

MODELO DE CLASIFICACIÓN AUTOMÁTICA DE TEXTO EN IDIOMA INDÍGENA WAYUUNAIKI QUE INCORPORA CARACTERÍSTICAS GRAMATICALES

Autor

Ing. Diana Milena Salazar Báez

Universidad Distrital Francisco José De Caldas

Maestría en Ciencias de la Información y las Comunicaciones

Énfasis en Ingeniería de Software

Bogotá, Colombia

2025



INVESTIGACIÓN __X__ PROFUNDIZACIÓN ____

MODELO DE CLASIFICACIÓN AUTOMÁTICA DE TEXTO EN IDIOMA INDÍGENA WAYUUNAIKI QUE INCORPORA CARACTERÍSTICAS GRAMATICALES

Tesis o trabajo de investigación presentada(o) como requisito parcial para optar al título de:
Magister en Ciencias de la Información y las Comunicaciones

Autor

Ing. Diana Milena Salazar Báez

Tutor

Jorge Enrique Rodríguez Rodríguez

MSc. Ingeniería de Sistemas

Esp. Ingeniería del Software

Universidad Distrital Francisco José De Caldas

Maestría en Ciencias de la Información y las Comunicaciones

Énfasis en Ingeniería de Software

Bogotá, Colombia

Junio 2025

Dedicatoria

A todas aquellas personas que persisten en mejorar el mundo con pequeñas acciones.

A todas aquellas personas que investigan para preservar la riqueza cultural de las comunidades indígenas.

A todas aquellas personas que se esfuerzan por alcanzar sus metas académicas y profesionales, superando los contratiempos y dificultades con perseverancia.

A mi familia por su amor y apoyo incondicional.

AGRADECIMIENTOS

Agradezco principalmente a Dios, fuente de amor y sabiduría, por permitirme culminar este trabajo a pesar de las dificultades.

A san José, por ser modelo de humildad, fortaleza, trabajo y confianza absoluta en Dios.

Agradezco profundamente a mi familia, por su apoyo emocional y económico durante este proceso, por ser la luz que ilumina el camino en momentos de incertidumbre.

A mi madre, por su amor incondicional, cuidados y apoyo en las noches más largas y oscuras.

Expreso mi gratitud a mi tutor de trabajo de grado, por su orientación, compromiso y contribución académica, pero especialmente por su paciencia y calidad humana.

A todos los profesores de la maestría y del Grupo de Investigación en Inteligencia Artificial IAFT, quienes con sus enseñanzas aportaron significativamente a mi crecimiento académico y profesional.

A los miembros de la familia Epiayú de la comunidad Wayuu, quienes acompañaron este trabajo de grado y me brindaron una perspectiva única sobre su riqueza cultural.

A la Universidad por brindarme los recursos académicos para el desarrollo de la investigación.

A todas aquellas personas que con sus palabras de aliento, me inspiraron a continuar y culminar este proceso.

Finalmente, quiero expresar un gran agradecimiento a todos aquellos investigadores del idioma Wayuunaiki, sin sus aportes este trabajo no hubiera sido posible.

CONTENIDO

RESUMEN 13

PALABRAS CLAVE 14

INTRODUCCIÓN 15

1. PROBLEMA DE INVESTIGACIÓN 16

1.1 PLANTEAMIENTO DEL PROBLEMA 16

1.2 FORMULACION DEL PROBLEMA 18

1.3 SISTEMATIZACIÓN DEL PROBLEMA 18

2. OBJETIVOS 19

2.1 OBJETIVO GENERAL 19

2.2 OBJETIVOS ESPECÍFICOS 19

3. JUSTIFICACIÓN 20

3.1 JUSTIFICACIÓN TEÓRICA 20

3.2 JUSTIFICACIÓN METODOLÓGICA 21

3.3 JUSTIFICACIÓN PRÁCTICA 21

4. MARCO DE REFERENCIA 23

4.1 MARCO TEÓRICO 23

4.1.1 LA COMUNIDAD INDIGENA WAYUÚ 23

4.1.2 EL IDIOMA WAYUUNAIKI 24

4.1.3 CONSERVACIÓN DEL IDIOMA WAYUUNAIKI. 25

4.1.4 MATERIALES DIGITALES EN WAYUUNAIKI Y ACCESO DE LOS WAYUÚ A LA INFORMACIÓN 26

4.1.5 PROCESAMIENTO DE LENGUAJE NATURAL 28

4.1.6 CLASIFICACIÓN AUTOMATICA DE TEXTO 30

4.1.7 REDES NEURONALES Y MACHINE LEARNIG PARA LA CLASIFICACIÓN DE TEXTO 31
4.1.8 TRABAJOS SIMILARES 36
4.2 MARCO CONCEPTUAL 38
4.2.1 IDIOMA WAYUUNAIKI 38
4.2.2 CARACTERISTICAS GRAMATICALES DEL WAYUUNAIKI 39
4.2.3 REDES NEURONALES ARTIFICIALES 44
4.2.4 MINERÍA DE DATOS 46
4.2.5 MACHINE LEARNING 47
4.2.6 DEEP LEARNING 48
4.2.7 REDES NEURONALES CONVOLUCIONALES (CNN). 50
4.2.8 DOC2VEC 50
4.2.9 RED NEURONALE RECURRENTE 51
4.2.10 LSTM 52
4.2.11 REDES BIDIRECCIONALES 53
4.2.12 MODELOS DE TRANSFORMERS 54
4.3 MARCO ESPACIAL 56
4.4 MARCO TEMPORAL 57
4.5 ALCANCE TEMÁTICO. 57
5. HIPÓTESIS 59
6. METODOLOGÍA 60
6.1 FASES DE LA METODOLOGÍA. 61
6.1.1 FASE 1. COMPRENSIÓN DEL NEGOCIO. 61
6.1.2 FASE 2. COMPRENSIÓN DE LOS DATOS. 61
6.1.3 FASE 3. PREPARACIÓN DE LOS DATOS. 62
6.1.4 FASE 4. MODELADO 63

6.1.4.1 DETERMINACIÓN DEL MODELO: 63
6.1.4.2 PROYECCIÓN DEL PLAN DE PRUEBAS: 63
6.1.4.3 DISEÑO DEL MODELO. 63
6.1.4.4 VALIDACIÓN DEL MODELO. 64
6.1.5 FASE 5. ANÁLISIS DE LOS RESULTADOS. 65
6.1.6 FASE 6. IMPLANTACIÓN. 65
7. DESARROLLO 66
7.1 FASE 1. COMPRENSIÓN DEL NEGOCIO. 66
7.1.1 OBJETIVOS DEL NEGOCIO. 66
7.1.2 VALORACIÓN DE LA SITUACIÓN. 67
7.1.3 DETERMINAR LOS OBJETIVOS DEL PLN. 67
7.1.4 REALIZAR EL PLAN DEL PROYECTO. 68
7.2 FASE 2. COMPRENSIÓN DE LOS DATOS. 69
7.2.1 PREPARACIÓN DE LOS DATOS. 69
7.2.2 DEFINICIÓN DEL TIPO DE TAREAS. 69
7.2.3 DEFINICIÓN DE LOS ATRIBUTOS DEL CONJUNTO DE DATOS. 69
7.2.4 REPORTES DE LOS DATOS. 70
7.2.5 ANÁLISIS ESTADÍSTICO 70
7.2.5.1 DIVISIÓN DEL CONJUNTO DE DATOS PARA EL ANÁLISIS EXPLORATORIO: 70
7.2.5.2 VALIDACIÓN Y VERIFICACIÓN DEL CONJUNTO DE DATOS PARA EL ANÁLISIS EXPLORATORIO: 70
7.2.5.3 RESULTADOS DEL ANÁLISIS EXPLORATORIO DE DATOS EN IDIOMA WAYUUNAIKI 71
7.2.6 REPORTE DE CALIDAD DE RECOLECCIÓN DE DATOS. 76
7.3 FASE 3. PREPARACIÓN DE LOS DATOS. 78
7.3.1 SELECCIONAR LOS DATOS. 78

7.3.2 PREPROCESAMIENTO DE DATOS. 79

LIMPIEZA DE DATOS: 79

7.3.3 REPRESENTACIÓN DEL TEXTO. 81

7.3.4 ENRIQUECIMIENTO DEL CONJUNTO DE DATOS: 83

7.3.5 DEFINICIÓN DE VARIABLES: 83

7.3.1 DEFINICIÓN DE CLASES. 84

7.3.1 CONJUNTO DE DATOS FINAL 85

7.4 FASE 4. MODELADO 89

7.4.1 DETERMINACIÓN DEL MODELO 89

7.4.1.1 MODELO CONCEPTUAL 89

7.4.1.2 COMPARACIÓN ARQUITECTURAS 90

7.4.1.2.1 EXPERIMENTOS INICIALES CON MACHINE LEARNING. 90

7.4.1.2.1.1 RANDOM FOREST. 90

7.4.1.2.1.2 NAIVE BAYES. 92

7.4.1.2.1.3 SUPPORT VECTOR MACHINES (SVMS) 92

7.4.1.2.1.4 K-NEAREST NEIGHBOR 93

7.4.1.2.1.5 KERNEL APROXIMATION 94

7.4.1.2.2 ANÁLISIS DE LOS RESULTADOS PRÁCTICOS DE EXPERIMENTOS CON TRES ARQUITECTURAS PARA CLASIFICACIÓN DE TEXTO EN WAYUUNAIKI. 95

7.4.1.2.2.1 RANDOM FOREST 95

7.4.1.2.2.2 CONVOLUTIONAL NEURAL NETWORKS - CNN 99

7.4.1.2.2.3 RED NEURONAL LSTM 102

7.4.1.2.2.4 RED TRANSFORMERS UTILIZANDO BERT MODIFICADO. 104

7.4.2 PROYECCIÓN DEL PLAN DE PRUEBAS. 106

7.4.3 DISEÑO DEL MODELO 106

7.4.4 DEFINICIÓN DE LA ARQUITECTURA 108
7.4.5 RECOLECCIÓN DE LOS DATOS 108
7.4.6 PREPROCESAMIENTO DE LOS DATOS 108
7.4.7 REPRESENTACIÓN DE LOS DATOS. 109
7.4.8 CONSTRUCCIÓN DEL MODELO LSTM 109
7.4.9 ENTRENAMIENTO DEL MODELO: 111
7.4.10 DEFINICIÓN DE LA FUNCIÓN DE PÉRDIDA 111
7.4.11 SELECCIÓN DE MÉTRICAS DE EVALUACIÓN 112
7.4.12 ENTORNO DE ENTRENAMIENTO. 112
8. VALIDACIÓN DEL MODELO 114
8.1 VALIDACIÓN DEL MODELO. 114
9. ANALISIS DE RESULTADOS 120
10. IMPLANTACIÓN DEL MODELO. 123
11. CONCLUSIONES. 125
12. PRINCIPALES APORTES 127
13. RECOMENDACIONES Y FUTUROS TRABAJOS. 129
13.1 RECOMENDACIONES 129
13.2 FUTUROS TRABAJOS 130
14. REFERENCIAS 131

LISTA DE ILUSTRACIONES

Figura 1. Información sociodemográfica del pueblo Wayúu. 23
Figura 2. Información del sistema de información Wayúu. 24
Figura 3. Fotografía sobre la entrega de tabletas en La Guajira. 27
Figura 4. Organización de la arquitectura de un modelo de Procesamiento del Lenguaje Natural (PLN). 28
Figura 5. Estructura gramatical del *Wayuunaiki.* 39
Figura 6. Capas básicas de una Red Neuronal Artificial 44
Figura 7. Etapas de la minería de datos. 46
Figura 8. Flujo de datos en una red LSTM. 53
Figura 9. Preentrenamiento y afinamiento en el modelo BERT 55
Figura 10. Dialectos de wayuunaiki. 56
Figura 11. Población que habla la lengua nativa de su pueblo. 57
Figura 12. Fases de la metodología CRISP-DM. 60
Figura 13. Plan de actividades. 68
Figura 14. Distribución de la longitud de las frases en Wayuunaiki. 72
Figura 15. Datos atípicos en las frases en Wayuunaiki. 73
Figura 16. Nube de palabras de frases en Wayuunaiki resultado del experimento con el conjunto de datos A. 74
Figura 17. Nube de palabras de las frases en Wayuunaiki prueba conjunto de datos B. 75
Figura 18, portada del principito traducido al wayuunaiki. 76
Figura 19. Base de datos ayonnajaabd. 78
Figura 20. Recursos Azure para el proyecto. 78
Figura 21. Estructura de la Base de datos en Azure para el proyecto. 79
Figura 22. Funcionamiento de la lematización. 81
Figura 23. Funcionamiento de Word2Vec. 82
Figura 24. Matriz de similitud de coseno de datos en Wayuunaiki. 83
Figura 25. Distribución de clases del modelo. 85
Figura 26. Distribución clases del modelo LSTM. 87
Figura 27. Modelo conceptual de clasificación automática de texto en idioma Wayuunaiki que incorpora características gramaticales. 90
Figura 28. Resultados obtenidos clasificador de texto wayuunaiki con Random Forest. . 91

Figura 29. Resultados obtenidos clasificador de texto wayuunaiki con Naive Bayes. 92
Figura 30. Resultados obtenidos clasificador de texto wayuunaiki con Support Vector Machines (SVMs). 93
Figura 31. Resultados obtenidos clasificador de texto wayuunaiki con K-Nearest Neighbor. 94
Figura 32. Resultados obtenidos clasificador de texto wayuunaiki con kernel Aproximation. 94
Figura 33. Resultados obtenidos clasificador de texto wayuunaiki con Random Forest. . 96
Figura 34. Métricas promedio y Hyperparametros para el clasificador multiclase. 96
Figura 35. Curva ROC para las métricas promedio del clasificador. 97
Figura 36. Predicciones correctas del modelo..... 97
Figura 37. Predicciones incorrectas del modelo..... 98
Figura 38. Entrenamiento red CNN con tasa de aprendizaje de 0.001..... 99
Figura 39. Validación red CNN con tasa de aprendizaje de 0.001. 100
Figura 40. Resultados red CNN con tasa de aprendizaje de 0.001. 100
Figura 41. Arquitecturas generadas para la red CNN para la clasificación..... 101
Figura 42. Resultados red LSTM con tasa de aprendizaje de 0.001 102
Figura 43. Entrenamiento red LSTM con tasa de aprendizaje de 0.001. 103
Figura 44. Entrenamiento red Transformers BERT con tasa de aprendizaje de 0.001. . 104
Figura 45. Resultados red Transformers BERT con tasa de aprendizaje de 0.001 104
Figura 46. Estructura del modelo clasificación automática de texto en idioma Wayuunaiki que incorpora características gramaticales. 107
Figura 47. Ejemplo de tokenización del texto en wayuunaiki..... 109
Figura 48. Histograma con longitud de palabras en las frases. 110
Figura 49. Entorno de entrenamiento del modelo para clasificación de texto en wayuunaiki. 112
Figura 50. Entrenamiento y validación red LSTM con tasa de aprendizaje de 0.0001... 115
Figura 51. Entrenamiento y validación red LSTM con tasa de aprendizaje de 0.001..... 116
Figura 52. Entrenamiento y validación red LSTM con tasa de aprendizaje de 0.01 116
Figura 53. Entrenamiento y validación red LSTM con tasa de aprendizaje de 0.001..... 117
Figura 54. Matriz de confusión para clasificación de texto en wayuunaiki. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. 118

Figura 55. Distribución clases del modelo LSTM. 120
Figura 56. Nube de palabras obtenida con el modelo LSTM. 121
Figura 57. Empaquetado del modelo en formato.mat. 123
Figura 58. Resultados prueba del modelo LSTM con tasa de aprendizaje de 0.0001. .. 124
Figura 59. Análisis de la red LSTM, generado por MATLAB. 124

LISTA DE TABLAS

Tabla 1. Fonemas wayuunaiki. 39
Tabla 2. Sufijos Determinativos wayuunaiki. (Captain y Captain, 2008). 40
Tabla 3. Los pronombres y prefijos personales. 40
Tabla 4. Atributos del conjunto de datos. 69
Tabla 5. Variables del modelo. 83
Tabla 6. Clases del modelo. 84
Tabla 7. Configuración de clases. 86
Tabla 8. Cantidad de datos por clases del modelo final. 87
Tabla 9. Comparación resultados experimentos iniciales modelos de clasificación. 95
Tabla 10. Resultados experimentos transformers 105
Tabla 11. Comparación arquitecturas de clasificación. 105
Tabla 12. Comparación de resultados de validación. 115
Tabla 13. Datos mejor resultado red LSTM. 117

RESUMEN

Las técnicas de procesamiento de lenguaje natural (PLN) aplicadas a la clasificación automática de texto, funcionan de manera óptima en la realización de tareas de ordenamiento, etiquetado y agrupación de textos escritos en idiomas muy utilizados como el inglés, chino y español, entre otros. Funcionamiento que se ha logrado gracias a los importantes avances en las arquitecturas de aprendizaje automático y aprendizaje profundo, las estrategias de representación semántica para preentrenamientos y a la existencia y acceso a grandes cantidades de datos. En el caso del procesamiento de lenguaje natural de idiomas de comunidades indígenas, existen pocos estudios que describan el procesamiento de un lenguaje indígena que tenga en cuenta las características gramaticales y la identidad cultural de la comunidad indígena. Esto, debido a las dificultades relacionadas con la existencia de pocos conjuntos de datos que contengan un adecuado numeró de registros y cuenten con un alto grado de calidad de los datos, igualmente no existen recursos lingüísticos como diccionarios, lematizadores o etiquetadores que fuera posibles utilizar para análisis gramaticales utilizando adaptaciones de otras soluciones de PLN.

Frente a este panorama, el presente trabajo describe la propuesta de un modelo de clasificación automática de texto en idioma indígena wayuunaiki, lengua autóctona de la comunidad indígena Wayuú que habita en Colombia y Venezuela. Este modelo se desarrolla utilizando técnicas de procesamiento de lenguaje natural (PLN) y la metodología CRISP-DM ((Cross Industry Standard Process for Data Mining), el modelo fundamentalmente integra las características gramaticales propias del wayuunaiki como las preposiciones, conjugaciones verbales marcadas por persona y género, y la morfología aglutinante del idioma, esto con la finalidad de realizar una clasificación más acertada que aporte a la ejecución de otras tareas de PLN. Este trabajo busca además de contribuir a procesos computacionales, aportar a investigaciones que fomenten la conservación y enseñanza del wayuunaiki que necesiten de un conjunto de datos de texto en wayuunaiki etiquetado de buena calidad.

PALABRAS CLAVE

Clasificación de textos, aprendizaje profundo, Idioma Indígena, características lingüísticas, procesamiento de lenguaje natural.

Key words: Text classification, Deep Learning, Indigenous language, Linguistic characteristics, Natural language processing.

INTRODUCCIÓN

La comunidad indígena wayuu afronta diversas problemáticas, entre ellas la desaparición progresiva de su idioma autóctono denominado wayuunaiki, idioma perteneciente a la familia lingüística Arawak con características gramaticales particulares y aunque existen variadas estrategias de conservación de este idioma, se presentan diferentes problemáticas para su adecuada conservación, entre estas problemáticas se encuentran la generación de textos en wayuunaiki y la traducción de textos de español a wayuunaiki y wayuunaiki a español que al ser revisados por traductores certificados presentan inconsistencias gramaticales, lo cual impide a los hablantes nativos del idioma entender correctamente la información de los textos y también afectando de manera significativa las estrategias de etnoeducación que busca que las nuevas generaciones de la comunidad wayuu se apropien de su lengua materna.

El uso de inteligencia artificial para tareas de procesamiento de lenguaje natural (PLN), muestra importantes logros en el campo de clasificación de textos en diferentes idiomas de uso común como el inglés, chino, francés, entre otros, pero se ha implementado muy poco para la clasificación de texto con características gramaticales tan particulares como las del idioma wayuunaiki, por tanto, la creación de un modelo que clasifique de manera correcta las diferentes características gramaticales del wayuunaiki aporta no solo a otras tareas de PLN de idiomas indígenas, sino también a la verificación de los textos escritos en wayuunaiki para su adecuada conservación y difusión.

Este trabajo de investigación propone un modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales particulares del idioma, mediante la adecuada recopilación de los datos, conceptualización y estructuración del modelo y la respectiva validación del funcionamiento del modelo. Por tanto, este documento de investigación se estructura en tres partes. La primera parte describe el planteamiento del problema, los objetivos y la justificación; en segundo lugar, encontramos el marco de referencia que contiene el estado del arte, el marco teórico, marco conceptual, marco espacial y marco temporal, desarrollo, conclusiones, principales aportes, recomendaciones y trabajo futuro.

1. PROBLEMA DE INVESTIGACIÓN

1.1 PLANTEAMIENTO DEL PROBLEMA

El wayuunaiki es la lengua materna de la comunidad indígena Wayuú, la cual se ubica en la parte norte de Colombia y Venezuela, es el idioma indígena con mayor cantidad de hablantes, quienes conservan sus tradiciones y cultura a través de la palabra, pero debido a los conflictos violentos de estos países y a los problemas de la etnoeducación, este idioma se encuentra en riesgo de extinción de acuerdo con lo informado por la ONU. (El Wayuunaiki, una lengua con una importante cantidad de hablantes en riesgo de extinción | Noticias ONU, 2019)

Las políticas del gobierno colombiano establecidas para la conservación y enseñanza del idioma wayuunaiki, han generado diferentes estrategias para la difusión y conservación del wayuunaiki (Wayús pueden leer en wayuunaiki sentencia que pide frenar desnutrición en la Guajira - Otras Ciudades - Colombia - ELTIEMPO.COM, 2023), pero las acciones propuestas por el gobierno y los investigadores de esta problemática no han evitado de manera significativa que cada día los pueblos indígenas como el wayuu y en especial las nuevas generaciones de esta comunidad utilicen menos su lengua materna, lo cual afecta su identidad cultural y adecuada conservación del idioma; Esta situación además se ha originado debido a que la mayoría de información a la que pueden acceder las comunidades wayuu, se encuentra escrita en medio físico o digital en español u otro idioma y es muy escaso el material escrito en wayuunaiki de alta calidad que incluya una correcta gramática e identidad cultural del pueblo wayuu, lo anterior debido al contexto cultural y a las características gramaticales propias del idioma (Pérez van Leenden, 2020).

Igualmente entre las acciones más destacables para conservar y enseñar el wayuunaiki se destacan varios trabajos de etnoeducación para la enseñanza del wayuunaiki (Pérez van Leenden, 2020) y proyectos de investigación con los cuales se busca fomentar la participación de las comunidades Wayuu en la sociedad del conocimiento (Suarez Suarez & López Uriana, 2022), permitiéndoles acceder, entender y difundir la información disponible digitalmente en línea y brindándoles dicha información en su propio idioma, todo

esto utilizando las tecnologías de la información (Maestre Salas, 2022), pero muchos de estos desarrollos tecnológicos presentan inconsistencias gramaticales y de contexto en las tareas de procesamiento de lenguaje natural PLN, como por ejemplo los traductores, los cuales al tener un procesamiento de palabras sin una adecuada clasificación generan resultados inesperados y con errores gramaticales y también altos costos de clasificación y bajo rendimiento de los modelos (Fernández et al., 2013).

Actualmente el uso de machine learning (aprendizaje automático) y deep learning (aprendizaje profundo) para el adecuado procesamiento de lenguaje Natural, ha demostrado que puede realizar de manera eficiente las tareas de PLN en idiomas como el inglés, chino, francés y español de acuerdo a su contexto y estructura gramatical (Abarna et al., 2022), pero en relación al tratamiento de datos de texto escrito en idioma indígena como el wayuunaiki que tiene una estructura gramatical con características aglutinantes y de yuxtaposición, aún se generan inconsistencias y bajo rendimiento en su procesamiento con técnicas de aprendizaje automático (Graichen et al., 2023).

La inadecuada clasificación de texto en wayuunaiki que no incluya las características propias del idioma indígena, ocasionaría problemas de inconsistencias gramaticales y bajo rendimiento en la realización de otras tareas de procesamiento de lenguaje natural, generando como resultado textos erróneos, los cuales se utilizarían para la conservación o difusión del idioma, produciendo corpus de baja calidad y sin contexto cultural, lo que también ocasionaría que los hablantes nativos no confíen en los textos resultantes y se desmotiven en el uso de herramientas tecnológicas para conservar, enseñar o consultar información en idioma wayuunaiki.

Para realizar adecuadamente las tareas de procesamiento de lenguaje natural del idioma indígena wayuunaiki, es importante generar un modelo de clasificación de texto que identifique en las palabras y frases las características gramaticales propias del wayuunaiki, en específico si contiene características como la aglutinación, multiplicidad de infinitivos, flexionadas de nombres y preposiciones, gradación verbal, sufijos temáticos, morfemas, yuxtaposición, para lo cual se propone el uso de modelos de Deep Learning para la clasificación de textos, los cuales han demostrado generar excelentes resultados de rendimiento con idiomas con estructuras gramaticales diferentes como el árabe (Elnagar et al., 2020) y el japonés (Chen et al., 2023). Igualmente se debe contar con conjuntos de

datos revisados por hablantes nativos o traductores de la comunidad indígena wayuu certificados por la Universidad de la Guajira.

1.2 FORMULACION DEL PROBLEMA

¿Cómo identificar en textos escritos en idioma Wayuunaiki las características gramaticales, para poder realizar una adecuada clasificación de texto?

1.3 SISTEMATIZACIÓN DEL PROBLEMA

- ¿Cómo se lograría identificar y recopilar un conjunto de datos de buena calidad que represente las características gramaticales del wayuunaiki relacionadas con la aglutinación, multiplicidad de infinitivos, flexionadas de nombres y preposiciones, gradación verbal, sufijos temáticos, morfemas, yuxtaposición?
- ¿Cuál es la mejor estrategia para elaborar un modelo conceptual de clasificación de texto automática que incluya las características gramaticales propias del idioma indígena wayuunaiki?
- ¿Cómo modelar una adecuada estructura de funcionamiento para realizar una clasificación automática de texto en idioma wayuunaiki que incluya las características gramaticales del idioma, utilizando un set de datos de calidad para entrenar y validar el modelo utilizando algoritmos de Deep Learning?
- ¿Cómo se puede determinar el mejor proceso de validación del modelo propuesto de clasificación de texto en wayuunaiki, que describa en específico el rendimiento y la efectividad de identificación de características gramaticales del idioma?

2. OBJETIVOS

2.1 OBJETIVO GENERAL

Crear un modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpore las características gramaticales particulares del idioma utilizando Deep Learning.

2.2 OBJETIVOS ESPECÍFICOS

- Recopilar y generar el conjunto de datos que represente la estructura gramatical y la terminología del idioma wayuunaiki.
- Establecer un modelo conceptual de clasificación de texto que permita incluir las principales características gramaticales del wayuunaiki.
- Estructurar el modelo para la clasificación automática de texto que permita incluir las principales características gramaticales del wayuunaiki utilizando Deep Learning.
- Validar el desempeño del modelo de clasificación de texto propuesto para determinar la efectividad de identificación de las principales características gramaticales del idioma wayuunaiki.

3. JUSTIFICACIÓN

El procesamiento de lenguaje natural (PLN) incluye la tarea de clasificación automática de textos, la cual se ha utilizado ampliamente en idiomas con grandes cantidades de datos y con estructuras gramaticales comunes como el inglés, el francés o el chino, entre otros, y actualmente gracias al uso de Deep Learning para tareas de PLN se han logrado aumentar la eficiencia de la clasificación de textos en idiomas que presentan características particulares en su estructura gramatical como el árabe o el japonés, por tanto, un modelo de clasificación automática de texto escrito en idioma indígena wayuunaiki que incorpora características gramaticales aportaría a otras tareas de PLN, lo cual contribuiría a promover la participación de las comunidades wayuu en la sociedad del conocimiento facilitándoles el acceso a la información disponible en internet en su propio idioma, esto debido a que las tareas de PLN se realizarían con textos clasificados correctamente que incluyan características gramáticas propias de la cultura wayuu y por otra parte cooperaria con las iniciativas y acciones de conservación de un idioma que está en peligro de desaparecer.

3.1 JUSTIFICACIÓN TEÓRICA

En el Procesamiento de Lenguaje Natural (NPL), la tarea de clasificación de texto ha sido ampliamente estudiada para ordenar, etiquetar y agrupar texto en idiomas como el inglés, chino, francés, español, entre otros (Rodrawangpai & Daungjaiboon, 2022), de igual manera, se han realizado trabajos de investigación de clasificación de texto con idiomas con características gramaticales peculiares como el árabe (Elnagar et al., 2020) o el japones (Chen et al., 2023), pero en el caso de los idiomas indígenas son pocos los estudios realizados y para el caso de un idioma indígena como el wayuunaiki, que su estructura gramatical presenta diferentes particularidades como la aglutinación, multiplicidad de infinitivos, flexionadas de nombres y preposiciones, gradación verbal, sufijos temáticos, morfemas, entre otras (Álvarez, 2008), son muy escasos los estudios de clasificación de textos que se especialicen en detectar estas particularidades gramaticales del idioma, por tanto una investigación que se centre en clasificar de manera correcta esta particularidades de la estructura gramatical del wayuunaiki aportaría a las teorías de procesamiento de

lenguaje natural en específico a la formulación de un adecuado modelo para la tarea de clasificación de textos en idiomas indígenas aglutinantes. Por tanto, esta investigación tiene como principal aporte caracterizar la estructura de un modelo que permita la correcta clasificación automática de texto, la cual permita identificar las principales características gramaticales del wayuunaiki utilizando Deep Learning, características relacionadas con la aglutinación, multiplicidad de infinitivos, flexionadas de nombres y preposiciones, gradación verbal, sufijos temáticos, morfemas, yuxtaposición.

3.2 JUSTIFICACIÓN METODOLÓGICA

El modelo para clasificación de texto automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales, luego se su respectiva experimentación y elaboración, además de proponer el uso de una metodología que aporte al desarrollo ágil del modelo utilizando técnicas de Deep Learning, también incluiría el adecuado tratamiento de información procedente del texto escrito por una cultura indígena, respetando sus creencias y tradiciones.

Además, la metodología propuesta de esta investigación aportaría a otras investigaciones de procesamiento de lenguaje natural de culturas indígenas colombianas y latinoamericanas que presenten similares características gramaticales.

3.3 JUSTIFICACIÓN PRÁCTICA

La presente investigación se realiza como aporte a la necesidad de una adecuada clasificación de texto que apoye con el modelo propuesto las demás tareas de procesamiento de lenguaje natural que se puedan realizar para un idioma indígena que posee características gramaticales particulares, tareas de PLN como traducción de texto, análisis de texto, análisis de sentimientos, así como sus distintas aplicaciones entre las cuales encontramos búsqueda avanzada de información, Chatbots. Detección de tópicos, Educación, similitudes o anomalías en los textos, etc. pueden aportar a la conservación de un idioma indígena en peligro de extinción, a la generación de textos de etnoeducación y a la divulgación de información en su propio idioma y con sus particularidades culturales.

Para el modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales, se utilizará Deep Learnig con la finalidad de que el modelo tenga mayor eficiencia y se mejore la mejoré la precisión del modelo, ya que, al implementar este tipo de redes neuronales profundas (Cho et al., 2014) El modelo puede aprender representaciones más ricas y abstractas de los textos en wayuunaiki. Igualmente, se busca que con el uso de Deep Learning, el modelo sea flexible y escalable al permitir realizar diferentes tareas de clasificación sin una reconfiguración significativas y al manejar desde pequeños a grandes volúmenes de datos, ya que se busca indagar sobre la efectividad del modelo utilizando una pequeña cantidad de datos, como lo han realizado otras investigaciones que utilizaron pequeños corpus para los experimentos realizados (Deng et al., 2024), lo que aportaría significativamente al procesamiento de lenguaje natural de un idioma que tiene limitados conjuntos de datos de buena calidad y avalados por traductores oficiales de la comunidad indígena.

