El análisis de similitud de documentos es útil para identificar las similitudes textuales que tiene un texto con respecto a otro. En este estudio, se propone emplear un escalamiento multidimensional y luego considerar un análisis semántico latente (ASL), ya que por medio de variables latentes se compacta la variabilidad textual de los documentos.

4.2.1 Escalamiento Multidimensional

En el escalamiento métrico, las diferencias entre cada uno de los elementos son valores definidos y se estiman a través de distancias. De este modo, los elementos se representan en un entorno métrico, las distancias se estiman y se contrastan con las diferencias. Luego, los elementos se reubican buscando que el ajuste sea lo mayormente preciso, hasta alcanzar la minimización de alguna función de costo (De Leeuw, 2011).

Siguiendo lo indicado por De Leeuw (2011), comúnmente se aplica la función de costo de mínimos cuadrados en las distancias, con el fin de medir el ajuste entre las diferencias y las distancias mediante la función *stress*, denotada como:

$$stress\ (X) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\delta_{ij} - d_{ij}(X))^2$$

Al minimizar la función *stress*, se obtiene la función *strain*, utilizada para en el algoritmo del escalamiento multidimensional.

4.2.2 Análisis Semántico Latente

El análisis semántico latente (ASL) es una técnica del procesamiento de lenguaje natural, la cual explora las relaciones entre un conjunto de documentos y los términos que

contienen (Gomede, 2024). En esta técnica se identifican patrones y conexiones entre los términos para comprender su significado y similitud.

En este análisis se construye una matriz con cada uno de los términos con el recuento en cada documento y se aplica la descomposición en valores singulares (DVS) para la reducción dimensional de los términos manteniendo la estructura de los documentos. Luego, los documentos se comparan mediante una medida de similitud textual, usualmente la similitud de coseno. Según Dumais (2005), en el ASL se siguen ciertos pasos, inicialmente la conformación de espacio de vectores de modelo mediante la matriz término-documento y la transformación de esta matriz. Posteriormente, la reducción del espacio mediante DVS para obtener un espacio reducido, en el cual los documentos como los términos se representan como vectores en el mismo espacio, las similitudes entre documentos, entre términos y términos-documento.

Al considerar lo expuesto por Dumais (2005), sea X una matriz de $t\ x\ d$, de términos y documentos. Usando la DVS, se obtiene que:

$$X = T^* S^* D^T$$

Donde T es una matriz $t\ x\ r$, con columnas ortonormales y D es una matriz $d\ x\ r$ ortonormal y S es una matriz diagonal $r\ x\ r$ con entradas ordenadas de forma descendente. El ASL utiliza una DVS truncada, manteniendo los mayores k valores singulares y sus vectores asociados, representada como:

$$X \approx T_k^* S_k^* D_k^T$$

La reducción trata de brindar la mejor aproximación por mínimos cuadrados a X con k parámetros y es lo que el ASL utiliza para su espacio semántico. Las filas de T_k son los vectores de términos en el espacio ASL y las filas de D son los vectores de documentos en el espacio. Las similitudes documento-documento, término-término y término-documento se calculan en la aproximación dimensional reducida a X.

4.3 Medidas de similitud textual

Para examinar la similitud textual de los documentos y los corpus lingüísticos que describe la proximidad semántica de las palabras textuales se consideran las medidas: la similitud de coseno y la similitud de Jaccard o índice de Jaccard.

4.3.1 Similitud de coseno

En el cálculo de esta medida de similitud los documentos o corpus se representan por medio de vectores de términos, la similitud corresponde a la correlación entre los vectores. En este se cuantifica el coseno del ángulo entre los vectores, en otras palabras, la similitud de coseno. Esta medida se utiliza comúnmente en la aplicación de análisis de texto en el procesamiento del lenguaje natural (Abbasi y Berrar, 2025).

Al seguir la representación matemática y al ser lo indicado por Huang (2008), la similitud de coseno se define de la siguiente forma: sean dos documentos d_a y d_b , y los vectores de los términos $\vec{t}_a$ y $\vec{t}_b$, la similitud viene dada por:

$$SC(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a \cdot \vec{t}_b}{|\vec{t}_a| \times |\vec{t}_b|}$$

En donde $\vec{t}_a$ y $\vec{t}_b$ son vectores m-dimensionales sobre el conjunto de términos $T = \{t_1, \dots, t_m\}$, en este caso cada dimensión representa un término con su peso en el documento, siendo de magnitud no negativa. Por lo que, la similitud de coseno es no negativa acotada entre $[0,1]$. Al final, esta métrica mide el coseno del ángulo entre dos vectores n-dimensionales proyectados en un espacio multidimensional. Si el valor de similitud de coseno es 1, significa que los dos vectores tienen la misma orientación. Un valor cercano a 0 indica una menor similitud entre los dos documentos. Además, una de las propiedades de la similitud de coseno es su independencia de la longitud del documento.

4.3.2 Similitud de Jaccard o índice de Jaccard

El índice o coeficiente de Jaccard mide la similitud como la intersección dividida por la unión de los documentos de texto, este compara la suma del peso de los términos compartidos en la intersección de los términos con la suma del peso de la unión de los términos que los dos documentos excluyen los términos que no son compartidos. Considerando, lo definido por Huang (2008), la expresión matemática se define como:

$$SJ(\vec{t}_a, \vec{t}_b) = \frac{\vec{t}_a \cdot \vec{t}_b}{\left|\vec{t}_a\right|^2 + \left|\vec{t}_b\right|^2 - \vec{t}_a \cdot \vec{t}_b}$$

La similitud de Jaccard oscila entre 0 y 1. Si dos documentos son idénticos, la puntuación de similitud de Jaccard es 1. Si no hay palabras en común entre ambos documentos, la puntuación de similitud de Jaccard es 0.

4.4 Modelos o clasificadores aplicados al análisis de texto

Según Munzert et al. (2014), los modelos estadísticos de clasificación comúnmente utilizados en minería de texto como ingenuo de Bayes, máquinas de soporte vectorial o bosques aleatorios, máxima entropía y asignación latente de Dirichlet. Seguidamente se describen los modelos aplicados.

4.4.1 Ingenuo de Bayes o Naive Bayes

Es un clasificador probabilístico simple con fuerte suposición de independencia. Este clasificador aprende de los datos de entrenamiento y luego predice la clase de la instancia de prueba con la mayor probabilidad posterior.

Matemáticamente según lo desarrollado por Jurafsky y Martin (2024), Naive Bayes es un clasificador probabilístico, lo que significa que para un documento *d*, de todas las clases $c \in C$ el clasificador devuelve la clase *c* que tiene la máxima probabilidad posterior dado el

documento. Se utiliza para referirse a «la estimación *argmax* de la clase correcta», y usamos *argmax* para referirnos a una operación que selecciona el argumento (en este caso la clase *c*) que maximiza una función (en este caso la probabilidad $P(c|d)$*).*

$$\hat{c} = \underset{c \epsilon C}{argmax}\ P(c|d)$$

Esta idea de la inferencia bayesiana se conoce desde los trabajos de Bayes (1763), la inferencia bayesiana, y fue aplicada por primera vez a la clasificación de textos por Mosteller y Wallace (1964). La intuición de la clasificación bayesiana consiste en utilizar la regla de Bayes para transformar la ecuación anterior en otras probabilidades que tengan algunas propiedades útiles. La regla de Bayes se puede descomponer en cualquier probabilidad condicional $P(x|y)$ en otras probabilidades:

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)}$$

Realizando la sustitución de la primera ecuación con esta descomposición se obtiene:

$$\hat{c} = \underset{c \epsilon C}{argmax}\ P(c|d) = \underset{c \epsilon C}{argmax}\ \frac{P(d|c)\,P(c)}{P(d)}$$

Se puede simplificar eliminando el denominador $P(d)$. Es posible para cada clase y se tiene que $P(d)$ no cambia para cada clase; siempre requiere la clase más probable para el mismo documento d, que debe tener la misma probabilidad $P(d)$. Por tanto, se puede elegir la clase que maximice esta fórmula de manera más sencilla, se denota:

$$\hat{c} = \underset{c \epsilon C}{argmax}\ P(c|d) = \underset{c \epsilon C}{argmax}\ P(d|c)\,P(c)$$

El clasificador de Naive Bayes se puede apreciar como una especie de suposición implícita sobre cómo se genera un documento, en donde primero se muestrea una clase a partir de $P(c)$ y luego se generan las palabras muestreando a partir de $P(d|c)$.

Para el análisis, se supone en primera instancia que la ubicación de la palabra al inicio o al final de texto tiene el mismo efecto en la clasificación, porque se asumen características $f_1, f_2, \ldots, f_n$ solo codifican la identidad de la palabra y no la posición. Por otra parte, se aplica la hipótesis ingenua de Bayes en la que se trata de la hipótesis de independencia condicional de que las probabilidades $P(f_i|c)$ son independientes dada la clase *c* y, por lo tanto, se pueden multiplicar «ingenuamente» de la siguiente manera:

$$P(f_1, f_2, \ldots, f_n\,|c) = P(f_1|c) \ldots P(f_n|c)$$

La ecuación final para la clase elegida por un clasificador ingenuo de Bayes se denota como:

$$c_{Naive\ Bayes} = \underset{c \epsilon C}{argmax}\ P(c) \prod_{f \epsilon F} P(f|c)$$

Para aplicar el clasificador Bayes ingenuo al análisis de texto, en este caso se utiliza cada palabra de los documentos como característica, y se considera que cada una de las palabras del documento recorre un índice por cada posición de palabra (conjunto *P*, posición de las palabras) en el documento, matemáticamente se tiene lo siguiente:

$$c_{Naive\ Bayes} = \underset{c \epsilon C}{argmax}\ P(c) \prod_{i \epsilon P} P(w_i|c)$$

4.4.2 Máquinas de soporte vectorial o Support vector machine (SVM)

Es uno de los más comunes en modelos de supervisión. Las máquinas de soporte vectorial fueron diseñadas para tratar con problemas binarios, sin embargo, la modificación

de la función objetivo de tal manera que, simultáneamente, permite el cálculo de un clasificador multiclase (Schölkopf y Smola,1998, p. 233). Este modelo actualmente es uno de los clasificadores más conocidos y más comúnmente aplicados en el aprendizaje supervisado (Munzert et al., 2014).

Seguidamente, se aborda el desarrollo matemático del modelo de máquinas de soporte vectorial. Según lo indicado por Isa et al. (2008), se considera el problema de separar un conjunto de vectores de entrenamiento en 3D pertenecientes a diferentes clases y se desea separar el conjunto de datos por medio de un hiperplano.

$$\boldsymbol{w}.\boldsymbol{x} + b = 0$$

Existe un número infinitos de hiperplanos de dividir el conjunto de datos en dos conjuntos, sin embargo, se pretende con las máquinas de soporte vectorial que solo habrá un hiperplano optimo donde se encuentra un margen máximo entre cada clase en que se divide el conjunto de datos. Las observaciones más cercanas determinan el hiperplano de separación óptimo, hay una manera de representarlos en el conjunto dado de puntos de entrenamiento, el margen máximo se encuentra minimizando, lo siguiente:

$$min\left\{\frac{1}{2}\ |\boldsymbol{w}|^2\right\}$$

Al minimizar la expresión anterior se encuentra el hiper plano óptimo bajo la restricción:

$$y_i.\,(x_i.\,\boldsymbol{w} + b) \geq 1\ , \forall\ i$$

El concepto de hiperplano de separación óptimo puede se puede generalizar para el caso no separable introduciendo un coste por violar las restricciones de separación y este coste puede hacerse introduciendo variables de holgura positivas en la restricción

$$y_i.(x_i.\boldsymbol{w}+b) \geq 1 - \xi_i, \forall\, i$$

Si se produce un error, el ξ_i correspondiente debe superar la unidad, por lo que la sumatoria de las variables holguras es un límite superior para el número de errores de clasificación. Por lo tanto, una forma lógica de asignar un coste adicional a los errores es cambiar la función objetivo que debe minimizarse a:

$$min\left\{\frac{1}{2}\,|w|^2 + C.\left(\sum\nolimits_i \xi_i\right)\right\}$$

Donde *C* es un parámetro elegido, un *C* mayor corresponde a asignar una mayor penalización a los errores de clasificación. Minimizando y bajo la restricción se obtiene el Hiperplano de Separación Óptimo Generalizado. Se trata de un problema de programación cuadrática, que puede resolverse aquí utilizando el método de los multiplicadores de Lagrange.

Después de realizar los cálculos necesarios, el problema de programación cuadrática se puede resolver mediante la búsqueda de los multiplicadores de Lagrange, *i*, que maximiza la función objetivo en:

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T x_j)$$

Sujeto a las restricciones:

$0 \leq \alpha_i \leq C\,, i = 1, \ldots, n$ y $\sum_{i=1}^{n} \alpha_i y_i = 0$

Se deduce a una nueva función objetivo que está en términos de los multiplicadores de Lagrange, α_i solamente. Esto se conoce como el problema dual, pues si se conoce $\boldsymbol{w}$, se conocen todos los α_i, si se conocen todos los α_i, se conoce $\boldsymbol{w}$. Algunos de los α_i son cero, y, por lo tanto, $\boldsymbol{w}$ es una combinación lineal de un número pequeño de puntos del conjunto de datos.

Sea $t_j \; con \; j = 1, \dots, s$ los índices de los soportes vectoriales, por lo que se denota la combinación lineal como:

$$w = \sum_{j=1}^{s} \alpha_{t_j} y_{t_j} x_{t_j}$$

Y se introduce la función *kernel*, como:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

En esta función no es necesario conocer explícitamente ϕ ya que el problema de optimización se puede traducir directamente a la versión más general del núcleo de la forma:

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, K(x_i, x_j)$$

Sujeta a:

$$C \geq \; \alpha_i \; \geq 0 \, , \sum_{i=1}^{n} \alpha_i y_i = 0$$

Este clasificador explora una representación espacial de los datos, en que las ocurrencias de términos se estructuran en las matrices término-documento y representan las ubicaciones espaciales de los documentos en espacios de alta dimensión. Con las máquinas de soporte vectorial se intenta ajustar los vectores entre las características de los documentos que mejor separan los documentos en los distintos grupos. En concreto, se seleccionan los vectores de forma que maximicen el espacio entre los grupos. Luego de la estimación se pueden clasificar nuevos documentos comprobando los vectores que caractericen los documentos y estimar la pertenencia.

4.4.3 Bosques aleatorios o Random Forest

Este clasificador crea múltiples árboles de decisión y toma la categoría predicha con mayor frecuencia de múltiples árboles de decisión como la clasificación que es más probable que sea precisa.

Matemáticamente desarrollando el modelo de bosques aleatorios y siguiendo los propuesto por Huang et al. (2021), se describe a continuación el modelo de bosques aleatorios.

Se supone un conjunto $V=\{x_1,\cdots,x_p\}$ de variables categóricas de entrada y una salida categórica y . Dada una muestra de entrenamiento S de n observaciones conjuntas de $x_1,\cdots,x_p$, y extraídas de $P=\{x_1,\cdots,x_p,\ y\}$, definamos para cualquier nodo interno t de un árbol de decisión construido a partir de S. Además, se define lo siguiente:

- El número de muestras de entrenamiento en t como n_t
- La proporción de muestras de entrenamiento en t como $p_r(t) = n_t/n$
- La impureza del nodo t como $i_p(t) = H(y|t)$
- La reducción de impurezas en el nodo t como $\Delta i_p(t) = i_p(t) - \left(\frac{n_{tL}}{n}\right) i_p(t_L) - \left(\frac{n_{tR}}{n}\right) i_p(t_R)$

En donde los subíndices L y R son el nodo izquierdo y el nodo derecho del nodo t. En un conjunto de árboles de decisión la importancia de la disminución media de la impureza de una variable de entrada x_m es la suma de las reducciones de impureza ponderadas $p_r(t)\Delta i(t)$, para todos los nodos t en los que se utiliza x_m, calculada como la media de todos los n_t árboles del conjunto, en el que se denota como:

$$Imp(x_m) = \frac{1}{n_T} \sum_{T_S} \sum_{t \in T_S} \sum_{v(s_t)=x_m} p_r(t)\Delta i_p(s_t, t)$$

Y en donde T_S es una estructura de árbol que representa un modelo de entrada-salida y $v(t)$ se adopta para dividir el nodo t. Un árbol de decisión completamente establecido y aleatorizado es aquel en el que cada nodo t se divide mediante una variable x_{iRF} seleccionada uniformemente al azar (de entre los nodos que no se han utilizado en los nodos padre) en $|\aleph_{iRF}|$ subárboles (es decir, uno por cada valor posible de $\aleph_{iRF}$); la construcción recursiva termina cuando cada una de las variables p ha sido utilizada a lo largo de la rama actual.

La importancia de la disminución media de la impureza de $x_m \in V$ para y calculada con un conjunto infinito de árboles totalmente aleatorios completamente desarrollados y una muestra de entrenamiento infinitamente grande, se denota como:

$$Imp(x_m) = \sum_{k_r}^{p-1} \frac{1}{C_p^{k_r}} \frac{1}{p - k_r} \sum_{B \in \mathcal{P}_k(V^{-m})}^{1} I(x_m; y|B)$$

Donde V^{-m} denota el subconjunto $V \backslash \{x_m\}$, y $\mathcal{P}_{kr}(V^{-m})$ es el conjunto de subconjuntos de V^{-m} de cardinalidad k_r, y $I(x_m;y|B)$ es la información mutua condicional de x_m y y dadas las variables en B. Para cualquier conjunto de árboles completamente desarrollados en condiciones asintóticas de un tamaño de muestra de aprendizaje, se tiene:

$$\sum_{m=1}^{p} Imp(x_m) = I(x_1, \dots, x_p, y)$$

Donde $x_i \in V$ es irrelevante para y con respecto a V si y solo si su importancia de tamaño de muestra infinita, calculada con un conjunto infinito de árboles totalmente aleatorios construidos sobre V para y, es 0.

Sea $V_R \in V$ el subconjunto de todas las variables en V que son relevantes para y con respecto a V. La importancia muestral infinita de cualquier variable $x_m \in V_R$ calculada con un conjunto infinito de árboles totalmente aleatorios completamente desarrollados

construidos sobre V_R para y es la misma que su importancia calculada en las mismas condiciones utilizando todas las variables en V.

4.4.4 Máxima entropía o maximum entropy

El clasificador de máxima entropía es análogo al modelo *logit* multinomial que es una generalización del modelo *logit*. El modelo *logit* predice la probabilidad de pertenecer a una de dos categorías, por otro lado, el modelo *logit* multinomial generaliza este modelo a una situación en la que la variable dependiente tiene más de dos categorías.

La clasificación de máxima entropía (ME) es una técnica alternativa que ha demostrado su eficacia en varias aplicaciones de procesamiento del lenguaje natural (Berger et al., 1996). Siguiendo lo descrito por Pang et al. (2002) se tiene que la estimación de probabilidad condicional en máxima entropía adopta la siguiente forma exponencial:

$$P(c\,|d\,) = \frac{1}{Z(d)}\, exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right)$$

En donde *Z(d)* es una función de normalización y $F_{i,c}$ es una función característica/clase para la característica f_i y la clase *c*, definida como:

$$F_{i,c}(d,c') = \begin{cases} 1\,, n_i(d) > 0\ y\ c' = c \\ 0\ en\ otros\ casos \end{cases}$$

En este caso, una función característica/clase concreta podría inestabilizarse si y solo si aparece algún caso extraño o no habitual dentro de la clase. La máxima entropía no hace suposiciones sobre las relaciones entre las características, por lo que podría funcionar mejor cuando no se cumplen las suposiciones de independencia condicional.

Los $\lambda_{i,c}$ son parámetros de ponderación de características; si se examina la definición de máxima entropía se observa que un $\lambda_{i,c}$ de alta magnitud significa que f_i se considera un

indicador fuerte para la clase *c*. Los valores de los parámetros se establecen para maximizar la entropía de la distribución inducida con la restricción de que los valores esperados de las funciones característica/clase con respecto al modelo sean iguales a sus valores esperados con respecto a los datos de entrenamiento. La filosofía subyacente es que se debe elegir el modelo en donde se asuman menos suposiciones sobre los datos sin dejar de ser coherente con ellos, lo que tiene un sentido intuitivo.

4.4.5 Asignación latente de Dirichlet o Latent Dirichlet Allocation (LDA)

Es un modelo generativo que permite que conjuntos de observaciones puedan ser explicados por grupos no observados que explican por qué algunas partes de los datos son similares. Además, en el análisis de textos, la ALD es un modelo probabilístico en donde se pretende captura la estructura tópica implícita de una colección de documentos o términos.