4. MARCO DE REFERENCIA

A Continuación, se presentan dos ejes principales, uno acerca de las comunidades wayúu y del idioma indígena wayuunaiki y otro lo relacionado a redes neuronales, minería de datos, Deep Learnig y procesamiento de lenguaje natural – PLN y finalmente se describen los antecedentes y conocimientos previamente construidos relacionados con la clasificación de textos en idiomas indígenas.

4.1 MARCO TEÓRICO

4.1.1 LA COMUNIDAD INDIGENA WAYUÚ

La comunidad indigena Wayuú representa alrededor del 20% de la población indigena en Colombia y se hubica principalmente en la alta juagira con un 88,5% de la problación Wayuú en el area rural disperso como lo muestra la figura 1, según la información sociodemográfica del pueblo Wayúu reportada por del DANE (Departamento Administrativo Nacional de Estadística - DANE, 2021).

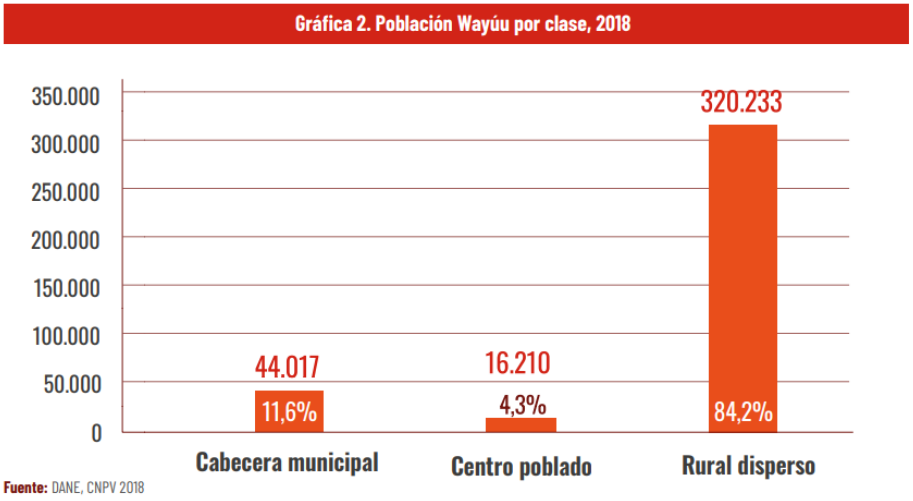


Figura 1. Información sociodemográfica del pueblo Wayúu.
Fuente: (DANE, 2021).

Los indigenas Wayuú son una población seminomada que viven en asentamientos dispersos denominados rancherias que se rehubican dependiendo de la temporada ambiental, necesidades de pastoreo o por vinculos familiares; El DANE identifico 21 resguardos de los cuales el 95% de los Wayuú en los resguardos se concentra en la Alta y Media Guajira, según Información sociodemográfica del pueblo Wayúu del DANE (Departamento Administrativo Nacional de Estadística [DANE], 2021).

4.1.2 El IDIOMA WAYUUNAIKI

El idioma wayuunaiki es el idioma oficial de la comunidad indigena wayuú, el wayuunaiki en Colombia es hablado por 334.939 indigenas Wayuú según la información sociodemográfica del pueblo Wayúu del DANE (Departamento Administrativo Nacional de Estadística - DANE, 2021) y pertenece a la familia lingüística Arawak dentro del grupo Maipuran, el wayuunaiki tiene dos dialectos el arribero que se habla en la Alta Guajira y el Abajero que se habla en la parte inferior de la península, estos dialectos presentan diferencias mininas que no intervienen en la comunicación de los indígenas Wayuú. En la figura 2, se muestra que en las rancherías de Manaure parte alta de la Guajira, se encuentran 89457 personas de la población indígena Wayuu, lo que indica su alta presencia en la Alta Guajira.

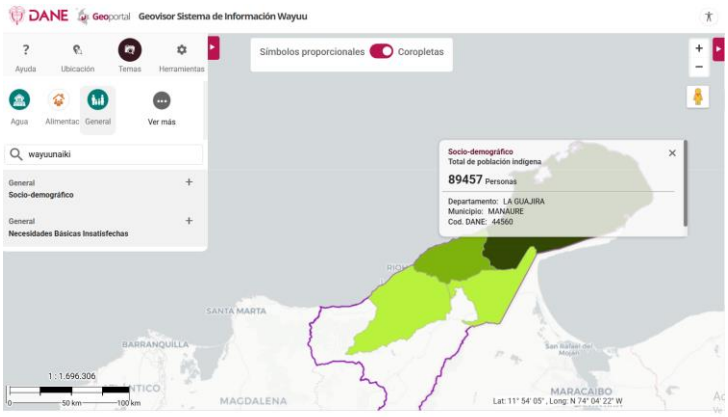


Figura 2. Información del sistema de información Wayúu.
Fuente: (DANE, 2025).

El idioma Wayuunaiki basa su escritura en el alfabeto latino modificado y es un tipo de lengua polisintética, es decir que una sola palabra puede expresar lo que en otros idiomas se expresaría con una oración compuesta por varias palabras, la polisintética se genera combinando en una sola palabra compleja múltiples unidades mínimas de significado denominadas morfemas.

El idioma Wayuunaiki está compuesto por 22 fonemas, divididos en 6 vocales y 16 consonantes y las palabras en una oración se organizan colocando primero el sujeto, luego el objeto y al final el verbo, esta organización puede presentar variaciones dependiendo del contexto de la oración.

4.1.3 CONSERVACIÓN DEL IDIOMA WAYUUNAIKI.

El 89,2% del total de la población de los indígenas Wayuú, entiende y habla el idioma wayuunaiki y un 7,1 % no lo entiende y no lo habla, estos porcentajes muestran que el wayuunaiki dentro del pueblo Wayuú tiene un alto porcentaje de conservación pero que una parte de la población está empezando a no utilizar el idioma wayuunaiki, sobretodo en la población joven.

El gobierno de Colombia ha fomentado importantes proyectos de etnoeducación para apoyar la conservación del wayuunaiki e igualmente las universidades tienen proyectos de investigación en compañía de organizaciones indígenas que buscan generar mecanismos que ayuden a la enseñanza y conservación del idioma, pero en general se presentan importantes problemas para estas estrategias, un problema muy recurrente es la poca cantidad de material de alta calidad física y digital escrito en idioma wayuunaiki existente, lo que genera poca disponibilidad para la enseñanza, estudio o análisis de datos (Pérez van Leenden, 2020). Otro problema es que muchos de estos materiales existentes generalmente no cuentan con un lingüista en wayuunaiki y/o con un hablante nativo del idioma.

4.1.4 MATERIALES DIGITALES EN WAYUUNAIKI Y ACCESO DE LOS WAYUÚ A LA INFORMACIÓN

La incursión de nuevas tecnologías en las comunidades indígenas es una iniciativa que el ministerio de las TIC quiere promover a través del uso de herramientas informáticas, la mayoría de estas herramientas se han desarrollado como materiales educativos computarizados o APP de traducción de palabras, pero hasta el momento no se ha desarrollado una herramienta que realice tareas de procesamiento de lenguaje natural como la clasificación de texto y traducción de palabras y frases especificas del idioma español a Wayuunaiki, que permita a las comunidades Wayuu acceder al conocimiento disponible en la web en su propio idioma.

El ministerio de las TIC de Colombia con su programa computadores para educar entre otros busca fomentar y apropiar a todo colombiano de las herramientas y aplicaciones digitales, por tal motivo ha entregado equipos con acceso a internet a las escuelas públicas del país y entre estas a las escuelas de las rancherías de la Guajira, beneficiando a los niños de la comunidad Wayuú.

En su Informe de gestión 2024, el programa Computadores para Educar (Computadores para educar, 2024) reportaba para la Guajira capacitaciones y acompañamiento a docentes, entrenamiento y acompañamiento a estudiantes y el desarrollo de 28 talleres para niños en el municipio de Maicao (Guajira). También Computadores para educar (2021) público en su página web oficial, la entrega de computadores portátiles para el departamento de la Guajira, en específico 1.500 para Riohacha, 160 para el Molino, 310 para Villanueva, 105 para Urumita y 60 para LaJagua del Pilar. También el Ministerio de Tecnologías de la Información y las Comunicaciones en Colombia – Mintic (2017), publico la entrega de soluciones fotovoltaicas y tablets a la ranchería del corregimiento El Estero, con la finalidad de apoyar la formación académica de los niños y jóvenes de la comunidad indígena Wayuu. Lo anterior representa una muestra importante del acceso que tiene la población de la Guajira y entre ellos las comunidades indígenas a las TIC.



Figura 3. Fotografía sobre la entrega de tabletas en La Guajira.
Fuente: (MinTIC, 2017).

Para promover un adecuado aprovechamiento de las TICs el gobierno nacional ha realizado proyectos de alfabetización, uno de estos proyectos denominado: "P*royecto de Alfabetización Digital realizado en el Resguardo Indígena de Guambía (Colombia)"*, en donde se realizaron capacitaciones de alfabetización digital; igualmente a través del ministerio de las TICs promueve la creación de aplicaciones para la Etnoeducación: "Es la que se ofrece a grupos o comunidades que integra la nacionalidad y que posee una cultura una lengua unas tradiciones y unos fueros propios y autóctonos. Educación que debe estar ligada al ambiente, al proceso productivo, al proceso social y cultural con el debido respeto a sus creencias y tradiciones." (Ley Nº 115, 1994).

En países como Venezuela también se trabaja en aplicativos específicos para la comunidad wayúu, entre ellos cabe destacar el trabajo de Quero y Madueño (2006), en el cual desarrollan el proyecto "Süchiki Walekerü: un ejemplo del uso de las tic en escuelas indígenas. Caso Wayuu", donde muestran que orientaron su investigación a la aplicación de tecnologías de la información y las comunicaciones (TIC), en específico el desarrollo y aplicación de un software educativo enfocado especialmente para la cultura wayuu, este software pretende fomentar la lectura y escritura del idioma Wayuunaiki y así apoyar el Régimen De Educación Intercultural Bilingüe del estado de Zulia en Venezuela.

Lo anterior ejemplifica las diversas propuestas que abordan la necesidad de contar con información en idioma wayuunaiki que tenga un correcto procesamiento de lenguaje natural y sus aplicaciones como búsqueda avanzada de información, Chatbots. Detección de tópicos, Educación, similitudes o anomalías en los textos, etc. PLN que aportaría a la conservación de un idioma indígena en peligro de extinción, a la generación de textos de etnoeducación y a la divulgación de información en wayuunaiki que hasta el momento cuenta con pocos textos de alta calidad como se muestra en la investigación realizada por Arroyave Ferreiro (2023), en donde se analizan diccionarios, manuales, artículos y libros revisados y recomendados en trabajos de análisis lexicográficos del idioma Wayuunaiki.

4.1.5 PROCESAMIENTO DE LENGUAJE NATURAL

El Procesamiento del Lenguaje Natural (PLN), según Vásquez et al. (2009) es una de las tareas más importantes de la inteligencia artificial, en la cual se busca procesar con herramientas de computación el lenguaje natural en formato de texto o voz, con la finalidad de que la maquina lo entienda y se puedan realizar tareas de traducción automática, recuperación de información, extracción de información, resúmenes de información, resolución de problemas, tutores inteligentes, clasificación de texto, detección de sentimientos, reconocimiento de voz.

El Procesamiento del Lenguaje Natural (PLN) según Vásquez et al. (2009) posee cinco niveles en su arquitectura: nivel fonológico, nivel morfológico, nivel sintáctico, nivel semántico y nivel pragmático.

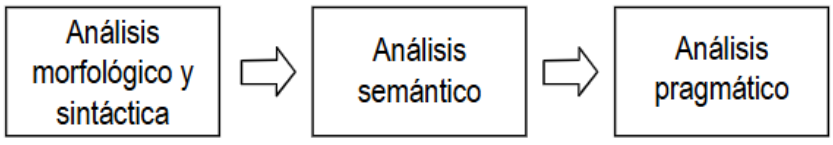


Figura 4. Organización de la arquitectura de un modelo de Procesamiento del Lenguaje Natural (PLN).
Fuente: (Vásquez et al., 2009)

Con relación a las tareas principales del PLN se encuentran:

Análisis de Sentimientos: Esta tarea es una de las más utilizadas actualmente y consiste en identificar las emociones en el texto escrito. Es muy utilizada para análisis de opinión por ejemplo.

Reconocimiento de Voz (speech-to-text): En esta tarea los datos de voz se transforman en texto escrito y se utiliza para aplicaciones que utilizan por ejemplo órdenes de voz o preguntas habladas. El principal desafío de estas tareas es las diferentes pronunciaciones que existen lo que genera errores al relacionarla con la gramática.

Extracción de Información: En esta tarea se identifica información de específica y sus relaciones con el texto con la finalidad de obtener la información más relevante de una temática establecida para análisis o para crear corpus de datos.

Traducción Automática: Esta tarea realiza la traducción entre un par de idiomas, pero presentan retos en relación a idiomas pocos conocidos.

Generación de Texto: En esta tarea se genera contenido de manera automática en respuesta a una petición como preguntas, como por ejemplo los chatbots y asistentes virtuales, también se utiliza para crear documentos, generar informes y en general variados textos.

Clasificación de texto es una de las tareas más importantes del Procesamiento del Lenguaje Natural (PLN), ya que es la base para el desarrollo de las otras tareas de PLN. Esta tarea identifica la categoría de un texto a partir de un conjunto de categorías previamente definidas de acuerdo a la temática establecida. Esta tarea se utiliza en general para organizar, analizar y extraer información útil de grandes volúmenes de texto de manera eficiente y tiene variadas aplicaciones dentro de las cuales se destacan el análisis de sentimientos, la categorización de noticias, el filtrado de spam, el preprocesamiento de datos, la Tokenización, la eliminación de palabras irrelevantes, la normalización de palabras y la eliminación de caracteres especiales y números.

4.1.6 CLASIFICACIÓN AUTOMATICA DE TEXTO

La clasificación de texto es una de las aplicaciones del procesamiento del lenguaje natural (PLN) más utilizadas, para esta tarea de clasificación de texto en la cual los datos se asignan a clases predefinidas (Muñoz Guillena, 2024), se han utilizado técnicas de machine learning, pero actualmente se utilizan técnicas de deep learning.

En un modelo de clasificación de texto se analizan las secuencias de entrada de datos de texto y de acuerdo a sus características se asignan a una clase (se etiquetan), un ejemplo de clases pueden ser análisis de sentimientos, análisis de opiniones, detección de enojos, detección de características lingüísticas, entre otras. Esta clasificación se puede realizar sobre palabra o sobre oraciones del texto.

Los principales métodos de clasificación de texto se dividen principalmente en tres categorías:

Basados en Reglas: Este tipo de método funciona utilizando reglas lingüísticas explícitas con la finalidad de determinar a qué clase corresponde un fragmento de texto de acuerdo a la temática seleccionada.

Basados en Aprendizaje Automático: En este método se busca que los algoritmos aprenden automáticamente a partir de datos etiquetados y se utilizan técnicas de machine learning como redes neuronales, SVM (máquinas de soporte vectorial) o árboles de decisión para clasificar el texto. Igualmente en este método en primer lugar se utiliza un conjunto de datos previamente entrenado o etiquetado para realizar el entrenamiento.

Híbridos: Algunos modelos de clasificadores de texto combinan reglas manuales como por ejemplo el etiquetado manual con técnicas de aprendizaje automático para obtener resultados de mejor calidad, aunque se generan mayores costos.

Las principales aplicaciones de los clasificadores de texto se enmarcan en los campos de identificación de información de acuerdo a temáticas, que necesiten un proceso automatizado de asignación de etiquetado al texto, como por ejemplo

en el campo de la medicina, las noticias, las empresas de análisis estadísticos, legal, entre otros.

En relación con las herramientas y lenguaje de Programación para el desarrollo de clasificadores se los más utilizados son Python y/o MATLAB, estás herramientas permiten de manera eficaz la utilización de librerías y módulos para el procesamiento de lenguaje natural utilizando algoritmos de machine learning o deep learning y en específico para la clasificación de texto, permitiendo la detección de las características relevantes para el análisis del texto.

4.1.7 REDES NEURONALES Y MACHINE LEARNIG PARA LA CLASIFICACIÓN DE TEXTO

En el estudio de RNAs para la traducción de textos se pueden encontrar documentos académicos como el propuesto por Casañ, y Castaño denominado: "Generación automática de codificaciones de vocabularios para Traducción Automática mediante Redes Neuronales", el cual aborda un modelo conexionista o Red Neuronal Artificial simple y utiliza ejemplos de frases del lenguaje objeto de estudio para luego codificar las palabras de dicho vocabulario, gracias a la experimentación estas codificaciones demostraron ser realmente útiles para la clasificación y posterior traducción automática de textos pero con la limitante de depender de la elección del tamaño de las codificaciones obteniendo una traducción que no es completamente automática, para resolver esta limitante los autores propusieron un mecanismo sencillo de poda para convertir el método totalmente en automático.

En síntesis (Casañ, y Castaño 2004) proponen la utilización de un Perceptrón Multicapa (PM) al cual se le aplicaron retrasos de tiempo para obtener mejores resultados de traducción texto a texto, específicamente el uso del traductor conexionista RECONTRA, el cual para disminuir el tamaño de la codificación selecciona las codificaciones de la capa oculta con una ventana de retrasos a la salida y un mecanismo de poda para encontrar el tamaño de la capa oculta y

aunque este mecanismo de poda no garantiza totalmente los resultados de traducción el mejorar constantemente el método de generación de codificaciones y aplicando nuevos algoritmos si podría garantizar dichos resultados.

En la tesis denominada “Traducción automatizada basada en algoritmos de redes neuronales artificiales” su autor (Escobar, 2009) busca encontrar un método para que las redes neuronales tengan mayor capacidad de aprendizaje y en específico al utilizar la red en perceptrón multicapa con backpropagation, lo anterior con la finalidad de obtener conclusiones lógicas gracias a un apropiado entrenamiento y a entradas adecuadas.

Escobar propone dos programas uno que para entrenar una RNA con una entrada fija de texto, y un segundo programa orientado a los usuarios el cual solicitara una traducción realizando consultas específicas, cada uno de los programas posee tres fases: conversión de datos, entrenamiento de datos y reconversión de datos.

El entrenamiento que propone escobar consta de más de 50 palabras por cada idioma y analizar de 1900 a 2000 frases, esto con la finalidad de aumentar la complejidad al aumentar el conjunto de entrada para identificar y corregir fallos secuencialmente, finalmente el autor puntualiza que el entrenamiento que brinda la posibilidad de verificar la conclusión lógica se logra al trabajar con una matriz de 1000 frases de entrada y 1000 frases de salida.

En general el trabajo de Escobar aporta ítems prácticos para el desarrollo del entrenamiento de una RNA para traducir texto del español al wayuunaiki, ya que describe de forma clara los conceptos fundamentales para realizar un entrenamiento que genere como resultado una traducción coherente.

En la investigación documental se abordaron también trabajos en otros idiomas que aportan una perspectiva más compleja y detalla del tema de redes Neuronales y entre ellos el trabajo de “Combining translation memories and statistical machine translation using sparse features” , describe los beneficios de combinar las memorias de traducción (TMS) y la traducción automática

estadística (SMT), esta combinación se produce gracias al uso de características aleatorias de tiempo de ejecución durante la decodificación, su trabajo se basa tanto en frases como en sintaxis y la aplicación de pruebas se efectuó realizando clasificaciones y traducciones de inglés a francés y de inglés al español obteniendo resultados considerablemente buenos (Li, L., Parra Escartín, C., Way, A., & Liu, Q. 2016).

En lo referente al trabajo de tesis “redes neuronales recurrentes para inferencia gramatical y traducción automática” realizado por Asunción Castaño (Castaño Álvarez, 1998), es de destacar el uso de Redes Neuronales Recurrentes (RNRs); el objetivo de utilizar estas redes es demostrar empíricamente que pueden resolver tareas específicas dentro del campo de la Inferencia Gramatical (IG), para proponer soluciones a problemas de Traducción Automática (TA) entre lenguajes e igualmente realizar de manera efectiva las tareas estadísticas que generan los datos de entrada.

Asunción Castaño menciona que para la buena compresión es necesario identificar las características sintácticas y semánticas para finalmente plantear un diseño conexionista para traducir naturalmente de un lenguaje a otro sin tener que pasar por una interlingua (sistema lingüístico de transición).

El articulo denominado “consideraciones sobre el desarrollo de un sistema experto para la traducción automatizada” de la revista de traductologia abarca los pasos necesarios para poder desarrollar un sistema experto para la traducción automática, estos pasos consisten en determinar en primer lugar el método para la captación de información, en segundo lugar la descripción del proceso de traducir, en tercer lugar la traducción automatizada y un apartado detalla de los factores a utilizar y las redes neuronales en el cual describe que existe la problemática de utilizar factores como palabras ya que el lenguaje cambia constantemente afectando el correcto funcionamiento del programa, teniendo en cuenta lo anterior los autores proponen utilizar una serie de redes neuronales dinámicas, las cuales solo tendrán en cuenta la red semántica y el valor del factor dependerá de los usos frecuente que utilice el sistema, también explican que los nodos de entrada, salida y enlaces dependerán del tamaño y

de la riqueza léxica del texto y enfatizan que la Red Neuronal Artificial no realiza una clasificación y traducción en la que se le entrega un texto como entrada y nos da una traducción como salida, lo que realiza la Red Neuronal Artificial es ajustar los factores de utilización para elegir la mejor clasificación y traducción entre las posibles clasificaciones y traducciones y finalmente que la Red Neuronal Artificial depende de su entrenamiento y experiencia para conseguir una clasificación y traducción más acertada (Alonso Berrocal, J. L., y otros 1996).

Es importante mencionar que Yoav Goldberg en su trabajo "A Primer on Neural Network Models for Natural Language Processing", explica detalladamente cómo se pueden utilizar diferentes modelos de RNAs para el procesamiento automático del lenguaje natural, como deben ser las codificaciones de entrada y la aplicación de redes feed-forward, recurrentes y recursivas, así como la aplicación del cálculo del gradiente automático, este trabajo destaca totalmente por el detalle con que desarrolla los temas en especial las delimitaciones para los datos de entrada y las representaciones de las palabras así como el proceso de entrenamiento y las limitaciones y regulaciones del mismo, también cabe destacar el esfuerzo del autor por motivar constantemente a utilizar las redes neuronales y la explicación matemática y estadística de las mismas.

El profesor Ernesto Llerena García de la universidad de córdoba en su artículo "Software traductor de español a lengua kuna" publicado en las memorias del XXVII Congreso Nacional y I Internacional de Lingüística, Literatura y Semiótica, describe un trabajo muy similar al que se quiere desarrollar con las redes neuronales para la traducción de textos del idioma español al idioma Wayuunaiki, pero con la puntual diferencia que utiliza un modelo de traducción automática estadística y un modelo basado en reglas, para realizar la traducción del idioma español al kuna, igualmente el trabajo desarrollado por Llerena (2012), se efectúan bajo procesos de traducción automática (MT) y afirma que existen dos tipos de traducción automática, la traducción automática estadística y la traducción automática basada en reglas. El primer modelo suele utilizarse para lenguas con una estructura gramatical similar y la segunda para estructuras gramaticales complejas con el objetivo de lograr mayor calidad en la coherencia

del traductor, puesto que se basa en la creación de unas reglas gramaticales sin importar la semejanza de las lenguas.

Asimismo, dentro de la traducción automática basada en reglas, existen dos modalidades, la basada en reglas por trasferencia y la basada en reglas interlingua. La traducción automática basada en reglas por trasferencia realiza un análisis de lo que se va a traducir y se encarga de remplazar palabra a palabra por un equivalente. Por otro lado la traducción automática basada en reglas interlingua utiliza una tercera lengua de referencia con estructura similar para la lengua de origen y destino, por lo que bajo esta modalidad las palabras pasan de una lengua de origen a un lengua intermedia, para finalmente llegar a la lengua destino (Llerena, 2012); Para concluir, es importante señalar que se considera como factor determinante para el éxito de un traductor la construcción de una base de datos completa y la efectividad del programa para buscar rápidamente las palabras en ella (Llerena, 2012).

Un trabajo verdaderamente interesante fue desarrollado por Fernández, Gamboa, Atencia y Bedoya (2013), estudiantes colombianos quienes en su artículo "*Design and implementation of an "Web API" for the automatic translation Colombia's language pairs: Spanish-Wayuunaiki case*" , describen como se diseñó e implemento un API que despliega servicios de traducción de español a wayuunaki a través de internet; este proyecto es muy similar a lo que se quiere trabajar pero con la diferencia que no utilizan redes neuronales artificiales y además no lograron una buena fidelidad del resultado ya que requerían un volumen grande de información para asegurar que las clasificaciones y traducciones se acerquen a la coherencia gramáticas del idioma wayuunaki que utilizan las comunidades wayúu.

Esta revisión documental se logró identificar la evolución de la aplicación de las redes neuronales para tareas de PLN como la clasificación, traducción y tratamiento de textos, las cuales han sido un reto interesante y para las cuales diferentes autores han abordado el tema con la finalidad de garantizar clasificaciones de textos de buena calidad.

Con respecto al aporte para el proyecto se puede concluir que en general los diferentes trabajos analizados dan una visión más clara y precisa de los pasos a seguir para diseñar e implementar una RNA para la clasificación automática del idioma español a wayuunaiki.

Finalmente se pudo establecer que las investigaciones realizadas hasta el momento no han considerado en entrenamiento de una Red Neuronal Artificial que utilice como entrada datos de un idioma indígena y que garantice que la salida cuenta con coherencia gramatical, lo cual genera la posibilidad de realizar un estudio del comportamiento y entrenamiento de las RNAs con respecto a un idioma con estructura gramatical poco común.

4.1.8 TRABAJOS SIMILARES

Se han utilizado modelos de Deep Learning para la clasificación de textos en con idiomas con estructuras gramaticales diferentes como el árabe (Elnagar et al., 2020) y el japones (Chen et al., 2023), los cuales han demostrado generar excelentes resultados de rendimiento.

En detalle para el caso de la clasificación de textos en idioma arabe, Los autores describen inicialmente que en internet se pueden encontrar una gran cantidad de texto no estructurado el cual es un gran recurso para el procesamiento y gestión de datos, los cuales se pueden utilizar para tareas automáticas de clasificación y agrupación de textos utilizando Deep Learning; igualmente se menciona la creciente necesidad de los clasificadores automáticos por ejemplo para la identificación de idiomas o dialectos, los cuales para su desarrollo necesitan que los documentos estructurados tengan un formato compatible con los algoritmos de aprendizaje automático y permitan la adecuada generación de vectores numéricos. En el caso del idioma árabe, el cual se encuentra en el listado de idiomas más utilizados, se registran pocas investigaciones relacionadas a la clasificación automática, las cuales se han realizado con procesos clásicos de aprendizaje automático para la clasificación y utilizando conjuntos de datos pequeños y en su mayoría de acceso restringido, por tanto, los autores destacan que en la investigación se crearan dos conjuntos grandes

de datos de libre acceso denominados: SANAD (conjunto de datos de artículos de noticias en árabe de etiqueta única) y NADiA (conjunto de datos de artículos de noticias de varias etiquetas en árabe), e igualmente informan que para la categorización de texto usando una sola etiqueta y usando etiquetas múltiples, se utilizará Deep Learning, en específico Convolution Neural Networks (CNN), Recurrent neural Networks (RNN), la técnica de atención y el uso de word2vec para mejorar el rendimiento; también indican que esta investigación es el primer estudio integral en utilizar Deep Learning para el procesamiento del lenguaje árabe y que se busca eliminar la necesidad de la fase de preprocesamiento y generar un clasificador robusto y altamente preciso para artículos de noticias en árabe.

Igualmente para el trabajo de investigación relacionado con la clasificación de texto en idioma japonés se describe que se centra en mejorar la clasificación de texto con clasificación múltiple en idioma japonés, utilizando un modelo de aprendizaje conjunto (ensemble learning), formado por la red neuronal convolucional TextCNN combinado con los modelos de procesamiento de lenguaje con preentrenamiento: ALBERT, RoBERTa, DistilBERT.

Inicialmente los autores introducen la importancia de la tarea de clasificación de texto en el procesamiento de lenguaje natural (Natural language processing-NLP), indicando que esta tarea es fundamental y esencial, además de ser muy utilizada para análisis de sentimientos, filtrado, clasificación de noticias, ChatGPT, entre otros usos. Igualmente detallan que constantemente se busca mejorar la clasificación de texto para logar mayor precisión y eficiencia y aunque en el principio fue muy común el uso de un conjunto de reglas establecidas manualmente, estas reglas se volvían muy difíciles de mantener debido a la gran cantidad de datos no estructurados lo que generaba un problema que actualmente al utilizar deep learning se podría solucionar, lo anterior teniendo en cuenta que estos métodos de aprendizaje profundo poseen funciones de aprendizaje automáticas a partir de datos enriquecidos, que trabajan bien en la tarea de clasificación de texto. Igualmente, los autores indican que la mayoría de los experimentos se han desarrollado con conjuntos de datos del idioma

inglés y muy poco con otros idiomas que no tienen disponible un gran conjunto de datos y difieren en su estructura gramatical, lo que genera la necesidad de investigar la clasificación de texto en un idioma como el japonés, el cual posee características diferentes al inglés.

4.2 MARCO CONCEPTUAL

4.2.1 IDIOMA WAYUUNAIKI

El wayuunaiki es hablado por los indígenas de los pueblos guajiros o Wayúus, este idioma hace parte de la familia Arawak o Arahuaca, específicamente dentro del grupo Maipuran. Álvarez González (2017) en su Manual de la lengua wayuú describe que el idioma de los indígenas Wayuús históricamente ha sido denominado como guajiro, pero actualmente se usa el autoglotónimo wayuunaiki, el concepto de autoglotónimo hace referencia al nombre de la lengua que es asignado por los hablantes de esa lengua.

El wayuunaiki en comparación con otras lenguas como el español presenta características gramaticales particulares del idioma como la formación de palabras compuestas uniendo varios morfemas lo que hace que este idioma se considere polisintético, también presenta un orden sintáctico diferenciador en su estructura gramatical, la más general inicia con el Sujeto, continua con el Objeto y termina con el Verbo, pero existen variadas formas de construir oraciones simples y complejas en wayuunaiki dependiendo del contexto según lo expuesto por Álvarez (2015).

4.2.2 CARACTERISTICAS GRAMATICALES DEL WAYUUNAIKI

El wayuunaiki es una lengua aglutinante, es decir que las palabras se forman uniendo monemas (Unidad mínima de significado de la lengua) independientes Según Goulet & Jusayú (1978). El wayuunaiki posee 22 fonemas, divididos en 6 vocales y 16 consonantes, los fonemas se muestran en la tabla No.1.

Tabla 1. Fonemas wayuunaiki.

CONSONANTES															
Ch	Sh	j	k	l	m	n	p	r	s	t	w	y	(´)	(v)	(ñ)
VOCALES															
a		e		i		o		u		Ü					

Vocales

Las vocales pueden ser dobles geminadas (aa, ee, ii, oo, uu, üü) y solamente pueden ser dobles o geminadas, las consonantes: mm, nn, tt.

La estructura gramatical del wayuunaiki está conformada por el siguiente esquema:

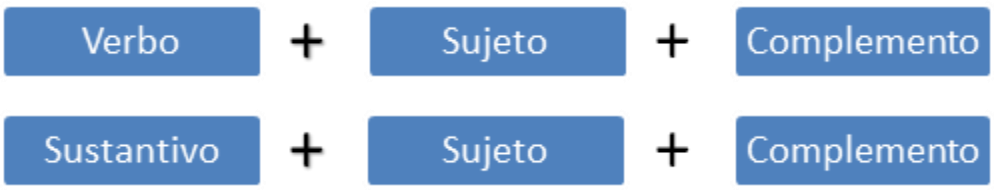


Figura 5. Estructura gramatical del *Wayuunaiki.*
Fuente: Elaboración propia.