En desarrollo matemático siguiendo lo indicado por Blei et. al (2003), en la ALD se establecen algunos supuestos como la dimensionalidad de la distribución de Dirichlet, se supine conocida y fija. También, se supone que las probabilidades de las palabras están parametrizadas por una estructura matricial, como la cantidad fija para estimar. Finalmente, se supone una distribución de Poisson asociada a la longitud de los documentos. Además, nótese que *N* es independiente de todas las demás variables generadoras de datos (θ y z). Por lo tanto, es una variable auxiliar y, en general, se ignora su aleatoriedad en el desarrollo posterior.

Una variable aleatoria Dirichlet *k*-dimensional θ puede tomar valores en el (k -*1*) (un *k*-vector θ se encuentra en el (k-*1*) si $\theta_i \geq 0$, $\sum k_i=1\ \theta_i = 1$), y tiene la siguiente densidad de probabilidad:

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}$$

En donde el parámetro α es un vector *k* con componentes $\alpha_i > 0$ y donde $\Gamma(x)$ es la función Gamma. Dirichlet es una distribución conveniente en el simplex y se encuentra en la familia exponencial, tiene estadísticas suficientes de dimensión finita, y es conjugada a la distribución multinomial. Dados los parámetros α y β, la distribución conjunta de una mezcla de temas θ, un conjunto de *N* temas *z* y un conjunto de *N* palabras *w* viene dada por:

$$p(\theta, z, w \,|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|z_n, \beta)$$

Donde $p(z_n \,|\theta)$ es simplemente θ_i para el único *i* tal que $z_n = 1$. Integrando sobre θ y sumando sobre *z*, obtenemos la distribución marginal de un documento:

$$p(w|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum p(z_n|\theta)\ \ p(w_n|z_n\,, \beta \right) d\theta$$

Por último, tomando el producto de las probabilidades marginales de los documentos individuales, obtenemos la probabilidad de un corpus:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum p(z_{dn}|\theta_d)\ \ p(w_{dn}|z_{dn}\,, \beta \right) d\theta_d$$

En este caso en el análisis de textos, siguiendo a Kulkarni y Rodd (2020), la aplicación de cada documento está compuesto por una estructura temática concreta y cada tema se basa en una distribución probabilística de palabras. Así pues, considerando un modelo jerárquico bayesiano de tres niveles. La jerarquía consta de un nivel de palabras, un nivel de temas y un nivel de documentos. Mediante el uso de los niveles y los documentos, se busca crear una estructura temática escondida. En los documentos o términos, si los contextos son estáticos o parcialmente observables, se puede construir el conocimiento latente mediante el uso de la ALD.

4.5 Medidas de ajuste para validación de la clasificación de los modelos

Para conocer la efectividad de los modelos aplicados, se calcularon algunas medidas de ajuste, las cuales brindan información sobre la clasificación efectuada en las categorías del modelo. Adicionalmente, estas medidas aportan información del error de clasificación del modelo. Las medidas se derivan de la matriz de confusión, esta es una tabla de contingencia que representa las categorías de la variable de respuesta en las filas y las predicciones del modelo de la variable de respuesta en las columnas.

Tabla 1. Estructura matriz de confusión

Obs/Pred	**1**	**0**	**Total**
1	Verdaderos Positivos (VP)	Falsos Negativos (FN)	Positivos Observados
0	Falsos Positivos (FP)	Verdaderos Negativos (VN)	Negativos Observados
Total	Positivos Predichos	Negativos Predichos	Total

Por medio de la matriz de confusión se pueden calcular las siguientes medidas:

Exactitud: Esta medida brinda la cercanía que se encuentra la predicción del valor verdadero. Esta se asocia al sesgo de una estimación. Es la proporción de los verdaderos positivos y verdaderos negativos entre el número total de los casos clasificados.

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN}$$

Precisión: Se refiere a la dispersión del conjunto de valores por medio de mediciones repetidas de una magnitud, entre menor es la dispersión mayor es la precisión. Es la proporción de los verdaderos positivos entre todos los resultados positivos verdaderos y falsos.

$$Presición = \frac{VP}{VP + FP}$$

Sensibilidad: Esta medida es proporción de los casos positivos que fueron correctamente identificadas por el modelo. Es la tasa de verdaderos positivos.

$$Sensibilidad = \frac{VP}{VP + FN}$$

Especificidad: Consta de los casos negativos que l modelo ha clasificado correctamente, es la tasa de verdaderos negativos.

$$Especificidad = \frac{VN}{VN + FP}$$

Precisión equilibrada: Es una métrica de rendimiento del modelo en las categorías del conjunto de datos, independiente de las proporciones de las categorías. Se utiliza cuando hay un desbalance en los datos.

$$Precisión\ equilibrada = \frac{Sensibilidad + Especificidad}{2}$$

CAPÍTULO V. METODOLOGÍA

5.1 Enfoque

Este estudio se desarrolla bajo el enfoque cuantitativo, con respecto al alcance del estudio este abarca un análisis estadístico descriptivo y mimería de texto. El análisis se desarrolla mediante técnicas y métodos matemáticos y estadísticos utilizados en el ámbito de la minería de textos.

5.2 Unidades de estudio

Las unidades de estudio serán las obras literarias latinoamericanas recomendadas por el Consejo Superior de Educación en el acuerdo N°04-36-2017 (CSE, 2017). Asimismo, algunas de estas formaron parte de la evaluación de las pruebas FARO y pruebas nacionales que realiza el Ministerio de Educación Pública.

Las obras literarias recomendadas por el CSE se detallan en el Anexo 1. Para el estudio se contemplan las siguientes variables:

- Nombre de la obra literaria
- Nombre de la persona autora
- Nacionalidad de la persona autora (costarricense o extranjero)
- Género literario de la obra literaria
- Nivel o ciclo educativo en que se recomienda la obra literaria
- Modalidad educativa en que se recomienda la obra literaria
- Obra literaria considerada en prueba FARO

En la creación de los corpus lingüísticos de las lecturas y los ítems de práctica de la PAA se construyen las siguientes variables:

- Palabras del corpus de la obra literaria
- Palabras del corpus de los ejercicios de práctica de la Prueba de Actitud Académica
- Índices de frecuencia o peso de palabras en el texto
- Índice de similitud de las obras
- Índice de similitud de los corpus

En las obras literarias recomendadas se limita a considerar obras de personas autoras latinoamericanas y obras en donde se tengan acceso para un fin académico y educativo, en algunos casos solo se tiene un fragmento parcial de la obra y en otros casos se tiene de manera completa. El material total o parcial se utiliza bajo el principio de la excepción académica, establecido en la Ley sobre derechos de autor y derechos conexos (N 6683) y su Reglamentos (N 24611-J) conforme a la legislación nacional.

Para este estudio se consideran 68 obras literarias, el listado se muestra en la Tabla 2.

Tabla 2. Obras literarias latinoamericanas consideradas, lecturas recomendadas por el CSE, 2017.

Autor	Título
Allende, Isabel	Cuentos de Eva Luna
Argüello Mora, Manuel	Elisa Delmar
Asturias, Miguel Ángel	El señor presidente
Azofeifa, Isaac Felipe	Invitación al diálogo de las generaciones
Benedetti, Mario	Inventarios I y II
Borges, Jorge Luis	El Aleph
Calvo, Yadira	La mujer víctima y cómplice
Campbell, Shirley	Rotundamente negra y otros poemas
Cañas Escalante, Alberto	Ni mi casa es ya mi casa
Carballido, Emilio	Estudio en blanco y negro
Contreras, Fernando	Cierto azul
Contreras, Fernando	Única mirando al mar
Cortázar, Julio	Bestiario
Cortés, Carlos	La última aventura de Batman
Darío, Rubén	Cantos de vida y esperanza
Darío, Rubén	Azul
De Sosa, Geovanny	Los dueños de la casa
de Vallbona, Rima	Los infiernos de la mujer y algo más
Debravo, Jorge	Los despiertos

Autor	Título
Debravo, Jorge	Nosotros los hombres
Dobles, Fabián	El sitio de las abras
Dobles, Fabián	¡Alerta, ustedes!
Dobles, Fabián	Cuentos de Tata Mundo
Echeverría, Aquileo	Concherías
Fallas, Carlos Luis	Mamita Yunai
Fernández Guardia, Ricardo	Magdalena
Fernández Guardia, Ricardo	Cuentos ticos
Fernández, Guillermo	Tu nombre será borrado del mundo
Fuentes, Carlos	El naranjo, o los círculos del tiempo
Gallegos, Daniel	La casa
Gallegos, Rómulo	Doña Bárbara
García Esperón, María	El disco del tiempo
García Lorca, Federico	La casa de Bernarda Alba
García Márquez, Gabriel	Yo no vengo a decir un discurso
García Márquez, Gabriel	Cien años de soledad
García Márquez, Gabriel	Crónica de una muerte anunciada
García Monge, Joaquín	El Moto
Garro, Elena	La culpa es de los tlaxcaltecas y otros relatos
González Zeledón, Manuel	Cuentos de Magón
Gutiérrez, Joaquín	Cocorí
Jiménez, Max	El jaúl
Lobo, Tatiana	Asalto al paraíso
Lobo, Tatiana	Calypso
Lyra, Carmen	Había una vez
Lyra, Carmen	En una silla de ruedas
Martí, José	Nuestra América
Mastretta, Ángeles	El mundo iluminado
Mastretta, Ángeles	Mujeres de ojos grandes
Méndez, Melvin	Un viejo con alas
Méndez, Melvin	Terminal del sueño
Naranjo, Carmen	Cinco temas en busca de un pensador
Neruda, Pablo	Veinte poemas de amor y una canción desesperada
Oreamuno, Yolanda	A lo largo del corto camino
Oreamuno, Yolanda	La ruta de su evasión
Parra, Nicanor	Poemas y antipoemas
Quiroga, Horacio	Cuentos de amor, de locura y de muerte
Quiroga, Horacio	Cuentos de la selva
Rossi, Anacristina	La loca de Gandoca
Rossi, Anacristina	Limón Blues
Roswell, Víctor	El canto de los quetzales
Rulfo, Juan	Pedro Páramo
Rulfo, Juan	El llano en llamas

Autor	Título
Salazar Herrera, Carlos	Cuentos de angustias y paisajes
Santa Ana, Antonio	Los ojos del perro siberiano
Sepúlveda, Luis	Historia de una gaviota y del gato que le enseñó a volar
Vargas Pizarro, Maureen	Danzas del bosque
Vodanovic, Sergio	El delantal blanco
Wohlstein, Harry	Piedra sobre piedra

Fuente: elaboración propia, 2025.

5.3 Procedimiento y estrategia de análisis

El procedimiento de análisis se fundamentará en un marco metodológico para el tratamiento de datos textuales no estructurados, adaptando las etapas clave propuestas por Gohil (2015). La ejecución secuencial de estas fases, que abarcan desde la recolección y preprocesamiento hasta la extracción de características, permitirá estructurar el corpus de datos final para el estudio. Es pertinente señalar que algunas de estas etapas conllevan una considerable demanda de recursos computacionales y un elevado tiempo de procesamiento.

Las etapas se describen a continuación:

Etapa I: Identificación de textos de las obras literarias

Se identificarán las obras literarias correspondientes para el análisis de texto. Se procede a realizar una búsqueda en diferentes navegadores web de archivos en formato texto de las obras. Estas obras son utilizadas para fines académicos y educativos.

Etapa II: Formato de textos de las obras literarias

Los archivos de las obras se ubican en formato de documento portátil (PDF, siglas en inglés). Estos archivos se convierten a archivos en texto plano para una mayor facilidad en los procesamientos del texto.

Etapa III: Extracción del texto y limpieza del corpus

Se centra en remover aspectos del texto innecesarios. En este caso se realiza una *tokenización*, un filtro de palabras (*stop word removal*), lematización, procedimientos lingüísticos, etiquetado de palabras, sentido de palabras no ambiguas (WSD) y estructura semántica. Además, se eliminan los signos de puntuación, espacios y símbolos no relevantes para la estructuración del corpus.

Etapa IV: Estructuración de bases de datos del corpus

Se realiza una transformación del texto a datos estructurados y se realiza una estructuración de corpus para la identificación de palabra frecuentes o no importantes. Adicionalmente, se estructura el corpus para identificar bigramas (asociación de palabras).

Etapa V: Aplicación de métodos, técnicas y modelos de minería de texto

En estos análisis descriptivos, considerando todo el corpus completo de todos los textos y los ítems de práctica de la PAA, se aplicaron diferentes métodos o técnicas de clasificación, resumen, temáticas, medidas de similitud, entre otras.

Entre los métodos comunes según Gaikwad et al (2014), que se utilizan en el análisis de minería de textos se encuentran:

- Método basado en términos (TBM): El término en el documento es una palabra que tiene un significado semántico.

- Método basado en conceptos (CBM): Los términos se analizan a nivel de oraciones y documentos. Las técnicas de minería de textos se basan principalmente en el análisis estadístico de palabras o frases. El análisis estadístico del término al considerar la frecuencia se puede obtener la importancia o peso de una palabra sino se encuentra en el documento.

- Método de taxonomía de patrones (PTM): Los documentos se analizan sobre patrones. Los patrones se pueden estructurar en taxonomías utilizando correlaciones.

En este estudio se consideran el método basado en términos y en conceptos, pues se determina la frecuencia de palabras, las presencia o no de la palabra en el corpus a la ahora de comparar y, adicionalmente, se estiman y calculan algunos índices de similitud textual. Luego de definir el corpus de las obras literarias, se realiza un análisis estadístico descriptivo. Además, se aplicarán técnicas de estadísticas para la clasificación y segmentación y similitud con otro corpus, en este caso el corpus asociado a los ítems de práctica de la Prueba de Aptitud Académica que aplica la Universidad de Costa Rica.

Entre las técnicas empleadas en la minería de texto, se implementaron las sugeridas por Gaikwad et al (2014) y Talib et al (2016), entre ellas:

- Extracción de información: Es una técnica para extraer automáticamente una pieza de información definida y estructurada a partir de datos no estructurados o semiestructurados en forma de texto mediante el procesamiento del lenguaje natural.

- Recuperación de información: Es un conjunto de métodos o enfoques para desarrollar metódicamente las necesidades de información en forma de consultas que se utilizan para obtener un documento de una colección de bases de datos. Esta ayuda a extraer patrones relevantes y asociados de acuerdo con un conjunto dado de palabras o frases.

- Categorización: Esta técnica implica la designación de categorías predeterminadas a documentos de texto libre que contienen información. El propósito de la clasificación de la categorización de texto es aumentar la detección de información de forma categorizada.

- Resumen de texto: El objetivo de esta técnica de extracción de texto es navegar a través de múltiples fuentes de texto para elaborar resúmenes que contienen una proporción considerable de información en un formato conciso, manteniendo el

significado general y la intención de los documentos originales esencialmente iguales.

- Análisis de conglomerados: Esta técnica se utiliza para buscar grupos de documentos con contenido similar. Hace uso y extracción de descriptores que son esencialmente conjuntos de palabras que describen los contenidos dentro del clúster.

Estas técnicas de minería de texto fueron implementadas en el análisis de los corpus, inicialmente realizando la extracción de los textos en palabras y luego en información. La recuperación de la información se basó en conocer las asociaciones y las métricas en la obtención de patrones como la aplicación de asignación latente de Dirichlet. La categorización de las palabras, el resumen y la agrupación se asocia a las técnicas descriptivas y de modelamiento aplicadas en el análisis de los corpus.

En cuanto al análisis de los modelos para la clasificación y segmentación de los corpus, se aplicaron cuatro modelos y luego se estimaron otros cuatro modelos adicionales considerando la estructura de Asignación Latente de Dirichlet (ALD). Adicionalmente, se estiman las medidas de ajuste de los modelos realizados para la clasificación de las palabras de los ítems de práctica de la PAA considerando el corpus de las obras literarias latinoamericanas de las lecturas recomendadas en el Tercer Ciclo y Educación Diversificada. En una primera instancia, se procede a estimar los modelos de forma usual y posteriormente se estiman los modelos considerando la estructura de ALD, la cual se encuentra conformada de variables latentes que se construyen por medio de la asociación de las palabras en el corpus.

Luego, en la estimación de los modelos se realiza una partición de los datos del 70% para el entrenamiento y 30% para la prueba del modelo de clasificación. Los modelos de clasificación pretenden clasificar la presencia o ausencia de una palabra en el corpus de ítems de práctica de la PAA, considerando la frecuencia y distribución del corpus de las obras literarias.

Además, para la estructuración de los textos de las obras literarias se utilizó para la estructuración de datos y el análisis el Lenguaje estadístico R (R Core Team, 2023) y Microsoft Office Excel. En el análisis en R se utilizaron librerías para la manipulación de datos y la visualización, se emplearon paquetes del ecosistema *tidyverse* (Wickham et al., 2019), principalmente *dplyr* (Wickham, et al., 2023) y *ggplot2* (Wickham, et al., 2023).

También, el preprocesamiento y análisis de texto se llevó a cabo utilizando un conjunto de herramientas que incluyen los paquetes *tm* (Feinerer y Hornik, 2023), *tidytext* (Silge y Robinson, 2022), *SnowballC* para lematización (Bouchet-Valat, 2020), *tokenizers* (Mullen et al., 2018), *NLP* (Hornik, 2023) y *pdftools* para la extracción de texto (Ooms, 2023). Adicionalmente, se utilizaron las librerías *lsa* para análisis semántico latente (Wild, 2015), *stringdist* para la comparación de cadenas de texto (van der Loo, 2023), *wordcloud* (Fellows, 2018) y *ggraph* (Pedersen, 2023) para la visualización de resultados. Finalmente, se emplearon los paquetes *openxlsx* (Walker y Schauberger, 2023) y *xlsx* (Dragulescu, 2023) para la gestión de archivos, *caret* para tareas de modelado (Kuhn, 2023), *scatterplot3d* (Ligges y Mächler, 2003) para gráficos tridimensionales y *hashr* para tablas hash (Brown, 2017).

En relación con la validación estadística de los resultados y el contraste de las hipótesis formuladas, se ha establecido un nivel de significancia (α) de 0.05. Este umbral, ampliamente aceptado como estándar en la investigación en ciencias sociales y de la educación, servirá como criterio para la estimación de los intervalos de confianza.

En general, el procedimiento y estrategia de análisis se articuló en varias fases secuenciales para garantizar la robustez de los resultados. Inicialmente, se ejecutó un análisis exploratorio de datos sobre el corpus textual para caracterizar sus propiedades fundamentales, tales como la distribución de frecuencias de términos, la longitud de los documentos y la identificación de palabras clave. Asimismo, se procedió a realizar un escalamiento multidimensional y un Análisis Semántico Latente (ASL), para conocer la similitud entre palabras y documentos y entre sí.

A continuación, se procedió al cálculo de la similitud textual, empleando métricas como la similitud del coseno y Jaccard, con el objetivo de cuantificar las relaciones semánticas y estructurales entre los documentos.

La etapa central del análisis consistió en la construcción y entrenamiento de modelos de clasificación, para los cuales se definió la variable dependiente como una variable binaria que indica la presencia o ausencia de términos específicos en el corpus. Como variables independientes o predictoras se utilizaron, por un lado, métricas de frecuencia de palabras y para los modelos que consideran la estructura temática como la Asignación Latente de Dirichlet (ALD), adicionalmente se emplearon las distribuciones de probabilidad de los tópicos latentes. A continuación, se presentan los modelos empíricos estimados:

Sea $\hat{Y}$ una variable binaria aleatoria dependiente, la cual el valor 0 indica ausencia de la palabra y el valor 1 indica presencia de la palabra en el corpus de los ítems de práctica de la PAA. Además, sea X una variable aleatoria independiente, relacionada al número de veces que la palabra se encuentra en el corpus de las obras literarias latinoamericanas y sea $Z = \{L_1, L_2, L_3, L_4\}$ el conjunto de variables latentes obtenidas de la ALD.

- Ingenuo de Bayes

$$\hat{Y} = argmax_k \left[\log\big(P = (Y = k)\big) + \sum_i \log\left(P(X_i|Y = k) + \sum_i \log\left(P(Z_i|Y = k)\right)\right)\right]$$

- Máquinas de soporte vectorial

$$\hat{Y} = sign\left[\sum_i \alpha_i * y_i * (\gamma * u_i^T * u + r)^d + b\right]$$

Donde:

$u = [X, Z]$

La suma se realiza sobre los vectores de soporte $u_i = [X_i, Z_i]$

Hiperparámetros d, γ, r

- Bosques aleatorios

$$\hat{Y} = I_{nodo}\{h_1([X,Z]), h_2([X,Z]), \dots, h_n([X,Z])\}$$

Donde:

n es el número de árboles en el bosque

$h_i([X,Z])$ es la función de predicción del i-ésimo árbol de decisión

- Máxima entropía

$$\hat{Y} = argmax_k[P = (Y = k) \mid X, Z]$$

Finalmente, el desempeño de cada modelo fue evaluado mediante el cálculo de un conjunto de indicadores de ajuste derivados de la matriz de confusión, incluyendo la exactitud, la precisión, la sensibilidad y la precisión equilibrada, para determinar su capacidad predictiva y seleccionar el modelo más idóneo.