En el wayuunaiki, el predicado suele proceder al sujeto de modo que se obtiene la información acerca de la acción y solo después acerca del que la ejecuta según lo explica en su manual Álvarez González (2017).

Sufijos determinativos

En el idioma wayuunaiki, los sufijos determinativos definen los nombres que identifican a personas y animales, estos sufijos determinativos pueden ser absolutos (no tiene relativos de dependencia), es decir que los nombres no se clasifican por género, si no por masculino o no masculino (Mujeres o animales hembras), Según Captain y Captain (2005).

Tabla 2. Sufijos Determinativos wayuunaiki. (Captain y Captain, 2008).

SUFIJOS DETERMINATIVOS		
	Singular	Plural
Masculino	-Kai	
No masculino	-Kat	
No especificado		-Kana

Igualmente, en wayuunaiki se pueden encontrar los nombres poseídos, los cuales son de distintas clases (Nombres de parentesco y partes del cuerpo) y su estructura requiere un prefijo que indica poseedor.

El pronombre personal

Los pronombres personales en Wayuunaiki tienen una estructura invariable, a diferencia de los pronombres en español, en la tabla 3, se detallan los 7 pronombres personales en Wayuunaiki.

Tabla 3. Los pronombres y prefijos personales.
(Captain y Captain, 2008).

PRONOMBRES Y PREFIJOS	Pronombre	Prefijo
Primera persona singular	taya	ta-
Segunda persona singular	pia	pü-
Tercera persona masculino singular	nia	nü-
Tercera persona no masculino singular	shia	sü-

PRONOMBRES Y PREFIJOS	Pronombre	Prefijo
Primera persona plural	waya	wa-
Segunda persona plural	jia	jü-
Tercera persona plural	naya	na-

La preposición

En el idioma español la preposición se utiliza como una unión entre elementos que conformar la oración y funcionan como complemento, pero la proposición en el wayuunaiki va después del complemento, esta particularidad del wayuunaiki causa que se cambie el orden esperado de la oración. Otra característica importante de las preposiciones en el idioma wayuunaiki es se acompañan de un prefijo personal, característica que algunas veces genera presenta aglutinación en el nombre que acompaña la preposición. (Captain y Captain, 2008).

Ejemplos:

- nüle'eru'u → "su-vientre-en" → "en su vientre"
- sulu'u tü apainkat → "ella-en la huerta" → "en la huerta"

El adverbio

En el idioma español el adverbio modifica el significado del verbo, en el idioma wayuunaiki el adverbio cumple la misma función, la mayoría de los adverbios en la estructura gramatical del wayuunaiki son invariables, pero existen los adverbios relativos los cuales necesitan un prefijo personal para que en la oración coincidan con el sustantivo de acuerdo en lo expuesto por Captain y Captain (2008).

Ejemplo:

- antüsü teikat süchukua'a → "llegó mi-mamá ella-otra vez" → "Mi mamá llegó otra vez"

Conjunciones

En el idioma español las conjunciones son palabras que enlazan ideas y oraciones, en el idioma wayuunaiki realizan la misma función de unir oraciones, frases o ideas en su mayoría usando palabras invariables pero en algunos casos existen las conjunciones subordinantes, las cuales tienen una estructura similar a las preposiciones y necesitan de un prefijo en tercera persona no masculina (Captain y Captain, 2008).

Ejemplo:
Conjunción aka ("porque"):

- süka ayuulin nia → "porque él está enfermo"

El verbo

En la gramática del idioma indígena wayuunaiki al igual que en el español, los verbos se clasifican como intransitivos y transitivos, pero existe una particularidad con los verbos denominados de estado, verbos en wayuunaiki que se emplean para expresar una característica del pronombre que en la estructura de la oración actúa como sujeto según lo expuesto por Captain y Captain (2008).

Por ejemplo en el idioma wayuunaiki de la palabra atsüin que significa ‘fuerza’, se pueden generar dos verbos de estado, el primero un verbo atributivo el cual se obtiene añadiendo el prefijo ka- generando la palabra katsüinwaa que significa ‘ser fuerte’ y el segundo un verbo negativo que se obtiene añadiendo el prefijo ma- a generando la **ma**tsüinwaa que significa estar ‘sin fuerza’. (Captain y Captain, 2008).

La gramática de los verbos en el idioma wayuunaiki está compuesta por una raíz más un prefijo (Persona) o un sufijo y su conjugación no tiene un patrón determinado. En general en el idioma wayuunaiki los verbos pueden presentar aglutinación y representan una importante expresión cultural.

En el idioma wayuunaiki también se presentan otras características importantes, las cuales se describen de manera general a continuación:

Ortografía y Fonética: El idioma wayuunaiki utiliza un alfabeto específico que representa los sonidos de la lengua, donde la pronunciación juega un papel muy importante para entender, comprender y comunicarse en wayuunaiki.

Morfosintaxis: El idioma indígena wayuunaiki posee una estructura gramatical que contiene sustantivos, adjetivos, pronombres y verbos desempeñan un papel fundamental en la construcción de frases y oraciones, donde por ejemplo es relevante para poder expresar modo, tiempo y/o aspecto la flexión verbal.

Los pronombres personales en el idioma wayuunaiki indican a los interlocutores espacialmente (geográficamente) en la acción de comunicar, lo cual genera una particularidad compleja al momento de realizar una clasificación de texto teniendo en cuenta la gramática, ya que estas palabras “hacen las veces del sustantivo o del sintagma nominal y que se emplean para referirse a las personas, los animales o las cosas sin nombrarlos”. (Viloria, et.al, 2021).

En el idioma wayuunaiki se entiende la yuxtaposición como una construcción de oraciones compuestas. Consiste en solapar oraciones sin usar nexos, aunque se suelen emplear signos de puntuación como comas o punto y coma.

La aglutinación en gramática es un tipo de morfología que se encuentra en ciertas lenguas. En estas lenguas, las palabras están compuestas por la combinación de varios morfemas (las unidades más pequeñas de significado) que se adhieren uno tras otro. Cada morfema tiene un significado específico y se puede utilizar solo o en combinación con otros. Al unirse, forman palabras más largas y complejas.

4.2.3 REDES NEURONALES ARTIFICIALES

Las redes neuronales (RNAs) son sistemas de información compuesto por neuronas artificiales interconectadas, que buscan imitar los procesos de un cerebro biológico, también se conocen como modelos conexionistas. En la Figura 6 se puede detallar las conexiones de una Red Neuronal Artificial y sus correspondientes capas.

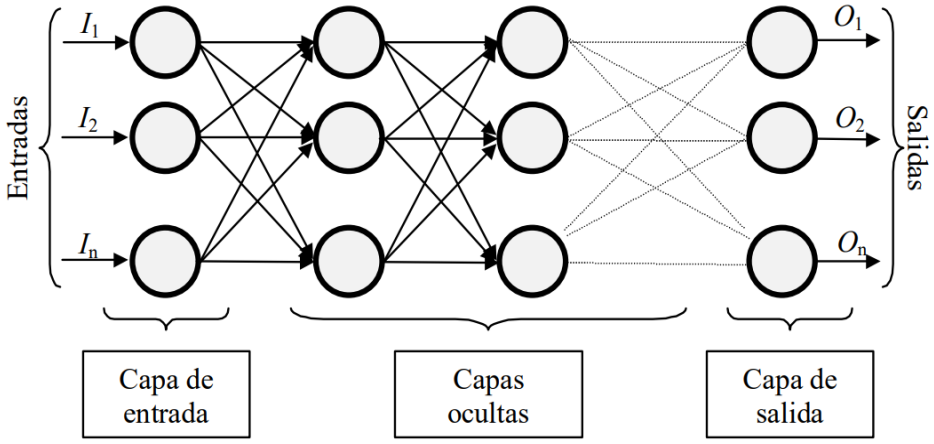


Figura 6. Capas básicas de una Red Neuronal Artificial
Fuente: (Matich, 2011).

La RNA consta generalmente de 3 capas:

- Capa Entrada: La cual recibe los datos de entrenamiento.
- Capas Ocultas: Realiza el procesamiento no lineal.
- Capa de salida: Envía las respuestas por cada entrada.

Entre las ventajas de las Redes Neuronales Artificiales se pueden encontrar:

- Aprendizaje Adaptativo*:* Las RNAs gracias a un entrenamiento aprenden a realizar tareas.

- Auto-organización: Tienen la habilidad en su etapa de aprendizaje de generar su propia organización de la información.

- Tolerancia a fallos: Si la RNA sufre algún fallo que conlleve a una pérdida parcial de su estructura se pueden conservar algunas capacidades de la red.

- Operación en tiempo real: Con una RNA es posible realizar cómputos neuronales en paralelo.

- Fácil inserción dentro de la tecnología existente*:* Existen chips especializados para RNAs que contribuyen a la mejor realización de las tareas.

Para la clasificación de texto las redes neuronales que más éxito han tenido son las redes neuronales convolucionales (Convolutional Neural. Networks, CNN), las cuales permiten la extracción automática de características importantes y asignan etiquetas de una forma más precisa, como se puede detallar en trabajos como el que propone un modelo de múltiples funciones para la clasificación de texto con la finalidad de reconocer la intención de habla, en específico la de los estudiantes en el campo del aprendizaje colaborativo en línea es de gran importancia para explorar las reglas de la construcción del conocimiento, como el desarrollo del conocimiento y la comunicación emocional en el proceso colaborativo. Pero la clasificación de texto que usa Deep Learning puede tener una identificación insuficiente de las relaciones semánticas contextuales y las palabras con características claves, que son características de la semántica contextual de la discusión colaborativa en línea. El modelo específicamente combina: BERT para representar el texto como un vector de palabras, BiLSTM para extraer las características contextuales del texto, CNN para extraer las características locales del texto y un modelo de agrupación promedio para extraer las características medias de representación del texto. (Li et al., 2023).

4.2.4 MINERÍA DE DATOS

La exploración de datos en búsqueda de grandes volúmenes de patrones de información se ha desarrollado utilizando la estadística, las bases de datos, la inteligencia artificial y el aprendizaje autónomo, las principales etapas de la minería de datos se observan en la Figura 7.

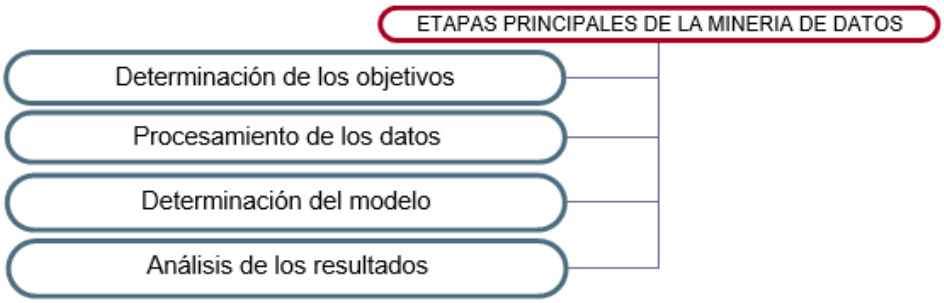


Figura 7. Etapas de la minería de datos.
Fuente: Elaboración propia.

Los algoritmos más destacados que se utilizan son de asociación, clústeres, árbol de decisión, regresión lineal, clasificador Bayesiano ingenuo, red neuronal, clústeres de secuencia y los de serie temporal.

"La minería de datos proporciona herramientas poderosas para descubrir patrones ocultos y relaciones en datos estructurados. Este proceso asume que los datos ya se encuentran almacenados en un formato estructurado. Por esta razón su pre-procesamiento consiste en la limpieza y normalización de los datos y la generación de numerosos enlaces entre las tablas de las bases de datos. La minería de datos usa técnicas y metodologías de RI, EI y corpus procesados con técnicas de lingüística computacional". (Hernández & Gómez, 2013)

Estas etapas de la minería de datos son un aporte importante para el desarrollo de una metodología que aporte al desarrollo de un modelo que utilice machine learning para la adecuada clasificación de texto en wayuunaiki.

4.2.5 MACHINE LEARNING

El concepto de aprendizaje automático (machine learning) tiene sus origenes en un articulo publicado en 1943 por los autores Walter Pitts y Warren McCulloch, titulado "Un cálculo lógico de las ideas inmanentes en la actividad nerviosa". En esta publicación, los autores explican que el funcionamiento de las neuronas en el sistema nervioso puede imitarse usando operaciones de lógica proposicional. Posteriormente en 1949 Donald Hebb publico el libro titulado "La organización del comportamiento", en el cual se describe que el aprendizaje a nivel neuronal (Ley de Hebb) se produce cuando las neuronas se activan de manera conjunta y refuerzan las sinapsis, esta y otras descripciones inspiraron directamente la idea de ajuste de pesos en las neuronas artificiales y en general varias teorías del aprendizaje automático. Más adelante en 1950, AlanTuring desarrollo la máquina de Turing y realizo la publicación científica titulada “Computing Machinery and Intelligence”, publicación en la que se describe el concepto de “máquinas de aprendizaje” que hace referencia a que estas máquinas pueden modificar sus procesos internos y/o comportamiento de acuerdo a la experiencia adquirida y en función de los datos ingresados (Firican, 2021). Desde estos aportes a hasta la actualidad el desarrollo del aprendizaje automático ha tenido grandes avances gracias a variados autores que han contribuido al desarrollo del machine learnig tal y como lo entendemos hoy en día.

El aprendizaje automático (machine learning) consiste en diseñar algoritmos que permitan a las maquinas aprender y mejorar continua y automáticamente en base a los datos que ingresan, sin necesidad de programarlas específicamente para cada tarea que deban realizar. Estos algoritmos deben ser capaces de identificar patrones y realizar acciones de decisión predictivas o descriptivas con la finalidad de ajustar parámetros de acuerdo a los datos que recibe. El machine learnig posee cuatro categorías de aprendizaje: supervisado, no supervisado, semisupervisado y por refuerzo y los algoritmos

más utilizados son redes neuronales, regresión lineal, regresión logística, agrupación en clústeres, arboles de decisión y Bosques aleatorios.

El machine learnig generalmente se usa para reconocimiento de voz, clasificación de imágenes y de texto, detección de fraudes, visión artificial, automatización de procesos, entre otras tareas. En el caso del procesamiento de texto el machine learnig permite automatizar de manera más sencilla y sin el proceso manual los datos de texto, adaptar los modelos a varios idiomas, un mejor rendimiento gracias al acceso actual a una mayor cantidad de datos, identificación de patrones semánticos y de contexto, pero también presenta desventajas como que requiere grandes cantidades de datos etiquetados, se presentan sesgos si los datos para el proceso de entrenamiento no son representativos, necesita bastantes recursos computacionales y más tiempo de entrenamiento, poco desarrollo de modelos de caja negra y necesitan una actualización contaste frente a los cambios del lenguaje.

Los avances en machine learnig se desarrollan a gran velocidad gracias a la nueva disponibilidad de hardware y a las mejoras en la capacidad de procesamiento y almacenamiento de datos, pero también se deben abordar los desafíos éticos que presentan estos avances y los posibles problemas que se generan debido a los sesgos en los set de datos o errores en los procesos de automatización, por tanto, es importante revisar de forma detallada la parte ética de los diseños de algoritmos y establecer límites y regulaciones para asegurar un desarrollo responsable.

4.2.6 DEEP LEARNING

La evolucion del aprendizaje profundo (machine learning) ha permitido la aparicion del subcampo denominado aprendizaje profundo (deep Learning) en el cual se utilizan redes neuronales con muchas capas profundas para aprender representaciones jerárquicas de los datos, en específico en cada capa se extraen y aprenden características cada vez más complejas que la anterior, identificando desde los patrones simples hasta los más complejos, lo que ha permitido a los modelos de deep learning realizar tareas como reconocimiento

de patrones económicos, detección de objetos en imágenes, segmentación semántica, procesamiento de lenguaje, generación de lenguaje creativo, identificación de características gramaticales complejas, entre otras tareas. (Gyanendra & Rajesh, 2023).

Actualmente el deep learning es más utilizado que el machine learning, esto debido a que presenta un mejor rendimiento para el procesamiento de conjuntos de datos de voz, imágenes y texto, estas mejoras en el rendimiento hacen posible abordar desafíos que antes no era posible abordar como por ejemplo en los casos de procesamiento y generación de texto, procesamiento de lenguaje natural (PLN), tratamiento de imágenes 3D y el problema de la aparición de varios conceptos o palabras aglutinadas en texto y voz.

En general al utilizar deep learning tambien se utilizan grandes cantidades de datos no estructurados con la finalidad de que las neuronas aprendan automáticamente, el aprendizaje se genera sin la necesidad de realizar de forma manual la identificación y etiquetado de los datos de acuerdo a las variables, esto debido a que el deep learning permite se puede extraer automáticamente los patrones y representaciones significativas de los datos generando de manera eficiente relaciones complejas y permitiendo realizar tareas avanzadas como reconocimiento de imágenes avanzado, procesamiento de lenguaje natural (PLN) y texto y predicciones de patrones complejos, entre otras.

Estos avances y características obtenidas gracias al uso del Deep Learnig ha permitido el mejor abordaje de problemas en campos como la robótica, la medicina, la gramática, la visión por computadora y marketing masivo entre otros; También el abordaje de variados problemas de análisis de conjunto de datos con pocos recursos de información se ha logrado realizar de manera más eficiente gracias a los avances en técnicas de Deep Learning. (Gegúndez & Pérez, 2021).

En el Deep Learnig la categorización de los datos de entrada hace referencia al conjunto de datos codificados con valores adecuados que permitan

representar las principales características del conjunto de datos y la operación en tiempo real hace referencia a procesamiento rápido de palabras de entrada sin necesidad de ajustar mucho los pesos y conexiones y realizar un entrenamiento mínimo.

4.2.7 REDES NEURONALES CONVOLUCIONALES (CNN).

Las redes neuronales convolucionales (CNN) son modelos planteados para procesar datos con estructura espacial, principalmente imágenes. Su principal fortaleza radica en tres propiedades clave: conectividad local, compartición de parámetros y extracción jerárquica de características. Esto permite reducir drásticamente el número de parámetros frente a una red totalmente conectada, manteniendo una alta capacidad de representación.

Para las tareas de procesamiento de lenguaje natural (PLN) y texto las redes neuronales convolucionales utilizan filtros de una dimensión que recorren secuencias de embeddings, con la finalidad de identificar patrones locales de combinaciones sintácticas o n gramas con la finalidad de facilitar la extracción de características relevantes en las secuencias de palabras y permitir un mejor análisis y clasificación de los datos. (Kim, 2014).

4.2.8 DOC2VEC

Doc2Vec es un mecanismo para representare en vectores densos de tamaño fijo fragmentos de texto (párrafos o documentos completos) como vectores densos de tamaño fijo, Doc2Vec nació de la idea de Word2Vec con la diferencia que Doc2Vec incorpora un “vector de documento” que ayuda a capturar información global sobre el contenido, más allá de las relaciones entre palabras.

Doc2Vec utiliza dos mecanismos, el primero PV-DM (Distributed Memory), en el cual se asocia un documento posee un vector fijo y único, vector que se combina con los embeddings de las palabras cercanas para predecir la palabra central, reforzando las relaciones en memoria del texto y su contexto; en la

estrategía PV-DBOW (Distributed Bag of Words), utiliza únicamente el vector del documento para predecir palabras muestreadas al azar dentro de ese mismo texto, prescindiendo de los embeddings contextuales (vectores de contexto), lo que simplifica el entrenamiento y la identificación eficaz de contextos globales. (Google Developers, 2025, sección “Cómo obtener embeddings”)

El Modelo de clasificación es un algoritmo o sistema que asigna etiquetas o categorías a textos basándose en características específicas. Puede ser supervisado (entrenado con datos etiquetados) o no supervisado (descubre patrones sin etiquetas previas) y las Redes de Palabras corresponden a Grafos que representan la co-ocurrencia de palabras. Se utilizan para capturar dependencias entre términos en el clasificador.

4.2.9 RED NEURONALE RECURRENTE

Las redes neuronales recurrentes (Recurrent neural networks - RNNs), aportann significativamente al procesamiento de lenguaje natural (PLN) al facilitar el análisis de secuencias de datos, como texto y voz, gracias a su estructura que incluye memoria interna. A diferencia de otras redes neuronales convencionales, las RNNs tienen la capacidad de retener información a lo largo del tiempo, lo que las convierte en una opción perfecta para tareas como la traducción automática, la generación de texto y la identificación y clasificación de sentimientos.

Las RNNs tienen una arquitectura donde las conexiones forman un ciclo interno, permitiendo que los nodos ocultos conservan los estados anteriores , los nodos se actualicen en cuanto reciben la nueva información. Su funcionamiento se basa en la propagación de datos a lo largo de una secuencia temporal. En las RNNs, cada uno de los nodos de la red procesa información de acuerdo a a la entrada actual y al estado previo.

Al principio, los modelos de clasificación de texto se basaban en la arquitectura del Perceptrón Multicapa (MLP), que funciona a través de capas

completamente conectadas y utiliza funciones de activación no lineales. Sin embargo, estos modelos tenían una limitación importante: la pérdida gradual de información en secuencias largas, lo que afectaba la capacidad de retener datos anteriores durante el proceso de clasificación. Este problema se debe al fenómeno del gradiente desvaneciente, que ocurre durante la retropropagación de la red neuronal. En este proceso, los gradientes que provienen de las funciones de activación tienden a disminuir exponencialmente al pasar por varias capas, lo que dificulta la actualización efectiva de los pesos en las capas iniciales de la red. Como resultado, el modelo pierde la habilidad de mantener información relevante en tareas de procesamiento de lenguaje natural. Para abordar este desafío, se han creado arquitecturas especializadas como las Redes Neuronales Recurrentes (RNN), Long Short-Term Memory (LSTM) y Gated Recurrent Units (GRU), que incluyen mecanismos de memoria diseñados para conservar el contexto a lo largo de una secuencia, evitando así la degradación de la información.

4.2.10 LSTM

Las redes neuronales recurrentes de tipo LSTM (Long Short Term Memory) son redes de larga memoria a corto plazo, son capaces de recordar dependencias a largo plazo y por tanto son una solución para mantener el contexto en secuencias de datos, algo que las redes neuronales tradicionales no pueden hacer bien. Lo que las hace especiales es que tienen una especie de memoria interna, lo que les permite recordar información relevante a lo largo del tiempo y usarla para mejorar su rendimiento en tareas futuras, según lo explicado en la documentación de Mathworks, (2024 –a). Como lo muestra la Figura 8, su funcionamiento se basa en un estado oculto y una serie de bucles que permiten que la información anterior siga influyendo en la salida de la red. Durante el entrenamiento, la red ajusta dos conjuntos de pesos: uno asociado al estado oculto y otro vinculado a las entradas. Esto le permite aprender patrones en los datos y tomar decisiones más acertadas.

Al momento de procesar datos, como una frase en un modelo de procesamiento de lenguaje, la salida no se basa solo en la palabra actual, sino también en el

contexto previo almacenado en el estado oculto. Esto es lo que hace que las LSTM sean tan efectivas en tareas como procesamiento de texto, reconocimiento de voz y predicción de series temporales.

El redescubrimiento de estas redes, aplicadas al PLN, ha supuesto una revolución que, pese a llevar varios años usándose, siguen consiguiendo algunos resultados cercanos al estado del arte.

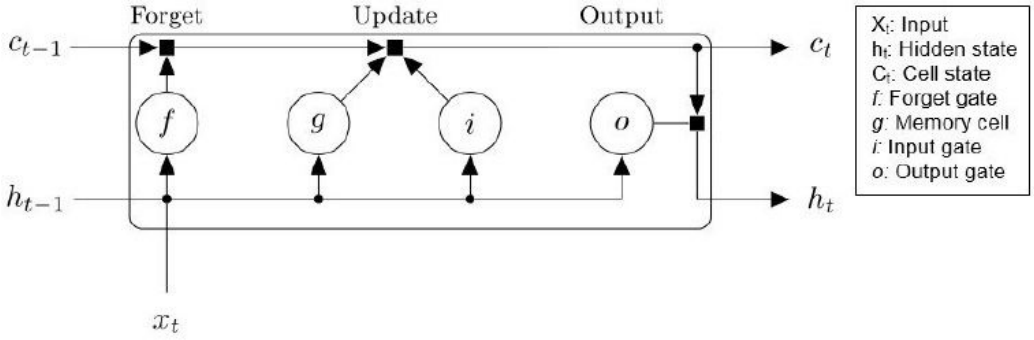


Figura 8. Flujo de datos en una red LSTM.
Fuente: (mathworks, 2024 –a).

4.2.11 REDES BIDIRECCIONALES

Las redes recurrentes bidireccionales es la unión de dos redes neuronales recurrentes, las entradas de una van en orden normal mientras que los de la otra van en orden inverso. Después, las salidas de ambas se concatenan (Datitos, 2021).

La estructura de las redes bidireccionales permite tener información en la secuencia de texto tanto hacia adelante como hacia atrás. Esta característica permite su utilización en tareas de procesamiento de lenguaje natural (PLN) como por ejemplo para la clasificación de texto e igualmente su uso como base en modelos como BERT(Bidirectional Encoder Representations from Transformers).

4.2.12 MODELOS DE TRANSFORMERS

En el año 2017 Vaswani et al. (2017) publican un interesante artículo titulado “Attention Is All You Need” (La atención es todo lo que se necesita) publicación en la cual se presenta un modelo que solo necesita mecanismos de atención y trabaja en paralelo sin necesidad de recurrencia, resolviendo así los problemas de las redes recurrentes. Los autores describen una arquitectura para trasformar las representaciones de entradas - secuencias (sequence-to-sequence) utilizando solo mecanismos de atención, arquitectura que denominaron transformers.

Según Vaswani et al. (2017) Los transformers poseen la siguiente arquitectura en cuatro bloques: (1) capa de embedding y codificación posicional, (2) capa de multi-head attention, (3) capa feed-forward totalmente conectada y (4) conexiones residuales y normalización de capa; en cada bloque se procesa la entrada en paralelo y se utiliza de manera eficiente la atención para capturar relaciones entre todos los elementos de la secuencia.

El modelo BERT (Bidirectional Encoder Representations from Transformers), está basado en la arquitectura transformers y fue presentando por Jacob Devlin en 2018 en el artículo titulado “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, en esta publicación Devlin et al. (2018), Se describe un preentrenamiento bidireccional enmascarando tokens y predijiendo la siguiente oración con la finalidad de capturar al mismo tiempo el contexto de una palabra a la izquierda y a la derecha.

El modelo BERT genero un importante avance en el procesamiento de lenguaje natural (PLN), ya que permitió un mejor procesamiento de la semántica, clasificación de texto, generación de respuestas y análisis de sentimientos. Como lo muestra la Figura 9, el modelo se preentrena primero con los set de datos y luego se afina (fine-tuned) para tareas específicas, el pre entrenamiento permite reutilizar conocimiento lingüístico sin entrenar modelos desde cero.

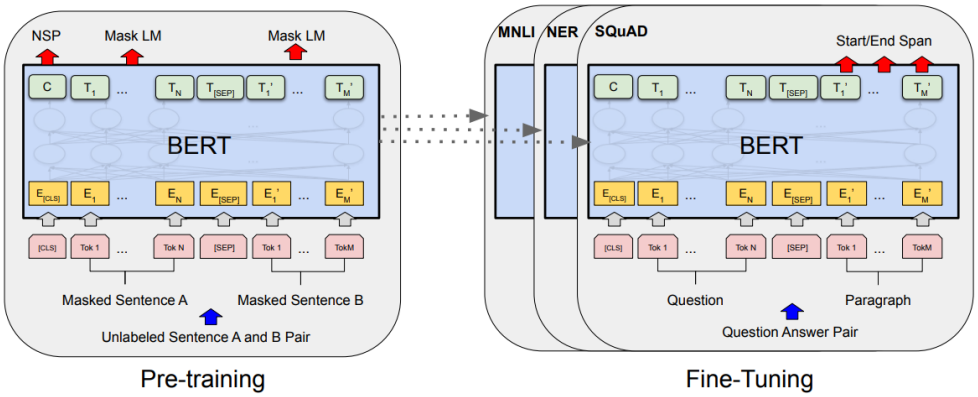


Figura 9. Preentrenamiento y afinamiento en el modelo BERT
Fuente: (Devlin et al. 2018).

Las principales aplicaciones de BERT son la clasificación de texto (sentiment analysis, spam detection), la generación automática de respuestas (chatbots, asistentes virtuales), la extracción de información (sistemas de recuperación de documentos), la traducción automática (optimización de modelos multilingües) y la generación de embeddings para representar significado de palabras en un espacio vectorial.

BERT enfrenta desafíos como la alta demanda computacional, lo que dificulta su implementación en dispositivos con recursos limitados, restricción en la longitud de entrada, ya que solo admite textos con un máximo de 512 tokens (Sun, Qiu, Xu y col ,2012); lo que obliga a segmentar documentos extensos y sobretodo presenta sesgo inherente en los datos de entrenamiento, que puede influir en la interpretación de resultados en tareas sensibles como clasificación de género o raza.

Existen investigaciones donde se realiza procesamiento de lenguaje natural (PLN) con pocos recursos de datos (Data Set pequeños), como por ejemplo el estudio presentado en la publicación “Impact of Using Bidirectional Encoder Representations from Transformers (BERT) Models for Arabic Dialogue Acts Identification.” donde los autores Gyanendra y Rajesh (2023) describen el desarrollo de un modelo para clasificación de actos de diálogo en textos en idioma Arabe usando distintas versiones de AraBERT, en donde cada versión se entrena con diferentes corpus de árabe y en donde se obtuvo

un rendimiento que anteriormente no se había obtenido con otros métodos de machine learnig y deep learnig.

4.3 MARCO ESPACIAL

El idioma indígena wayuunaiki maneja principalmente dos dialectos, los cuales son utilizados dependiendo de la ubicación geográfica de las comunidades wayuu, los habitantes de la parte superior de la península (la Alta Guajira) poseen el dialecto denominado arribero (Winpumüin) y los que viven en la parte central o inferior poseen el dialecto denominado abajero en Wopumüin, como lo muestra la Figura 10, estos dialectos no presentan una diferencia notable en la parte gramatical relacionada con el índice de la 3ª persona singular no masculino.



Figura 10. Dialectos de wayuunaiki
Fuente: (Mejía Rodríguez, 2011)

Este proyectó utilizará textos producidos en wayuunaiki arribero, teniendo en cuenta que es dialecto con la mayor cantidad de hablantes, en específico 142649 hablantes

como se indica en el visor del DANE (DANE, 2024) y en el informe del DANE de 2021 basado en datos obtenidos en el año 2021 (DANE, 2021), Datos que se muestras en la gráfica de población que habla la lengua nativa de su pueblo (Figura 11).

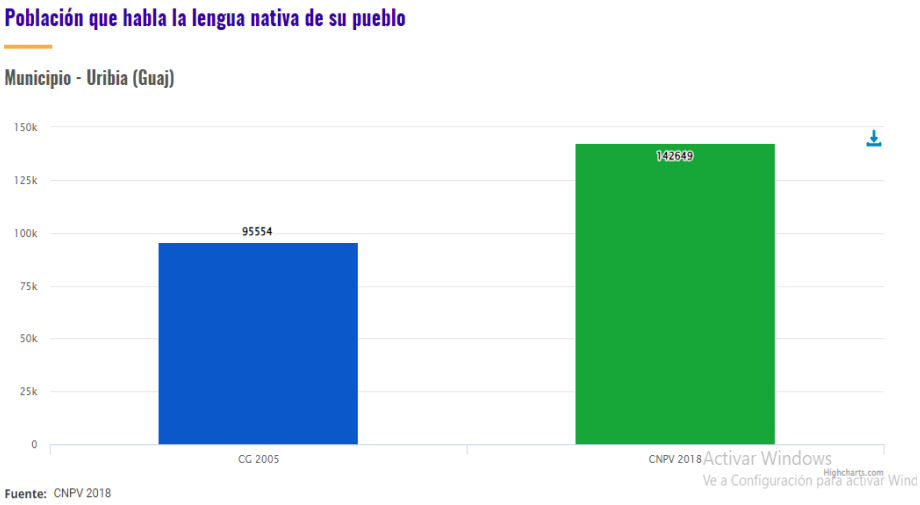


Figura 11. Población que habla la lengua nativa de su pueblo.
Fuente: (DANE, 2024)

4.4 MARCO TEMPORAL

Se tomará un marco temporal relacionado con la generación de texto en wayuunaiki entre los años 2005 y 2023 y en relación con las fuentes documentales de 2006 en adelante, ya que corresponde al mayor auge de investigación de clasificación de textos.