CAPÍTULO VI. RESULTADOS

En este capítulo se describe el análisis descriptivo de los corpus de las obras literarias latinoamericanas consideras y el corpus de ítems de práctica de la PAA.

6.1 Análisis exploratorio

En este apartado se presenta el análisis descriptivo exploratorio de las obras analizadas, el análisis descriptivo textual de los corpus lingüísticos de las obras literarias latinoamericanas consideradas y el corpus de ítems de práctica de la PAA y las medidas de similitud léxica de los documentos y los corpus.

6.1.1 Descriptivos de las lecturas analizadas

En este análisis se consideran 68 obras literarias, según la Tabla 3, estas se caracterizan por leerse en el nivel educativo de noveno año (27,9%), pertenecen mayoritariamente al género literario de novela (25%) y época vanguardista (20,6%). Asimismo, predominan obras de personas autoras nacionales (57,4%), 70,6% son obras consideradas en la modalidad técnica y solo 7,4% son obras que se encuentran en temario de la prueba FARO. Otro aspecto por considerar es que en algunos casos no se analizó la obra completa sino de forma parcial, este corresponde al 26,5% del total de las obras. En el Anexo 2, se muestra el listado de las obras analizadas y enumeradas para identificarlas en la interpretación de estos resultados.

Tabla 3. Número y porcentaje de las características de las obras literarias analizadas

Característica	**Frecuencia**	**Porcentaje**
Total	68	100%
Nivel educativo		
Noveno	19	27,9%
Octavo	16	23,5%
Undécimo	15	22,1%
Sétimo	13	19,1%
Décimo	5	7,4%

Característica	Frecuencia	Porcentaje
Género literario		
Novela	17	25,0%
Época vanguardista	14	20,6%
Cuento	12	17,6%
Lírica	7	10,3%
Drama	6	8,8%
Época modernista	6	8,8%
Drama	3	4,4%
Ensayo	3	4,4%
Nacionalidad persona autora		
Nacional	39	57,4%
Extranjero	29	42,6%
Modalidad técnica		
Sí	48	70,6%
No	20	29,4%
Lectura en prueba FARO		
No	63	92,6%
Sí	5	7,4%
Obra analizada		
Completa	50	73,5%
Parcial	18	26,5%

Fuente: elaboración propia con información de las Obras recomendadas por el CSE (2008).

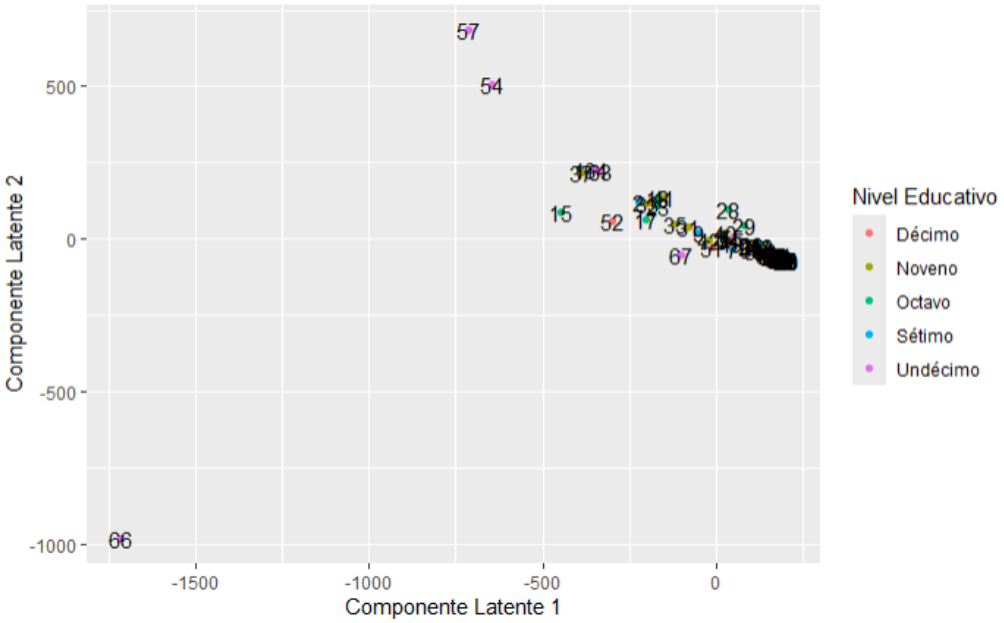
Por otro lado, el CSE recomienda otras obras más, sin embargo, este análisis se considera representativo y aproximado a lo que mayoritariamente se considera en los programas de español del MEP.

En las obras literarias se realizó un análisis semántico latente (ASL) entre ellas y según sus características para identificar asociaciones. A continuación, se describe los resultados de los análisis semánticos latentes por cada característica de las obras.

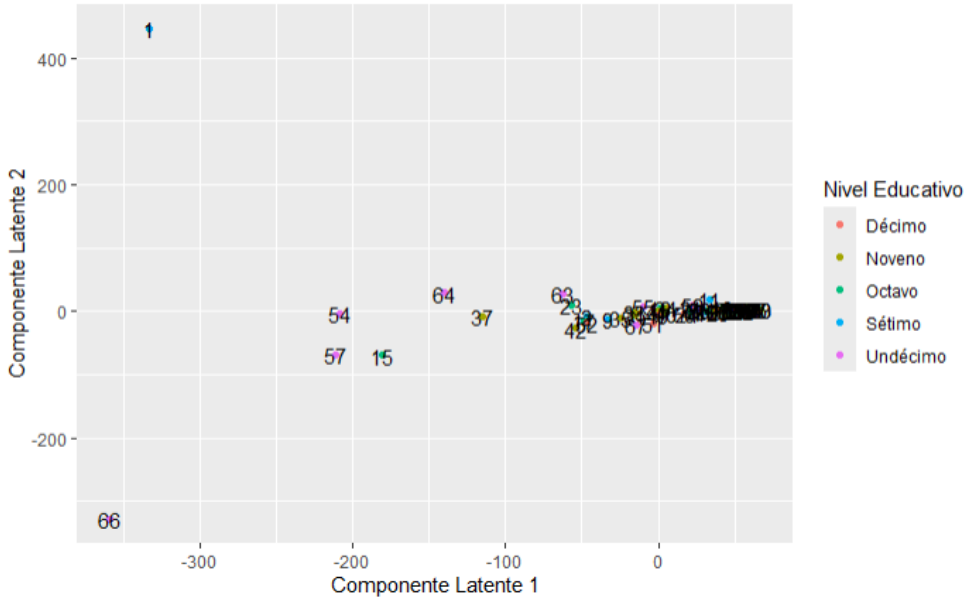
De acuerdo con la Figura 1, se muestra en escalamiento algunas obras distantes de otras, específicamente la obra con identificador 66, la cual es “Cien años de soledad”, igual existe un distanciamiento de la 54 “El señor presidente” y 57 “Doña Barbara”. En relación con el nivel educativo, se muestra que estas obras se leen en el nivel de undécimo, uno de los últimos niveles educativos. Si se considera el análisis semántico latente, se muestra una

separación de la obra 1 “Cuentos de Magón”, además de las obras 15 “Cuentos de Eva Luna”, 37 “El sitio de las abras” y 64 “Mamita Yunai”.

Figura 1. Análisis de escalamiento multidimensional de las obras literarias según nivel educativo



a) Escalamiento multidimensional

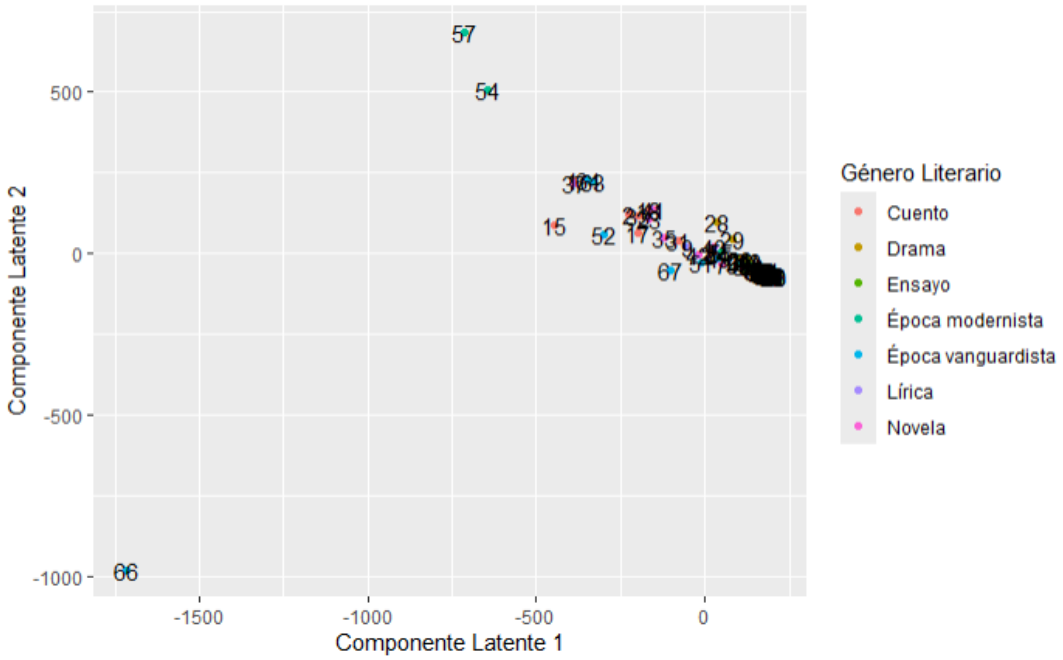


b) Escalamiento multidimensional considerando ASL

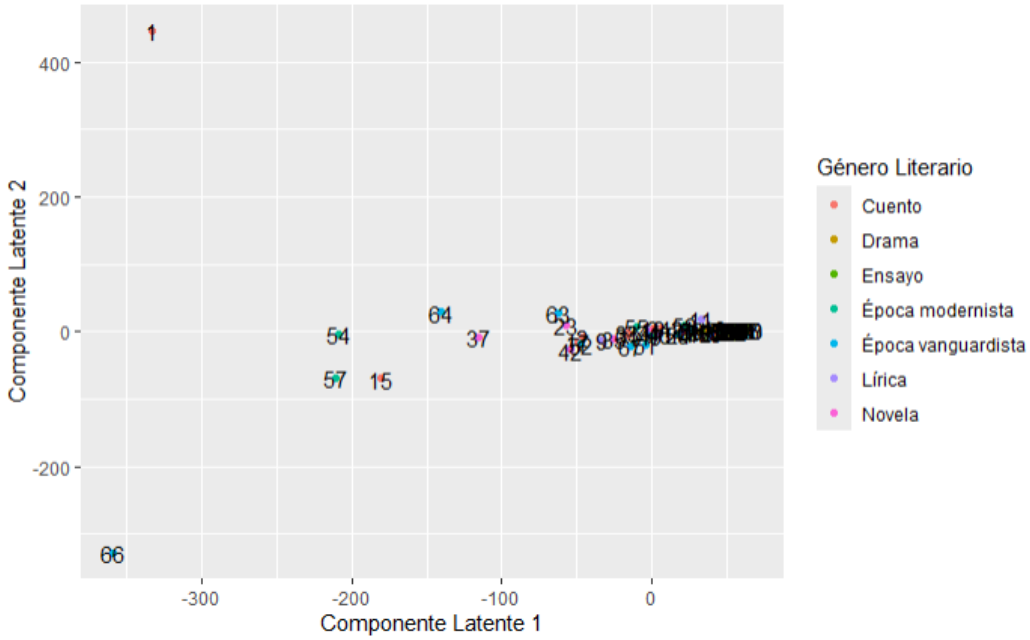
Fuente: elaboración propia con información de las obras analizadas.

Ahora, al considerar en el escalamiento el género literario, se muestra en la Figura 2 que las obras distantes forman parte de la época modernista y en el escalamiento con la estructura semántica latente se mantiene el género modernista, se incorpora dos obras del género cuento y una obra de la época vanguardista y novela.

Figura 2. Análisis de escalamiento multidimensional de las obras literarias según género literario



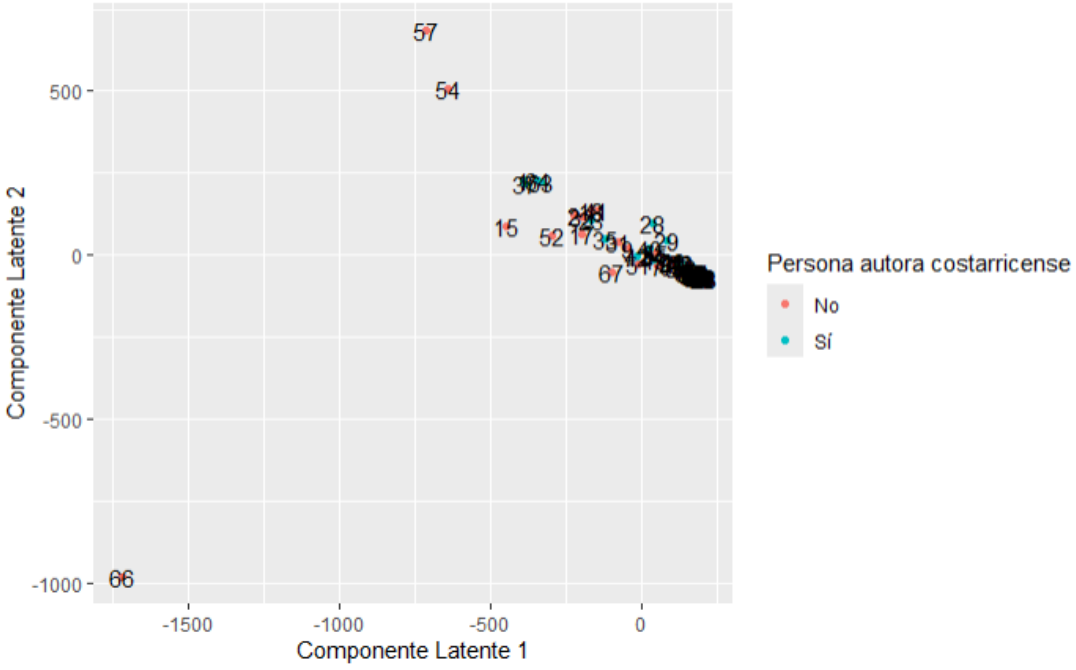
a) Escalamiento multidimensional



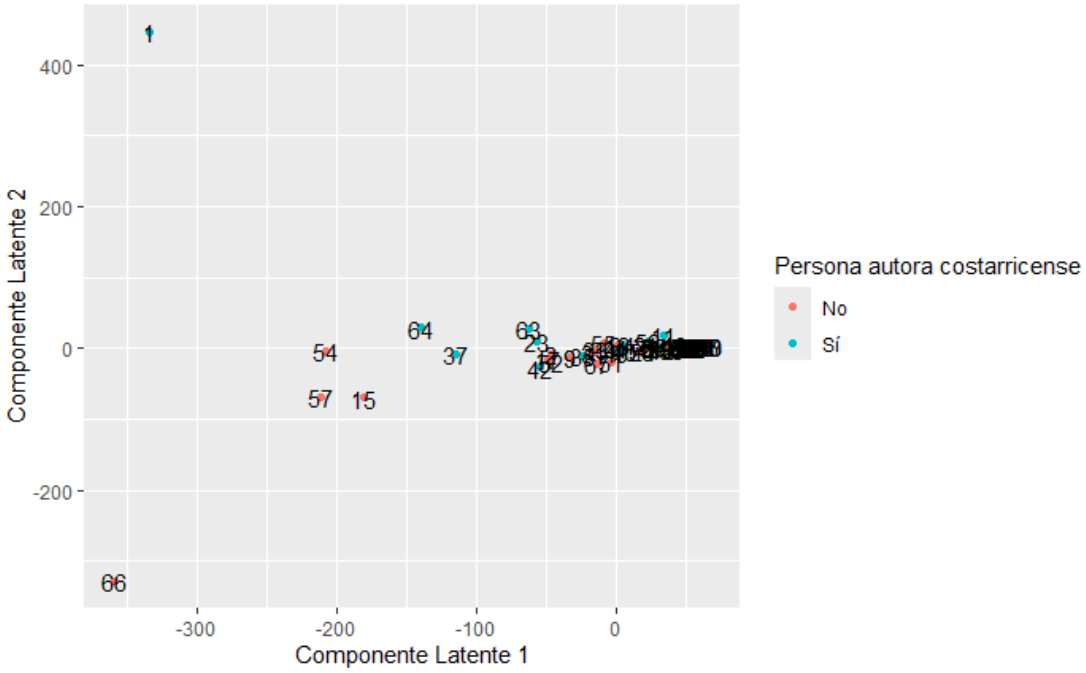
b) Escalamiento multidimensional considerando ASL

Fuente: elaboración propia con información de las obras analizadas.

Figura 3. Análisis de escalamiento multidimensional de las obras literarias según nacionalidad de la persona autora



a) Escalamiento multidimensional



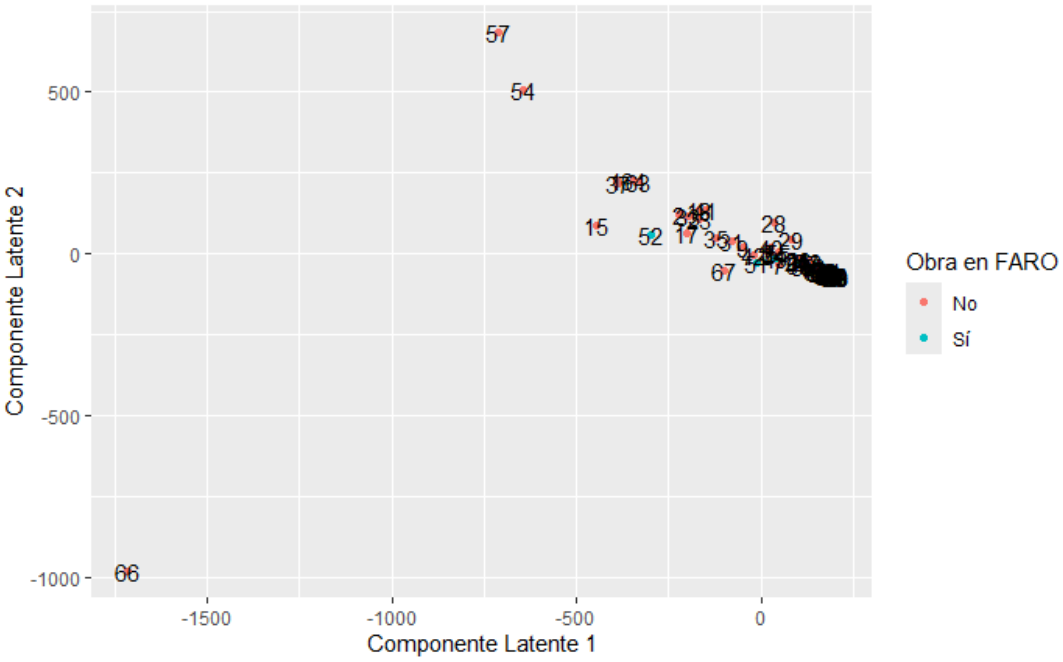
b) Escalamiento multidimensional considerando ASL

Fuente: elaboración propia con información de las obras analizadas.

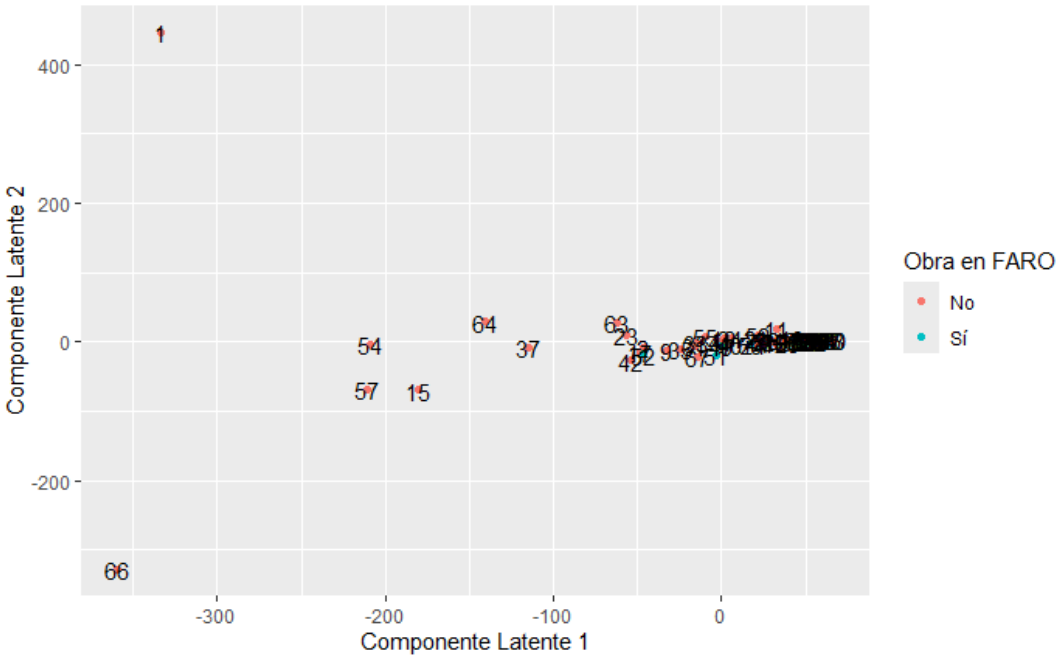
Según la Figura 3, en el escalamiento multidimensional de las obras se muestra en primera instancia que las obras distantes pertenecen a personas autoras extranjeras, luego aplicando el ASL, se incorporan tres obras de personas autoras nacionales, entre ellas 1 “Cuentos de Magón”, 37 “El sitio de las abras” y 64 “Mamita Yunai”.

Posteriormente, se considera en el escalamiento de las obras literarias si estas se encuentran evaluadas en la prueba FARO, según la Figura 4, se observa que las obras distantes en ambos escalamientos no forman parte de la evaluación de la prueba. Solamente, en la distribución del escalamiento, se muestra que la única obra evaluada en la prueba FARO, que dista un poco de las demás obras, es la identificada con el 52 “Mujeres de ojos grandes”, perteneciente al género vanguardista, su lectura se recomienda a nivel de educativo de décimo y su autora es de nacionalidad extranjera.

Figura 4. Análisis de escalamiento multidimensional de las obras literarias según evaluación en prueba FARO



a) Escalamiento multidimensional

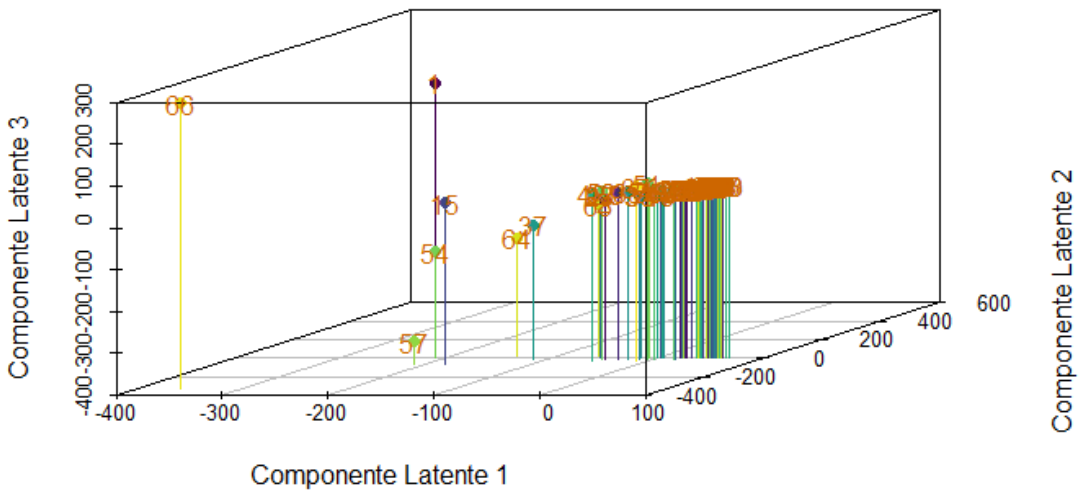


b) Escalamiento multidimensional considerando ASL

Fuente: elaboración propia con información de las obras analizadas.

Finalmente, en la Figura 5, se visualiza el espacio multidimensional del análisis semántico latente de las obras literarias analizadas, según los tres componentes latentes. Existe una alta concentración entre las lecturas y solo una dista de la mayoría las siete obras anteriormente indicadas.

Figura 5. Análisis de escalamiento multidimensional de las obras literarias en las tres componentes latentes



Fuente: elaboración propia con información de los corpus.

6.1.2 Descriptivos de los corpus de las lecturas

En el análisis, es importante considerar que previamente se aplicó las etapas descritas en la metodología para la construcción del corpus lingüístico de las obras literarias latinoamericanas consideradas por el CSE. Inicialmente, se realizó un análisis descriptivo en el cual se identificaron las palabras más frecuentes en las obras literarias analizadas, para ello se aplicó una nube de palabras. Es importante mencionar que este corpus está constituido por más de setenta mil palabras y tiene un porcentaje de esparcimiento del 95%.

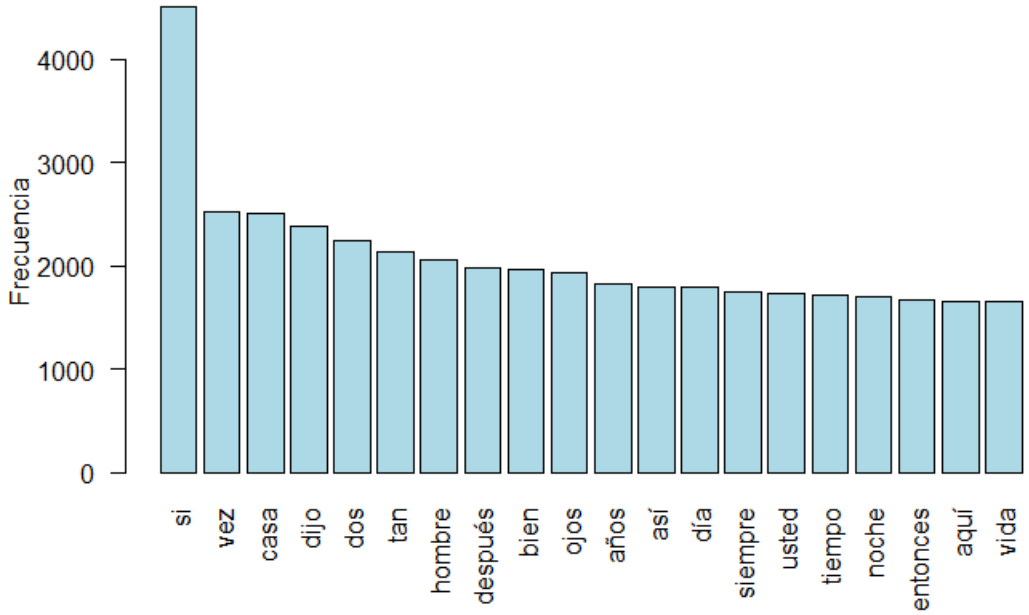
En la Figura 6, se observa que el término más frecuente es “si”, luego se destacan las palabras “casa”, “dijo”, “dos”, “tan”, “hombre”, después” y “bien”. Estas palabras se destacan por ser sustantivos y en menor medida adjetivos y adverbios.

Figura 6. Nube de las palabras más frecuentes en corpus de las obras literarias

Fuente: elaboración propia con información de los corpus.

Según el Gráfico 1, se visualiza que la conjunción "si" presenta la mayor frecuencia entre todas las obras analizadas se muestra 4508, seguidamente predominan los términos "vez" (2528), "casa" (2505), "dijo" (2376), "dos" (2242), "tan" (2137), "hombre" (2059), "después" (1970) y "bien" (1970). Estas palabras presentan una frecuencia de más de mil apariciones en las obras literarias.

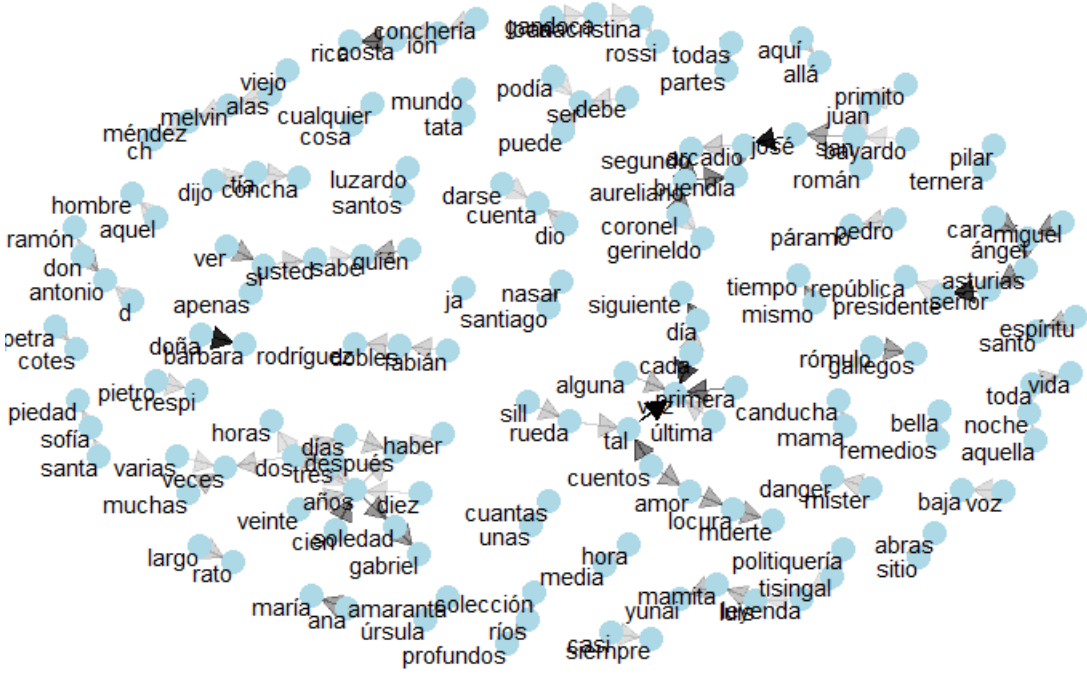
Gráfico 1. Número de frecuencia de las principales palabras del corpus de las obras literarias



Fuente: elaboración propia con información de los corpus.

Posteriormente, se realiza un análisis de la asociación de biagramas, es decir, la asociación entre dos palabras y entre las palabras. De acuerdo con la Figura 7, esta asociación se encuentra enfocada por frases sobresalientes en el corpus de las obras literarias y también se destaca la asociación de los personajes de las obras. Además, se muestran asociaciones entre palabras que en conjunto muestran un significado o palabras que se caracterizan entre sí.

Figura 7. Biagramas de palabras asociadas del corpus de las obras literarias



Fuente: Elaboración propia con información de los corpus.

Adicionalmente, la asociación de estos biagramas muestra asociaciones de las obras más relevantes en el corpus, como los es “Cien años de soledad”, “Señor presidente”, “Mamita Yunai” y “La loca de Gandoca”.

6.1.3 Descriptivos del corpus de ítems de práctica de la PAA

Para la construcción del corpus de los ítems de práctica de la PAA, se procedió al igual que con el corpus de obras literarias a aplicar las etapas para elaborarlo. En el corpus se enfatiza que se encuentra constituido por más de mil setecientas palabras y este presenta un esparcimiento del 21%, es decir, un corpus muy compacto de las palabras utilizadas para elaborar los ítems.

En la Figura 8, se muestra la nube de algunas de las palabras que se destacan en el corpus de los ítems de práctica de la PAA. En esta se observa que sobresalen términos utilizados en la construcción de los enunciados de los ítems, se destacan las palabras “solo”,

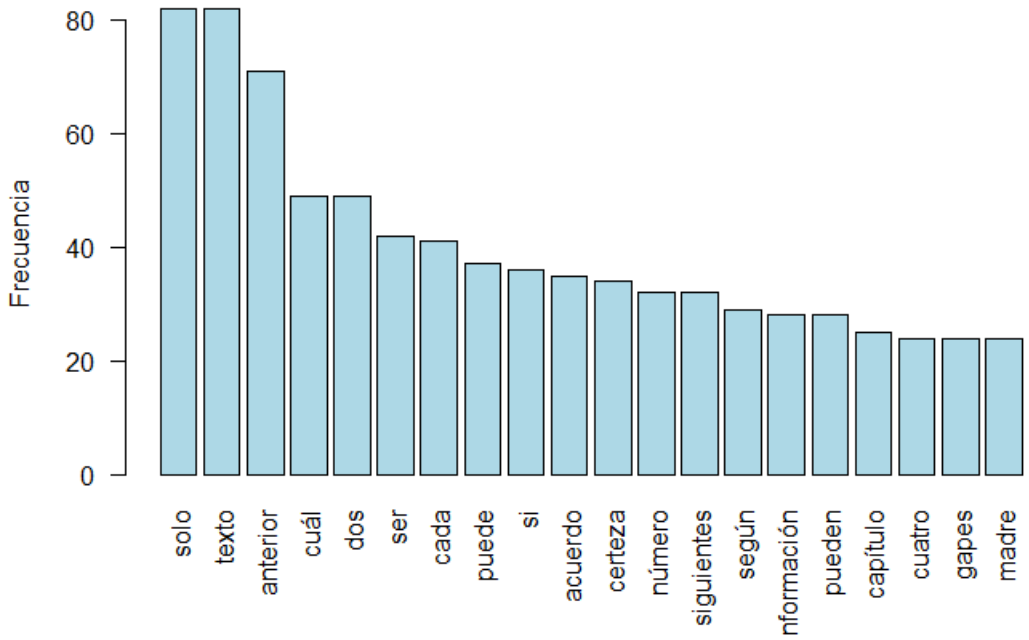
"texto", "anterior", "cuál", "ser" y "dos". Las palabras destacadas hacen referencia en dirigir la instrucción del ítem, como por ejemplo cuando se enfatiza en la lectura de un texto, validar la certeza de afirmaciones y en indicar suposiciones de algunas situaciones en los ítems.

Figura 8. Nube de las palabras más frecuentes en corpus de los ítems de práctica de la PAA

Fuente: elaboración propia con información de los corpus.

Asimismo, en el Gráfico 2, se muestra la frecuencia de las principales palabras, el término más frecuente en el corpus de los ítems de práctica de la PAA es "solo", al igual que el término "texto" se presenta en 82 ocasiones. También, se destacan palabras como "anterior" (71), "cuál" (49), "dos" (49), "ser" (42) y "cada" (41). En este caso predominan sustantivos y algunos adjetivos.

Gráfico 2. Número de frecuencia de las principales palabras del corpus de los ítems de práctica de la PAA

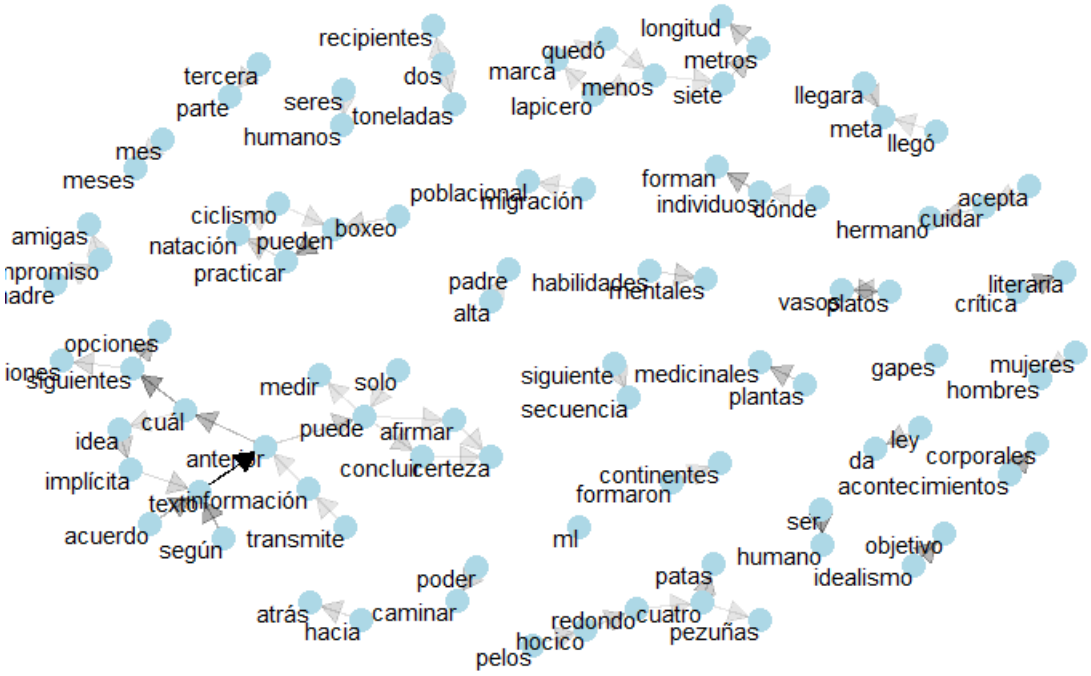


Fuente: elaboración propia con información de los corpus.

En la Figura 9, se visualizan los biagramas entre las palabras más comunes en el corpus de ítems de práctica, en estas asociaciones de términos se destacan entre sí los más comúnmente utilizados en la elaboración de los enunciados del ítem palabras como “texto” y “anterior” se encuentran altamente asociadas. Al igual que palabras como “afirmar”, “certeza” y “concluir”.

Por otro lado, se destacan la asociación de palabras que describen alguna situación o circunstancia, se destaca que palabras vinculas a animales existe una cohesión como el grupo de términos “patas”, “pezuñas”, “hocico” y “pelos”. Esto similarmente ocurren en conjunto de palabras asociada a los deportes se correlacionan vocablos como “ciclismo”, “natación”, “boxeo” y “practicar”. Estos biagramas muestran las asociaciones predominantes que se describen en el corpus de los ítems de práctica de la PAA.

Figura 9. Biagramas de palabras asociadas del corpus de los ítems de práctica de la PAA



Fuente: elaboración propia con información de los corpus.

6.1.4 Análisis de similitud textual de los documentos y los corpus

En el análisis de textos, con el fin de conocer la asociación de similitud entre documentos o corpus, se estiman algunas métricas para este objetivo. Algunas comúnmente usadas con la similitud de coseno y el índice de similitud de Jaccard, estas métricas se aplican al corpus de las obras literarias y al corpus de ítems de práctica de la PAA.

Inicialmente, de manera exploratoria, se aplican las métricas entre las obras analizadas y se destaca que entre ellas existe alta asociación, pues superan las magnitudes de 0.9. Luego se calculan las métricas de similitud entre las obras literarias analizadas y el corpus de los ítems de práctica de la PAA, los resultados se muestran en la Tabla 4. Es importante indicar que se muestran similitudes con magnitud igual a 1, esto se debe a que solo se tiene un breve fragmento de la obra por lo que existen una alta coincidencia entre los términos de la obra y el corpus de los ítems.

Tabla 4. Similitud de coseno y Jaccard entre las obras literarias y el corpus de ítems de práctica de la PAA

# Obra	Nombre de la obra	Coseno	Jaccard
1	Cuentos de Magón	0,9920954	0,9988225
2	Cuentos de amor, de locura y de muerte	0,9966648	0,9985525
3	Cuentos de la selva	0,9950338	0,9979737
4	Elisa Delmar	0,996633	0,9995771
5	Cocorí	0,9990575	0,9990897
6	El canto de los quetzales *	0,9943131	0,9978056
7	Los ojos del perro siberiano	0,9943195	0,9982895
8	La mujer víctima y cómplice *	1	1
9	Inventarios I y II	0,9948609	0,99907
10	Rotundamente negra y otros poemas *	0,9970877	0,9994777
11	Concherías	0,9858542	0,9991432
12	Poemas y antipoemas	0,998022	0,9992189
13	Había una vez	0,998919	0,999125
14	Cinco temas en busca de un pensador *	0,9988727	0,9993831
15	Cuentos de Eva Luna	0,9917847	0,9987445
16	La última aventura de Batman *	0,996734	0,9993704
17	El mundo iluminado	0,989811	0,9982163
18	El llano en llamas	0,9959061	0,9986946
19	Danzas del bosque *	0,9941756	0,9983865
20	Cierto azul *	0,9953859	0,9987222
21	El jaúl	0,9971841	0,9989868
22	Asalto al paraíso *	0,9990587	0,9992382
23	En una silla de ruedas	0,9955116	0,9987327
24	Historia de una gaviota y del gato que le enseñó a volar	0,9938837	0,9986614
25	Los despiertos	0,9919795	0,999274
26	Veinte poemas de amor y una canción desesperada	0,999439	0,9998219
27	Ni mi casa es ya mi casa	0,9954597	0,9990097
28	Magdalena	0,9960929	0,9989406
29	Un viejo con alas	0,9989186	0,9988905
30	A lo largo del corto camino *	0,9916502	0,9988306
31	Bestiario	0,9902079	0,998901
32	Cuentos ticos *	0,9955577	0,9990659
33	La culpa es de los tlaxcaltecas y otros relatos	0,9957576	0,9987659
34	Cuentos de angustias y paisajes	0,9968372	0,99832
35	Única mirando al mar	0,98921	0,9984053
36	Los dueños de la casa *	0,9956448	0,9989699
37	El sitio de las abras	0,9921862	0,9984721
38	El disco del tiempo *	1	1
39	Calypso *	0,9978536	0,9986555

# Obra	Nombre de la obra	Coseno	Jaccard
40	La loca de Gandoca	0,9959675	0,9985991
41	Pedro Páramo	0,9933793	0,9983946
42	Piedra sobre piedra	0,9872263	0,9982922
43	Nosotros los hombres *	1	1
44	Estudio en blanco y negro	0,9991062	0,9995708
45	La casa *	0,9996918	0,9997576
46	La casa de Bernarda Alba	0,9982198	0,999055
47	Terminal del sueño	0,9937913	0,9979898
48	El delantal blanco	0,9981589	0,9994291
49	Los infiernos de la mujer y algo más	0,9948305	0,9986282
50	El naranjo, o los círculos del tiempo *	0,9970161	0,9988387
51	Yo no vengo a decir un discurso	0,9877253	0,9977779
52	Mujeres de ojos grandes	0,9928895	0,9983152
53	La ruta de su evasión *	0,9945256	0,9988575
54	El señor presidente	0,9966356	0,9988421
55	Azul	0,999132	0,9993448
56	Cantos de vida y esperanza	0,9996857	0,9997933
57	Doña Bárbara	0,9978528	0,9993092
58	El Moto	0,9985024	0,9988662
59	Nuestra América	0,9986417	0,9990817
60	Invitación al diálogo de las generaciones	0,9863951	0,9977273
61	El Aleph	0,9992684	0,9993271
62	¡Alerta, ustedes!	0,9920715	0,9986802
63	Cuentos de Tata Mundo	0,9931838	0,998539
64	Mamita Yunai	0,9950589	0,9988798
65	Tu nombre será borrado del mundo	0,9959169	0,9988558
66	Cien años de soledad	0,9940043	0,9988896
67	Crónica de una muerte anunciada	0,9974247	0,9988666
68	Limón Blues *	0,9954966	0,9988085

Nota: * Solo se cuenta con un breve fragmento de la obra literaria.
Mayores magnitudes se muestran de color azul y menores magnitudes tienden a color celeste.
Fuente: elaboración propia con información de las Obras analizadas.