4.5 ALCANCE TEMÁTICO.

En relación con el alcance temático, en este proyecto se utilizaran conocimientos de gramática del idioma wayuunaiki, en específico las siguientes características gramaticales:

- Sustantivos que se expresan con un sufijo que indica el número plural (-kana) o el singular según sea del género masculino (-kai) o no masculino (-kat).
- Pronombres demostrativos, los cuales poseen una raíz que indica si es masculino singular (*chi*-), no masculino singular (*tü*-) o plural (*na*-).

- Los interrogativos.
- Los pronombres y prefijos personales.
- Verbos, sufijos e infijos.

Lo anterior con la finalidad de identificar las principales características gramaticales particulares del idioma wayuunaiki relacionadas con la aglutinación, multiplicidad de infinitivos, flexionadas de nombres y preposiciones, gradación verbal, sufijos temáticos, morfemas y yuxtaposición (Pérez van Leenden, 2020).

5. HIPÓTESIS

A partir de un modelo de clasificación automático de texto en wayuunaiki usando Deep Learning se puede identificar las principales características gramaticales del idioma indígena wayuunaiki relacionadas con la aglutinación, multiplicidad de infinitivos, flexionadas de nombres, preposiciones, gradación verbal, sufijos temáticos, morfemas y yuxtaposición.

6. METODOLOGÍA

La investigación realizada en este trabajo será de tipo descriptiva – exploratoria con enfoque experimental, teniendo en cuenta que se caracterizara la clasificación automática de frases y palabras en idioma wayuunaiki, cuyos rasgos diferenciadores serán que dichas frases y palabras pertenecen a un conjunto de datos que presentan características gramaticales, igualmente atendiendo al proceso formal de investigación, se utilizará un método hipotético - deductivo formulando una hipótesis la cual se validará de una manera empírica.

El proyecto busca alcanzar los objetivos propuestos mediante el uso de la metodología CRISP-DM ((Cross Industry Standard Process for Data Mining). Las fases de la metodología se muestran en la Figura 12.

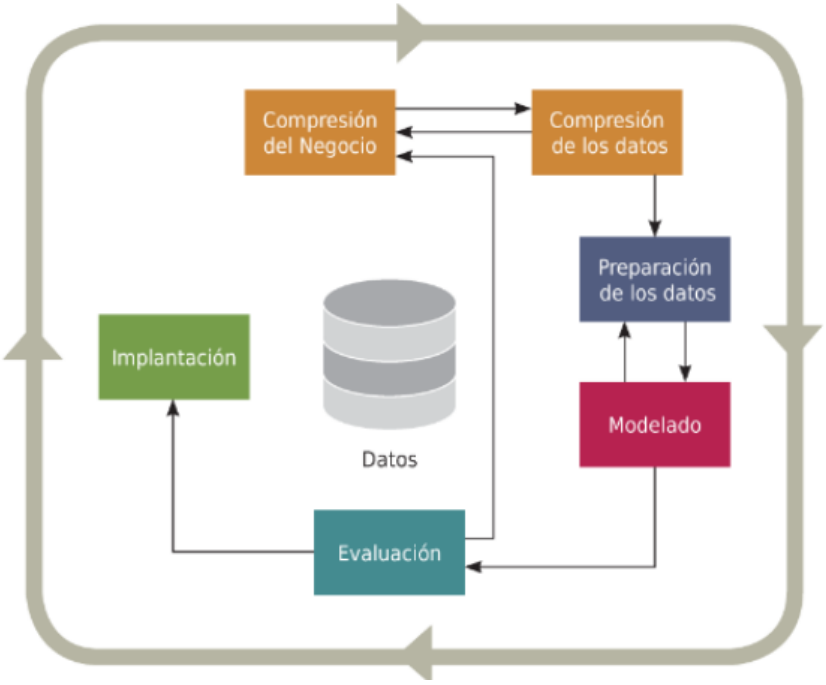


Figura 12. Fases de la metodología CRISP-DM.
Fuente: (Castillo Romero, 2019).

6.1 FASES DE LA METODOLOGÍA.

Las fases de la metodología según Gallardo (2009), Castillo Romero (2019) e IBM (2021), se describen a continuación:

6.1.1 Fase 1. Comprensión del Negocio.

En la fase 1, es crucial conocer y comprender la organización para determinar los objetivos y la pregunta adecuada para el proyecto de Procesamiento de lenguaje natural que también se pueda aplicar a la estructura de la organización. Igualmente la adecuada formulación del problema aportara de manera significativa a la comprensión de los datos y al análisis de los resultados.

Las actividades que se desarrollara en esta fase son:

- Determinar los objetivos del negocio.
- Valoración de la situación.
- Determinar los objetivos del PLN.
- Realizar el plan del proyecto.

6.1.2 Fase 2. Comprensión de los datos.

En la fase 2, en base a la hipótesis y pregunta se determinan las variables, fuente y medios de recolección de información. Se utilizará un método empírico - estadístico, donde la etapa empírica permitirá la aplicación de técnicas adecuadas para la recolección de los datos y la etapa de análisis estadísticos desarrollara enfoques descriptivos y exploratorios, con la finalidad identificar la distribución y patrones presentes en la información recopilada en el conjunto de datos que contiene frases y palabras en idioma wayuunaiki.

Las actividades que se desarrollara en esta fase son:

- Recolección de datos iniciales (listado).
- Descripción de los datos.
- Reportes de los datos (Descriptiva y exploratoria).

- Reporte de calidad de los datos.

Se obtiene principalmente el Conjunto de datos de frases y palabras en idioma wayuunaiki. (Sin preprocesamiento).

6.1.3 Fase 3. Preparación de los datos.

En la fase 3, se preparan los datos para el modelo. Los datos se seleccionan, limpian, enriquecen y normalizan.

Las actividades que se desarrollara en esta fase son:

- *Seleccionar los datos:* Se seleccionan y/o incluyen los datos que aporten a la tarea de PLN a realizarse y se eliminan los datos que generen ruido.
- *Limpiar los datos (Preprocesamiento de datos):* Para garantizar un alto grado de confiabilidad del modelo, se aplican procesos de normalización y estandarización para mejorar la calidad de los datos.
- *Construcción de nuevos datos:* los nuevos datos de pueden integrar mediante la adición de columnas (Derivación de atributos) o filas (Otros registros).
- *Integrar los datos:* En esta actividad se pueden integrar los datos mediante la estrategia de fusión de datos, que combina conjuntos con registros similares pero atributos diferentes, adicionando columnas; o mediante la estrategia de adición de datos, que integra conjuntos con atributos similares pero registros diferentes, aumentando las filas del conjunto de datos.
- Seleccionar las características relevantes para el modelo. (Definición final de variables).
- Formateo de los datos (Representación del Texto): En esta actividad se realiza la tokenización (dividir el texto en palabras o subunidades) y se propone un modelo para generar vectores numéricos.

De esta fase se obtiene el Conjunto de datos de frases y palabras en idioma wayuunaiki (preprocesado) y su representación en vectores numéricos de palabras.

6.1.4 Fase 4. Modelado

En la fase 4, se realiza principalmente la determinación y el diseño el modelo para resolver el problema planteado, realizando las siguientes actividades:

6.1.4.1 Determinación del modelo:

Se utiliza un método hipotético-deductivo para realizar las predicciones de una primera aproximación del modelo. También en esta fase se utiliza el método de modelación con la finalidad de obtener una extracción del problema.

Actividades:

- Teniendo en cuenta el objetivo principal se establece un modelo conceptual de clasificación de texto que permita incluir las características gramaticales del wayuunaiki.
- Se comprará el funcionamiento de Random Forest y diferentes arquitecturas de red neuronal, realizando experimentos con el conjunto de datos de frases en wayuunaiki.

De esta actividad se obtiene el modelo conceptual y la arquitectura a utilizar.

6.1.4.2 Proyección del plan de pruebas:

En el plan de pruebas se establece las actividades a realizar para comprobar el funcionamiento del modelo (método de validación).

6.1.4.3 Diseño del modelo.

Se utiliza el método de modelación para obtener la estructura del modelo de clasificación de texto en idioma wayuunaiki, idioma que posee características gramaticales particulares. Realizando las siguientes actividades:

- El diseño de la estructura, el cual se genera en base al modelo conceptual, utilizando la arquitectura de red neuronal elegida en la fase de determinación del modelo de clasificación.
- Utilizando el método empírico de experimentación y medición de variables, se entrena el modelo de clasificación y se verifica el rendimiento esperado del modelo. Para esta actividad de entrenamiento del modelo se realizan las siguientes actividades:
  - Se definen los porcentajes para entrenamiento y validación.
  - Se realizará el ajuste de hiperparámetros y el establecimiento de la función de pérdida y las mejoras apropiadas.
  - Se revisa la arquitectura, hiperparámetros y técnicas de regularización con la finalidad de realizar ajustes.

Como productos se obtienen la arquitectura del modelo y las estadísticas de rendimiento utilizando el conjunto de datos de entrenamiento.

6.1.4.4 Validación del modelo.

Esta fase se utilizará un método empírico en específico de experimentación y medición de variables, con la finalidad de demostrar la veracidad de la hipótesis planteada sobre el modelo de clasificación automático de texto en wayuunaiki que usando Deep Learning pueda identificar las principales características gramaticales del idioma indígena wayuunaiki.

En esta fase se realizarán las siguientes actividades:

- Se utiliza el porcentaje del conjunto de datos para la validación.
- Se miden las métricas de determinadas en la fase de determinación del modelo

En esta fase se obtendrán las estadísticas de rendimiento del modelo de clasificación de texto en wayuunaiki utilizando el conjunto de datos de validación y se verificara su funcionamiento utilizando los datos del conjunto de datos sin etiquetar para pruebas.

6.1.5 Fase 5. Análisis de los resultados.

En la fase 5, se utiliza el método análisis de los resultados particulares de los diferentes experimentos con la finalidad de llegar a una síntesis de los diferentes resultados.

En esta fase se realizarán las siguientes actividades:

- Se verifican los resultados obtenidos, estos deben ser los esperados, deben tener coherencia.
- Se deben comparar con los resultados de los análisis estadísticos, los resultados de validación
- Se identifican los trabajos futuros resultado de la investigación.

Como producto de esta fase se obtendrá el análisis de los resultados del modelo.

6.1.6 Fase 6. Implantación.

En la fase 6, se describirán algunas nociones para implementar el modelo en un entorno para su utilización y se realizan las siguientes actividades:

Plan de Implementación: Para garantizar que el modelo esté disponible para los usuarios que puedan aprovechar los resultados, se garantiza la accesibilidad al modelo guardándolo en formato (.mat) para futuras ejecuciones y se genera una función para recibir texto y devolver la clasificación.

Plan de monitoreo y mantención: Se debe establecer las estrategias para monitorear el modelo, como la evaluación de precisión con datos nuevos sin etiquetar y la comparación de métricas.

Informe final: Se generan las conclusiones del proyecto, principales aportes, recomendaciones y trabajos futuros.

7. DESARROLLO

7.1 FASE 1. COMPRENSIÓN DEL NEGOCIO.

El wayuunaiki es la lengua materna de la comunidad indígena Wayuu y aunque el wayuunaiki es considerado un idioma altamente utilizado, este idioma se encuentra en peligro de desaparecer. Las acciones para la conservación del wayuunaiki están enfocadas a su preservación, análisis y accesibilidad, pero el desarrollo de modelos computacionales que apoyen estas acciones presenta varios desafíos, como la escasez de datos de calidad, variabilidad dialectal y características gramaticales particulares del idioma.

El procesamiento de lenguaje natural (PLN), es un gran progreso en la manera en que se procesa e interactúa con la información en diferentes idiomas y su constante perfeccionamiento ha permitido realizar significativos avances en diferentes tareas relacionadas con el procesamiento de lenguaje natural, como por ejemplo la tarea de clasificación automática de texto, en la cual se realizan de manera óptima tares de ordenamiento, etiquetado y agrupación de texto, pero en el caso del idioma wayuunaiki, son muy pocos los estudios desarrollados. Por tanto, se genera la pregunta ¿Cómo identificar en textos escritos en idioma Wayuunaiki las características gramaticales, para poder realizar una adecuada clasificación de texto?, pregunta que debe resolverse teniendo en cuenta que no existen conjuntos de datos de gran tamaño, calidad y de acceso abierto.

7.1.1 Objetivos del negocio.

- Creación de conjuntos de datos de alta calidad que apoye la conservación de textos en idioma wayuunaiki.
- Desarrollo de modelos computacionales que aporten a la preservación, análisis y accesibilidad a textos en idioma wayuunaiki.

7.1.2 Valoración de la situación.

El modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales puede presentar las siguientes limitaciones:

- Limitación en la obtención de información estadística actualizada del acceso a computadores en específico de las comunidades Wayuu, pero esto no influirá en el alcance del proyecto.
- Complejidad morfología y fonéticas del idioma puede generar dificultades en la identificación de patrones y la generalización del clasificador automático.

7.1.3 Determinar los objetivos del PLN.

- Facilitar la integración del wayuunaiki en aplicaciones de PLN, comunicación y educación, entre otras.
- Aportar a otras investigaciones de estudios lingüísticos y culturales que requieran usar el análisis de texto.

7.1.4 Realizar el plan del proyecto.

EL plan de proyecto se realizara de acuerdo a las actividades y productos establecidos en la metodología. El cronograma de actividades se muestra en la Figura 13.

Actividad	2024												2025				Producto
	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Enero	Febrero	Marzo	Abril	
Preparación de los datos.																	–
Recolección de datos.																	–
Creación del conjunto de datos																	Conjunto de datos 31270 frases
Procesamiento de datos.																	–
Eliminación de caracteres																	Conjunto de datos 3000 frases
Remplazo de palabras																	
Se aplica la lematización																	
Representación del Texto.																	–
Conversión numérica																	Vectores de palabras
Determinación del modelo.																	–
Análisis estadísticos.																	–
Diseño del modelo conceptual.																	Modelo conceptual
Comparación de arquitecturas.																	
Diseño del modelo.																	–
Estructuración del modelo																	Diseño de la estructura del modelo
Entrenamiento del modelo																	Modelo entrenado
Ajustes																	Estadísticas de rendimiento.
Revisión de arquitectura																	Estadísticas de entrenamiento.
Validación del modelo.																	–
Prueba con el Set de datos de validación.																	Estadísticas de rendimiento
Medición métricas																	Graficas de resultados.
Análisis de los resultados.																	–
Verificación																	Análisis del modelo, conclusiones y trabajo futuro.
Comparación																	
Identificación trabajo futuro.																	

Figura 13. Plan de actividades.
Fuente: Elaboración propia.

Como conclusión parcial de esta fase, se puede indicar que un adecuado análisis del negocio permite establecer un marco sólido para el desarrollo del modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales, asegurando que las decisiones técnicas estén alineadas con los objetivos lingüísticos y culturales del proyecto.

7.2 FASE 2. COMPRENSIÓN DE LOS DATOS.

7.2.1 Preparación de los datos.

En la fase de preparación de datos se describe la creación y el análisis exploratorio de un conjunto de datos compuesto por frases en idioma Wayuunaiki, realizado para determinar la adecuada construcción y tratamiento de los datos para el procesamiento de lenguaje natural*. Para el análisis exploratorio, se propone una metodológica específica para la construcción de conjuntos de datos en idioma Wayuunaiki, implementando la preparación de datos para construcción de modelos propuesta por Provost y Fawcett (2013), complementada con lo descrito en la página de la IABAC (International Association of Business Analytics Certifications) (2023). A continuación de describe la propuesta metodológica compuesta por las siguientes etapas:

7.2.2 Definición del tipo de tareas.

En la primera etapa se estableció que la construcción del conjunto de datos en idioma Wayuunaiki está planificado para utilizarse en tareas de Procesamiento de Lenguaje Natural (PLN), en específico para esta investigación en la tarea de clasificación de texto en la cual se incorporen las particularidades gramaticales del idioma Wayuunaiki.

7.2.3 Definición de los atributos del conjunto de datos.

En la Segunda etapa, se definió que el conjunto de datos contará con los siguientes atributos:

Tabla 4. Atributos del conjunto de datos.

Nombre de la variable:	Idfrase	Wayuunaiki phrase	Grammar	Cgwayuunaiki
Tipo de dato:	Numérico	texto	texto	texto
Valores posibles:	Número entero único	Frases o palabras en idioma Wayuunaiki.	Categórico	Categórico
Restricciones:	Sin duplicados, solo números.	Longitud máxima de 30 palabras.	Solo características principales.	Solo una categoría gramatical.

7.2.4 Reportes de los datos.

En esta fase se presentan las predicciones de una primera aproximación del modelo y también se muestran los resultados del método de modelación con la finalidad de obtener una extracción del problema de clasificación de texto en idioma Wayuunaiki.

7.2.5 Análisis estadístico

Para comprender mejor la estructura de los datos en idioma Wayuunaiki, detectar datos atípicos, datos faltantes patrones y/o las tendencias de las características gramaticales del idioma, se realizó un análisis exploratorio de datos (EDA), el cual incluye una visualización de la distribución de los datos, correlaciones y métricas adecuadas para el cálculo de estadísticas.

7.2.5.1 División del conjunto de datos para el análisis exploratorio:

Para la división del conjunto de datos, se tomaron los resultados de la etapa de preprocesamiento y se dividió el conjunto en dos: Conjunto A y Conjunto B. De acuerdo a lo expuesto por Radford et al. (2020), para en análisis exploratorio de los datos se asignaron los siguientes porcentajes: Para el conjunto A un (70%) y para el conjunto B (30%). Posteriormente se realiza un análisis para evaluar si la cantidad de los datos frente al rendimiento son adecuados.

7.2.5.2 Validación y Verificación del conjunto de datos para el análisis exploratorio:

La validación y verificación del conjunto de datos, se realizó siguiendo las recomendaciones de la investigación de Abedjan, Golab,y Naumann (2015) , en la cual se analiza el estado actual de las herramientas de detección de errores en los conjuntos de datos y muestran estrategias para mejorar la eficacia de las herramientas. Puntualmente, para el conjunto de datos en idioma wayuunaiki se verifico que las variables representen en gran medida las principales características gramaticales del wayuunaiki, esto con la finalidad de mejorar el rendimiento del modelo.

7.2.5.3 Resultados del análisis exploratorio de datos en idioma Wayuunaiki

En esta investigación para la realización del análisis exploratorio del conjunto de datos en idioma Wayuunaiki, se utilizó MATLAB, herramienta que posee librerías con funciones que facilitan la exportación de los datos, la generación de resúmenes estadísticos, la visualización de los datos, detección de datos atípicos, entre otras funcionalidades las cuales se describen en The MathWorks Inc. (2024 - b).

En este trabajo de investigación, se realizó el análisis exploratorio de un conjunto de datos en idioma Wayuunaiki que contiene 3127 filas, conjunto que se exporto a MATLAB en formato CVS mediante la función *readtable;* también se utilizó la función *head* para verificar la estructura de los datos y la función *summary* para verificar los valores de las variables. Los resultados obtenidos se presentan a continuación:

```
Variables:

    Idfrase: 3127 ×1 double

        Properties:
            Description:  Id_Frase
        Values:
            Min               1
            Min               1
            Median         1564
            Max            3127
            NumMissing     3127

    WayuunaikiPhrase: 3127×1 cell array of character vectors

        Properties:
            Description:  Wayuunaikiphrase
    Grammar: 3127×1 cell array of character vectors

        Properties:
            Description:  Grammar
    GgWayuunaiki: 3127×1 cell array of character vectors

        Properties:
            Description: Cgwayuunaiki
```

Para la fase de transformación de los datos en una representación numérica, el análisis exploratorio muestra información adicional para mejorar los valores de dimensionalidad en la técnica Word Embeddings denominada Word2Vec.

Para la mejor visualización de la distribución de los datos se generaron los siguientes gráficos**:**

Histograma de longitud de frases: En la Figura 14, se presenta un histograma con la distribución de la longitud de las frases en idioma Wayuunaiki, en el cual se observa que 389 frases tienen entre 1 y 3 palabras, 298 frases contienen entre 3 y 5 palabras, 16 frases contienen entre 13 y 15 palabras, 4 frases tienen entre 18 a 20 palabras, mientras que solo 1 frase contiene de 25 a 27 palabras. A partir de estos datos, se puede concluir que la mayoría de las frases en el conjunto de datos tienen un máximo de 5 palabras, y que la longitud más extensa registrada alcanza 27 palabras.

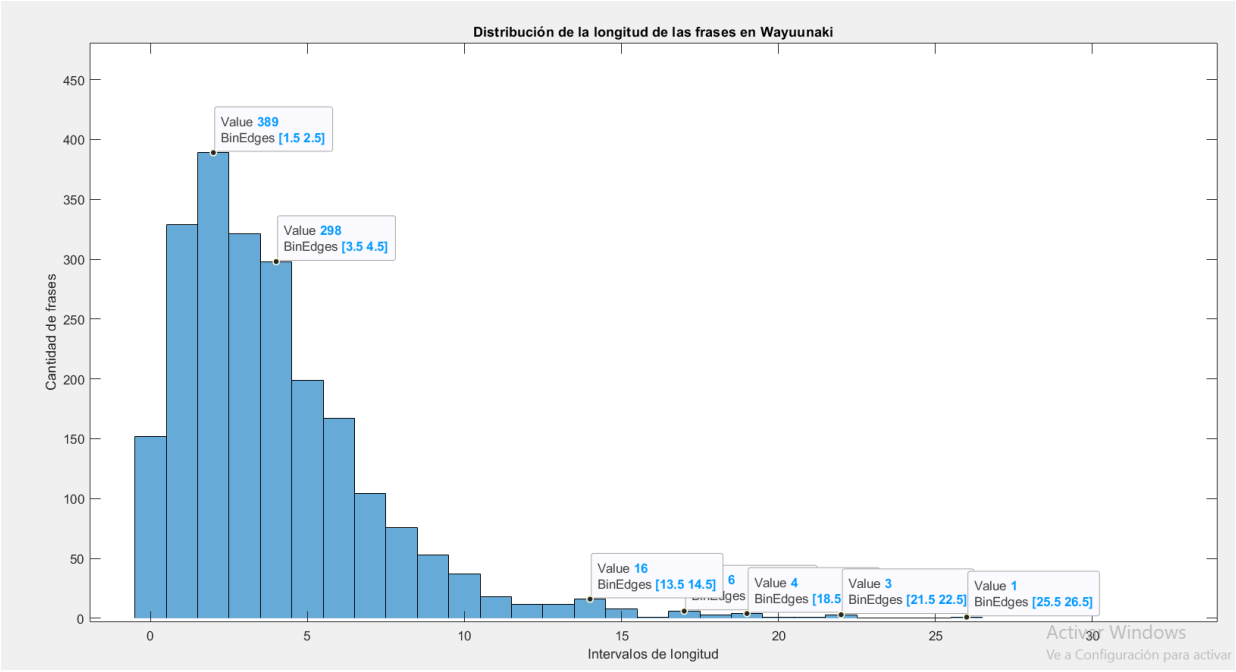


Figura 14. Distribución de la longitud de las frases en Wayuunaiki.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB.

En la Figura 14 también se observa que La distribución de los datos parece estar sesgada hacia la derecha, lo que indica que una distribución normal no parece adecuada para este conjunto de datos y que se debe considerar utilizar modelos robustos a la variabilidad como las redes neuronales o Random Forest.

Diagramas de caja: En la Figura 15, el diagrama de cajas (boxplot) muestra los cuartiles 0,25, 0,5 y 0,75. La cola superior larga y los signos más indican la falta de simetría en los valores de los datos. A partir de estos resultados se confirma que para un conjunto de datos en idioma Wayuunaiki se deben utilizar modelos robustos a la variabilidad.

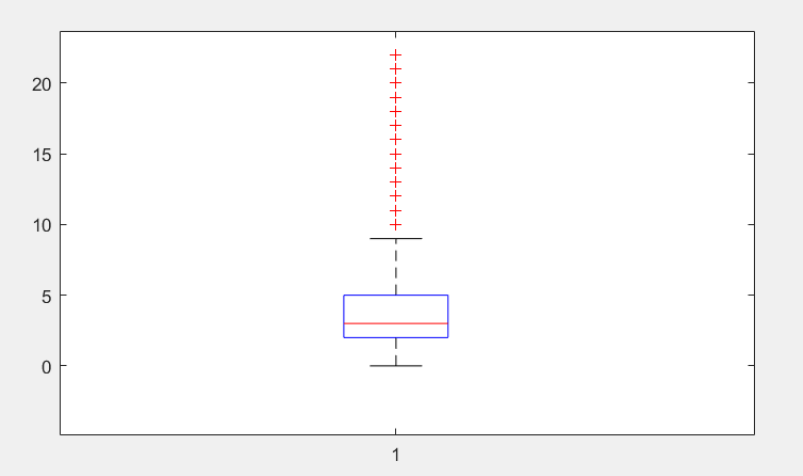


Figura 15. Datos atípicos en las frases en Wayuunaiki.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB.

En este sentido, dado que el conjunto de datos no presenta una distribución normal, es importante considerar enfoques estadísticos alternativos para su análisis como la matriz de similitud de coseno, lo anterior teniendo en cuenta que no se pueden normalizar los datos debido a que los datos que se consideran como atípicos aportan a la clasificación del idioma Wayuunaiki.

Nubes de palabras: Las nubes de palabras en este análisis exploratorio se utilizaron para identificar las palabras en Wayuunaiki que aparecen con más frecuencia y que pueden ser un identificador de característica gramatical de la frase.

Para la generación de las nubes de palabras, el conjunto de datos se dividió en dos grupos. El conjunto A, compuesto por 2189 frases y el conjunto B, conformado por 938 frases. En la Figura 16. se observa la nube de palabras generada con los datos del conjunto A, la cual muestra palabras relacionadas con nombres y prefijos personales masculinos y femeninos como por la palabra en Wayuunaiki “taya”, que corresponde a la primera persona singular; la palabra “pia”, que pertenece a la segunda persona del singular, la palabra “Shia” que hace referencia a la tercera persona no masculino singular, la palabra “kasa” que corresponde a una pregunta y por tanto indica interrogación; la palabra “wanee” que hace referencia a algo indeterminado (artículo indefinido) y palabras relacionadas con los verbos más usados en la novela del Principito, como hablar, responder, preguntar, tener, dibujar, etc. (Esto depende del contexto del texto).



Figura 16. Nube de palabras de frases en Wayuunaiki resultado del experimento con el conjunto de datos A.
Fuente: (Salazar y Rodríguez, 2024).

En la Figura 17, se presenta la nube de palabras generada a partir de los datos del conjunto B, la cual refleja resultados similares a los obtenidos con el conjunto A. Destacan palabras en Wayuunaiki como wanne, atunk, ju, chi, nia, shia, taya entre otras, que correspondientes a los nombres y prefijos personales masculinos y femeninos, así como a los verbos más utilizados en los textos, dependiendo del contexto lingüístico.

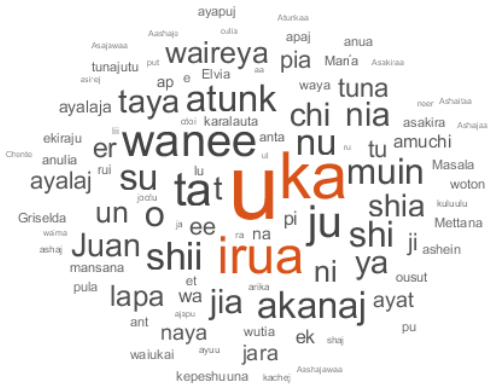


Figura 17. Nube de palabras de las frases en Wayuunaiki prueba conjunto de datos B. Fuente: (Salazar y Rodríguez, 2024).

Los principales hallazgos del análisis exploratorio del conjunto de datos en idioma Wayuunaiki, se describen a continuación:

- El idioma Wayuunaiki presenta una estructura gramatical rica y elaborada que posee varias variables relacionadas con las características gramaticales.
- Para la construcción de oraciones en idioma Wayuunaiki, las características como los sufijos determinativos y anunciativos, los pronombres y prefijos personales, los interrogativos, adverbios y conjunciones desempeñan un papel fundamental.
- Se identificó que en el idioma Wayuunaiki no es muy frecuente que se presente la yuxtaposición, pero es importante identificar en las frases cuando se solapan oraciones.
- En el idioma Wayuunaiki las palabras se formar mediante la combinación de morfemas (Unidades más pequeñas de significado). Esta característica hace que el idioma sea altamente aglutinante, generando la formación de frases complejas a partir de la unión de varios elementos lingüísticos.
- En el idioma Wayuunaiki una palabra puede tener un significado diferente de acuerdo a como se pronuncie (tono), por tanto, es importante identificar las palabras que tengan la misma estructura pero que su significado cambia de acuerdo al contexto de la frase.

- En el conjunto de datos se deben repetir los registros de palabras que tengan diferentes significados y que representen diferentes características gramaticales.

7.2.6 Reporte de calidad de recolección de datos.

El conjunto de datos se elaboró con información recolectada de los diccionarios, manuales, artículos y libros revisados y recomendados en trabajos de análisis lexicográficos del idioma Wayuunaiki propuestos en el trabajo de investigación de Arroyave Ferreiro (2023) y también se recolectaron datos de los siguientes documentos: El Diccionario Básico Ilustrado wayuunaiki-español Captain y Captain (2005), El artículo de Ramírez González sobre Los pronombres personales en Wayuunaiki Ramírez González y Dieck Novial (2019), el Compendio de la gramática de la lengua wayuu Álvarez González, J. R. (2017), elaborado por el profesor José Álvarez y la mayor parte de los datos se tomó de la traducción del Principito (Pürinsipechonkai: Una traducción de “El principito” de Antoine de Saint-Exupéry a la lengua wayuu) Saint-Exupéry (2016), traducción elaborada por el profesor José Ramón Álvarez González. El profesor Álvarez tiene una significativa trayectoria en materia de Lingüística del idioma Wayuunaiki.

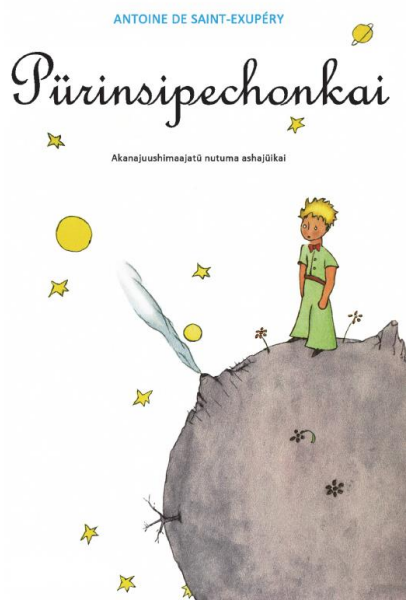


Figura 18, portada del principito traducido al wayuunaiki.
Fuente: Saint-Exupéry (2016, trad. Álvarez, Dorado, Fernández & Montiel).

En esta etapa, se identificó la relevancia de contar con un hablante nativo del idioma Wayuunaiki, acompañado por un traductor oficial en lo posible certificado, para llevar a cabo la revisión de los datos (frases y palabras). El acompañamiento del traductor y del hablante nativo garantiza que el conjunto de datos conserve la estructura gramatical del idioma Wayuunaiki y además respete la identidad cultural y la tradición oral de la comunidad indígena Wayuu, en línea con lo señalado en Mejía Rodríguez (2011).

Como conclusión parcial de esta fase, gracias al análisis estadístico se obtienen las principales características para la creación del conjunto de datos, el cual debe estar compuesto por frases y palabras en wayuunaiki, datos que deben estar etiquetados de acuerdo con las características gramaticales del wayuunaiki, relacionadas con la aglutinación, multiplicidad de infinitivos, flexionadas de nombres y preposiciones, gradación verbal, sufijos temáticos, morfemas, yuxtaposición. Al finalizar la recolección de los datos, se obtuvo un conjunto de datos compuesto por 3850 frases, etiquetadas de acuerdo a las categorías gramaticales definidas.

7.3 FASE 3. PREPARACIÓN DE LOS DATOS.

7.3.1 Seleccionar los datos.

El conjunto de datos obtenido sin estructurar, se almacena en la plataforma de Azure, en un servidor SQL creado para este proyecto denominado wayuu2025 y en una base de datos MySQL denominada *ayonnajaabd* (Figura 19, 20 y 21). Esta base de datos se crea con el propósito de almacenar los datos obtenidos en idioma wayuunaiki de una manera más estructurada, segura, escalable y accesible.

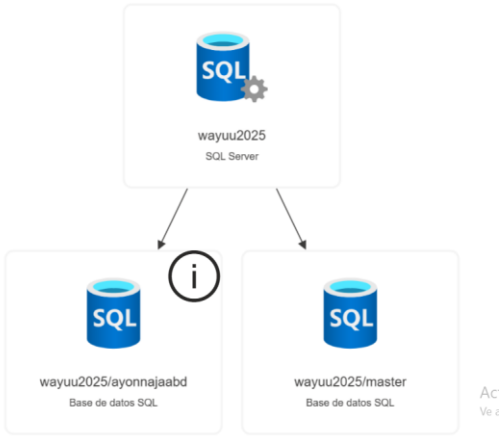


Figura 19. Base de datos ayonnajaabd.
Fuente: La imagen fue generada de la implementación de la Base de Datos en Azure.

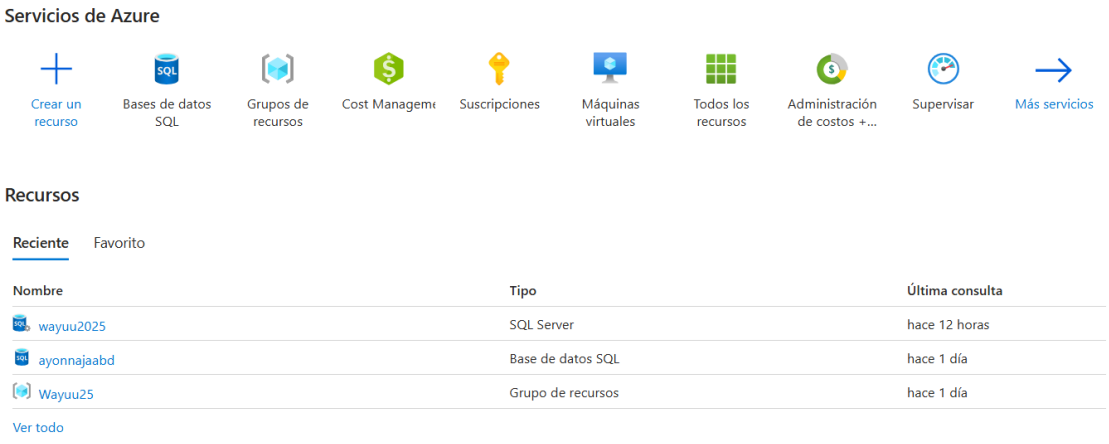


Figura 20. Recursos Azure para el proyecto.
Fuente: La imagen fue generada de la implementación de la Base de Datos en Azure.

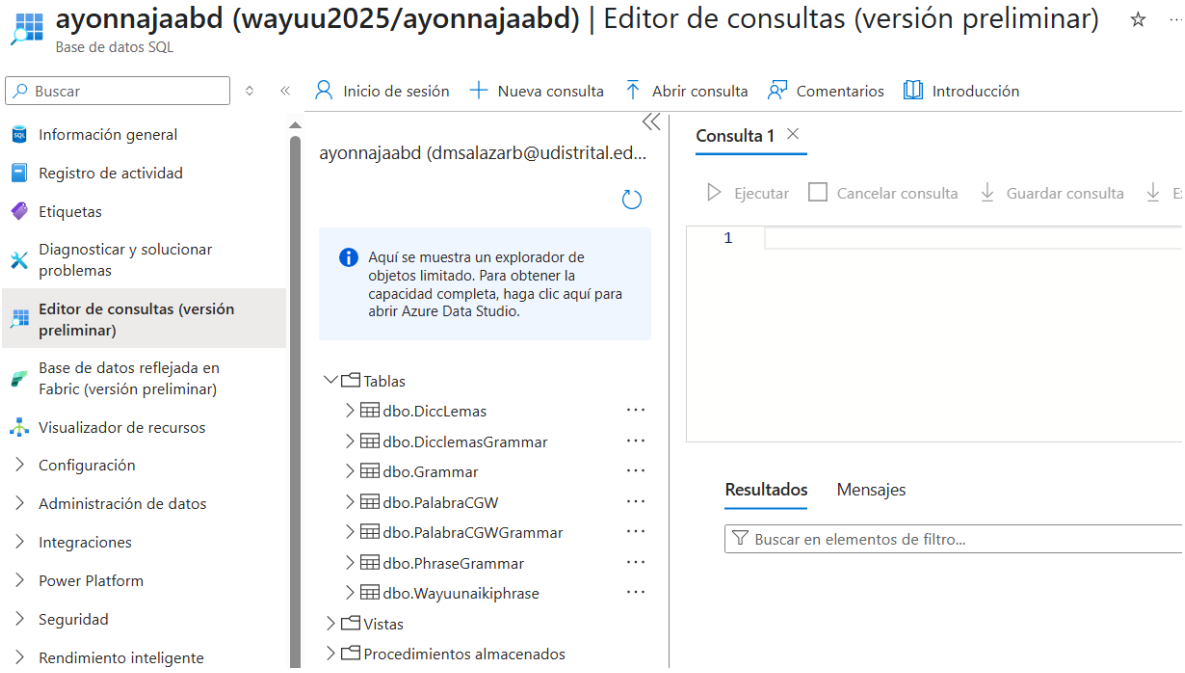


Figura 21. Estructura de la Base de datos en Azure para el proyecto.
Fuente: La imagen fue generada de la implementación de la Base de Datos en Azure.

7.3.2 Preprocesamiento de datos.

En esta etapa, se describirá el preprocesamiento de los datos utilizando la plataforma de programación y cálculo numérico MATLAB (MATrix LABoratory).

Limpieza de Datos: Utilizando MATLAB se realizó la limpieza de los datos recolectados en idioma Wayuunaiki, corrigiendo los datos erróneos, duplicados que no aportan a la gramática del idioma o irrelevantes, de acuerdo a lo explicado por Abedjan, Golab,y Naumann (2015). En específico para el idioma Wayuunaiki se llevó a cabo el siguiente tratamiento de datos:

- **Caracteres especiales:** Utilizando la función regexprep de MATLAB, se eliminaron los caracteres especiales como signos ortográfico (« » ·), guiones (-), comillas dobles (“”) y guiones largos (—).

- **Remplazo de caracteres especiales:** Modificando la función regexprep, se remplazó la u con diéresis (Ü) por la letra (x) la cual no está en el alfabeto del idioma Wayuunaiki, la tilde (´) se remplazó por la comilla sencilla (‘) y aunque el acento no se presenta comúnmente en el idioma Wayuunaiki, las letras con acento (á, é, í, ó, ú) se reemplazaron por letras sin acento (a, e, i, o, u).

- **Conversión a minúsculas:** Utilizando la función newStr = lower(str), se realizó la conversión de palabras en mayúscula a minúscula.

- **Proceso de lematización:** El proceso de lematización que se muestra en la Figura No. 22 consistió en crear un diccionario de lemas en Wayuunaiki (frases comunes) modificando el diccionario (lemmaDict), tokenizar el texto y utilizar la función normalizeWords de MATLAB para convertir las palabras en Wayuunaiki en sus formas raíz con la finalidad de identificar por ejemplo la multiplicidad de infinitivos, la cual presenta un gran reto para la creación del diccionario y para la identificación de la forma raíz, debido a que los verbos pueden presentar diversas formas infinitivas dependiendo de los morfemas que los acompañan según lo descrito por (Álvarez, 2008).

- Un ejemplo de la lematización en Wayuunaiki sería la gradación verbal de sufijos temáticos como: G1: e-jime-ja-a, G2: e-jime-ja-a, G3: e-jime-ja-a-waa, glosa: ‘pescar’. Salazar y Rodríguez (2024).

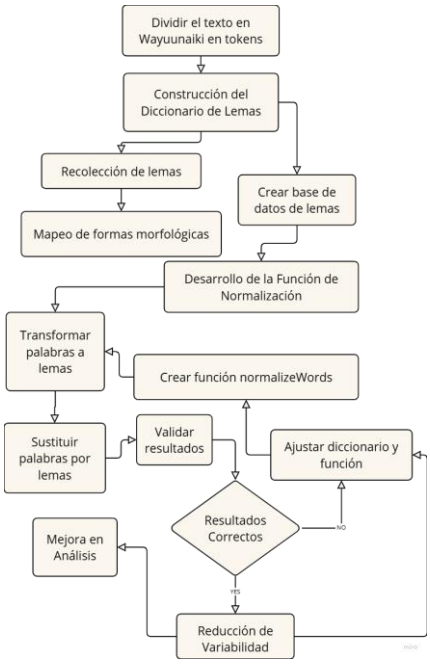


Figura 22. Funcionamiento de la lematización.
Fuente: Fuente: (Salazar y Rodríguez, 2024).

7.3.3 Representación del Texto.

Transformación de Datos: En esta etapa se realizó la trasformación de las palabras en idioma Wayuunaiki a representaciones numéricas (vectores de palabras), utilizando un tipo de vectorización denominado Word Embeddings y en específico la técnica Word2Vec como lo muestra la Figura 23 y según lo explicado por Jatnika, Bijaksana y Ardiyanti (2019) se realizó la evaluación de similitudes semánticas en palabras Wayuunaiki, generando diagramas de calor (heatmaps).

El funcionamiento del modelo Word2Vec para la vectorización de palabras en Wayuunaiki se describe a continuación:

- Se divide el texto en palabras o subunidades (Tokenización).
- Utilización complementaria de dos arquitecturas para entrenar el modelo: Continuous Bag of Words (CBOW) en el cual se predice la palabra actual en función de su

característica gramatical y el Skip-Gram en el cual se predice las palabras de la característica gramatical dada la palabra actual.

- Transformación de datos numéricos -Vectorización de los datos.
- Codificación de variables categóricas.
- Extracción de los vectores de palabras del modelo entrenado.
- Reducción de dimensionalidad con PCA (Análisis de Componentes Principales) y con SNE (t-distributed Stochastic Neighbor Embedding) para identificar estructuras y agrupaciones de las palabras en Wayuunaiki.
- Generación de diagramas heatmap.
-

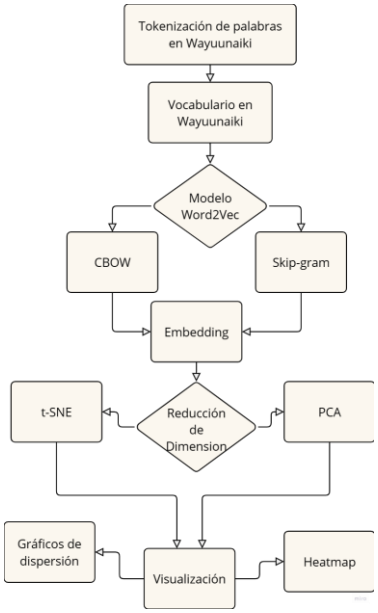


Figura 23. Funcionamiento de Word2Vec.
Fuente: (Salazar y Rodríguez, 2024).

Los diagramas de calor (heatmap) obtenidos, se utilizan para identificar y analizar la similitud entre palabras en wayuunaiki usando una matriz de similitud de coseno. Estos gráficos permiten identificar relaciones significativas en el Wayuunaiki proporcionando información importante sobre contextos similares, significado o uso similar, anomalías en palabras, deducir temáticas e información para ajustar la calidad de los embeddings. Un ejemplo de diagrama obtenido para la clase PÜ se presenta en la Figura 24.

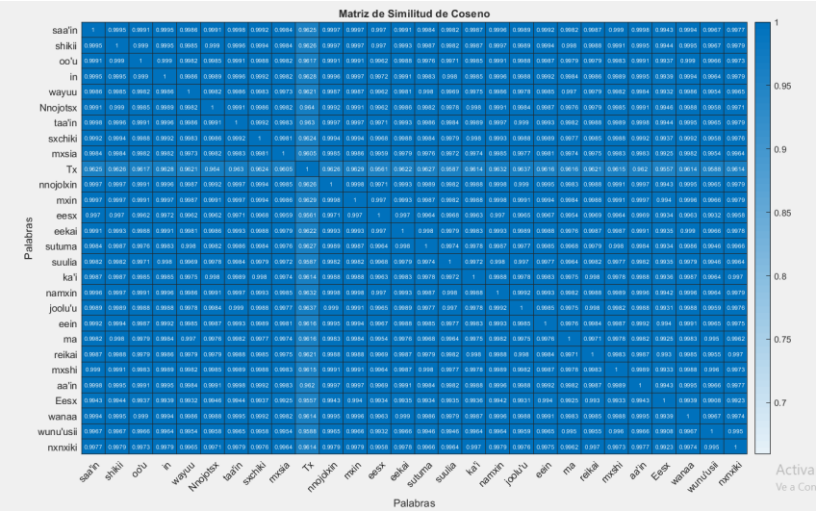


Figura 24. Matriz de similitud de coseno de datos en Wayuunaiki. Fuente: (Salazar y Rodríguez, 2024).

7.3.4 Enriquecimiento del conjunto de datos:

En esta etapa para el enriquecimiento del conjunto de datos, se incluyen las características gramaticales que aporten a la clasificación del texto, identificadas en el análisis exploratorio.

7.3.5 Definición de variables:

Tras realizar un exhaustivo análisis de correlación entre los distintos factores, la estadística exploratoria de los datos y evaluar la frecuencia de utilización de las variables en estudios afines, se lleva a cabo una selección preliminar de variables. En este proceso, se excluyen aquellas que presentan una afinidad baja con el rendimiento o que podrían obstaculizar el desempeño general del modelo predictivo. Las variables preseleccionadas son:

Tabla 5. Variables del modelo.

Idfrase	**Wayuunaikiphrase**	**Grammar**	**Cgwayuunaiki**
Numérico	texto	texto	texto
Es un valor único asignado a cada registro (identificador).	Representa las frases y palabras escritas en idioma wayuunaiki.	Representa la descripción de la característica gramatical más relevante en la oración.	Representa la letra y/o palabra en wayuunaiki relacionada con la característica gramatical.

7.3.1 Definición de clases.

En el análisis estadístico realizado se identificaron las clases más relevantes que representa una gran parte de las características gramaticales más representativas del idioma Wayuunaiki, las cuales se presentan en la tabla 6 y en la Figura 25.

Tabla 6. Clases del modelo

Gramática en español	Gramática en Wayuunaiki
sufijo_ determinativo_MS	–kai
sufijo_ determinativo_FS	–kalü
sufijo_ determinativo_P	–kana
sufijo_anunciativo_MS	chi–
sufijo_anunciativo_FS	tü–
sufijo_anunciativo_P	na–
Interrogativo	
Interrogativo	jarai, jaralü, jaralii
Interrogativo	kasa
Interrogativo	je'raa
Interrogativo	jalaa yaa
Interrogativo	jalapüshii
Interrogativo	jouja
Interrogativo	jamaa
Interrogativo	je'tshii
Interrogativo	jana
Pronombres personales	
1a Singular (Pro1PS)	taya
2a Singular (Pro2PS)	pia
3a Singular Masculina (Pro3PSM)	nia
3a Singular Femenina (Pro3PSF)	shia/jia
1a Plural (Pro1PP)	waya
2a Plural (Pro2PP)	aya/jia
3a Plural (Pro3PP)	naya
Prefijos personales	
1a Singular (Pre1PS)	ta-
2a Singular (Pre2PS)	pü-
3a Singular Masculina (Pre3PSM)	nü-
3a Singular Femenina (Pre3PSM)	sü-/jü-
1a Plural (Pre1PP)	wa-
2a Plural (Pre2PP)	ja-/jü-
3a Plural (Pre3PP)	na-
Infijos	

Gramática en español	Gramática en Wayuunaiki
Adjetivo	jiatta'a
Adverbio	Ej. aipa'a , akatsa'a, joolu'u …
Aglutinación	Ej. joolu'ule'ella, Aja'ttineematakaja'a …
conjunción	Ej. akatsa'a …
Indeterminado	wanee
Infijo	Ej. chon
Interjección	Ej. Atee, Tajak, Ajaa…
Multiplicidad PC1	Ej. aa → a, ee → e, ii → i …
Multiplicidad SC2	waa
Sufijo	Ej. shaana
Sustantivo	Ej ka'i, kotpo, Jülatapa …
Verbo	Ej. Wane'ella, a'yalajüshi, anteechi…
Verbo_Aux_no	nnojoluu

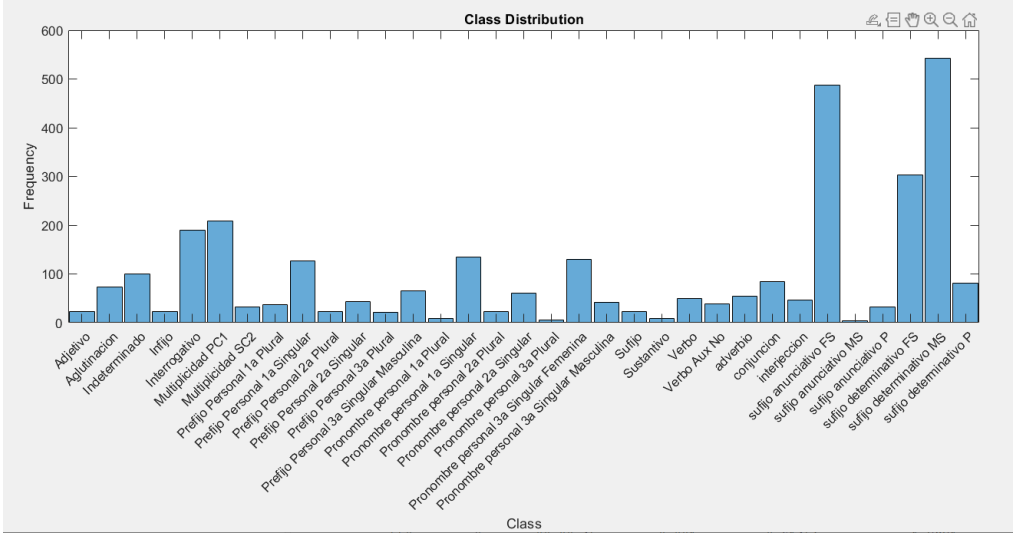


Figura 25. Distribución de clases del modelo.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante el análisis exploratorio en MATLAB.
Fuente: Elaboración propia.

7.3.1 Conjunto de datos final.

Para este proyecto luego de utilizar Fine-tunnig y embedings preentrenados se obtuvo que el conjunto de datos debería contener alrededor de 300 a 500 frases por clase, como se muestra a continuación:

En primer lugar se realizó la siguiente estimación: 120,000 (Cantidad de parámetros estimados) / 33 (clases) ≈ 3,636 parámetros por clase.

Entonces:
3,636 × 10 ≈ 36,360 ejemplos totales, es decir ~1,100 ejemplos por clase

Pero al tener pocos textos de alta calidad en idioma wayuunaiki revisados por expertos, se procede a utilizar embeddings preentrenados y fine-tuning parcial, para reducir la cantidad de ejemplos por clase necesarios y se obtiene como resultado la siguiente configuración:

Tabla 7. Configuración de clases.

Clases	33
Ejemplos por clase	300–500
Total de textos	9,900–16,500
Arquitectura sugerida	LSTM + Embedding (Word2Vec) + Fine-tunnig (parcial).
Regularización sugerida	Dropout.

Lo anterior teniendo en cuenta que para la capacidad de generalización del modelo es importante el número de parámetros y el tamaño del conjunto de datos, para este conjunto de datos en idioma wayuunaiki limitado por la inexistencia de textos o bases de datos ya etiquetadas, se procede a utilizar vectores de representación semántica como Word2Vec (Mikolov et al., 2013) con la finalidad de incorporar conocimiento lingüístico previo para disminuir la cantidad de grandes volúmenes de datos y mejorar que las representaciones distribuidas logren capturar relaciones semánticas profundas y mejorar el rendimiento en tareas supervisadas con este conjunto de datos reducido.

Igualmente, se optó por aplicar fine-tuning parcial como estrategia de entrenamiento, tal como lo plantean Howard y Ruder (2018) en su propuesta ULMFiT, y Devlin et al. (2019). Esta técnica permite reutilizar embeddings preentrenados, preservando el conocimiento gramatical, mientras se ajustan únicamente las capas superiores del modelo para adaptarlo a la tarea específica de clasificación gramatical. Esta aproximación resulta especialmente útil cuando se dispone de un conjunto de datos moderado, como señala

Chollet (2018) en Deep Learning with Python *"Cuando se dispone de pocos datos, lo más efectivo es usar lo aprendido y ajustar solo las capas superiores. Esto permite aprovechar el conocimiento aprendido previamente y adaptarlo a la nueva tarea con una cantidad moderada de ejemplos*", quien indica que el fine-tuning parcial puede producir resultados eficaces con tan solo 300 a 500 ejemplos por clase, gracias a la reutilización eficiente de los parámetros previamente entrenados.

Finalmente se obtiene un conjunto de datos para entrenamiento y prueba del modelo compuesto por un aproximado de 400 frases por clase como lo muestra la siguiente Figura 26 y en la tabla 8.

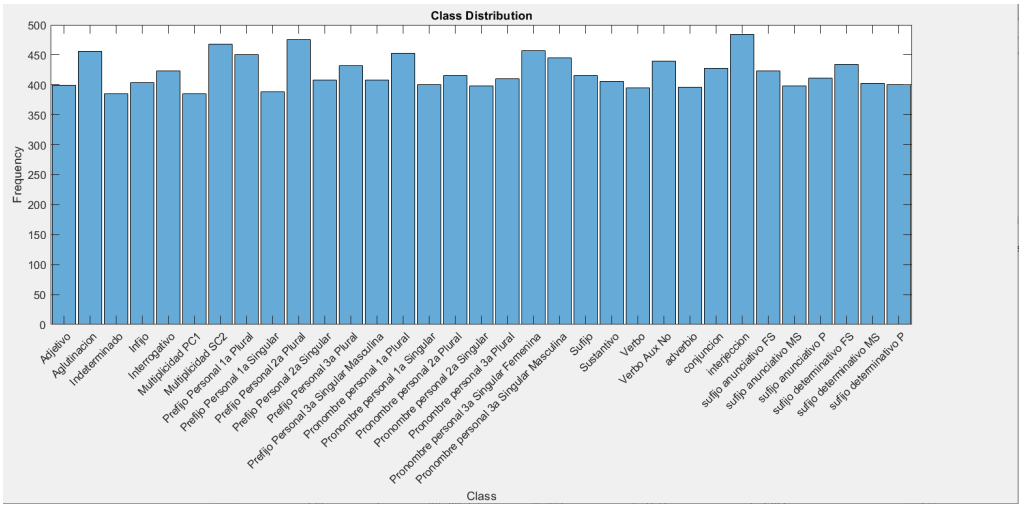


Figura 26. Distribución clases del modelo LSTM.

Tabla 8. Cantidad de datos por clases del modelo final.

Clase	Cantidad de frases
Adjetivo	399
adverbio	396
Aglutinación	456
conjunción	428
Indeterminado	385
Infijo	404
interjección	484
Interrogativo	423
Multiplicidad PC1	385
Multiplicidad SC2	468

Clase	Cantidad de frases
Prefijo Personal 1a Plural	450
Prefijo Personal 1a Singular	388
Prefijo Personal 2a Plural	476
Prefijo Personal 2a Singular	408
Prefijo Personal 3a Plural	432
Prefijo Personal 3a Singular Masculina	408
Pronombre personal 1a Plural	453
Pronombre personal 1a Singular	400
Pronombre personal 2a Plural	416
Pronombre personal 2a Singular	398
Pronombre personal 3a Plural	410
Pronombre personal 3a Singular Femenina	457
Pronombre personal 3a Singular Masculina	445
Sufijo	416
sufijo enunciativo FS	423
sufijo enunciativo MS	398
sufijo enunciativo P	411
sufijo determinativo FS	434
sufijo determinativo MS	403
sufijo determinativo P	400
Sustantivo	406
Verbo	395
Verbo Aux No	440
Total Frases	**13895**
Total Palabras	**57138**

Como conclusión parcial de la fase de preparación datos, se puede indicar que en análisis desarrollado ha permitido definir estrategias de preprocesamiento y representación de los datos, que aportaran al modelado y ejecución del clasificador de texto en idioma wayuunaiki. La próxima fase se enfocará en la preparación del modelo, optimizando la arquitectura seleccionada e implementando técnicas de regularización para mejorar la capacidad de generalización.

7.4 FASE 4. MODELADO

7.4.1 Determinación del modelo

7.4.1.1 Modelo conceptual

En esta fase se realizó la definición del modelo conceptual asociado a la clasificación de texto en idioma Wayuunaiki que incorpora características gramaticales. La Figura 27 muestra que la etapa de *entrenamiento* comienza con la construcción del conjunto de datos recopilando frases y/o palabras extraídas de documentos escritos en Wayuunaiki (Documentos elaborados por lingüistas en Wayuunaiki); El conjunto de datos etiquetados se preprocesa realizando una la limpieza de los datos y un proceso de lematización para identificar las características gramaticales que afecten la identificación de la forma raíz. Posteriormente el conjunto de datos se procesa para realizar la representación de los datos en vectores numéricos. A continuación, se identifican las variables para el modelo (Principales características gramaticales del Wayuunaiki) y se procede a entrenar y validar el modelo para obtener las métricas de desempeño las cuales se analizan para verificar la eficacia del modelo de clasificación y determinar si se requieren ajustes a los datos y/o a las variables. Luego de completar el entrenamiento, se ejecuta la etapa de *pruebas,* donde se evalúa el rendimiento del modelo clasificador ingresando un conjunto de datos nuevo, preprocesado pero no etiquetado; Finalmente se obteniendo texto clasificado de acuerdo a las características gramaticales del Wayuunaiki.

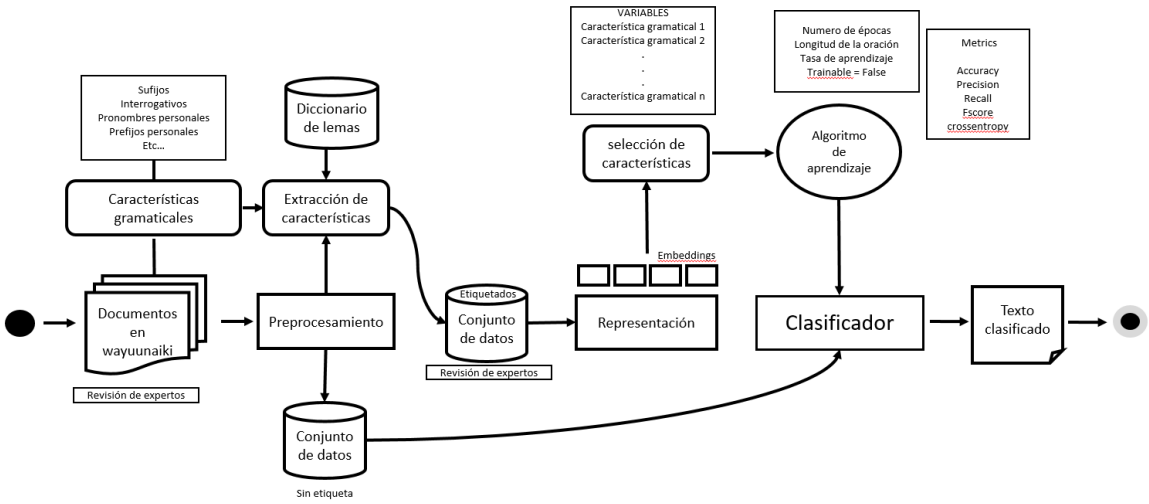


Figura 27. Modelo conceptual de clasificación automática de texto en idioma Wayuunaiki que incorpora características gramaticales.
Fuente: Elaboración propia.

7.4.1.2 Comparación arquitecturas

Esta sección se describe la comparación de los diferentes modelos para la clasificación automática de texto descritos en el marco conceptual de este documento, esto con la finalidad de determinar el modelo a utilizar. Inicialmente se lleva a cabo una comparación de literatura de otras investigaciones en las cuales realizan clasificación de texto y a continuación, se procede con la evaluación experimental de los modelos elegidos.

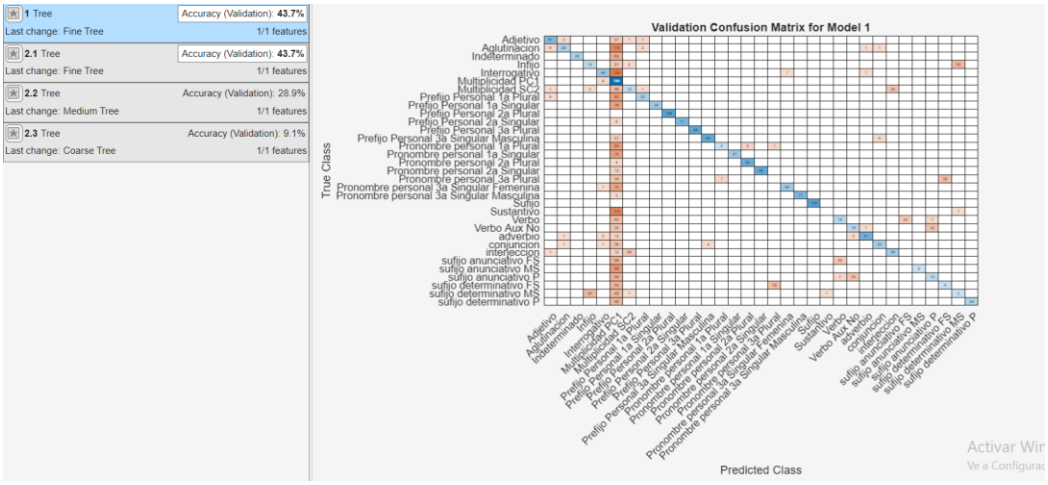
7.4.1.2.1 Experimentos iniciales con machine learning.

Utilizando MATLAB para la clasificación de texto en wayuunaiki con varios modelos de machine learning, se realizaron inicialmente los siguientes experimentos:

7.4.1.2.1.1 Random Forest.

Inicialmente se realizó los primeros experimentos con árboles de decisión, en la Figura 28, se muestran los resultados obtenidos de los experimentos, en los cuales se observa

que el valor de accuaracy (exactitud) de entrenamiento fue 43,7%, 28,9%, 9,1 % evidenciándose que el modelo no aprende correctamente, lo que se confirma con la matriz de confusión y el grafica ROC del modelo.