Adicionalmente, en la Tabla 4, se observa que la obra 11 “Concherías”, es la obra que presenta la menor similitud de coseno con el corpus de los ítems de la práctica de la PAA. Esta obra corresponde al género de lírica, se analiza en el nivel educativo de sétimo y es escrita por un autor nacional. Por otro lado, la obra con menor similitud del índice de Jaccard es obra “Invitación al diálogo de las generaciones”, esta corresponde al género literario de la época vanguardista, se lee a nivel de undécimo y fue escrita por un autor nacional.

Finalmente, se obtiene que el corpus de las obras literarias latinoamericanas de las lecturas recomendadas en el Tercer Ciclo y Educación Diversificada y el corpus los ítems de práctica de la PAA presentan una similitud de coseno de 0.9760 y del índice de Jaccard de 0.9994. Este resultado muestra que ambos corpus presentan asociación.

6.2 Modelos de estadísticos de clasificación

En el análisis de texto se utilizaron modelos de clasificación, en este estudio se aplican ocho modelos, se utilizan cuatro tipos de modelos clasificadores, entre ellos: Ingenuo de Bayes, Máquina de soporte vectorial con kernel polinomial, bosques aleatorios y máxima entropía. Luego a cada uno de estos se considera la estructura de la asignación latente de Dirichlet, esta es una estructura de componente latentes que agrupa por los términos según características que tengas las palabras entre sí.

Para este análisis se cuenta con aproximadamente setenta y tres mil palabras, sin embargo, se considera al menos un porcentaje de dispersión de los términos en ambos corpus de al menos 95%, ya que se presentan términos que aparecen una sola vez o dos veces en algunos de los documentos de obra literaria, con esto se pretende analizar en relación con la presencia de dispersión de la palabra en el corpus. Al final, se analizan más de dieciocho mil palabras.

6.2.1 Medidas de ajuste de los modelos

Las medidas de ajuste de los modelos contribuyen en cuantificar la calidad de la clasificación correcta o incorrecta de las categorías de la variable de respuesta. En este caso se considera la precisión, la sensibilidad, la especificidad y la precisión equilibrada. Estas métricas son consideradas en la matriz de confusión para saber la correcta clasificación de los modelos.

De acuerdo con los resultados de la Tabla 5, se muestran presiones mayores a 0.90 en todos los modelos estimados, en particular el modelo de bosques aleatorios presenta la mayor

magnitud de precisión de 0.9320 y modelo con la menor magnitud es ingenuo de Bayes (0.9061). Como en los modelos predomina la categoría de la no presencia de la palabra en el corpus de los ítems de práctica de la PAA, adecuadamente se considera la precisión equilibrada, específicamente de igual manera el modelo de bosques aleatorios presenta la mayor magnitud con 0.6684 y en menor medida el modelo de máquinas de soporte vectorial con kernel polinomial (0.51459).

Tabla 5. Medidas de ajuste de los modelos usados en la clasificación

Modelo o clasificador	Precisión	IC 95%	Sensibilidad	Especificidad	Precisión equilibrada
Ingenuo de Bayes	0.9061	(0.8981, 0.9137)	0.99342	0.04706	0.52024
Modelo Máquinas de soporte vectorial (SVM) con kernel polinomial	0.9086	(0.9007, 09161)	0.99781	0.03137	0.51459
Bosques aleatorios	0.9320	(0.9250, 0.9385)	0.99160	0.34510	0.66840
Máxima entropía	0.9079	(0.900, 0.9154)	0.99522	0.04902	0.52212

Fuente: elaboración propia con información de los corpus.

Posteriormente, se estiman los modelos considerando la estructura de ALD. En la Tabla 6 se muestran las medidas de ajuste, la precisión en todos los modelos es superior a 0.9. Los modelos de ingenuos de Bayes, máquina de soporte vectorial y máxima entropía poseen magnitudes de precisión similares. Con respecto a los modelos donde no se consideró la estructura latente, se muestra una leve mejoría en la precisión en el modelo de bosques aleatorios, con una magnitud de 0.9329. En los modelos de ingenuo de Bayes y máxima entropía no se muestra mejoría en la precisión.

Por otra parte, dado que predomina una de las categorías, se valora la precisión equilibrada y se muestra mejoría en la clasificación en los modelos ingenuo de Bayes, máquina de soporte vectorial y bosques aleatorios. En el modelo de bosques aleatorios se logra una mejora de 1,4% en la precisión equilibrada.

Tabla 6. Medidas de ajuste de los modelos usados en la clasificación considerando ALD

Modelo o clasificador	Precisión	IC 95%	Sensibilidad	Especificidad	Precisión equilibrada
Ingenuo de Bayes con ALD	0.9056	(0.8976, 0.9132)	0.99223	0.05294	0.52258
Modelo Máquinas de soporte vectorial (SVM) con kernel polinomial y con ALD	0.9086	(0.9007, 0.9161)	0.99761	0.03333	0.51547
Bosques aleatorios con ALD	0.9329	(0.9260, 0.9393)	0.99060	0.36470	0.67770
Máxima entropía con ALD	0.9074	(0.9074, 0.9149)	0.99482	0.04706	0.52094

Fuente: elaboración propia con información de los corpus.

Con estos resultados, se muestra que considerando la estructura de ALD se evidencia una leve mejoría en la precisión equilibrada en algunos de los modelos, puede que la estructura latente no esté detectando algún patrón sólido y excluyente en las palabras de los corpus, por lo que puede que la estructura no esté aportando alta información para la clasificación de los modelos. Sin embargo, los modelos en los que se consideró la estructura ALD presentan la mejor métrica en la precisión equilibrada.

CAPÍTULO VII. CONCLUSIONES

Según los desarrollado en este estudio de análisis de texto y cumpliendo con los objetivos establecidos, seguidamente se abordarán las conclusiones.

De acuerdo con el escalamiento multidimensional con el análisis semántico latente, existen obras que distan en su asociación con otras obras, se obtiene que las obras “Cien años de soledad”, “El señor presidente” y “Doña Barbara”, se distinguen de gran manera de las demás obras. En particular, son obras de personas autoras no costarricenses, que se estudian en niveles de educación diversificada y forman parte del género literario de la época vanguardista y modernista.

Con respecto a la construcción de los corpus lingüísticos, se construye un corpus de las obras literarias latinoamericanas de las lecturas recomendadas en el Tercer Ciclo y Educación Diversificada en el sistema educativo costarricense, en este se aplicó una metodología de construcción de corpus ejecutando las diferentes etapas. Está constituido por más de setenta mil palabras y tiene un porcentaje de esparcimiento del 95%.

Entre las palabras más comunes se encuentran adverbios y sustantivos como: “si”, “vez”, “casa”, “dos”, “tan” y “hombre”. Adicionalmente, se analizan biagramas entre las palabras y se observa una asociación marcada por las frases de lecturas predominantes en el corpus y algunos personajes de las obras. Sin embargo, también se muestran asociaciones entre palabras que en conjunto tienen un significado

Por otro lado, el corpus de los ítems de práctica de la Prueba de Aptitud Académica, al igual que el corpus de las lecturas recomendadas, siguió la misma metodología de construcción. Este corpus está constituido por más de mil setecientas palabras, presenta un esparcimiento del 21%.

Entre las palabras más frecuentes se encuentran: “solo”, “texto”, “anterior”, “cuál” y “dos”. Al igual que corpus de lecturas prevalecen sustantivos y adverbios. Las palabras más

frecuentes se encuentran relacionadas en la redacción de los enunciados de los ítems, específicamente cuando el ítem va dirigido a la lectura de un texto o de afirmaciones anteriormente expuestas.

De igual manera, en la asociación de los biagramas, se evidencia una conexión entre las palabras utilizadas en los enunciados, también se asocian palabras con ciertas temáticas relacionadas a los ítems, es decir, si algunas palabras se asocian a animales estas se encuentran asociadas.

Con relación al análisis de la estructura léxico semántica, se calcularon medidas de similitud textual entre los documentos y entre los corpus, las medidas utilizadas fueron la similitud de coseno y el índice de Jaccard. En relación con las asociaciones de los documentos de las obras literarias, entre ellos se encuentran asociaciones con magnitudes muy altas superiores al 0.9, lo que indica que las lecturas presentan alta asociación a nivel léxico. En algunos casos se estiman asociaciones con magnitud de 1, pues solo se obtuvo una pequeña parte de la obra, por lo que el conjunto de palabras se encuentra completamente en otras obras.

Por otra parte, con respecto a los documentos asociados a los ítems de práctica de la PAA, estos presentan una similitud de coseno y Jaccard de magnitud alta, ambos documentos mantienen una estructura léxica de alta asociación.

Ahora bien, en el análisis de la similitud de ambos corpus, se obtiene en la similitud de coseno y en el índice de Jaccard asociaciones con magnitud muy alta. En ambas medidas, resulta que ambos corpus tienen una estructura léxica semántica altamente asociada, por lo que el corpus lingüístico basado en las obras literarias y el corpus asociado a los ítems de práctica de la Prueba de Aptitud Académica comparten palabras entre sí.

Por otra parte, con respecto a los modelos de clasificación, estos presentan una alta precisión y detectan una buena clasificación en las palabras que no se encuentran el corpus

de los ítems de práctica de la PAA, por lo que en este caso se considera la precisión equilibrada.

Se obtiene que el modelo con la mejor precisión en la clasificación de las palabras del corpus lingüístico de las obras literarias y el corpus de los ítems de práctica de la PAA, es el modelo de bosques aleatorios. De igual forma, en relación con los modelos en donde se considera la estructura ALD, se observa que la mejor precisión equilibrada la obtiene igualmente en el modelo de bosques aleatorios. Por lo que, se sugiere considerar el modelo de bosques aleatorios como una opción para clasificar las palabras.

En general, en este estudio se concluye que el corpus lingüístico basado en las obras literarias y el corpus asociado a los ítems de práctica de la Prueba de Aptitud Académica presentan alta asociación léxica y se pueden aplicar modelos estadísticos y matemáticos para la clasificación de las palabras, los cuales pueden contribuir en la incorporación de nuevas palabras y si previamente existe una asociación.

En conclusión, una persona estudiante de secundaria que comprenda los términos de las obras recomendadas en los diferentes niveles educativos podría comprender de gran manera los términos que se encuentren en la redacción de los ítems de práctica de la PAA para ingresar a la Universidad de Costa Rica. No obstante, el hábito y la actitud hacia la lectura son fundamentales para la adquisición de vocabulario y, adicionalmente, las vivencias y las creencias son importantes en la comprensión lectura. Así pues, depende de la persona estudiante secundaria y qué tanto la persona docente en su proceso de enseñanza en la materia de español expuso y analizó las obras recomendadas y adicionalmente haya abordado la función gramatical de las palabras en el texto.

Algunas limitaciones que surgieron en el desarrollo de este análisis se encuentran en que no se tuvo acceso al total de las obras literarias recomendadas por el CSE, solo se analizó una parte por lo que pudo haber obras literarias que se ajustaran de buena forma a al estudio. Otro aspecto importante es que este análisis de corpus lingüísticos, por volumen y densidad de los corpus analizados, requiere un alto recurso y rendimiento computacional.

Finalmente, para futuros análisis similares se recomienda considerar no solamente obras literarias sino valorar otro tipo de texto recomendados. Por otro lado, en relación con las palabras de los corpus se recomendaría agregar adicionalmente la categoría gramatical o léxica, según la función de la palabra en el texto, pues esto contribuiría a identificar no solamente la palabra sino también su función a nivel de sintaxis, ya que por ejemplo algún término puede actuar en función de sustantivo o adjetivo en el texto. En los modelos de clasificación se sugiere efectuar una posible estratificación de las palabras del corpus de las obras literarias en donde se considere el género literario, época o características de la autoría de la obra.

Además, para futuros análisis similares se sugiere abordar algunas otras técnicas avanzadas del procesamiento de lenguaje natural y considerar otras funciones de enlace o kernel para verificar algunas variantes adicionales al modelamiento realizado, ya que este análisis fue un primer paso de comparación lingüística para contribuir en la construcción de ítems en una prueba estandarizada.

REFERENCIAS

Abarca-Petitjean, M., y Romero-Zúñiga, M. (1991). *Comprensión de Lectura en estudiantes de Primer año universitario*. https://repositoriotec.tec.ac.cr/bitstream/handle/2238/679/comprension_lectura_1.pdf?sequence=1

Abbasi, A., y Berrar, D. (2025). Explainable artificial intelligence for transparent medical image analysis. *Journal of Big Data, 12*, Article 127. https://doi.org/10.1186/s40537-025-01227-1

APPF.es (2020). *La importancia de la comprensión lectora en el aprendizaje*. https://www.appf.edu.es/la-importancia-de-la-comprension-lectora-en-el-aprendizaje/

Arancibia-Gutiérrez, B., y Leiva, F. (2022). Fluidez lectora, reconocimiento de palabras y velocidad lectora en escolares de 3° y 4° año de enseñanza básica. *Literatura y Lingüística, 46*, 367-388. https://doi.org/10.29344/0717621X.46.2673

Arancibia-Gutiérrez, B., Castro-Yáñez, G. G., Sáez-Carrillo, K., Barrientos, F., y Toloza, M. (2022). Comprensión de lectura, reconocimiento de palabras y fluidez lectora en escolares de sexto año básico. *Onomázein, 55*, 156-173. https://doi.org/10.7764/onomazein.55.05

Arias Orozco, G., y Vargas Valverde, L. (2016). Los textos literarios que asigna el MEP para el tercer ciclo de la educación general básica en Costa Rica: algunas reflexiones al respecto. *Actualidades investigativas en educación*, *16*(2), 1-17. https://www.scielo.sa.cr/pdf/aie/v16n2/1409-4703-aie-16-02-00022-gt1.pdf

Arias, G., y Vargas, L. (2016). Los textos literarios que asigna el MEP para el tercer ciclo de la educación general básica en Costa Rica: algunas reflexiones al respecto. *Actualidades Investigativas en Educación*, *16*(2), 1-17. https://doi.org/10.15517/aie.v16i2.23565

Banco Interamericano de Desarrollo [BID]. (2019). *PISA 2018 En América Latina: ¿Cómo nos fue en lectura?* https://publications.iadb.org/publications/spanish/document/Nota_PISA_18_PISA_2018_en_Am%C3%A9rica_Latina_C%C3%B3mo_nos_fue_en_lectura_es.pdf

Bayes, T. (1763), An Essay Towards Solving a Problem in the Doctrine of Chances, *Philosophical Transactions of the Royal Society* 53, 370-418. Reimpreso en in *Biometrika* 45 (1958), 296-315.Appendix, pp. 122-49

Bennett, K., Seashore, G. y Wesman, G. (1992). Tests de Aptitudes Diferenciales, DAT. *Manual Forma* T. Buenos Aires: Paidós.

Berger, A., Della Pietra, S. A., y Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, *22*(1), 39-71.

Blei, D., Ng, A., y Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*, 993-1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Bouchet-Valat, M. (2020). *SnowballC: Snowball stemmers based on the C 'libstemmer' UTF-8 library* (Versión 0.7.0) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=SnowballC

Brizuela-Rodríguez, A., Pérez, N., y Rojas, G. (2018). Respuestas guiadas por el experto: Validación de las inferencias basadas en los procesos de respuesta. *Actualidades Investigativas en Educación, 18*(3), 1-21. https://doi.org/10.15517/aie.v18i3.33456

Brizuela-Rodríguez, A., Rodríguez-Villagra, O. A., y Villalobos-Cardozo, L. M. (2021). Aportes de la psicología cognitiva al estudio y mejoramiento de la comprensión lectora en la educación superior. *Comunicación*, *30*(2), 4-17. https://www.scielo.sa.cr/scielo.php?pid=S1659-38202021000200004&script=sci_arttext

Calvo, K., Rojas, G., Pérez, N., y Ríos, A. J. (2019). Identificación de elementos irrelevantes para la comprensión de ítems de una prueba de razonamiento. *Revista Mexicana de Investigación Educativa*, *24*(81), 463-480. https://www.scielo.org.mx/scielo.php?pid=S1405-66662019000200463&script=sci_arttext

Carpio, M., y Méndez, E. (2016). Validación de contenido léxico de los textos para la comprensión lectora del Test de Lectura y Escritura en Español (LEE) para su aplicación en Costa Rica. *Actualidades Investigativas en Educación*, *16*(2), 74-102. https://doi.org/10.15517/aie.v16i2.23923

Carreras, M. A., Brizzio, A., Darricarrere, M. A., y Fernández Liporace, M. (2009). *La evaluación de las habilidades de razonamiento verbal y abstracto en estudiantes de diferentes carreras de la Universidad de Buenos Aires*. En I Congreso Internacional de Investigación y Práctica Profesional en Psicología XVI Jornadas de Investigación Quinto Encuentro de Investigadores en Psicología del MERCOSUR, Facultad de Psicología-Universidad de Buenos Aires. https://www.aacademica.org/000-020/749

Consejo Superior de Educación [CSE]. (2017). *Lista de Lecturas Recomendadas 2018*. https://sibeycra.mep.go.cr/wp-content/uploads/2019/09/literatura_recomendada_2018.pdf

Corella Esquivel, K. (2018). *Análisis de las estrategias de comprensión de lectura y las características de las redacciones de los estudiantes de primer año de la carrera de Ingeniería en Computación del Instituto Tecnológico de Costa Rica, Sede Regional*

San Carlos: una propuesta didáctica para el fortalecimiento de las habilidades para la compresión y la producción de documentos académicos [Tesis de maestría, Universidad Estatal a Distancia]. https://core.ac.uk/download/pdf/160496996.pdf

Crais, E. (1990). World knowledge to word knowledge. *Topics in language disorders*, *10*(3), 45-62.

De Mier, M. V., Borzone, A. M., y Cupani, M. (2012). La fluidez lectora en los primeros grados: Relación entre habilidades de decodificación, características textuales y comprensión. Un estudio piloto con niños hablantes de español. *Neuropsicologia Latinoamericana, 4*(spe), 18-33. https://doi.org/10.5579/rnl.2012.0079

Dragulescu, A. A. (2023). *xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files* (Versión 0.6.5) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=xlsx

Feinerer, I., y Hornik, K. (2023). *tm: Text mining package* (Versión 0.7-11) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=tm

Fellows, I. (2018). *wordcloud: Word clouds* (Versión 2.6) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=wordcloud

Gaikwad, S.V., Chaugule, A., y Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications, 85*(17). https://research.ijcaonline.org/volume85/number17/pxc3893507.pdf

García, E. (1993). La comprensión de textos. Modelo de procesamiento y estrategias de mejora. *Didáctica*, *5*(1), 87-113. https://revistas.ucm.es/index.php/DIDA/article/view/DIDA9393110087A/20216

Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied linguistics*, *25*(1), 1-37.