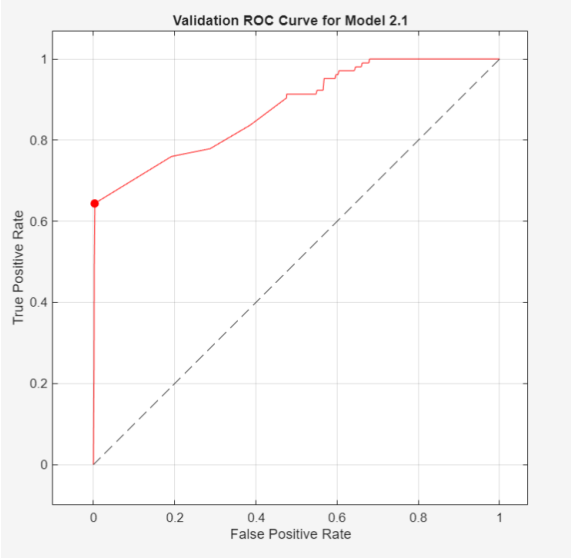


Figura 28. Resultados obtenidos clasificador de texto wayuunaiki con Random Forest. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

7.4.1.2.1.2 Naive Bayes.

En la Figura 29, se muestran los resultados obtenidos de los experimentos con clasificadores Naive Bayes, en los cuales se observa que el valor de accuaracy (exactitud) de entrenamiento fue 38,5 % evidenciándose que el modelo no aprende correctamente, lo que se confirma con la matriz de confusión del modelo.

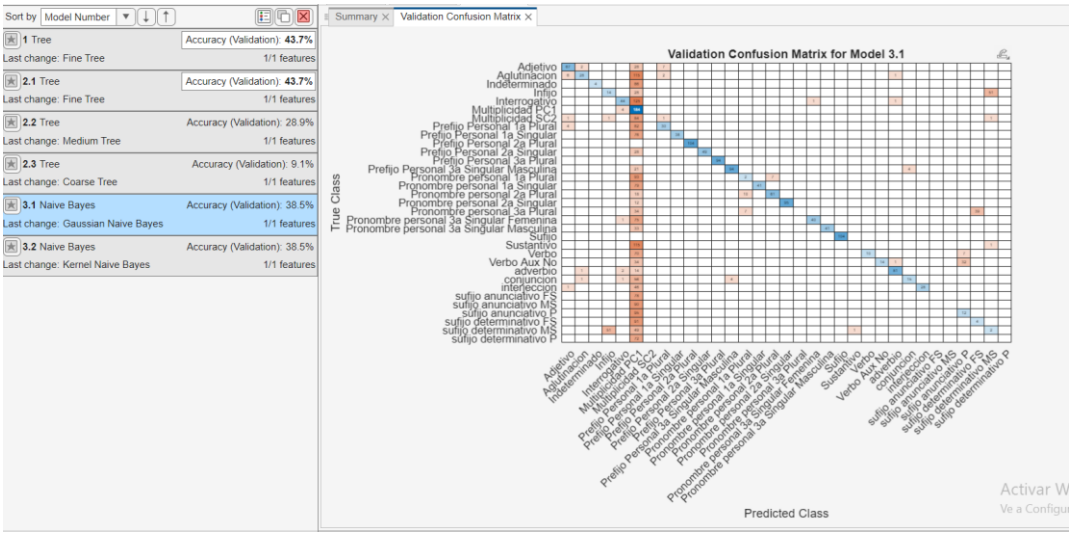


Figura 29. Resultados obtenidos clasificador de texto wayuunaiki con Naive Bayes. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

7.4.1.2.1.3 Support Vector Machines (SVMs)

En la Figura 30, se muestran los resultados obtenidos de los experimentos con clasificadores basados en Máquinas de vectores de soporte (MVS), en los cuales se observa que el valor de accuaracy (exactitud) de entrenamiento fue 38,8 %, 43,4% y 16,8% evidenciándose que el modelo no aprende correctamente, lo que se confirma con la matriz de confusión del modelo.

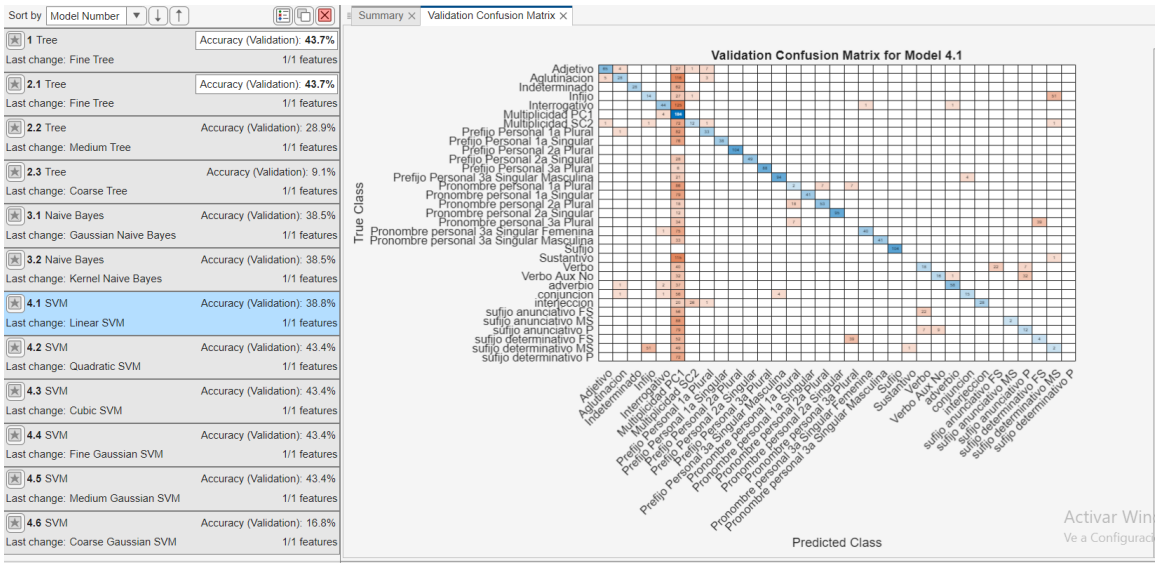


Figura 30. Resultados obtenidos clasificador de texto wayuunaiki con Support Vector Machines (SVMs).
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

7.4.1.2.1.4 K-Nearest Neighbor

En la Figura 31, se muestran los resultados obtenidos de los experimentos con clasificadores basados en K-Nearest Neighbor, en los cuales se observa que el valor de accuaracy (exactitud) de entrenamiento fue 41,0 %, 16,9%, 3,3% y 42,1% evidenciándose que el modelo no aprende correctamente, lo que se confirma con la matriz de confusión del modelo.

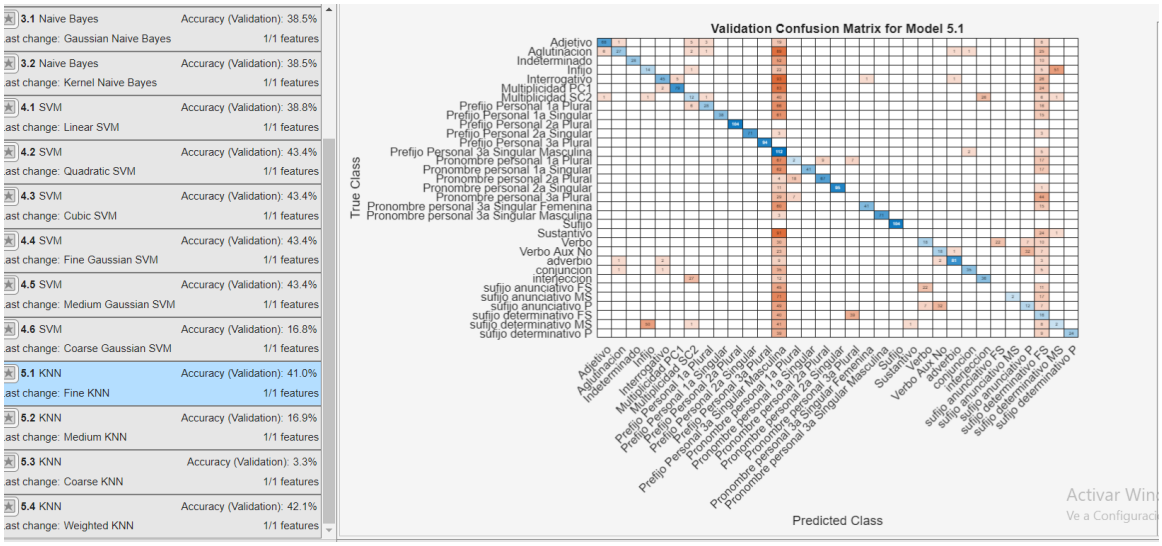


Figura 31. Resultados obtenidos clasificador de texto wayuunaiki con K-Nearest Neighbor.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

7.4.1.2.1.5 Kernel Aproximation

En la Figura 32, se muestran los resultados obtenidos de los experimentos con clasificadores basados en Kernel Aproximation, en los cuales se observa que el valor de accuaracy (exactitud) de entrenamiento fue 17,6% evidenciándose que el modelo no aprende correctamente.

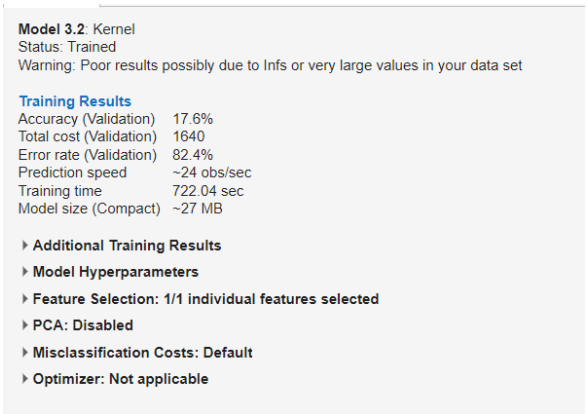


Figura 32. Resultados obtenidos clasificador de texto wayuunaiki con kernel Aproximation.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

Los experimentos generaron resultados que se muestran en la tabla 9. Resultados en los cuales se evidencia que el modelo que presento el mejor comportamiento es el de Random Forest, por tanto establece que se incluirá como uno de los modelos a comparar con modelos de Deep Learning.

Tabla 9. Comparación resultados experimentos iniciales modelos de clasificación.

Modelo	Validation Accuracy
Random Forest	**43,7%**
Naive Bayes	38,5 %
Support Vector Machines (SVMs)	43,4%
K-Nearest Neighbor	42,1 %
kernel Aproximation	17,6 %

7.4.1.2.2 Análisis de los resultados prácticos de experimentos con tres arquitecturas para clasificación de texto en wayuunaiki.

Para evaluar el comportamiento de diferentes arquitecturas, mediante el uso de MATLAB, se dividió el conjunto de datos en idioma wayuunaiki en un 80 % para entrenamiento y un 20% para validación y se llevaron a cabo los siguientes experimentos:

7.4.1.2.2.1 Random Forest

El modelo Random Forest se entrenó utilizando la aplicación "Classification Learner" de MATLAB. Para este experimento se eligió el algoritmo “bagged tres” disponibles en el set de modelos que combinan múltiples clasificadores “classification ensembles”.

En la Figura 33, se muestran los resultados obtenidos de los experimentos, en los cuales se observa que el valor de accuaracy (exactitud) de entrenamiento fue 49,1 % con una tasa de error del 50, 9 %, evidenciándose que el modelo no aprende correctamente.

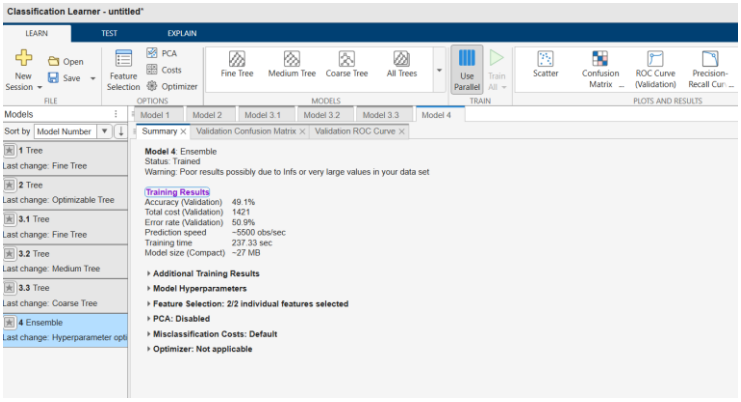


Figura 33. Resultados obtenidos clasificador de texto wayuunaiki con Random Forest. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

La aplicación genera el cálculo de las métricas promedio para el clasificador multiclase de frases en wayuunaiki. En la Figura 34, se muestran las métricas obtenidas y en la Figura 35, se presenta el rendimiento del clasificador y las curvas de las métricas promedio y los puntos para cada clase en su problema binario. En el comportamiento de las curvas se observa que la curva dela métrica micro de color café es la que más se acerca al eje superior, mientras, que la curva de la métrica macro de color azul está más cerca de la diagonal. Este comportamiento identificado en las gráficas, reitera que el modelo no aprende correctamente.

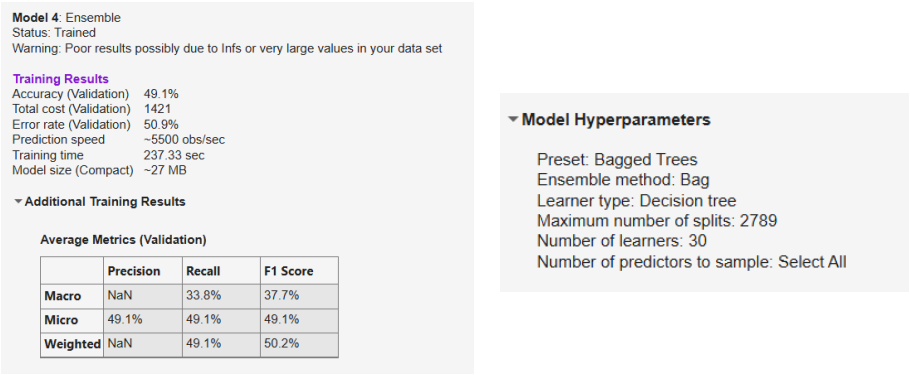


Figura 34. Métricas promedio y Hyperparametros para el clasificador multiclase. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

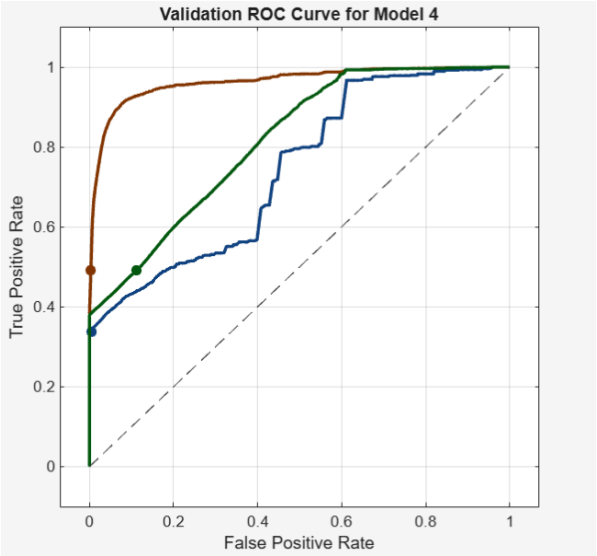


Figura 35. Curva ROC para las métricas promedio del clasificador.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

En la Figura 36, se muestran las predicciones correctas generadas por el modelo, mientas que en la Figura 37, se muestran las predicciones incorrectas. El análisis de los gráficos confirmar los resultados obtenidos de exactitud de 49,1 % y una tasa de error del 50, 9 %. Nuevamente, los datos presentados en las gráficas, confirman que el modelo no aprende correctamente.

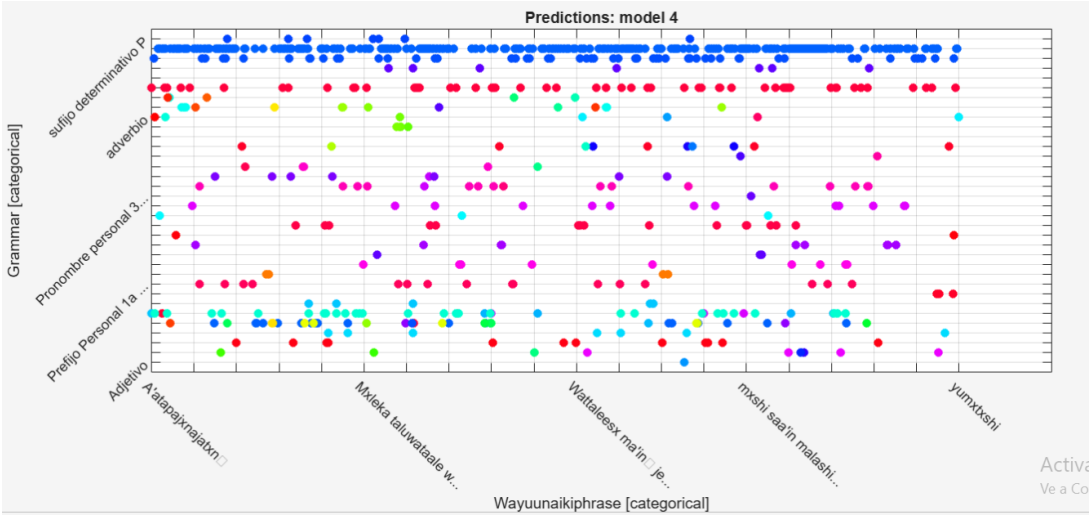


Figura 36. Predicciones correctas del modelo.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

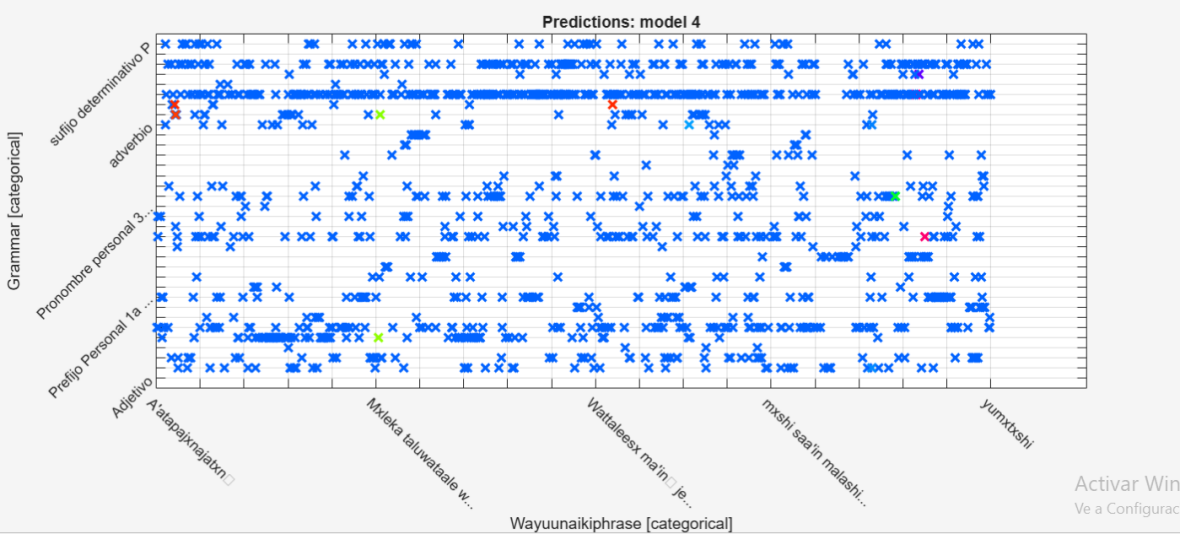


Figura 37. Predicciones incorrectas del modelo.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

En conclusión, el modelo de clasificación Random Forest y utilizando el algoritmo “bagged tres”, no es adecuado para la clasificación de texto en wayuunaiki.

7.4.1.2.2.2 Convolutional Neural Networks - CNN

En este experimento se implementó una Red Neuronal Convolucional – CNN, utilizando un pequeño subconjunto de los datos de entrenamiento (mini-batch), este subconjunto se usó en cada iteración del backpropagation. Esta red se adaptó de acuerdo a lo descrito en MathWorks (2021) y MathWorks (2024-c), en donde se explica cómo entrenar una red Red Neuronal Convolucional – CNN que utiliza filtros convolucionales unidimensionales, filtros que poseen diferentes tamaños de anchos (longitud n-grama).

El primer resultado obtenido de este experimento, índico que el conjunto de datos presentaba valores similares (palabras en wayuunaiki se repiten en el set de datos al representar algo deferente en la gramática) y por tanto no era posible ejecutar el experimento. Por lo tanto, se procedió a eliminar las palabras repetidas con la finalidad de observar el comportamiento de la red CNN.

En las Figuras 38 y 39 se muestra el comportamiento de la red durante 30 épocas, red con un tamaño de n-gramas de [2 3 4 5 6] y una tasa de aprendizaje de 0.001 (Experimento a). Los resultados obtenidos del experimento son: valor de accuaracy (exactitud) de entrenamiento: 48,44 %, y un valor de perdida en el entrenamiento del 1,5583, como se muestra en la Figura 40. Evidenciándose que el modelo no aprende correctamente.

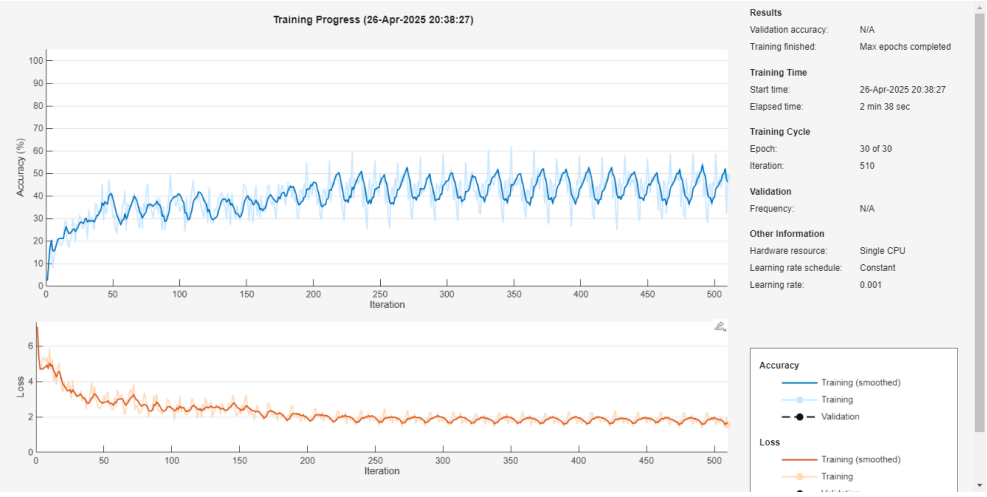


Figura 38. Entrenamiento red CNN con tasa de aprendizaje de 0.001

Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

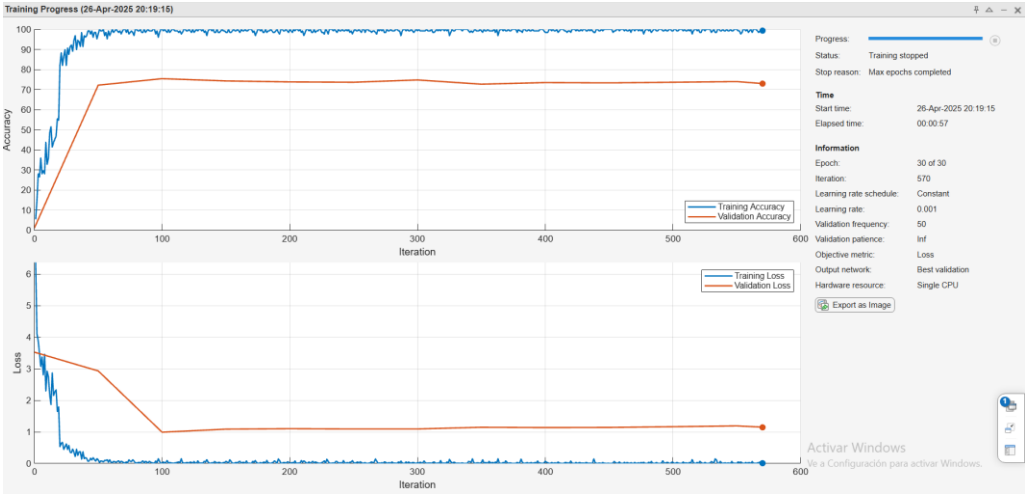


Figura 39. Validación red CNN con tasa de aprendizaje de 0.001.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

```
Training on single CPU.
|=========================================================================================|
|  Epoch  |  Iteration  |  Time Elapsed  |  Mini-batch  |  Mini-batch  |  Base Learning  |
|         |             |   (hh:mm:ss)   |   Accuracy   |     Loss     |      Rate       |
|=========================================================================================|
|       1 |           1 |       00:00:02 |        2.34% |       7.0923 |          0.0010 |
|       3 |          50 |       00:00:18 |       30.47% |       3.1931 |          0.0010 |
|       6 |         100 |       00:00:35 |       32.81% |       2.4316 |          0.0010 |
|       9 |         150 |       00:00:50 |       42.19% |       2.3903 |          0.0010 |
|      12 |         200 |       00:01:05 |       45.31% |       1.9144 |          0.0010 |
|      15 |         250 |       00:01:20 |       49.22% |       1.8406 |          0.0010 |
|      18 |         300 |       00:01:35 |       42.19% |       1.7626 |          0.0010 |
|      21 |         350 |       00:01:49 |       46.09% |       1.9132 |          0.0010 |
|      24 |         400 |       00:02:05 |       42.97% |       1.8314 |          0.0010 |
|      27 |         450 |       00:02:20 |       56.25% |       1.6532 |          0.0010 |
|      30 |         500 |       00:02:35 |       44.53% |       2.0327 |          0.0010 |
|      30 |         510 |       00:02:38 |       48.44% |       1.5583 |          0.0010 |
|=========================================================================================|
Training finished: Max epochs completed.
```

Figura 40. Resultados red CNN con tasa de aprendizaje de 0.001.
Fuente: Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

En la ilustraciones de la figura 41, se observa la arquitectura de la red generada por MATLAB para una red con un tamaño de n-gramas de [2 3 4 5 6], en la cual se conectan la capa de concatenación con las capas de máxima agrupación global.

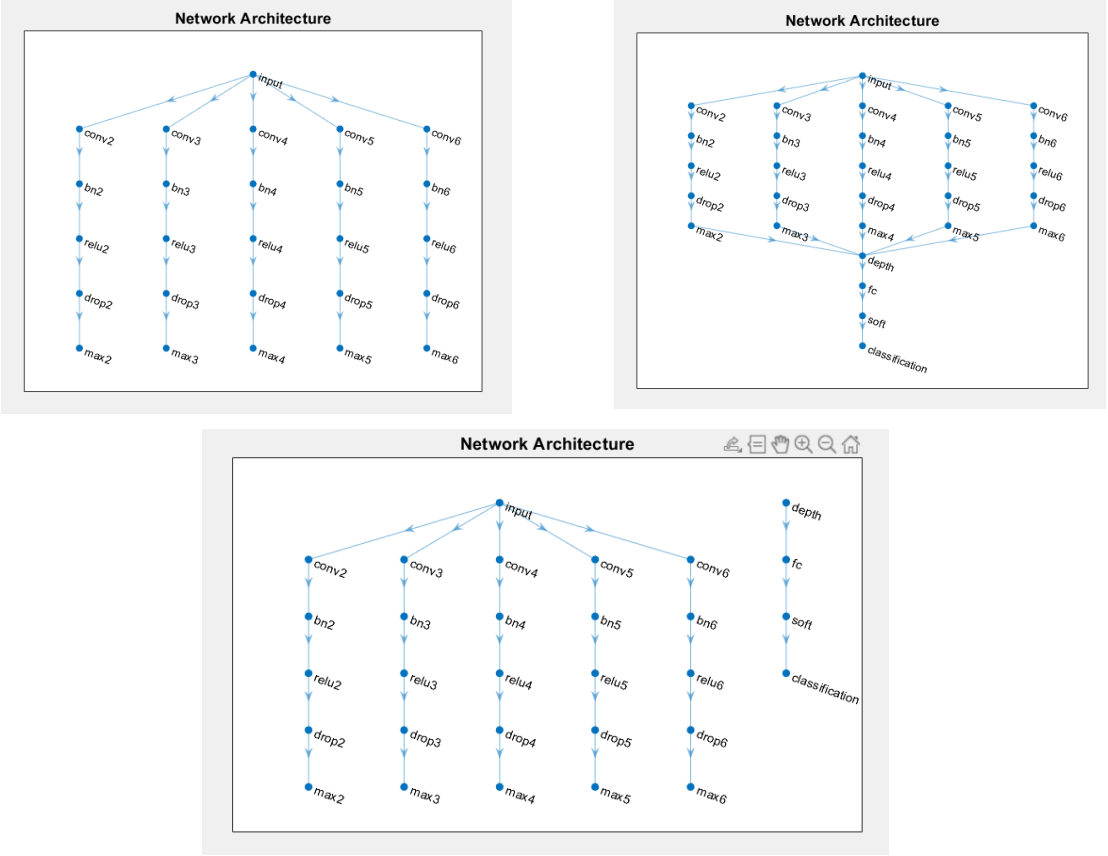


Figura 41. Arquitecturas generadas para la red CNN para la clasificación.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

En conclusión, el modelo para clasificación de texto utilizando una red Neuronal Convolucional - CNN, no es adecuado para la clasificación de texto en wayuunaiki. Lo anteriormente expuesto se fundamente en los resultados de exactitud obtenidos en el experimento; Además, se identificó que el modelo presenta errores en el contexto global de las frases debido a que no puede abarcar los datos duplicados que aportan a la identificación de la gramática del idioma wayuunaiki.

7.4.1.2.2.3 Red Neuronal LSTM

En este experimento se implementó una red de memoria a corto-largo plazo (LSTM), adaptando lo descrito en MathWorks (2024-d). La red LSTM adaptada, entrenada y probada genero los resultados listados en Figura 33 y en la Figura 42 se muestra el comportamiento de la red para 30 épocas.

En la Figura 43, se muestra que para el conjunto de datos de validación se obtuvo un valor de exactitud (accuracy) de 68,50% y para el conjunto de datos de entrenamiento un valor de exactitud (accuracy) de 93,75%. Estos resultados muestras un sobreajuste en el modelo, esto debido a los valores similares presentes en el conjunto de datos que están relacionados con las palabras en wayuunaiki que se repiten porque tienen un significado diferente).

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Validation Accuracy	Mini-batch Loss	Validation Loss	Base Learning Rate
1	1	00:00:02	0.00%	19.55%	3.5256	3.5160	0.0010
1	50	00:00:07	18.75%	17.45%	2.7852	2.9116	0.0010
1	100	00:00:11	18.75%	27.63%	2.9813	2.6767	0.0010
1	150	00:00:15	50.00%	35.22%	2.2946	2.5280	0.0010
2	200	00:00:20	37.50%	36.83%	2.1104	2.2559	0.0010
2	250	00:00:24	50.00%	45.07%	1.8688	2.0475	0.0010
2	300	00:00:29	68.75%	53.15%	1.5115	1.8461	0.0010
3	350	00:00:33	81.25%	54.28%	0.8768	1.7463	0.0010
3	400	00:00:38	81.25%	58.32%	1.0673	1.5983	0.0010
3	450	00:00:42	93.75%	61.87%	0.4868	1.5046	0.0010
4	500	00:00:47	87.50%	61.71%	0.3885	1.5908	0.0010
4	550	00:00:51	93.75%	62.68%	0.3254	1.5499	0.0010
4	600	00:00:56	81.25%	66.07%	0.5782	1.4450	0.0010
5	650	00:01:00	87.50%	63.81%	0.2672	1.4589	0.0010
5	700	00:01:05	93.75%	66.56%	0.2733	1.4557	0.0010
5	750	00:01:09	100.00%	67.53%	0.0408	1.4061	0.0010
6	800	00:01:14	93.75%	67.53%	0.1966	1.3976	0.0010
6	850	00:01:18	93.75%	68.01%	0.1354	1.4499	0.0010
6	900	00:01:23	81.25%	67.69%	0.5682	1.4325	0.0010
7	950	00:01:28	100.00%	67.04%	0.0453	1.4496	0.0010
7	1000	00:01:32	100.00%	67.69%	0.0524	1.4590	0.0010
7	1050	00:01:37	100.00%	68.17%	0.0523	1.4540	0.0010
8	1100	00:01:41	100.00%	67.37%	0.1167	1.4787	0.0010
8	1150	00:01:46	100.00%	67.85%	0.0281	1.5168	0.0010
8	1200	00:01:52	93.75%	67.37%	0.1277	1.5131	0.0010
9	1250	00:01:57	100.00%	68.66%	0.0210	1.4691	0.0010
9	1300	00:02:02	100.00%	69.63%	0.0536	1.4887	0.0010
9	1350	00:02:06	100.00%	68.66%	0.0463	1.5279	0.0010
10	1400	00:02:11	93.75%	67.21%	0.2731	1.5501	0.0010
10	1450	00:02:16	100.00%	68.01%	0.0082	1.5091	0.0010
10	1500	00:02:20	93.75%	69.31%	0.1193	1.5303	0.0010
10	1550	00:02:25	93.75%	68.50%	0.1355	1.5146	0.0010

Training finished: Max epochs completed.