Georgia Department of Education (2008). *Accommodations manual: A guide to selecting, administering, and evaluating the use of test administration accommodations for students with disabilities.* Georgia: Georgia Department of Education.

Gohil, L. (2015) Text Mining: Process and Techniques. *International Journal of Innovative Research in Computer Science y Technology, 3*(3). https://ijircst.org/DOC/16_irp380906f6fff-4b9f-4686-a4a5-cfa526dd4c0b.pdf

Gomede, E. (2024). *Latent semantic analysis: Unveiling the hidden context of words and documents.* AI in Plain English. https://ai.plainenglish.io/latent-semantic-analysis-unveiling-the-hidden-context-of-words-and-documents-b782343b1df5

Gómez, L., y Silas, J. (2012). Impacto de un programa de comprensión lectora. *Revista Latinoamericana de Estudios Educativos*, *42*(3), 35-63. https://www.redalyc.org/pdf/270/27024686003.pdf

Hornik, K. (2023). *NLP: Natural language processing infrastructure* (Versión 0.2-1.1) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=NLP

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), 4*, 9-56. https://www.researchgate.net/publication/228743668_Similarity_measures_for_text_document_clustering

Huang, C-Y., Yang, C-L., y Hsiao, Y-H. (2021). A Novel Framework for Mining Social Media Data Based on Text Mining, Topic Modeling, Random Forest, and DANP Methods. *Mathematics*, 9(17), 2041. https://doi.org/10.3390/math9172041

Instituto Nacional de Estadística y Censos [INEC]. (2017). *Encuesta Nacional de Cultura 2016: Principales Resultados*. https://admin.inec.cr/sites/default/files/media/reenc2016-27092017_2.pdf

Isa, D., Lee, L. H., Kallimani, V. P., y Rajkumar, R. (2008). Text document preprocessing with the Bayes formula for classification using the support vector machine. *IEEE Transactions on Knowledge and Data engineering*, *20*(9), 1264-1272.

Jiménez, K., Rojas, G., Brizuela, A., y Pérez, N. (2018). Validación de un modelo de cuatro estrategias de resolución de ítems de razonamiento en una prueba estandarizada de selección. *Revista Costarricense de Psicología*, *37*(1), 77-88. http://dx.doi.org/10.22544/rcps.v37i01.04

Jurafsky, D., y Martin, J. H. (2024). Naive Bayes, Text Classification, and Sentiment. *Speech and Language Processing*, *3*(4), 60-94.

Justicia de la Torre, M.C. (2017). *Nuevas técnicas de minería de textos: Aplicaciones*. Universidad de Granada. http://hdl.handle.net/10481/46975

Kuhn, M. (2023). *caret: Classification and regression training* (Versión 6.0-94) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=caret

Kulkarni, S., y Rodd, S. F. (2020). Context Aware Recommendation Systems: A review of the state of the art techniques. *Computer Science Review*, *37*(2), 100255. http://dx.doi.org/10.1016/j.cosrev.2020.100255

La Gaceta. (2019). *La Gaceta N° 114 del 19 de junio del 2019. "Reforma de los artículos 44° y 61° el Capítulo V e inclusión de transitorios del Reglamento de Evaluación de*

los Aprendizajes - Decreto Ejecutivo Nº 40862-MEP". https://www.imprentanacional.go.cr/pub/2019/06/19/COMP_19_06_2019.html

La Gaceta Universitaria (2003). *Alcance a La Gaceta Universitaria 01-2003. "Reglamento del Proceso de Admisión Mediante Prueba de Aptitud Académica"*. https://www.cu.ucr.ac.cr/gacetas/2003/a01-2003.pdf

Leighton, J. P., y Sternberg, R. J., (2004). *The Nature of Reasoning.* United States: Cambridge University Press.

Ligges, U., y Mächler, M. (2003). *scatterplot3d: 3D scatter plot* (Versión 0.3-44) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=scatterplot3d

Linguamatics (2021). What is Text Mining, Text Analytics and Natural Language Processing?

Martínez, A. (2023). El Ministerio de Educación Pública (MEP) destaca necesidad de redoblar esfuerzos en zonas con notas deficientes. *Delfino*. https://delfino.cr/2023/07/pruebas-diagnosticas-del-mep-senalan-poca-diferencia-entre-centros-educativos-privados-y-publicos

Martínez, A. (2023). El Ministerio de Educación Pública (MEP) destaca necesidad de redoblar esfuerzos en zonas con notas deficientes. https://delfino.cr/2023/07/pruebas-diagnosticas-del-mep-senalan-poca-diferencia-entre-centros-educativos-privados-y-publicos

McKeown, M.G. y Curtis, M.E. (eds) (1987) *The Nature of Vocabulary Acquisition.* Lawrence Erlbaum Associates, Hillsdale, N.J.

Medellín, A., y Rodríguez, I. (2014). Propuesta metodológica para la evaluación de vocabulario académico a través de la lingüística de corpus. *RLA. Revista de lingüística teórica y aplicada, 52*(2), 41-63. http://dx.doi.org/10.4067/S0718-48832014000200003

Ministerio de Educación Pública [MEP]. (2017). *Programa de Español Comunicación y comprensión lectora. Tercer Ciclo y Educación Diversificada*. https://www.mep.go.cr/sites/default/files/media/espanol3ciclo_diversificada.pdf

Ministerio de Educación Pública [MEP]. (2021). *Guía técnica 1: FARO Secundaria. I Semestre 2021*. Dirección de Gestión y Evaluación de la Calidad. https://coned.ac.cr/images/documentos/decimo/faro/guia-tecnica-faro-secundaria.pdf

Ministerio de Educación Pública [MEP]. (2022). *MEP comparte resultados de las Pruebas Nacionales FARO en Secundaria*. https://www.mep.go.cr/noticias/mep-comparte-resultados-pruebas-nacionales-faro-secundaria

Ministerio de Educación Pública [MEP]. (2024). *Marco de Especificaciones Prueba Nacional Estandarizada Secundaria*. https://dgec.mep.go.cr/wp-content/uploads/2024/03/MARCO-DE-ESPECIFICACIONES-SECUNDARIA-2024_FINAL.pdf

Ministerio de Educación Pública [MEP]. (s.f.). *Información técnica de los resultados de las Pruebas Nacionales FARO*. https://dgecold.mep.go.cr/sites/all/files/dgec_mep_go_cr/documentos/enlace_de_informacion_tecnica_del_analisis_de_los_resultados.pdf

Montero, E., Rojas, S., y Zamora, E. (2014). *Quinto informe del Estado de la Educación: Costa Rica En Las Pruebas Pisa 2012 (Programa Internacional Para La Evaluación De Los Estudiantes)*. https://repositorio.conare.ac.cr/server/api/core/bitstreams/47937d91-c78c-43ae-adfc-a72ff8b96ab6/content

Montero-Rojas, E., Rojas-Rojas, G., Negrín-Hernández, M., y Francis-Salazar, S. (2015). Efecto de una capacitación sobre los puntajes de la prueba de admisión de la Universidad de Costa Rica: una aproximación bayesiana. *Actualidades en Psicología*, *29*(119), 115-130. https://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S2215-35352015000200115

Moreira-Mora, T. (2021). Propiedades Psicométricas de una prueba de admisión universitaria. *Revista Evaluar, 21(1)*, 73-93. https://revistas.unc.edu.ar/index.php/revaluar/article/download/32833/33554/111897

Mosteller, F., y Wallace, D. (1964). *Inference and disputed authorship: the federalist*. Massachusetts: Addison-Wesley.

Mullen, L., Brumfield, S., y Rinker, T. W. (2018). *tokenizers: Fast, consistent tokenization of natural language text* (Versión 0.3.0) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=tokenizers

Munzert, S., Rubba, C., Meißner, P., y Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining.* John Wiley & Sons.

Muñoz, L., y Díaz-Soucy, M.C. (2021). *Igualdad de oportunidades en el sistema educativo costarricense. Principios, estrategias y resultados*. Instituto de Investigación en Educación. https://repositorio.inie.ucr.ac.cr/server/api/core/bitstreams/b35337f9-224e-4c77-9517-bcb5836b153d/content

Murillo, M. (2009). Diversidad de vocabulario en los preescolares. Aportes para valorar su competencia léxica. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, *35*(1), 123-138. https://doi.org/10.15517/rfl.v35i1.1271

Murillo, M. (2012). La evaluación del vocabulario en la educación preescolar: El Tevopreesc. *Revista Káñina*, *36*(2), 151-162. https://www.redalyc.org/pdf/442/44249254016.pdf

Murillo, M., Araya, J., y Barquero, K. (2023). *Situación de la pobreza de los aprendizajes en la población estudiantil de 10 años en el período 2019-2022*. [Disponible en Noveno Informe Estado de la Educación 2023]. https://repositorio.conare.ac.cr/server/api/core/bitstreams/271fd61f-1e26-410f-a509-d07533045b8e/content

Naciones Unidas. (2007). Declaración de las Naciones Unidas sobre los Derechos de los Pueblos Indígenas. Resolución 61/295. https://www.un.org/esa/socdev/unpfii/documents/DRIPS_es.pdf

Naciones Unidas (1979). Convención sobre la eliminación de todas las formas de discriminación contra la mujer. Resolución 34/180 https://www.ohchr.org/sites/default/files/cedaw_SP.pdf

Naciones Unidas (1948). Declaración Universal de Derechos Humanos, 10 de diciembre de 1948. Resolución 217 A (III), A/RES/217(III). https://www.un.org/es/about-us/universal-declaration-of-human-rights

Naciones Unidas. (1959). Declaración de los Derechos del Niño (Resolución 1386 (XIV)). https://digitallibrary.un.org/record/195831?ln=es&v=pdf

Ooms, J. (2023). *pdftools: Text extraction, rendering and converting of PDF documents* (Versión 3.4.0) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=pdftools

Organización para la Cooperación y el Desarrollo Económico [OCDE]. (2010). *Resultados de Pisa 2009: Lo que los estudiantes saben y pueden hacer: Rendimiento estudiantil en lectura, matemáticas y ciencias (Volumen I)*. OCDE. https://www.oecd.org/content/dam/oecd/es/publications/reports/2010/12/pisa-2009-results-what-students-know-and-can-do_g1g114e2/9789264174900-es.pdf

Organización para la Cooperación y el Desarrollo Económico [OCDE]. (2021). *Programme for International Student Assessment. PISA en español*. https://www.oecd.org/en/about/programmes/pisa.html

Pang, B., Lee, L., y Vaithyanathan, S. (2002). *Thumbs up? Sentiment classification using machine learning techniques* [arXiv preprint cs/0205070]. https://doi.org/10.48550/arXiv.cs/0205070

Pedersen, T. L. (2023). *ggraph: An implementation of grammar of graphics for graphs and networks* (Versión 2.1.0) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=ggraph

Perera, G. G., y Segura, J. M. (2004). *La comprensión lectora como pilar esencial para el aprendizaje del alumnado en todas las áreas curriculares*. CEIP Santa Lucía.

Programa Permanente de la Prueba de Aptitud Académica-IIP (2025a). *Práctica para la Prueba de Aptitud Académica. Folletos de práctica. Publicaciones.* https://www.paa.iip.ucr.ac.cr/folletosdepractica/

Programa Permanente de la Prueba de Aptitud Académica-IIP. (2025b). *Preguntas frecuentes*. https://www.paa.iip.ucr.ac.cr/preguntasfrecuentes/

Puyuelo, M., Rondal, J. A., y Wiig, E. (2000). Evaluación del lenguaje. *vol*, *1*, 9-17.

R Core Team. (2023). R: A language and environment for statistical computing [Software]. R Foundation for Statistical Computing. https://www.R-project.org/

Ramírez, C. (2018). Un compás que hace cuadrados: La escritura y la lectura en el currículum de séptimo año en Costa Rica. *Actualidades Investigativas en Educación*, *18*(1), 3-32. https://doi.org/10.15517/aie.v18i1.30704

Regueyra, M. G., y Argüello, S. (2018). Superando mitos sobre la comprensión lectora en la población estudiantil universitaria. *Káñina*, *42*(1), 33-49. https://doi.org/10.15517/rk.v42i1.32941

Rodríguez, L. (2019). *Noticias MEP. País mantiene promedio en los resultados de PISA 2018*. https://www.mep.go.cr/noticias/pais-mantiene-promedio-resultados-pisa-2018

Rojas, L. (2014). Evidencias de validez de la Prueba de Aptitud Académica de la Universidad de Costa Rica basadas en su estructura interna. *Actualidades en psicología*, *28*(116), 15-26. https://doi.org/10.15517/ap.v28i116.14889

Rojo, G. (2014). Hispanic corpus linguistics. *The Routledge handbook of Hispanic applied linguistics* (pp. 371-387). Routledge.

Schölkopf, B., y Smola, A.J. (1998). *Learning With Kernels* (Vol. 4). GMD-Forschungszentrum Informationstechnik.

Silge, J., y Robinson, D. (2022). *tidytext: Text mining using 'dplyr', 'ggplot2', and other tidy tools* (Versión 0.4.1) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=tidytext

Talib, R., Hanif, M. K., Ayesha, S., y Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications,* *7*(11), 414-418. https://thesai.org/Downloads/Volume7No11/Paper_53-Text_Mining_Techniques_Applications_and_Issues.pdf

Taylor, J. R. (2017). Lexical Semantics. In B. Dancygier (Ed.), *The Cambridge Handbook of Cognitive Linguistics* (pp. 246–261). Cambridge University Press.

The glossary of education reform. (2015). *Standardized Test*. https://www.edglossary.org/standardized-test/

Tristán-López, A., y Pedraza, N. (2017). La objetividad en las pruebas estandarizadas. *Revista Iberoamericana de evaluación educativa*, *10*(1), 11-31. https://revistas.uam.es/riee/article/view/7592

van der Loo, M. P. J. (2023). *stringdist: Approximate string matching and string distance functions* (Versión 0.9.10) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=stringdist

Vilarino, D., Tovar, M., Beltrán, B., y León, S. (2014). Un modelo para detectar la similitud semántica entre textos de diferentes longitudes. *Research in Computing Science, 85*, 57-64. https://www.rcs.cic.ipn.mx/2014_85/Un%20modelo%20para%20detectar%20la%20similitud%20semantica%20entre%20textos%20de%20diferentes%20longitudes.pdf

Walker, A., y Schauberger, P. (2023). *openxlsx: Read, write and edit xlsx files* (Versión 4.2.5.2) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=openxlsx

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., y Dunnington, D. (2023). *ggplot2: Create elegant data visualisations using the grammar of graphics* (Versión 3.4.4) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=ggplot2

Wickham, H., François, R., Henry, L., y Müller, K. (2023). *dplyr: A grammar of data manipulation* (Versión 1.1.3) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=dplyr

Wild, F. (2015). *lsa: Latent semantic analysis* (Versión 0.73-1) [Paquete de R]. CRAN. https://CRAN.R-project.org/package=lsa

ANEXOS

Anexo 1. Lista de textos literarios, y su correspondiente dosificación, para Tercer Ciclo (rama académica y técnica) y Educación Diversificada (académica) recomendada por el Consejo Superior de Educación

Séptimo (9 textos)		
Cuento		
Autor	**Título**	**Dosificación**
Chang, Guiselle (compiladora)	Cuentos tradicionales afrolimonenses	3 textos
González Zeledón, Manuel	Cuentos de Magón	
Poe, Edgar Allan	Narraciones extraordinarias	
Quiroga, Horacio	Cuentos de amor, de locura y de muerte	
Quiroga, Horacio	Cuentos de la selva	
Lyra, Carmen	Narrativa de Carmen Lyra	
Stoker, Bram	El huésped de Drácula	
Novela		
Argüello Mora, Manuel	Elisa Delmar	1 texto
Gutiérrez, Joaquín	Cocorí	
Ortega Rodríguez, Manuel	Los peces de Cooper	
Roswell, Víctor	El canto de los quetzales	
Rovinski, Yanini	Una montaña de aserrín	
Santa Ana, Antonio	Los ojos del perro siberiano	
Swift, Jonathan	Los viajes de Gulliver	
Verne, Julio	Viaje al centro de la Tierra	
Lírica		
Benedetti, Mario	Inventarios I y II	2 textos
Campbell, Shirley	Rotundamente negra y otros poemas	

Echeverría, Aquileo	Concherías	
Fajardo Korea, Miguel	Casa Guanacaste	
Marchena, Julián	Alas en fuga	
Parra, Nicanor	Poemas y antipoemas	
Drama		
Arroyo, Jorge	La tea fulgurante	1 texto
Lyra, Carmen	Había una vez	
Tagore, Rabindranath	El cartero del rey	
Ensayo		
Naranjo, Carmen	Cinco temas en busca de un pensador	1 texto
Relato autobiográfico		
Frank, Ana	Diario	1 texto
Octavo (10 textos)		
Cuento		
Allende, Isabel	Cuentos de Eva Luna	3 textos
Conan Doyle, Arthur	Aventuras de Sherlock Holmes	
Cortés, Carlos	La última aventura de Batman	
Duncan, Quince	Cuentos escogidos	
Faingezicht, Vilma	Cuentos de la niña judía	
Fernández, Guillermo	Efecto invernadero	
Mastretta, Ángeles	El mundo iluminado	
Rulfo, Juan	El llano en llamas	
Vargas Pizarro, Maureen	Danzas del bosque	
Novela		
Boyne, John	El niño con el pijama de rayas	2 textos
Contreras, Fernando	Cierto azul	

Hosseini, Khaled	Cometas en el cielo	
Jiménez, Max	El jaúl	
León Sánchez, José	La isla de los hombres solos	
Lobo, Tatiana	Asalto al paraíso	
Lyra, Carmen	En una silla de ruedas	
Mora, Jairo	El oscuro corazón de Talamanca	

Sepúlveda, Luis	Historia de una gaviota y del gato que le enseñó a volar	
Verne, Julio	Veinte mil leguas de viaje submarino	
Zúñiga Arias, Ana Yolanda y Masis Olivas, Sergio	Más abajo del aire	
Lírica		
Bernard, Eularia	Ciénaga	2 textos
Debravo, Jorge	Los despiertos	
Dobles, Julieta	Los trabajos de Pandora	
Gallegos, Mía	Los sueños y los días	
García Lorca, Federico	Romancero Gitano	
Neruda, Pablo	Veinte poemas de amor y una canción desesperada	
Drama		
Cañas Escalante, Alberto	Ni mi casa es ya mi casa	2 textos
Chejov, Anton	Petición de mano	
Fernández Guardia, Ricardo	Magdalena	
Méndez, Melvin	Eva, sol y sombra	
Méndez, Melvin	Un viejo con alas	
Ensayo		
Oreamuno, Yolanda	A lo largo del corto camino	1 texto
Noveno (10 textos)		
Cuento		
Cortázar, Julio	Bestiario	
Fernández, Guillermo	Hagamos un ángel	2 textos

Fernández Guardia, Ricardo	Cuentos ticos	
Garro, Elena	La culpa es de los tlaxcaltecas y otros relatos	

Oreamuno, Yolanda	Relatos	
Ovares, flora y Rojas, Margarita	La ciudad imaginada	
Oviedo, Marietta	Fue en abril	
Porras, Santiago	Cuentos Guanacastecos	
Salazar Herrera, Carlos	Cuentos de angustias y paisajes	
Novela		
Carrol, Lewis	Alicia en el país de las maravillas	2 textos
Chavarría Gómez, Rafael Ángel	Los ojos de abril	
Contreras, Fernando	Única mirando al mar	
De Sosa, Geovanny	Los dueños de la casa	
Dobles, Fabián	El sitio de las abras	
Faingezicht, Aída	Azulejos blancos	
García Esperón, María	El disco del tiempo	
Gutiérrez, Joaquín	Puerto Limón	
Hernández, Luisa Josefina	Una noche para Bruno	
Levine, Karen	La maleta de Hannah	
Lobo, Tatiana	Calypso	
Pinto, Julieta	El eco de los pasos	
Porras, Santiago	Avancari	
Rossi Anacristina	La loca de Gandoca	
Rulfo, Juan	Pedro Páramo	
Soto, Rodrigo	El país de la lluvia	
Wohlstein, Harry	Piedra sobre piedra	
Lírica		
Debravo, Jorge	Nosotros los hombres	2 textos
Gelman, Juan	Hechos y relaciones	
Mc Donald, Delia Woolery	Todas las voces que canta el mar	
Sauma, Osvaldo	Retrato en familia	
Sauma, Osvaldo	El libro del adiós	
Drama		

Carballido, Emilio	Estudio en blanco y negro	2 textos
Gallegos, Daniel	La casa	

García Lorca, Federico	La casa de Bernarda Alba	
Istarú, Ana	Madre nuestra que estás en el cielo	
Méndez, Melvin	Terminal del sueño	
Shakespeare, William	El mercader de Venecia	
Vodanovic, Sergio	El delantal blanco	
Ensayo		
Calvo, Yadira	La mujer víctima y cómplice	2 textos
Montero, Rosa	Historias de mujeres	
Quesada, Juan Rafael	Clarín patriótico	
Vargas, Armando	El lado oculto del presidente Mora	
Décimo (10 textos)		
Época Clásica		
Esquilo	Orestíada	2 textos
Homero	Odisea	
Sófocles	Edipo Rey	
Época renacentista y Siglo de Oro		
Calderón de la Barca, Pedro	La vida es sueño	2 textos
Cervantes, Miguel	El ingenioso hidalgo don Quijote de La Mancha (I y II parte)	
Shakespeare, William	Sueño de una noche de verano	
de Vega, Lope	Fuenteovejuna	
Siglo XIX (épocas romántica, realista y naturalista)		
Austin, Jane	Orgullo y prejuicio	2 textos
Brontë, Emily	Cumbres borrascosas	
Dostoyevski, Mijáilovich	Crimen y castigo	
Dumas, Alexandre	El conde de Montecristo	
Flauberth, Gustavo	Madame Bovary	
Shelley, Mary	Frankenstein o el moderno Prometeo	
Stoker, Bram	Drácula	

Época vanguardista		
McDonald, Delia y Campbell, Shirley	Palabras indelebles de poetas negras	4 textos (dos ensayos y otros dos textos)
de Vallbona, Rima	Los infiernos de la mujer y algo más	
Duncan, Quince	Un mensaje de Rosa	
Fernández, Guillermo	Babelia	
Fuentes, Carlos	El naranjo, o los círculos del tiempo	
García Márquez, Gabriel	Yo no vengo a decir un discurso	
Mastretta, Ángeles	Mujeres de ojos grandes	
Oreamuno, Yolanda	La ruta de su evasión	
Pacheco, Abel	Más abajo de la piel	
Prifer Friedman, Gustavo	El quinto mandamiento	
Quijano Vincenzi, Laura	Señora del tiempo	
Varios (compilación)	Antología poética de la generación del 27	
Víquez, Alí	El coraje de leer	
Undécimo (10 textos)		
Época neoclásica y posromántica		
Baudelaire, Charles	Las flores del mal	2 textos
Defoe, Daniel	Robinson Crusoe	
Moliere	Don Juan	
Whitman, Walt	Hojas de hierba	
Época modernista		
Asturias, Miguel Ángel	El Señor Presidente	2 textos
Darío, Rubén	Azul	
Darío, Rubén	Cantos de vida y esperanza	
Gallegos, Rómulo	Doña Bárbara	
García Monge, Joaquín	El Moto	
Martí, José	Nuestra América	
Orwell, George	Rebelión en la granja	
Woolf, Virginia	Orlando	
Época vanguardista		

Azofeifa, Isaac Felipe	Invitación al diálogo de las	6 textos
	generaciones	(dos ensayos y otros cuatro textos)
Borges, Jorge Luis	El Aleph	
Caamaño, Sonia	500 años después	
Casona, Alejandro	Prohibido suicidarse en primavera	
Chase, Alfonso	Mirar con inocencia	
Dobles, Fabián	¡Alerta, ustedes!	
Dobles, Fabián	Cuentos de Tata Mundo	
Fallas, Carlos Luis	Mamita Yunai	
Fernández, Guillermo	Tu nombre será borrado del mundo	
García Márquez, Gabriel	Cien años de soledad	
García Márquez, Gabriel	Crónica de una muerte anunciada	
González, Edelmira	Yo soy Marlín	
Gutiérrez, Joaquín	Murámonos, Federico	
Kafka, Franz	Metamorfosis	
Rossi, Anacristina	Limón Blues	
Salaverry, Arabella	Impúdicas	
Süskind, Patrick	El perfume	

Lista de textos literarios, y su correspondiente dosificación, para Educación Diversificada (modalidad técnica)

Décimo (8 textos)		
Época Clásica		
Esquilo	Orestíada	2 textos
Homero	Odisea	
Sófocles	Edipo Rey	
Época renacentista y Siglo de Oro		
Calderón de la Barca, Pedro	La vida es sueño	2 textos
Cervantes, Miguel	El ingenioso hidalgo don Quijote de La Mancha (I y II parte)	
Shakespeare, William	Sueño de una noche de verano	
Vega, Lope de	Fuenteovejuna	

Época vanguardista		
McDonald, Delia y Campbell, Shirley	Palabras indelebles de poetas negras	4 textos (dos ensayos y
de Vallbona, Rima	Los infiernos de la mujer y algo más	otros dos textos)
Duncan, Quince	Un mensaje de Rosa	
Fernández, Guillermo	Babelia	
Fuentes, Carlos	El naranjo, o los círculos del tiempo	
García Márquez, Gabriel	Yo no vengo a decir un discurso	
Mastretta, Ángeles	Mujeres de ojos grandes	
Oreamuno, Yolanda	La ruta de su evasión	
Pacheco, Abel	Más abajo de la piel	
Prifer Friedman, Gustavo	El quinto mandamiento	
Quijano Vincenzi, Laura	Señora del tiempo	
Varios (compilación)	Antología poética de la generación del 27	
Víquez, Alí	El coraje de leer	
Undécimo (8 textos)		
Siglo XIX (épocas romántica, realista y naturalista)		
Austin, Jane	Orgullo y prejuicio	2 textos
Brontë, Emily	Cumbres borrascosas	
Dostoyevski, Mijáilovich	Crimen y castigo	
Dumas, Alexandre	El conde de Montecristo	
Flauberth, Gustavo	Madame Bovary	
Shelley, Mary	Frankenstein o el moderno Prometeo	
Stoker, Bram	Drácula	
Época neoclásica y posromántica		
Baudelaire, Charles	Las flores del mal	2 textos
Defoe, Daniel	Robinson Crusoe	
Moliere	Don Juan	
Whitman, Walt	Hojas de hierba	
Época modernista		
Asturias, Miguel Ángel	El Señor Presidente	2 textos

Darío, Rubén	Azul	
Darío, Rubén	Cantos de vida y esperanza	
Gallegos, Rómulo	Doña Bárbara	
Martí, José	Nuestra América	
Orwell, George	Rebelión en la granja	
Woolf, Virginia	Orlando	
Época vanguardista		
Borges, Jorge Luis	El Aleph	2 textos (un ensayo y otro texto)
Casona, Alejandro	Prohibido suicidarse en primavera	
Chase, Alfonso	Mirar con inocencia	
Dobles, Fabián	¡Alerta, ustedes!	
Fernández, Guillermo	Tu nombre será borrado del mundo	
García Márquez, Gabriel	Crónica de una muerte anunciada	
González, Edelmira	Yo soy Marlín	
Kafka, Franz	Metamorfosis	
Rossi, Anacristina	Limón Blues	
Süskind, Patrick	El perfume	
Duodécimo (4 textos)		
Época vanguardista		
Azofeifa, Isaac Felipe	Invitación al diálogo de las generaciones	4 textos (un ensayo y otros tres textos)
Caamaño, Sonia	500 años después	
Dobles, Fabián	¡Alerta, ustedes!	
Dobles, Fabián	Cuentos de Tata Mundo	
Fallas, Carlos Luis	Mamita Yunai	
García Márquez, Gabriel	Cien años de soledad	
Gutiérrez, Joaquín	Murámonos, Federico	
Salaverry, Arabella	Impúdicas	

Anexo 2. Listado de las obras literarias analizadas

# Obra	Autor	Titulo	Nivel educativo	Género Literario	Autor de Costa Rica
1	González Zeledón, Manuel	Cuentos de Magón	Sétimo	Cuento	Sí
2	Quiroga, Horacio	Cuentos de amor, de locura y de muerte	Sétimo	Cuento	No
3	Quiroga, Horacio	Cuentos de la selva	Sétimo	Cuento	No
4	Argüello Mora, Manuel	Elisa Delmar	Sétimo	Novela	Sí
5	Gutiérrez, Joaquín	Cocorí	Sétimo	Novela	Sí
6	Roswell, Víctor	El canto de los quetzales	Sétimo	Novela	Sí
7	Santa Ana, Antonio	Los ojos del perro siberiano	Sétimo	Novela	No
8	Calvo, Yadira	La mujer víctima y cómplice	Noveno	Ensayo	Sí
9	Benedetti, Mario	Inventarios I y II	Sétimo	Lírica	No
10	Campbell, Shirley	Rotundamente negra y otros poemas	Sétimo	Lírica	Sí
11	Echeverría, Aquileo	Concherías	Sétimo	Lírica	Sí
12	Parra, Nicanor	Poemas y antipoemas	Sétimo	Lírica	No
13	Lyra, Carmen	Había una vez	Sétimo	Drama	Sí
14	Naranjo, Carmen	Cinco temas en busca de un pensador	Sétimo	Ensayo	Sí
15	Allende, Isabel	Cuentos de Eva Luna	Octavo	Cuento	No
16	Cortés, Carlos	La última aventura de Batman	Octavo	Cuento	Sí
17	Mastretta, Ángeles	El mundo iluminado	Octavo	Cuento	No
18	Rulfo, Juan	El llano en llamas	Octavo	Cuento	No
19	Vargas Pizarro, Maureen	Danzas del bosque	Octavo	Cuento	Sí
20	Contreras, Fernando	Cierto azul	Octavo	Novela	Sí
21	Jiménez, Max	El jaúl	Octavo	Novela	Sí
22	Lobo, Tatiana	Asalto al paraíso	Octavo	Novela	Sí
23	Lyra, Carmen	En una silla de ruedas	Octavo	Novela	Sí
24	Sepúlveda, Luis	Historia de una gaviota y del gato que le enseñó a volar	Octavo	Novela	No
25	Debravo, Jorge	Los despiertos	Octavo	Lírica	Sí
26	Neruda, Pablo	Veinte poemas de amor y una canción desesperada	Octavo	Lírica	No

# Obra	Autor	Titulo	Nivel educativo	Género Literario	Autor de Costa Rica
27	Cañas Escalante, Alberto	Ni mi casa es ya mi casa	Octavo	Drama	Sí
28	Fernández Guardia, Ricardo	Magdalena	Octavo	Drama	Sí
29	Méndez, Melvin	Un viejo con alas	Octavo	Drama	Sí
30	Oreamuno, Yolanda	A lo largo del corto camino	Octavo	Ensayo	Sí
31	Cortázar, Julio	Bestiario	Noveno	Cuento	No
32	Fernández Guardia, Ricardo	Cuentos ticos	Noveno	Cuento	Sí
33	Garro, Elena	La culpa es de los tlaxcaltecas y otros relatos	Noveno	Cuento	No
34	Salazar Herrera, Carlos	Cuentos de angustias y paisajes	Noveno	Cuento	Sí
35	Contreras, Fernando	Única mirando al mar	Noveno	Novela	Sí
36	De Sosa, Geovanny	Los dueños de la casa	Noveno	Novela	Sí
37	Dobles, Fabián	El sitio de las abras	Noveno	Novela	Sí
38	García Esperón, María	El disco del tiempo	Noveno	Novela	No
39	Lobo, Tatiana	Calypso	Noveno	Novela	Sí
40	Rossi, Anacristina	La loca de Gandoca	Noveno	Novela	Sí
41	Rulfo, Juan	Pedro Páramo	Noveno	Novela	No
42	Wohlstein, Harry	Piedra sobre piedra	Noveno	Novela	Sí
43	Debravo, Jorge	Nosotros los hombres	Noveno	Lírica	Sí
44	Carballido, Emilio	Estudio en blanco y negro	Noveno	Drama	No
45	Gallegos, Daniel	La casa	Noveno	Drama	No
46	García Lorca, Federico	La casa de Bernarda Alba	Noveno	Drama	No
47	Méndez, Melvin	Terminal del sueño	Noveno	Drama	Sí
48	Vodanovic, Sergio	El delantal blanco	Noveno	Drama	No
49	de Vallbona, Rima	Los infiernos de la mujer y algo más	Décimo	Época vanguardista	Sí

# Obra	Autor	Titulo	Nivel educativo	Género Literario	Autor de Costa Rica
50	Fuentes, Carlos	El naranjo, o los círculos del tiempo	Décimo	Época vanguardista	No
51	García Márquez, Gabriel	Yo no vengo a decir un discurso	Décimo	Época vanguardista	No
52	Mastretta, Ángeles	Mujeres de ojos grandes	Décimo	Época vanguardista	No
53	Oreamuno, Yolanda	La ruta de su evasión	Décimo	Época vanguardista	Sí
54	Asturias, Miguel Ángel	El señor presidente	Undécimo	Época modernista	No
55	Darío, Rubén	Azul	Undécimo	Época modernista	No
56	Darío, Rubén	Cantos de vida y esperanza	Undécimo	Época modernista	No
57	Gallegos, Rómulo	Doña Bárbara	Undécimo	Época modernista	No
58	García Monge, Joaquín	El Moto	Undécimo	Época modernista	Sí
59	Martí, José	Nuestra América	Undécimo	Época modernista	No
60	Azofeifa, Isaac Felipe	Invitación al diálogo de las generaciones	Undécimo	Época vanguardista	Sí
61	Borges, Jorge Luis	El Aleph	Undécimo	Época vanguardista	No
62	Dobles, Fabián	¡Alerta, ustedes!	Undécimo	Época vanguardista	Sí
63	Dobles, Fabián	Cuentos de Tata Mundo	Undécimo	Época vanguardista	Sí
64	Fallas, Carlos Luis	Mamita Yunai	Undécimo	Época vanguardista	Sí
65	Fernández, Guillermo	Tu nombre será borrado del mundo	Undécimo	Época vanguardista	Sí
66	García Márquez, Gabriel	Cien años de soledad	Undécimo	Época vanguardista	No

# Obra	Autor	Titulo	Nivel educativo	Género Literario	Autor de Costa Rica
67	García Márquez, Gabriel	Crónica de una muerte anunciada	Undécimo	Época vanguardista	No
68	Rossi, Anacristina	Limón Blues	Undécimo	Época vanguardista	Sí

Anexo 3. Código R del análisis

Librerías utilizadas

```{r,include=FALSE}
knitr::opts_chunk$set(echo=TRUE)
#install.packages("pdftools")
#install.packages("tidyverse")
#install.packages("tokenizers")
#install.packages("stringdist")
#install.packages("hashr")
#install.packages("flextable")
#installklippyforcopy-to-clipboardbuttonincodechunks
#install.packages("remotes")
#remotes::install_github("rlesur/klippy")
#install.packages("tm")
#install.packages("ggplot2")
#install.packages("lsa")
#install.packages("scatterplot3d")
#install.packages("SnowballC")
#install.packages("NLP")
#install.packages("xlsx")
#install.packages("wordcloud")
#install.packages("caret")
#install.packages("corrplot")
#install.packages("viridis")
#install.packages("openxlsx")
#install.packages("dplyr")
#install.packages("tidytext")
#install.packages("ggraph")
library(igraph)
library(dplyr)
library(grid)
library(ggraph)
library(ggraph)
```

```
library(tidytext)
library(tidyr)
library(wordcloud)
library(NLP)
library(tm)
library(ggplot2)
library(scatterplot3d)
library(SnowballC)
library(lsa)
library(pdftools)
library(tidyverse)
library(tokenizers)
library(stringdist)
library(hashr)
library(tidyverse)
library(flextable)
library(readxl)
library(caret)
library(xlsx)
library(ggplot2)
library(foreign)
library(corrplot)
library(viridis)
library(openxlsx)
#activateklippyforcopy-to-clipboardbutton
klippy::klippy()

rem_dup_word<-function(x){
paste(unique(trimws(unlist(strsplit(x,split="",fixed=F,perl=T)))),collapse=
"")
}
```

Descriptivos de Lecturas

```{r}

Datos_Lecturas<-read_excel("C:/Users/Meli/Documents/Trabajo Final
Maestria/Datos_Lecturas.xlsx")
View(Datos_Lecturas)

t1<-table(Datos_Lecturas$Nivel)
sum(t1)
```

```
t1<-
data_frame(Nivel=names(t1),Frecuencia=as.numeric(t1),Porcentaje=round(as.numeric(t1)*
100/sum(t1),1))
t1
arrange(t1,desc(Frecuencia))

t2<-table(Datos_Lecturas$Genero_Literario)
t2
t2<-
data_frame(`Géneroliterario`=names(t2),Frecuencia=as.numeric(t2),Porcentaje=round(as.n
umeric(t2)*100/sum(t2),1))
t2
arrange(t2,desc(Frecuencia))

t3<-table(Datos_Lecturas$Autor_CostaRica)
t3
t3<-
data_frame(`Autorcostarricense`=names(t3),Frecuencia=as.numeric(t3),Porcentaje=round(a
s.numeric(t3)*100/sum(t3),1))
t3
arrange(t3,desc(Frecuencia))

t5<-table(Datos_Lecturas$Tecnica)
t5
t5<-
data_frame(`Modalidadtécnica`=names(t5),Frecuencia=as.numeric(t5),Porcentaje=round(as
.numeric(t5)*100/sum(t5),1))
t5
arrange(t5,desc(Frecuencia))

t4<-table(Datos_Lecturas$Lectura_Faro)
t4
t4<-
data_frame(`ObraenFARO`=names(t4),Frecuencia=as.numeric(t4),Porcentaje=round(as.nu
meric(t4)*100/sum(t4),1))
t4
arrange(t4,desc(Frecuencia))

```
Cargar de Corpusde Lecturas e itemsde Práctica PAA

```{r,warning=FALSE,message=FALSE}



#setwd(path)
```

```
archivosLec<-list.files(path="C:/Users/Meli/Documents/Trabajo Final Maestria/Lecturas recomendada/Lecturas", pattern='*.pdf')
archivosLec


setwd("~/Trabajo Final Maestria/Lecturas recomendada/Lecturas")
text=sapply(archivosLec,function(x) paste0(pdf_text(x),collapse=''))

archivosLectextDF<-data.frame(Document=archivosLec, text=text)


archivosLectextDF$Cod<-substring(archivosLectextDF$Document,1,4)

write.xlsx(archivosLectextDF,'datos.lecturasPDF.xlsx')

dfLec<-merge(archivosLectextDF,Datos_Lecturas,by.x="Cod",by.y="COD")


archivositems<-list.files(path="C:/Users/Meli/Documents/Trabajo Final Maestria/Items Practica PAA", pattern='*.pdf')
archivositems

setwd("~/Trabajo Final Maestria/Items Practica PAA")
text1=sapply(archivositems,function(x) paste0(pdf_text(x),collapse=''))
archivositemstextDF<-data.frame(Document=archivositems,text=text1)

dfPAA<-archivositemstextDF

```


##LimpezadecorpusdeLecturaseitemsdeprácticaPAA


```{r,warning=FALSE,message=FALSE}

corpusLec<-Corpus(VectorSource(dfLec$text))
corpusLec<-tm_map(corpusLec,tolower)
corpusLec<-tm_map(corpusLec,function(x)removeWords(x,stopwords("spanish")))
corpusLec<-tm_map(corpusLec,function(x)removePunctuation(x,
preserve_intra_word_contractions=TRUE,
preserve_intra_word_dashes=TRUE,
ucp=TRUE))

corpusLec<-tm_map(corpusLec,function(x)removePunctuation(x,
preserve_intra_word_contractions=TRUE,
```

```
preserve_intra_word_dashes=TRUE,
ucp=TRUE))
corpusLec<-tm_map(corpusLec,removeNumbers)
corpusLec<-tm_map(corpusLec,stripWhitespace)
```

```{r,warning=FALSE,message=FALSE}

corpusPAA<-Corpus(VectorSource(dfPAA$text))
corpusPAA<-tm_map(corpusPAA,tolower)
corpusPAA<-tm_map(corpusPAA,function(x)removeWords(x,stopwords("spanish")))
corpusPAA<-tm_map(corpusPAA,function(x)removePunctuation(x,
preserve_intra_word_contractions=TRUE,
preserve_intra_word_dashes=TRUE,
ucp=TRUE))

corpusPAA<-tm_map(corpusPAA,removeWords,c("a"," a ","a ","
a","b","c","d","g","i","ii","iii","m","n","n ","," n","n
","p","pd","pdx","pdz","pm","pq","q","r","s"," s" ," s ","s ","t","wundt","x","xix","z","n","pd
"," pd "," pd", "práctica", "aspirantes","s"))



corpusPAA<-tm_map(corpusPAA,removeNumbers)
corpusPAA<-tm_map(corpusPAA,stripWhitespace)

corpusPAA<-tm_map(corpusPAA,removeWords,c("a"," a ","a ","
a","b","c","d","g","i","ii","iii","m","n","n ","," n","n
","p","pd","pdx","pdz","pm","pq","q","r","s"," s" ," s ","s ","t","wundt","x","xix","z","n","pd
"," pd "," pd", "práctica", "aspirantes","s"))

corpusPAA<-tm_map(corpusPAA,function(x)removePunctuation(x,
preserve_intra_word_contractions=TRUE,
preserve_intra_word_dashes=TRUE,
ucp=TRUE))
```
Descriptivos corpus Lecturas

```{r,warning=FALSE,message=FALSE}

lista_corpusLec<-list()
lista_aux_corpusLec<-list()


for(j in 1:68){
```

```
lista_aux_corpusLec[j]<-corpusLec$content[[j]]
lista_corpusLec<-c(lista_corpusLec,lista_aux_corpusLec[j])
}
lista_corpusLec<-unlist(lista_corpusLec)

palabrasLec<-unlist(tokenize_words(lista_corpusLec))
length(palabrasLec)
tabla1<-table(palabrasLec)
tabla1
tabla1<-data_frame(Palabra=names(tabla1),Frecuencia=as.numeric(tabla1))
tabla1
arrange(tabla1,desc(Frecuencia))
```

```
```{r}
wordcloud(tabla1$Palabra,tabla1$Frecuencia,
#min.freq=500,
#max.words=1500,
random.order=FALSE,
colors=brewer.pal(name="Dark2",n=100)
)
```
```

```
```{r}
tabla1<-arrange(tabla1,desc(Frecuencia))
barplot(tabla1[1:20,]$Frecuencia, las = 2, names.arg = tabla1[1:20,]$Palabra,
        col ="lightblue",
        ylab = "Frecuencia")
```
```

```
##Descriptivos corpus PAA