Figura 42. Resultados red LSTM con tasa de aprendizaje de 0.001
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

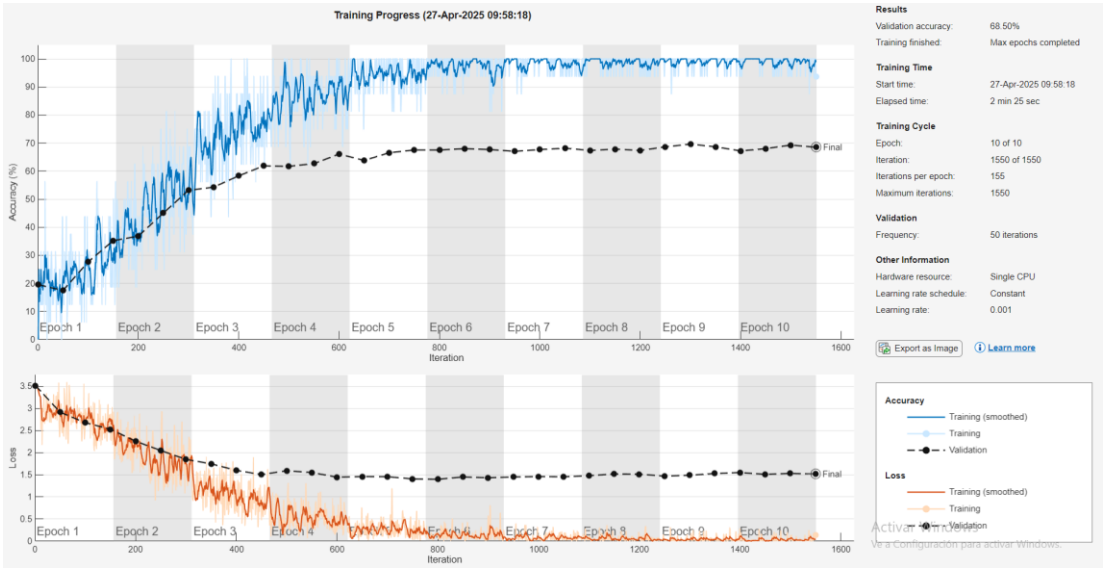


Figura 43. Entrenamiento red LSTM con tasa de aprendizaje de 0.001.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

7.4.1.2.2.4 Red transformers utilizando BERT modificado.

En este experimento se implementó una red neuronal de tipo transformer, modificando BERT para datos específicos en idioma wayuunaiki.

En la Figura 44, se muestra el comportamiento de la red durante 30 épocas, con una tasa de aprendizaje de 0.001 (Experimento a). Los resultados obtenidos del experimento son: valor de accuaracy (exactitud) de entrenamiento: 65,62 %, valor de accuaracy (exactitud) de validación: 55,09% y un valor de perdida en la validación del 1,5936, como se muestra en la Figura 45.

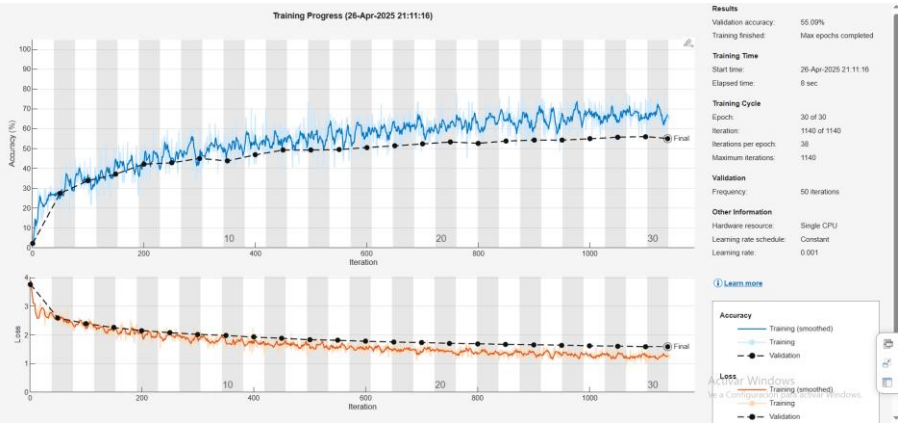


Figura 44. Entrenamiento red Transformers BERT con tasa de aprendizaje de 0.001. Fuente: Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

```
Training on single CPU.
|=========================================================================================================|
|  Epoch  |  Iteration  |  Time Elapsed  |  Mini-batch  |  Validation  |  Mini-batch  |  Validation  | Base Learning |
|         |             |    (hh:mm:ss)  |   Accuracy   |   Accuracy   |     Loss     |     Loss     |     Rate      |
|=========================================================================================================|
|       1 |           1 |       00:00:03 |       1.56% |       2.26% |       3.8316 |       3.7739 |        0.0010 |
|       2 |          50 |       00:00:03 |      26.56% |      27.46% |       2.5840 |       2.5963 |        0.0010 |
|       3 |         100 |       00:00:04 |      35.94% |      33.76% |       2.1379 |       2.4012 |        0.0010 |
|       4 |         150 |       00:00:04 |      37.50% |      37.16% |       2.2179 |       2.2575 |        0.0010 |
|       6 |         200 |       00:00:04 |      35.94% |      42.16% |       2.1317 |       2.1567 |        0.0010 |
|       7 |         250 |       00:00:04 |      54.69% |      42.81% |       1.9778 |       2.0806 |        0.0010 |
|       8 |         300 |       00:00:05 |      50.00% |      45.07% |       1.8025 |       2.0255 |        0.0010 |
|      10 |         350 |       00:00:05 |      51.56% |      43.78% |       1.6558 |       1.9842 |        0.0010 |
|      11 |         400 |       00:00:05 |      67.19% |      46.85% |       1.4666 |       1.9275 |        0.0010 |
|      12 |         450 |       00:00:05 |      50.00% |      49.27% |       1.8027 |       1.8868 |        0.0010 |
|      14 |         500 |       00:00:06 |      67.19% |      49.27% |       1.2804 |       1.8360 |        0.0010 |
|      15 |         550 |       00:00:06 |      57.81% |      49.43% |       1.4061 |       1.8167 |        0.0010 |
|      16 |         600 |       00:00:06 |      53.12% |      50.57% |       1.6783 |       1.7819 |        0.0010 |
|      18 |         650 |       00:00:06 |      48.44% |      51.37% |       1.7500 |       1.7575 |        0.0010 |
|      19 |         700 |       00:00:06 |      62.50% |      52.34% |       1.3470 |       1.7376 |        0.0010 |
|      20 |         750 |       00:00:07 |      64.06% |      53.31% |       1.4464 |       1.7110 |        0.0010 |
|      22 |         800 |       00:00:07 |      56.25% |      52.67% |       1.4489 |       1.6954 |        0.0010 |
|      23 |         850 |       00:00:07 |      68.75% |      53.80% |       1.3510 |       1.6722 |        0.0010 |
|      24 |         900 |       00:00:07 |      56.25% |      54.28% |       1.4707 |       1.6599 |        0.0010 |
|      25 |         950 |       00:00:07 |      70.31% |      54.28% |       1.2203 |       1.6465 |        0.0010 |
|      27 |        1000 |       00:00:08 |      65.62% |      54.93% |       1.3478 |       1.6214 |        0.0010 |
|      28 |        1050 |       00:00:08 |      67.19% |      55.74% |       1.3756 |       1.6127 |        0.0010 |
|      29 |        1100 |       00:00:08 |      75.00% |      56.06% |       0.9963 |       1.5912 |        0.0010 |
|      30 |        1140 |       00:00:08 |      65.62% |      55.09% |       1.2641 |       1.5936 |        0.0010 |
|=========================================================================================================|
Training finished: Max epochs completed.

accuracy =

    0.5509
```

Figura 45. Resultados red Transformers BERT con tasa de aprendizaje de 0.001 Fuente: Fuente: La imagen fue generada a partir de los resultados obtenidos mediante experimentación en MATLAB. Elaboración propia.

Los resultados del experimento sugieren que el modelo puede estar sobre ajustado. Por lo tanto, se realizan seis experimentos adicionales, cambiando la tasa de aprendizaje y ajustando *dropout rate* en la capa final para reducir el sobreajuste.

En la tabla 10, se listan los resultados obtenidos al realizar diferentes experimentos cambiando la tasa de aprendizaje de la red. Estos resultados muestran que la red mejoro su desempeño utilizando una tasa de aprendizaje de 0.01; los resultados obtenidos con esta tasa de aprendizaje son 84,38% de exactitud en el entrenamiento y 64,94% en la validación, lo cual resulta bajo y sugiere que el modelo puede estar sobre ajustado.

Tabla 10. Resultados experimentos transformers

Data modification	Validation Loss	Train Accuracy	Validation Accuracy
InitialLearnRate= 0.0002 (2.0000e-05)	2.6603	31.25%	26.82%
InitialLearnRate= 0.001	1.5912	65.62%	55.09%
InitialLearnRate= 0.01	1.5417	84.38%	64.94%
InitialLearnRate= 0.1	6.4146	65.62%	52.02%
InitialLearnRate= 0.05	3.6816	56.25%	60.10%
InitialLearnRate= 0.02	1.6454	78.12%	62.36%

Los experimentos generaron resultados que se muestran en la tabla 11. Resultados en los cuales se evidencia que el modelo que presento el mejor comportamiento es la Red Neuronal LSTM.

Tabla 11. Comparación arquitecturas de clasificación.

Modelo	Train Loss	Train Accuracy	Validation Loss	Validation Accuracy	Tasa de aprendizaje
Random forest	–	49,1 %	–	–	0.001
CNN	1.5583	48,44%	1.6714	39,12%	0.001
LSTM	0,1355	93,75 %	1.5146	**68,50 %**	0.001
Red transformers	1.2641	65,62%	1.5936	55,09 %	0.0010

Tabla. Comparación de resultados de arquitecturas para clasificación de texto en idioma wayuunaiki.

Como conclusión en esta fase se obtuvo una definición de variables y una aproximación inicial del modelo (modelo conceptual) que muestra la incorporación de las principales características de modelo de clasificación automática de texto en idioma indígena Wayuunaiki que incorpora características gramaticales. Igualmente, en la comparación de arquitecturas se logró establecer que la más adecuada para las particularidades del conjunto de datos, es la Red Neuronal LSTM.

7.4.2 Proyección del plan de pruebas.

Para evaluar el diseño del modelo para clasificación automática de texto en wayuunaiki que incorpora características gramaticales, se validara la robustez utilizando los siguientes enfoques:

Validación Hold-Out: Para la validación, el conjunto de datos en idioma wayuunaiki se divide en datos de entrenamiento y prueba, para evaluar el rendimiento (precisión) del modelo con datos nuevos no etiquetados.

Validación K-Fold Cross-Validation: Para esta validación, el conjunto de datos se dividirá en k particiones las cuales tendrán una segmentación iterativa. Lo anterior con la finalidad de entrenar y validar el modelo con múltiples configuraciones y así disminuir la variabilidad y logar la estabilidad del modelo.

7.4.3 Diseño del modelo

En esta sección se describe la estructuración del modelo para la clasificación automática de texto en wayuunaiki que incorpora características gramaticales, de acuerdo con los resultados obtenidos en la determinación del modelo.

Diversos estudios han explorado la aplicación de CRISP-DM en su metodología de desarrollo. Por ejemplo, Sanango y Hallo (2024) describen que las actividades de las etapas de CRISP-DM se pueden organizar en capas. Para este proyecto de investigación, se adaptaran las capas de la siguiente manera:

Capa de datos: En esta capa se encuentran los textos seleccionados en wayuunaiki, el diccionario de lemas y las versiones del conjunto de datos en formato CSV (Valores Separados por Comas).

Capa lógica: Esta capa contiene el preprocesamiento de los datos (limpieza, lematización, etc.), transformación de los datos, modelado, el entrenamiento del modelo de clasificación del texto en wayuunaiki y la integración de los resultados al conjunto de datos.

Capa de presentación: Visualización de las métricas clave como exactitud, precisión, recall y F1-score y los resultados de la prueba con datos sin etiquetar.

La arquitectura del sistema diseñada para este proyecto de investigación, se muestra en la Figura 46.

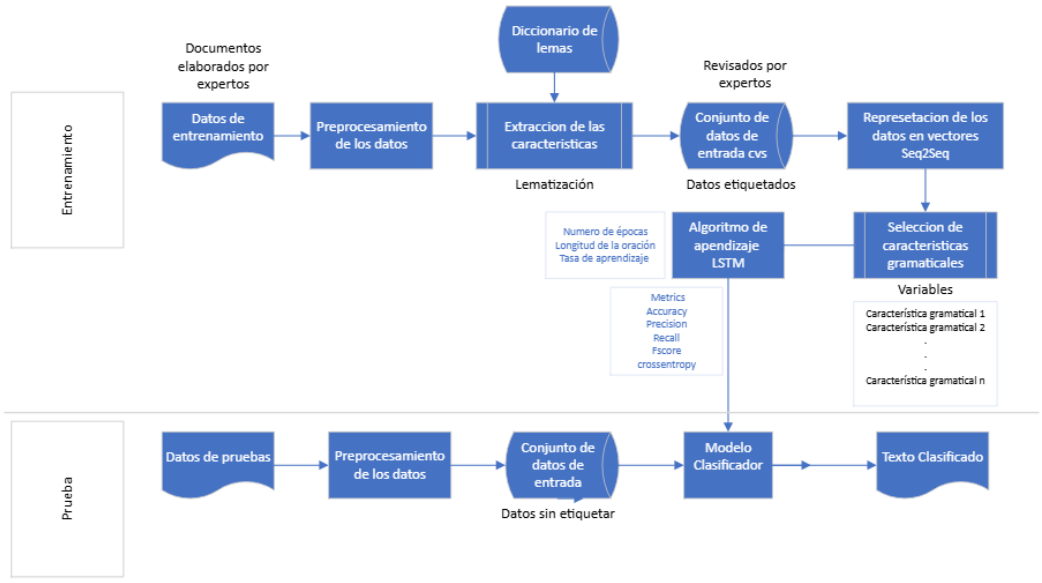


Figura 46. Estructura del modelo clasificación automática de texto en idioma Wayuunaiki que incorpora características gramaticales.
Fuente: Elaboración propia.

Las actividades desarrolladas para la estructuración del modelo para la clasificación de texto en idioma wayuunaiki que incorpora características gramaticales, se describen a continuación:

7.4.4 Definición de la arquitectura

De acuerdo a los resultados obtenidos en la sección de comparación de arquitectura, el diseño del modelo para la clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales, se basa en la modificación de la arquitectura de red neuronal recurrente de memoria a corto-largo plazo (LSTM) con mecanismo de atención.

7.4.5 Recolección de los datos

De acuerdo en lo establecido en la fase 1: Preparación de los datos, los documentos en wayuunaiki son recolectados de fuentes reconocidas en el análisis lingüístico del idioma wayuunaiki y posteriormente el conjunto de datos se crea, se revisa y se etiqueta con ayuda de un experto en el idioma.

Posteriormente, para dividir los datos en clases, se establecen las etiquetas de características gramaticales como categóricas.

7.4.6 Preprocesamiento de los datos

Para el preprocesamiento de los datos, se realiza el procedimiento descrito en la fase 1. Preparación de los datos, en la cual se seleccionan, limpian, enriquece y se normalizan los datos del conjunto de frases y palabras de acuerdo a las características del idioma wayuunaiki.

Posteriormente, se procede a realizar el proceso de tokenización los datos sin eliminar ninguna palabra, el resultado de este proceso se muestra en la Figura 47.

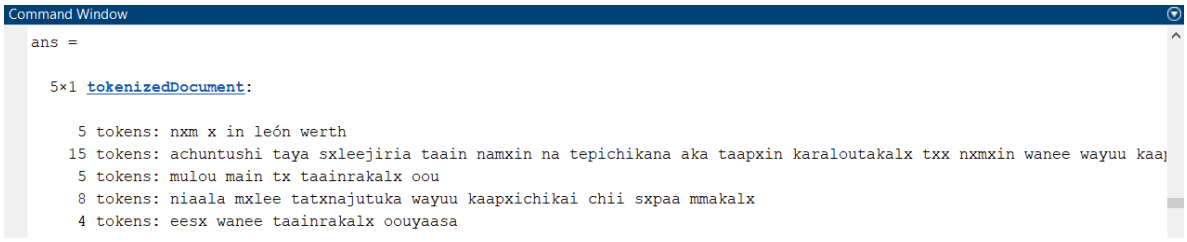


Figura 47. Ejemplo de tokenización del texto en wayuunaiki.
Fuente: Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

7.4.7 Representación de los datos.

Los resultados obtenidos en la fase de determinación del modelo, establecen que la vectorización de los datos en wayuunaiki, mejora significativamente con la utilización del modelo Seq2Seq (Sequence to Sequence). Este modelo permite una mejor representación de los datos secuenciales, beneficiando la identificación de relaciones de acuerdo al contexto del idioma (Sutskever et al, 2014).

7.4.8 Construcción del Modelo LSTM

En esta sección, se describen el funcionamiento y las características principales del modelo de clasificación automática de texto en idioma wayuunaiki que incorpora características gramaticales, implementando una red LSTM en MATLAB.

a) Se instalan Database Toolbox en MATLAB, para conectar con el conjunto de datos (Figura x), que se encuentra alojado en un servidor en la plataforma Azure.
b) Se utiliza la tabla relacional que contiene las frases en wayuunaiki y las etiquetas que corresponden a las clases (características gramaticales).
c) Para Los datos que corresponden a las etiquetas, se debe verificar que sean de tipo categóricos.
d) Se dividen los datos en clases de acuerdo a su categoría.
e) Utilizando Hold out, se dividen los datos de la siguiente manera: 80% para entrenamiento y 20 % para la validación, de ese porcentaje de validación se puede establecer un 5% para pruebas.

f) Se extrae el texto en wayuunaiki y las etiquetas de las tablas particionadas.

g) Se utiliza una función para tokenizar y preprocesar el texto.

h) Convertir las palabras en índices numéricos con la función de codificación de palabras *wordEncodingfunción*.

i) Para rellenar y truncar los documentos con la finalidad de que todos tengan la misma longitud. Se selecciona una longitud objetivo (ideal que sea corta) y posteriormente se truncan los documentos más largos y se rellenan los más cortos desde el margen izquierdo, esto se con la finalidad de que el entrenamiento sea mejor. El valor de la longitud objetivo se puede establecer generando un histograma de longitudes como lo muestra la Figura 48. en la cual se muestra que la longitud máxima son 24 palabras, por lo cual serían 24 tokens para la longitud objetivo.

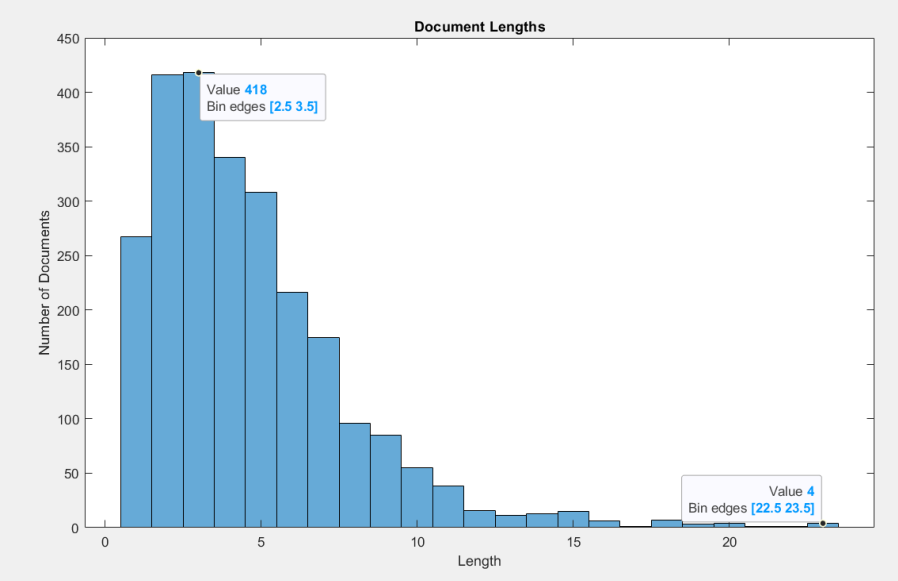


Figura 48. Histograma con longitud de palabras en las frases.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

j) Utilizar Seq2Seq (Sequence to Sequence), para convertir los datos en secuencia de palabras (vectores de palabras).

k) Se crea la estructura de la red LSTM con los siguientes características: Una capa de entrada de secuencia (Tamaño de entrada = 1), una capa de incrustación de palabras (Tamaño = 25), Una capa LSTM (Unidades ocultas: 100), se establece

el modo de salida para la tarea de clasificación de secuencia de etiqueta (Modo de salida = "*last"*) y al final se incluye una capa totalmente conectada (Tamaño = número de clases).

l) Defina la arquitectura de red LSTM. Para introducir datos de secuencia en la red, incluya una capa de entrada de secuencia y establezca el tamaño de entrada en 1. A continuación, incluya una capa de incrustación de palabras de dimensión 50 y el mismo número de palabras que la codificación de palabras. A continuación, incluya una capa LSTM y establezca el número de unidades ocultas en 80. Para usar la capa LSTM en un problema de clasificación de secuencia a etiqueta, establezca el modo de salida en "last". Finalmente, agregue una capa completamente conectada con el mismo tamaño que el número de clases y una capa softmax.

m) Se configuran las parámetros en la función t*rainingOptions*, así: Se establece el algoritmo de optimización "Adam"; el formato de los datos de entrada se establece como "CBT"; el número de muestras procesadas en cada iteración en 16; se mezclan los datos en cada época; se genera una gráfica *Plotsopción,* para observar el progreso de entrenamiento; se indica el conjunto de datos para la validación y finalmente se puede establecer con la el parámetro *Verboseopción* , si se desea imprimir la información sobre el progreso del entrenamiento.

7.4.9 Entrenamiento del Modelo:

Para el entrenamiento del modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales, se utiliza la función *trainnet* de la siguiente manera:

```
[net,info] = trainnet(XTrain,YTrain,layers,"crossentropy",options);
info.ValidationHistory
```

7.4.10 Definición de la función de pérdida

La definición de la función de perdida, se establece a partir de la validación de la entropía cruzada, lo anterior para asegurar que el modelo generaliza bien los datos nuevos. Para

indicar que la función de perdida que se utilizará es entropía cruzada, se define como "crossentropy":

[net,info] = trainnet(XTrain,YTrain,layers,"crossentropy",options).

7.4.11 Selección de métricas de evaluación

Para seleccionar las métricas de evaluación, se identificó en los trabajos relacionados cuales eran las más usadas y se determinó utilizar las siguientes métricas: Exactitud, precisión, recall, F1-score. Para la evaluación del modelo, se seleccionó la validación cruzada de acuerdo por lo expuesto por Hernández, Ramírez y Ferri (2005), en el capítulo 17: *Técnicas de evaluación – evaluación de clasificadores*.

7.4.12 Entorno de entrenamiento.

El entorno de entrenamiento para el modelo de clasificación en idioma wayuunaiki, realizado en MATLAB y configurado con una tasa de aprendizaje de 0.001 y un máximo de 30 iteraciones, se muestra en la Figura 49.

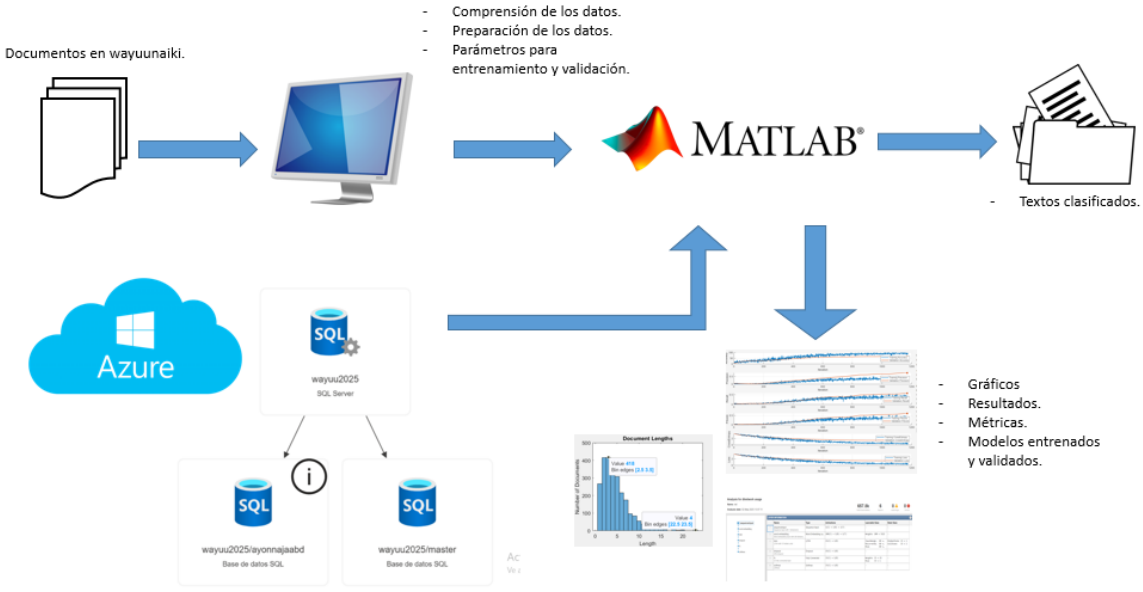


Figura 49. Entorno de entrenamiento del modelo para clasificación de texto en wayuunaiki.
Fuente: Elaboración propia.

Como conclusión parcial de la fase del modelado, se puede indicar que es fundamental para el desarrollo del modelo, ya que en esta fase se determina el modelo, se proyecta el modelo conceptual, la comparación de arquitecturas, la definición de arquitectura y la construcción del modelo. En específico para el modelado de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales, se definió una estructura para la representación de los datos y la adaptación y/o modificación de la red neuronal recurrente LSTM Las LSTM (Long Short-Term Memory).

8. VALIDACIÓN DEL MODELO

En esta sección se describe la fase de validación del modelo para clasificación de texto en wayuunaiki que incorpora características gramaticales, utilizando un método empírico en específico de experimentación y medición de variables, con la finalidad de demostrar la veracidad de la hipótesis planteada sobre el modelo de clasificación automático de texto en wayuunaiki que usando Deep Learning pueda identificar las principales características gramaticales del idioma indígena wayuunaiki.

En específico del conjunto de datos utilizó el 20% para la validación y 10% para la pruebas con datos nuevos sin etiquetar se midieron las métricas de precisión utilizando una matriz de confusión y calculando la precisión y el alcance (precision and recall) y los valores de exactitud (accuracy) y f1- Score para el modelo. Para estos cálculos se realizaron tres pruebas con la finalidad de validar los resultados con diferentes hiperparametros y disminuir el sobreajuste del modelo identificado en la fase de determinación.

8.1 VALIDACIÓN DEL MODELO.

Para evitar el sobreajuste del modelo de clasificación de texto en idioma indígena wayuunaiki, se eligió la técnica de validación cruzada (k-fold) para evaluar el rendimiento. En la técnica k-fold se utiliza un único parámetro K (K= # subconjuntos de igual tamaño en los que se divide el conjunto de datos), el cual se utiliza para posteriormente entrenar el modelo k veces.

Comparación de resultados de validación.

Como lo muestra la tabla 12, para una tasa de aprendizaje de 0,0001, el modelo obtuvo una precisión del 66,72% en validación, lo que significa que el modelo clasificó correctamente el 66% de los datos de prueba, mientras que el 34% de las predicciones fueron incorrectas. Para una tasa de aprendizaje de 0,001, el modelo obtuvo una precisión del 71,56% en validación, lo que significa que el modelo clasificó correctamente el 71% de los datos de prueba, mientras que el 29% de las predicciones fueron incorrectas. Para una tasa de aprendizaje de 0,01, el modelo obtuvo una

precisión del 68,82% en validación, lo que significa que el modelo clasificó correctamente el 69% de los datos de prueba, mientras que el 31% de las predicciones fueron incorrectas.

Tabla 12. Comparación de resultados de validación.

Data modification	Loss	Accuracy	precision	Recall	FScore	CrossEntropy
InitialLearnRate=0.0001	1.8306	66.721	0.37573	0.56542	0.54929	1.8306
InitialLearnRate=0.001	1.5894	71.567	0.71676	0.63511	0.60322	1.5293
InitialLearnRate=0.01	1.7777	68.821	0.43257	0.61652	0.59178	1.7776

En la Figuras 50, 51 y 52, se observa el comportamiento del modelo y la convergencia entre las épocas 20 y 30.