```{r,warning=FALSE,message=FALSE}

lista_corpusPAA<-list()
lista_aux_corpusPAA<-list()

for(i in 1:2){
lista_aux_corpusPAA[i]<-corpusPAA$content[[i]]
lista_corpusPAA<-c(lista_corpusPAA,lista_aux_corpusPAA[i])
```
```

```
}
lista_corpusPAA<-unlist(lista_corpusPAA)

palabrasPAA<-unlist(tokenize_words(lista_corpusPAA))
length(palabrasPAA)
tabla2<-table(palabrasPAA)
tabla2
tabla2<-data_frame(Palabra=names(tabla2),Frecuencia=as.numeric(tabla2))
tabla2
arrange(tabla2,desc(Frecuencia))
```

```
```{r}
wordcloud(tabla2$Palabra,tabla2$Frecuencia,
#min.freq=2000,
#max.words=150,
random.order=FALSE,
colors=brewer.pal(name="Dark2",n=100)
)
```
```

```
```{r}
tabla2<-arrange(tabla2,desc(Frecuencia))
barplot(tabla2[1:20,]$Frecuencia, las = 2, names.arg = tabla2[1:20,]$Palabra,
        col ="lightblue",
        ylab = "Frecuencia")
```
```

Biagramas Lecturas

```
```{r}

texto_tidy <- tibble(line = 1:length(unlist(corpusLec)), text = unlist(corpusLec)) %>%
unnest_tokens(bigram, text, token = "ngrams", n = 2)

bigramas <- texto_tidy %>%
   count(bigram, sort = TRUE)

cat("\n\nBigramas más frecuentes:\n")
print(head(bigramas, 20))
```
```

```
bigrams_separated <-  bigramas %>%
  separate(bigram, c("word1", "word2"), sep = " ")


bigram_graph <- bigrams_separated %>%
  filter(n > 50) %>%
  graph_from_data_frame()

bigram_graph

set.seed(2020)


a <- grid::arrow(type = "closed", length = unit(.15, "inches"))

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()

bigramas

```


Biagramas Items Practica PAA

```{r}


texto_tidy <- tibble(line = 1:length(unlist(corpusPAA)), text = unlist(corpusPAA)) %>%
unnest_tokens(bigram, text, token = "ngrams", n = 2)

bigramas <- texto_tidy %>%
    count(bigram, sort = TRUE)

cat("\n\nBigramas más frecuentes:\n")
print(head(bigramas, 20))


bigrams_separated <-  bigramas %>%
  separate(bigram, c("word1", "word2"), sep = " ")
```

```
bigram_graph <- bigrams_separated %>%
  filter(n > 5) %>%
  graph_from_data_frame()

bigram_graph

set.seed(2020)



a <- grid::arrow(type = "closed", length = unit(.15, "inches"))

ggraph(bigram_graph, layout = "fr") +
  geom_edge_link(aes(edge_alpha = n), show.legend = FALSE,
                 arrow = a, end_cap = circle(.07, 'inches')) +
  geom_node_point(color = "lightblue", size = 5) +
  geom_node_text(aes(label = name), vjust = 1, hjust = 1) +
  theme_void()



```


Matrices de corpus Lecturas y escalamiento

```{r,warning=FALSE,message=FALSE}

td.matLec<-as.matrix(TermDocumentMatrix(corpusLec))
td.matLec
dist.matLec<-dist(t(as.matrix(td.matLec)))
dist.matLec


fitLec1<-cmdscale(dist.matLec,eig=TRUE,k=2)
points<-data.frame(x=fitLec0$points[,1],y=fitLec0$points[,2])
ggplot(points,aes(x=x,y=y))+geom_point(data=points,aes(x=x,y=y,
color=dfLec$Nivel))+geom_text(data=points,aes(x=x,y=y-0.2,label=
row.names(dfLec)))

fitLec2<-cmdscale(dist.matLec,eig=TRUE,k=2)
points<-data.frame(x=fitLec0$points[,1],y=fitLec0$points[,2])
ggplot(points,aes(x=x,y=y))+geom_point(data=points,aes(x=x,y=y,
```

```
color=dfLec$Genero_Literario))+geom_text(data=points,aes(x=x,y=y-0.2,label=
row.names(dfLec)))

fitLec3<-cmdscale(dist.matLec,eig=TRUE,k=2)
points<-data.frame(x=fitLec0$points[,1],y=fitLec0$points[,2])
ggplot(points,aes(x=x,y=y))+geom_point(data=points,aes(x=x,y=y,
color=dfLec$Autor_CostaRica))+geom_text(data=points,aes(x=x,y=y-0.2,label=
row.names(dfLec)))

fitLec4<-cmdscale(dist.matLec,eig=TRUE,k=2)
points<-data.frame(x=fitLec0$points[,1],y=fitLec0$points[,2])
ggplot(points,aes(x=x,y=y))+geom_point(data=points,aes(x=x,y=y,
color=dfLec$Lectura_Faro))+geom_text(data=points,aes(x=x,y=y-0.2,label=
row.names(dfLec)))

#MDS con LSA
td.matLec.lsa<-lw_bintf(td.matLec)*gw_idf(td.matLec)
lsaSpaceLec<-lsa(td.matLec.lsa)
dist.matLec.lsa<-dist(t(as.textmatrix(lsaSpaceLec)))
dist.matLec.lsa

#MDS
fitLec5<-cmdscale(dist.matLec.lsa,eig=TRUE,k=2)
points<-data.frame(x=fitLec1$points[,1],y=fitLec1$points[,2])
ggplot(points,aes(x=x,y=y))+geom_point(data=points,aes(x=x,y=y,
color=dfLec$Nivel))+geom_text(data=points,aes(x=x,y=y-0.2,label=row.names(dfLec)))

fitLec6<-cmdscale(dist.matLec.lsa,eig=TRUE,k=2)
points<-data.frame(x=fitLec1$points[,1],y=fitLec1$points[,2])
ggplot(points,aes(x=x,y=y))+geom_point(data=points,aes(x=x,y=y,
color=dfLec$Genero_Literario))+geom_text(data=points,aes(x=x,y=y-
0.2,label=row.names(dfLec)))

fitLec7<-cmdscale(dist.matLec.lsa,eig=TRUE,k=2)
points<-data.frame(x=fitLec1$points[,1],y=fitLec1$points[,2])
ggplot(points,aes(x=x,y=y))+geom_point(data=points,aes(x=x,y=y,
color=dfLec$Autor_CostaRica))+geom_text(data=points,aes(x=x,y=y-
0.2,label=row.names(dfLec)))

fitLec8<-cmdscale(dist.matLec.lsa,eig=TRUE,k=2)
points<-data.frame(x=fitLec1$points[,1],y=fitLec1$points[,2])
ggplot(points,aes(x=x,y=y))+geom_point(data=points,aes(x=x,y=y,
color=dfLec$Lectura_Faro))+geom_text(data=points,aes(x=x,y=y-
0.2,label=row.names(dfLec)))
```

```
fitLec9<-cmdscale(dist.matLec.lsa,eig=TRUE,k=3)
scatterplot3d(fitLec2$points[,1],fitLec2$points[,2],fitLec2$points[,3],color=viridis(68),
pch=16,main="Espacio Semántico Escalado 3D",xlab="x",ylab="y",
zlab="z",type="h")
```

```

## Matrices de corpus items práctica PAA

```{r}

td.matPAA<-as.matrix(TermDocumentMatrix(corpusPAA))
td.matPAA
dist.matPAA<-dist(t(as.matrix(td.matPAA)))
dist.matPAA

```


Distancias de similitud textual Lecturas

###JaccardyCoseno

```{r}

datos.distJac<-matrix(NA,nrow=68,ncol=68)
datos.distCos<-matrix(NA,nrow=68,ncol=68)

for(i in 1:68){
for(j in 1:68){
jacLec<-
seq_dist(hash(strsplit(corpusLec$content[i],"\\s+")),hash(strsplit(corpusLec$content[j],"\\s+
")),method="jaccard",q=2)
cosLec<-
seq_dist(hash(strsplit(corpusLec$content[i],"\\s+")),hash(strsplit(corpusLec$content[j],"\\s+
")),method="cosine",q=2)
datos.distJac[i,j]<-jacLec
datos.distCos[i,j]<-cosLec
}
}

```
```

Jaccard

```{r}
datos.distJac
```

Coseno

```{r}
datos.distCos
```

Distancias de similitud textual Lecturas y items de práctica PAA

Jaccard y Coseno

```{r}
datos.distJac2<-matrix(NA,nrow=68,ncol=2)
datos.distCos2<-matrix(NA,nrow=68,ncol=2)

for(i in 1:68){
for(j in 1:2){
jacLecPAA<-
seq_dist(hash(strsplit(corpusLec$content[i],"\\s+")),hash(strsplit(corpusPAA$content[j],"\\s
+")),method="jaccard",q=2)
cosLecPAA<-
seq_dist(hash(strsplit(corpusLec$content[i],"\\s+")),hash(strsplit(corpusPAA$content[j],"\\s
+")),method="cosine",q=2)
datos.distJac2[i,j]<-jacLecPAA
datos.distCos2[i,j]<-cosLecPAA
}
}
```

```{r}
datos.distJac3<-matrix(NA,nrow=68,ncol=1)
datos.distCos3<-matrix(NA,nrow=68,ncol=1)

for(i in 1:68){
```

```
for(j in 1:1){
jacLecPAA3<-
seq_dist(hash(strsplit(corpusLec$content[i],"\\s+")),hash(unlist(tokenize_words(lista_corpu
sPAA))),method="jaccard",q=2)
cosLecPAA3<-
seq_dist(hash(strsplit(corpusLec$content[i],"\\s+")),hash(unlist(tokenize_words(lista_corpu
sPAA))),method="cosine",q=2)
datos.distJac3[i,j]<-jacLecPAA3
datos.distCos3[i,j]<-cosLecPAA3
}
}
```

Jaccard

```{r}
datos.distJac2
```

Coseno

```{r}
datos.distCos2
```

Jaccard todo el corpus

```{r}
jacLecTodo<-
seq_dist(hash(unlist(tokenize_words(lista_corpusLec))),hash(unlist(tokenize_words(lista_c
orpusPAA))),method="jaccard",q=2)
jacLecTodo
```

Coseno todo el corpus

```{r}
cosLecTodo<-
seq_dist(hash(unlist(tokenize_words(lista_corpusLec))),hash(unlist(tokenize_words(lista_c
orpusPAA))),method="cosine",q=2)
cosLecTodo
```

```

##Distancias de similitud textual Lecturas


##Modelación

```{r,include=FALSE}

course_dtm<-DocumentTermMatrix(corpusLec)
course_dtm

course_dtm1<-DocumentTermMatrix(corpusPAA)
course_dtm1


dense_course_dtm_p<-removeSparseTerms(course_dtm,.95)
dense_course_dtm1_p<-removeSparseTerms(course_dtm1,.95)


df_terminosLecp <- as.data.frame(as.matrix(course_dtm))
df_terminosPAAp <- as.data.frame(as.matrix(course_dtm1))


df_terminosLecp <- as.data.frame(as.matrix(dense_course_dtm_p))
df_terminosPAAp <- as.data.frame(as.matrix(dense_course_dtm1_p))


sumar_columnas <- function(dataframe) {
  if (!is.data.frame(dataframe)) {
  }
  sumas <- colSums(dataframe[sapply(dataframe, is.numeric)])
  resultado <- data.frame(
    Palabra = names(sumas),
    Frecuencia = as.numeric(sumas)
  )

  return(resultado)
}


resultado_finalLecp <- sumar_columnas(df_terminosLecp)

resultado_finalPAAp <- sumar_columnas(df_terminosPAAp)
```
```

```
resultado_finalp <- union(resultado_finalLecp$Palabra, resultado_finalPAAp$Palabra)

resultado_finalLecp$Palabra<-as.character(resultado_finalLecp$Palabra)

resultado_finalPAAp$Palabra<-as.character(resultado_finalPAAp$Palabra)



datos_modelop <- data.frame(
Palabra = resultado_finalp,
frecuencia1 = sapply(resultado_finalp, function(p){ifelse(p %in% resultado_finalLecp$Palabra,resultado_finalLecp$Frecuencia[resultado_finalLecp$Palabra == p],0)}),
frecuencia2 = sapply(resultado_finalp, function(p){ifelse(p %in% resultado_finalPAAp$Palabra,resultado_finalPAAp$Frecuencia[resultado_finalPAAp$Palabra == p],0)})
)

  # Añadir etiqueta de clasificación
datos_modelop$categoria <- ifelse(datos_modelop$frecuencia2 > 0,"Si","No")
datos_modelop <- datos_modelop[-c(18430,18431),]


set.seed(123)
indices <- createDataPartition(datos_modelop$categoria ,p = 0.7, list = FALSE)

datos_entrenamientop <- datos_modelop[indices, ]
datos_pruebap <- datos_modelop[-indices, ]



```


Modelo Naive Bayes

```{r,include=FALSE}


x_trainp<- as.data.frame(datos_entrenamientop[, c("frecuencia1")])
names(x_trainp)[1] <-"frecuencia1"
y_trainp <- datos_entrenamientop$categoria

mod_nbp<-train(x_trainp,y_trainp,method="nb")
mod_nbp
```

```
x_testp <- as.data.frame(datos_pruebap[, c("frecuencia1")])
names(x_testp)[1] <-"frecuencia1"
y_testp <- datos_pruebap$categoria



mod_nb_predictionsp<-predict(mod_nbp,x_testp)

confusionMatrix(table(mod_nb_predictionsp,as.factor(y_testp)))

```


Modelo Máquinas de soporte vectorial (SVM)con kernel polinomial


```{r,include=FALSE}


mod_svmPp<-train(x_trainp,y_trainp,method="svmPoly")
mod_svmPp

mod_svmP_predictionsp<-predict(mod_svmPp,x_testp)

confusionMatrix(table(mod_svmP_predictionsp,as.factor(y_testp)))




```


Modelo Bosques aleatorio

```{r,include=FALSE}


mod_rfp<-train(x_trainp,y_trainp,method="rf")
mod_rfp

mod_rf_predictionsp<-predict(mod_rfp,x_testp)

confusionMatrix(table(mod_rf_predictionsp,as.factor(y_testp)))

```

Modelo Máxima entropía

```{r,include=FALSE}

objControl <- trainControl(method = "boot",
                           number = 2,
                           returnResamp = 'none',
                           summaryFunction = twoClassSummary,
                           classProbs = TRUE,
                           savePredictions = TRUE)


mod_mxep<-train(x_trainp,y_trainp,method="glm",trControl = objControl, metric = "ROC")
mod_mxep

mod_mxe_predictionsp<-predict(mod_mxep,x_testp)

confusionMatrix(table(mod_mxe_predictionsp,as.factor(y_testp)))

```

Construcción LDA

```{r}
# Función para crear n-gramas de caracteres para cada palabra
create_char_ngrams <- function(words, min_n = 2, max_n = 4) {
  ngram_list <- list()

  for (i in 1:length(words)) {
    word <- words[i]
    word_ngrams <- c()

    for (n in min_n:max_n) {
      if (nchar(word) >= n) {
        for (j in 1:(nchar(word) - n + 1)) {
          ngram <- substr(word, j, j + n - 1)
          word_ngrams <- c(word_ngrams, ngram)
        }
      }
    }

    ngram_list[[i]] <- word_ngrams
  }
```

```
  # Obtener todos los n-gramas únicos
  all_ngrams <- unique(unlist(ngram_list))

  # Crear matriz de documentos-términos
  dtm <- matrix(0, nrow = length(words), ncol = length(all_ngrams))
  colnames(dtm) <- all_ngrams

  for (i in 1:length(words)) {
    word_ngrams <- ngram_list[[i]]
    for (ngram in word_ngrams) {
      if (ngram %in% all_ngrams) {
        col_idx <- which(all_ngrams == ngram)
        dtm[i, col_idx] <- dtm[i, col_idx] + 1
      }
    }
  }

  return(list(dtm = dtm, ngrams = all_ngrams))
}

# Función para aplicar LDA a las palabras
n_topics = 3
  # Crear n-gramas de caracteres para cada palabra
  ngram_result <- create_char_ngrams(datos_modelop$Palabra)
  dtm <- ngram_result$dtm

  # Asegurarse de que hay suficientes datos para el número de temas
  n_topics <- min(n_topics, ncol(dtm) - 1, nrow(dtm) - 1)
  if (n_topics < 2) n_topics <- 2

  # Aplicar LDA
  lda_model <- LDA(dtm, k = n_topics, control = list(seed = 42))

  # Extraer las distribuciones de temas
  topic_distributions <- posterior(lda_model)$topics

  # Añadir características LDA al dataframe
  topic_cols <- paste0("topic_", 1:n_topics)
  topic_df <- as.data.frame(topic_distributions)
  colnames(topic_df) <- topic_cols

  # Combinar con el dataframe original
  final_df <- cbind(datos_modelop, topic_df)
```

```
 # Aplicar LDA a las palabras
 df_lda <- lda_result$df

 # Preparar datos para el modelo
 X <- df_lda[, grep("topic_", colnames(df_lda))]
 y <- df_lda$categoria

 # Dividir datos en entrenamiento y prueba (70/30)
 set.seed(123)
 train_indices <- createDataPartition(y, p = 0.7, list = FALSE)
 X_train <- X[train_indices, ]
 X_test <- X[-train_indices, ]
 y_train <- y[train_indices]
 y_test <- y[-train_indices]

 # Escalar características
 preproc <- preProcess(X_train, method = c("center", "scale"))
 X_train_scaled <- predict(preproc, X_train)
 X_test_scaled <- predict(preproc, X_test)
```

```

###Modelo Naive Bayes considernado LDA

```{r,include=FALSE}

# Entrenar SVM
nb_model_lda <-train(X_train_scaled, y_trainplda,method="nb")

# Evaluar modelo
y_predplda_nb <- predict(nb_model_lda, X_test_scaled)

# Calcular métricas
confusionMatrix(table(y_predplda_nb,as.factor(y_testplda)))


```

###Modelo Máquinas de soporte vectorial (SVM) con kernel polinomial y LDA

```{r,include=FALSE}

# Entrenar SVM
```
```

```
svm_model_lda <- train(X_train_scaled, y_trainplda,method="svmPoly")

# Evaluar modelo
y_predplda_svm <- predict(svm_model_lda, X_test_scaled)


# Calcular métricas
confusionMatrix(table(y_predplda_svm,as.factor(y_testplda)))
```

```
###Modelo Bosques aleatorio considerando LDA

```{r,include=FALSE}

  # Entrenar SVM
rf_model_lda <- train(X_train_scaled, y_trainplda,method="rf")


  # Evaluar modelo
y_predplda_rf <- predict(rf_model_lda, X_test_scaled)


  # Calcular métricas
confusionMatrix(table(y_predplda_rf,as.factor(y_testplda)))

```

```
###Modelo Máxima entropía considerando LDA

```{r,include=FALSE}

objControl2 <- trainControl(method = "boot",
                        number = 2,
                        returnResamp = 'none',
                        summaryFunction = twoClassSummary,
                        classProbs = TRUE,
                        savePredictions = TRUE)



  # Entrenar SVM
mod_mxep_model_lda <-train(X_train_scaled, y_trainp, method="glm",trControl = objControl2, metric = "ROC")

  # Evaluar modelo
```

```
y_predplda_mod_mxep<-predict(mod_mxep_model_lda , X_test_scaled)

# Calcular métricas

confusionMatrix(table(y_predplda_mod_mxep,as.factor(y_testp)))
```

```

---
```