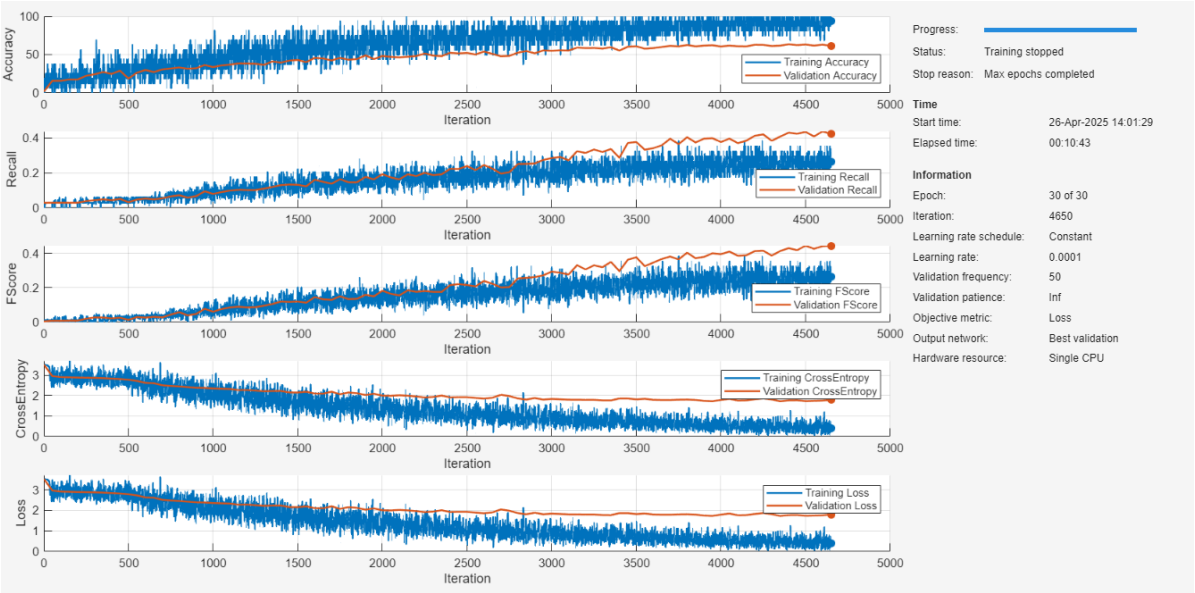


Figura 50. Entrenamiento y validación red LSTM con tasa de aprendizaje de 0.0001. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

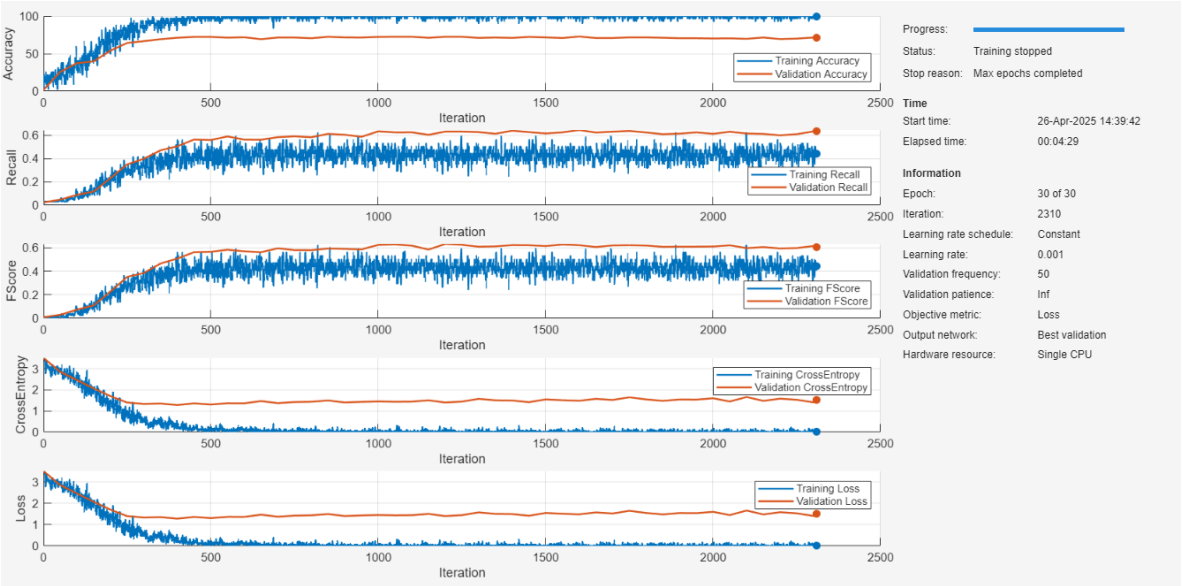


Figura 51. Entrenamiento y validación red LSTM con tasa de aprendizaje de 0.001. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

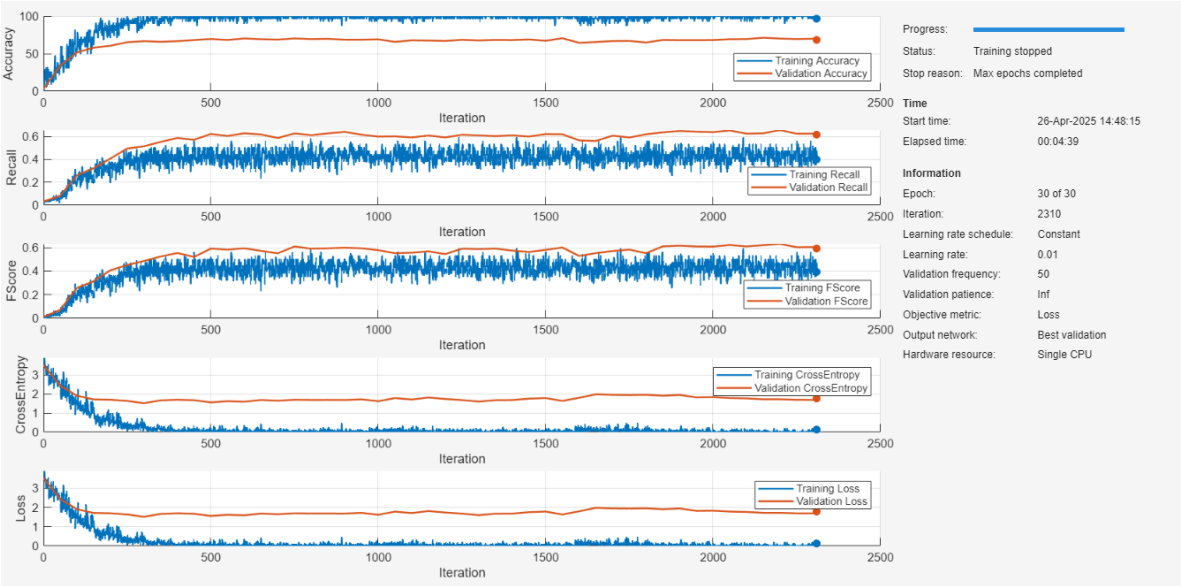


Figura 52. Entrenamiento y validación red LSTM con tasa de aprendizaje de 0.01 Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

De acuerdo a lo observado anteriormente, se puede establecer que el modelo funciona mejor para una tasa de aprendizaje de 0.001, por tanto para corregir el sobreajuste del modelo de clasificación, se procede a: disminuir a 60 las unidades ocultas del modelo LSTM; incluir la función *dropoutLayer* con la finalidad de desactivar aleatoriamente el 30% de las unidades ocultas del modelo de clasificación durante el entrenamiento, esto evita que la red neuronal recurrente dependa demasiado de ciertas neuronas y también se disminuye la cantidad de épocas a 15, para evitar que el modelo memorice en lugar de aprender los patrones. Los resultados obtenidos se muestran a continuación:

Tabla 13. Datos mejor resultado red LSTM.

Data modification	Loss	Accuracy	precision	Recall	FScore	CrossEntropy
InitialLearnRate=0.001 numHiddenUnits = 60 dropoutLayer(0.3) MaxEpochs=15	1.2393	**73.021**	0.72233	0.67178	0.67269	1.2393

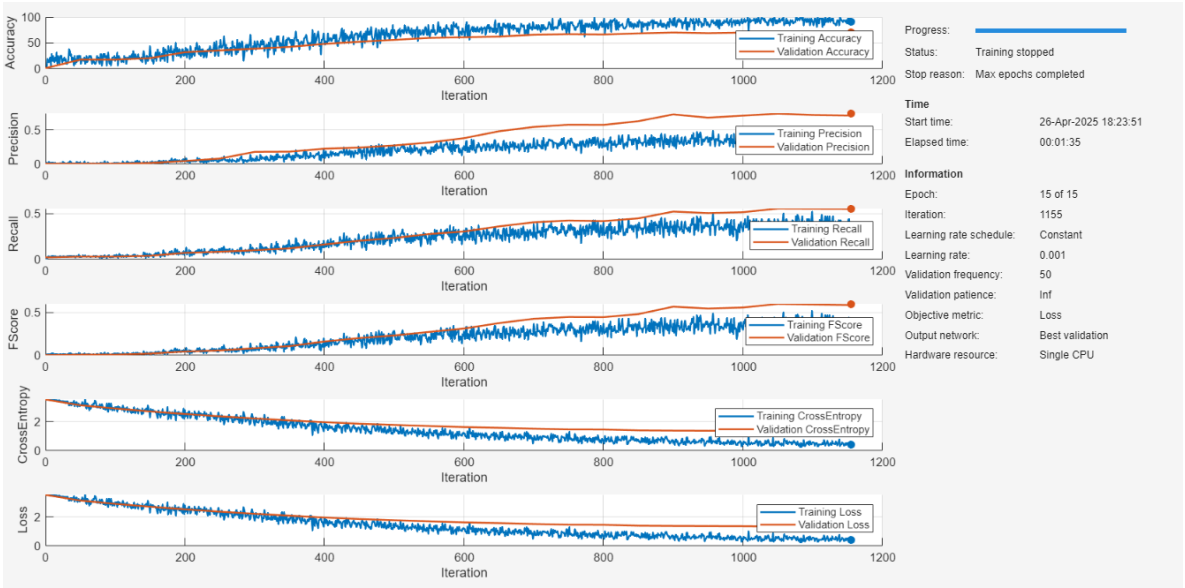


Figura 53. Entrenamiento y validación red LSTM con tasa de aprendizaje de 0.001. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

Como lo muestra la tabla 13, el modelo obtuvo una precisión del 73% en validación, lo que significa que el modelo clasificó correctamente el 73% de los datos de prueba, mientras que

el 27% de las predicciones fueron incorrectas; También disminuyó el valor de la entropía cruzada acercándose más a 1 y por tanto disminuyeron los errores en la predicción y finalmente se disminuyó en sobreajuste como se muestra en la Figura 53. Lo anterior indica que el modelo generaliza razonablemente bien, pero aún hay posibilidades de mejora. También se evidencia que las pérdidas en la validación disminuyeron, lo cual apoya la hipótesis de que un modelo de clasificación automática de texto en idioma indígena wayuunaiki puede incorporar características gramaticales.

En la Figura 54, la matriz de confusión muestra los aciertos y errores para cada clase, lo cual evidencia los patrones claves sobre la precisión de las predicciones por clase. En general se observó que las clases mayoritarias como la multiplicidad singular C2 el prefijo personal en segunda persona, pronombre personal el primera personal y las interjecciones, presentan altos valores, lo que indica un buen desempeño de la clasificación. Lo anterior teniendo en cuenta que se muestra una cantidad considerable de datos clasificados correctamente y la distribución de los errores se presentó con altos valores solo en dos clases.

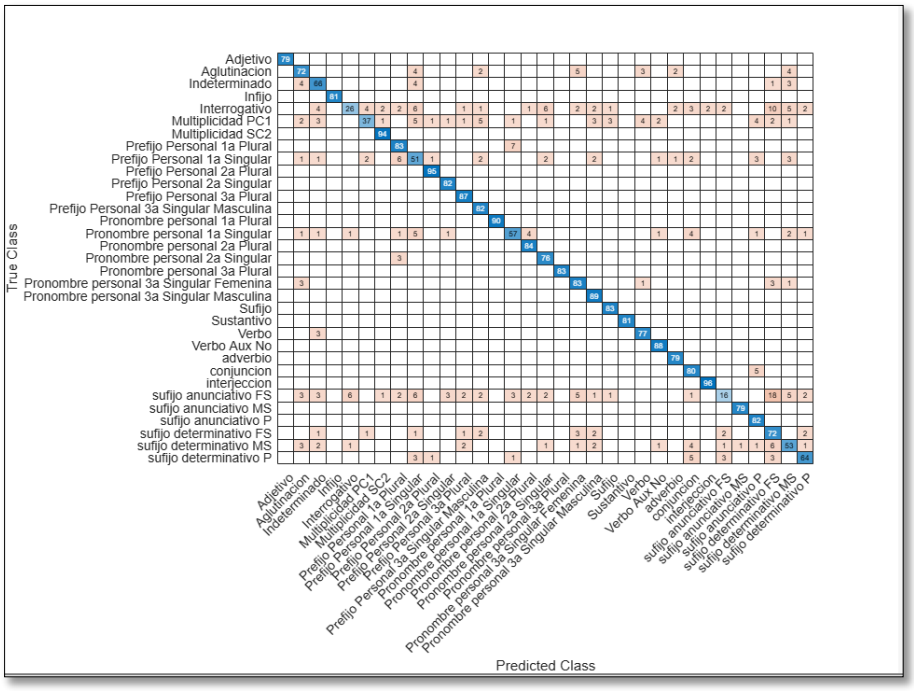


Figura 54. Matriz de confusión para clasificación de texto en wayuunaiki. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB.
Fuente: Elaboración propia.

Como conclusión parcial, en esta etapa se evidencio que para el conjunto de datos de validación y prueba, los resultados de la medición de métricas resultan buenos y admisibles, lo que demuestra un desempeño adecuado del modelo de clasificación automática de texto en idioma indígena wayuunaiki puede incorporar características gramaticales, pero también muestra que aún se puede mejorar.

9. ANALISIS DE RESULTADOS

En esta sección se describe el análisis de los resultados comparando los resultados del análisis exploratorio con los resultados de validación y de prueba con la finalidad de mostrar el análisis del modelo de clasificación de texto en idioma indígena wayuunaiki que incorpora características gramaticales.

En el análisis exploratorio utilizando un diagrama de calor (heatmap) se mostró la similitud entre palabras con la finalidad de identificar información sobre contextos similares, significado o uso similar, anomalías en palabras, deducción de temáticas, de esos análisis se logró identificar la representación de algunas características gramaticales, las cuales coinciden con los resultados obtenidos en la fase de validación y pruebas del modelo. Estos resultados obtenidos reflejan la distribución de las clases (características gramaticales) y la frecuencia con la que aparecen en el conjunto de datos. La Figura 55, presenta esta distribución, la cual aporta para la identificación de patrones y tendencias relevantes para el análisis.

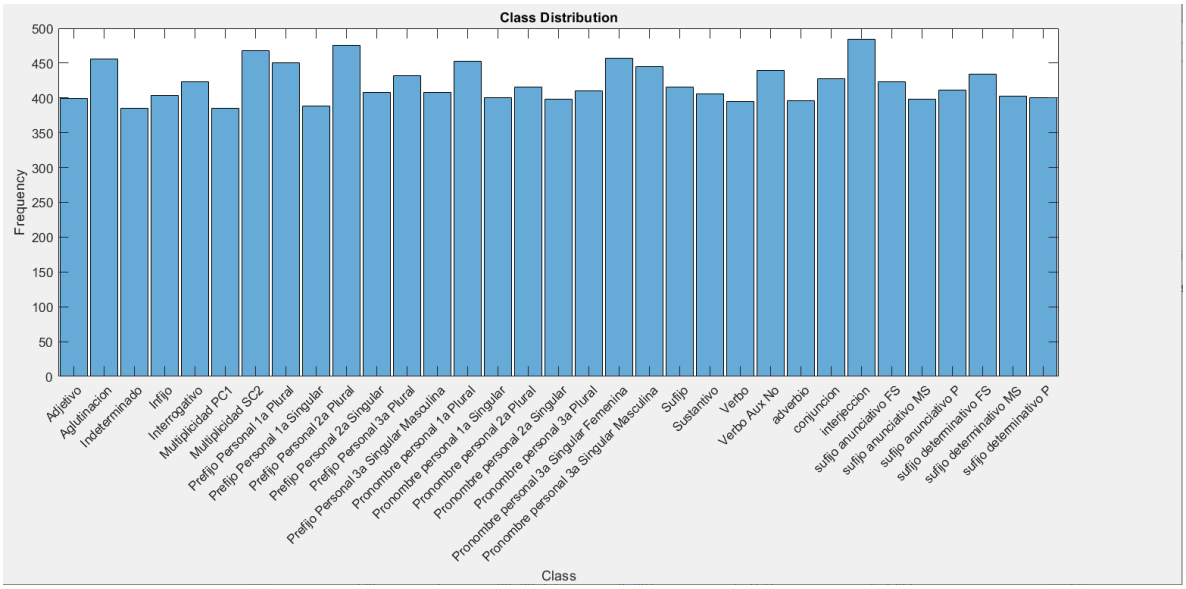


Figura 55. Distribución clases del modelo LSTM.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

En la Figura 17, se muestran la nube de palabras generada en el análisis exploratorio y en la Figura 56, se muestra la nube de palabras generada con el modelo de clasificación automática de texto; En las dos nubes de palabras se evidencia la similitud de palabras que corresponden a algunas características gramaticales del idioma Wayuunaiki.



Figura 56. Nube de palabras obtenida con el modelo LSTM.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

Los resultados obtenidos en la etapa de validación del modelo, muestran que gracias a los ajustes realizados a la arquitectura de la red, con la finalidad de disminuir el sobreajuste, generaron que el desempeño del modelo sea mucho mejor. Igualmente se evidencia que las pérdidas en la validación disminuyeron y aumento la exactitud del modelo, reflejando que el modelo de clasificación automática de texto en idioma indígena wayuunaiki puede incorporar características gramaticales. Funcionamiento que se comprueba en la fase de pruebas al ingresar datos nuevos sin etiquetar al modelo y obteniendo como respuesta las etiquetas para el nuevo texto.

La red Neuronal Recurrente (RNN) de tipo LSTM (Long short-term memory) para el problema de clasificación de texto donde los patrones secuenciales pueden tener distintas probabilidades de ocurrir a través del tiempo. Igualmente el entrenamiento de la red LSTM puede tomar bastante tiempo de procesamiento, pero logra compensar velocidad en relación con la cantidad de máximos y mínimos que ignora.

Como conclusión parcial, es posible afirmar que en general el modelo de clasificación de texto en idioma wayuunaiki que incorpora características gramaticales que utiliza una Red

de tipo LSTM, muestra una buena efectividad para detectar, clasificar y aprender nuevos patrones en texto en idioma wayuunaiki sin etiquetar, lo anterior de acuerdo a la comparación de resultados del análisis exploratorio con los resultados de validación y de pruebas.

10. IMPLANTACIÓN DEL MODELO.

En esta fase, se describirán algunas nociones para implementar el modelo en un entorno para su uso:

Plan de Implementación: Para garantizar que el modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales, esté accesible para los usuarios que puedan aprovechar su funcionalidad, se guarda el modelo en formato (.mat) para futuras ejecuciones (save('modeloClasificacion.mat', 'net', 'enc') como lo muestra la Figura 57 y también se genera una función para recibir texto y devolver la clasificación.

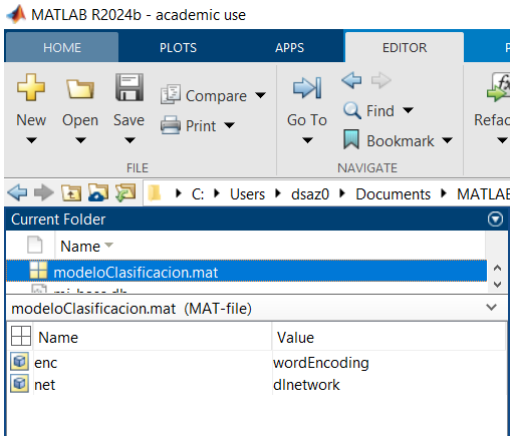


Figura 57. Empaquetado del modelo en formato.mat.
Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

Plan de monitoreo y mantención: Para monitorear el modelo de clasificación automática, evaluar su precisión con datos nuevos sin etiquetar y comparar las métricas, se propone realizar experimentos de prueba del modelo, por ejemplo, se pueden ingresan datos nuevos sin etiquetar procedentes del 10% del conjunto de datos que se estableció para pruebas y analizar los resultados obtenidos. En la Figura 58 se muestra los datos ingresados y en la Figura 59 se muestran los resultados de etiquetas para esos datos.

```
91
92      phraseNew = [ ...
93          "nnojotsü süpülain eme'erajawaa"
94          "Kasa shaitaa kamalainka sümüin"
95          "Nnojolu pu'unuin."];
96
97      documentsNew = preprocessText(phraseNew);
98
```

```
Command Window
  labelsNew =

    3×1 categorical array

       sufijo anunciativo FS
       Interrogativo
       Verbo Aux no
```

Figura 58. Resultados prueba del modelo LSTM con tasa de aprendizaje de 0.0001. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

También, en la Figura 59, se muestra el análisis de la red utilizada para el modelo, generado por MATLAB, el cual se puede incluir en el plan de monitoreo.

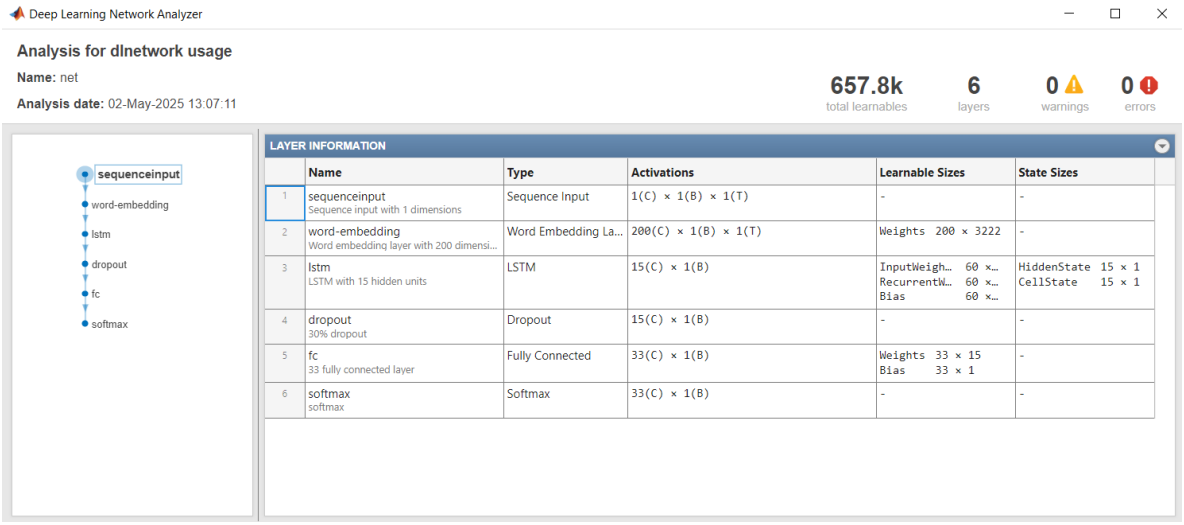


Figura 59. Análisis de la red LSTM, generado por MATLAB. Fuente: La imagen fue generada a partir de los resultados obtenidos mediante un modelo LSTM en MATLAB. Elaboración propia.

Informe final: Las conclusiones del proyecto, principales aportes, recomendaciones y trabajos futuros, se describen en los próximos capítulos de este documento.

11. CONCLUSIONES.

A través del desarrollo metodológico, la investigación logro alcanzar su objetivo general, culminando el diseño de un modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales. Este modelo utilizando Deep Learnig en específico la red neuronal recurrente (RNN) denominada LSTM (Long Short-Term Memory), red que demostró su efectividad para clasificar texto que presenta características gramaticales propias del idioma indígena wayuunaiki.

En la fase de preparación de los datos, la recolección de documentos en idioma wayuunaiki, elaborados por investigadores lingüistas en wayuunaiki (fuentes confiables), permitió la creación de un conjunto de datos que represente la estructura gramatical y la terminología del idioma wayuunaiki. Es importante que el conjunto de datos sea revisado por un hablante del idioma Wayuunaiki y/o por un traductor oficial de preferencia un lingüista, esto con la finalidad de validar el etiquetado del texto de acuerdo a las características gramaticales y la terminología que refleje la cultura y el conocimiento ancestral de los wayuu. También es importante para lograr una buena representación e incorporación de las características gramaticales realizar una adecuada limpieza, trasformación, enriquecimiento y representación del conjunto de datos antes y después del análisis exploratorio.

En la fase de determinación del modelo, se comprobó que la vectorización de los datos en wayuunaiki, mejora en gran medida cuando se utiliza el modelo Seq2Seq (Sequence to Sequence). Igualmente, se realizó una primera aproximación al modelo de clasificación, mediante el análisis de los gráficos generados en el análisis estadístico, los gráficos del análisis exploratorio, el diseño conceptual del modelo y la comparación de diferentes arquitecturas. En el modelo conceptual propuesto, muestra la integración de las características gramaticales al clasificador de texto en idioma wayuunaiki y los experimentos realizados para la comparación de arquitecturas, mostraron que la red STML es la mejor opción para el diseño del modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales.

En la fase de diseño de modelo, se aplicó la metodología de desarrollo CRISP-DM y se adaptaron las capas de datos, lógicas y de presentación, con la finalidad de obtener un

enfoque más estructurado de modelo de clasificación. También en esta etapa se elaboró el diseño de la estructura del modelo para la clasificación automática de texto y se establecieron los parámetros e hiperparametros que permitan incluir las principales características gramaticales del idioma wayuunaiki en la Red LSTM.

En la fase de validación y pruebas, para demostrar la veracidad de la hipótesis se realizaron tres experimentos cambiando la tasa de aprendizaje con la finalidad de mejorar el desempeño y disminuir el sobreajuste del modelo con STLM, ajustando parámetros e hiperparametros. Del conjunto de datos, se utilizó el 70% para entrenamiento, 20% para la validación y 10% para las pruebas. Se analizaron los resultados de entrenamiento y validación de las métricas de Exactitud, precisión, recall, F1-score e igualmente se utilizó la validación cruzada para la validación. Para la prueba del modelo se ingresaron frases en wayuunaiki sin etiquetar y se obtuvo como resultado la clase gramatical a la que pertenecen.

El modelo propuesto para la clasificación de texto en idioma wayuunaiki que incorpora las características gramaticales, obtuvo una exactitud del 73% en validación y un valor de entropía cruzada muy cerca 1. Lo anterior muestra que el modelo generaliza bien y que las pérdidas en la validación disminuyeron, sin embargo, existe la posibilidad de mejora del sobreajuste. Este sobreajuste sucede debido a las palabras que se repiten (Palabras iguales con significado distinto) que no se aconseja retirar del conjunto de datos.

El análisis de métricas comprobó la efectividad del modelo LSTM de clasificación de texto en idioma wayuunaiki que incorpora características gramaticales, para detectar, clasificar y aprender nuevos patrones. Sin embargo, se identificaron oportunidades para mejorar el modelo que se consideraran como trabajo futuro.

12. PRINCIPALES APORTES.

La adecuada creación de un conjunto de datos de alta calidad en idioma indígena Wayuunaiki ofrece un importante aporte para el adecuado Procesamiento de Lenguaje Natural e igualmente para el análisis de datos, la investigación científica, el desarrollo de aplicaciones y entrenamiento de modelos de machine learning y deep learnig entre otros. La ciencia de datos con adecuados conjuntos de datos que respeten la gramática, identidad cultural y la tradición oral del pueblo wayuu contribuyen significativamente a la conservación y divulgación del idioma Wayuunaiki, idioma que con el paso del tiempo se encuentra en peligro de desaparecer.

Este modelo podrá contribuir a otras investigaciones relacionadas con el Procesamiento de Lenguaje Natural del idioma indígena wayuunaiki, como por ejemplo detección de sentimientos, traducciones automáticas, etc. Lo anterior teniendo en cuenta que este modelo incorpora las principales características del idioma wayuunaiki que hasta el momento no se han tenido en cuenta en otras investigaciones. En este sentido se podría considerar que la utilidad de este proyecto puede categorizarse en apropiación social del conocimiento, ya que con el modelo de clasificación automática de texto en idioma indígena wayuunaiki que incorpora características gramaticales, se espera que los investigadores de PLN desarrollen estudios más rigurosos y precisos, que faciliten la integración de las comunidades Wayuu en la sociedad del conocimiento en su propio idioma.

Socialización de la investigación del proceso de creación y análisis exploratorio del conjunto de datos en idioma indígena wayuunaiki, en un evento científico de carácter internacional. Esto ayudo a la visualización de la importancia de generar un conjunto de datos de alta calidad gramatical en idioma wayuunaiki para el adecuado procesamiento de lenguaje natural (PLN) y en general para la ciencia de datos, la investigación científica, el desarrollo de aplicaciones y entrenamiento de modelos de machine learning y deep learnig.

Para la adecuada preparación de los datos, se creó para el proceso de lematización, un diccionario de lemas en Wayuunaiki, El cual, hasta ahora, no ha sido desarrollado en investigaciones previas. Este diccionario se elaboró con la finalidad de identificar por

ejemplo la multiplicidad de infinitivos, característica gramatical que presenta un gran reto al momento de identificar de la forma raíz de la palabra, debido a que los verbos pueden presentar diversas formas infinitivas dependiendo de los morfemas que la componen.

El diseño del modelo propone para el Procesamiento de Lenguaje Natural del idioma wayuunaiki, la adecuación de la arquitectura de red neuronal recurrente de memoria a corto-largo plazo (LSTM) con mecanismo de atención y para la representación del texto en wayuunaiki en vectores de palabras, propone la adecuación de modelo Word2Vec y el modelo Seq2Seq (Sequence to Sequence). Estas adecuaciones proponen incluir en su estructura las particularidades de la gramática del idioma indígena wayuunaiki.

El modelo de clasificación automática de texto en idioma indígena wayuunaiki, posee una estructura que permite identificar las principales características gramaticales del idioma wayuunaiki utilizando Deep Learning, características como: sustantivos que se expresan con un sufijo, pronombres demostrativos, interrogativos, pronombres y prefijos personales, diminutivos, indeterminados, verbos auxiliares, etc., relacionadas con la aglutinación, multiplicidad de infinitivos, flexionadas de nombres y preposiciones, gradación verbal, sufijos temáticos, morfemas y en algunos casos exepcionales yuxtaposición.

La metodología propuesta de esta investigación puede ser aplicable a otras investigaciones de procesamiento de lenguaje natural de culturas indígenas colombianas y latinoamericanas que presenten similares características gramaticales.

El modelo para clasificación de texto en idioma wayuunaiki elaborado en este proyecto de investigación permite la flexibilidad y escalabilidad para la realización de diferentes tareas de clasificación de texto sin una reconfiguración significativas, lo que aportar considerablemente al procesamiento de lenguaje natural de un idioma que tiene limitados conjuntos de datos de buena calidad y avalados por traductores oficiales de la comunidad indígena.

13. RECOMENDACIONES Y FUTUROS TRABAJOS.

13.1 RECOMENDACIONES

Para la clasificación de texto y el desarrollo de otras tareas de procesamiento de lenguaje natural del idioma indígena wayuunaiki que incorpore características gramaticales se recomienda lo siguiente:

- Para una futura implementación se recomienda realizar un análisis estadístico y exploratorio de acuerdo a la tarea de procesamiento de lenguaje natural que este proyectada para realizar.
- Utilizar fuentes de datos elaboradas y/ o revisadas por expertos en el idioma wayuunaiki.
- No eliminar palabras repetidas, consonantes o vocales, estas aportan a la estructura gramatical del idioma wayuunaiki.
- Identificar las palabras que tengan la misma estructura morfológica pero que su significado cambia de acuerdo al contexto de la frase.
- Contar con el acompañamiento de un experto en el idioma (traductor y/o lingüista). Para recibir una orientación sobre gramática y aspectos culturales.
- Tener en cuenta la complejidad morfología y fonéticas del idioma al momento de identificar patrones.
- Usar GPUs (Unidades de Procesamiento Gráfico), para acelerar realizar tareas de forma más rápida y eficiente.
- El algoritmo de aprendizaje automático, permita un adecuado procesamiento de conjuntos de datos pequeños.
- Generar representaciones semánticas eficientes para conjuntos de datos pequeños, por ejemplo utilizar embeddings preentrenados de palabras como Word2Vec y FastText.
- Para un procesamiento eficiente, se recomienda utilizar técnicas de generalización adecuada, como por ejemplo la validación cruzada (K-Fold), esta técnica para el

entrenamiento divide en conjunto de datos en subconjuntos, mejorando la robustez y una técnica de regularización como Dropout.

13.2 FUTUROS TRABAJOS

Los futuros trabajos relacionados con esta investigación, se presentan a continuación:

- Investigación de otras tareas de procesamiento de lenguaje natural del idioma indígena wayuunaiki.
- Análisis exploratorio de los conjuntos de datos en idioma wayuunaiki, de acuerdo a las necesidades de la tarea de procesamiento de lenguaje natural que se decida abordar.
- Adición de una variable para las frases que contengan palabras hibridas compuestas por el idioma español y Wayuunaiki.
- Diseño de una red hibrida de LSTM con Transformer, para identificar los detalles específicos del contexto de las frases escritas en wayuunaiki.
- Expansión del conjunto de datos con nuevos textos elaborados por investigadores de la gramática del idioma wayuunaiki.
- Aplicación del modelo para un conjunto de datos compuesto por de audios de frases en wayuunaiki.

14. REFERENCIAS

Abarna, S., Sheeba, J. I., & Devaneyan, S. P. (2022). An ensemble model for idioms and literal text classification using knowledge-enabled BERT in deep learning. *Measurement: Sensors*, *24*, 100434. https://doi.org/10.1016/J.MEASEN.2022.100434

Alonso Berrocal, J. L., Domínguez Ollero, A., Figuerola Paniagua, C. G., Vázquez de Aldana, E. R., & Zazo Rodríguez, Á. F. (1996). Consideraciones sobre el desarrollo de un sistema experto para la traducción automatizada. TRANS: revista de traductología, (1), 59–72.

Álvarez González, J. R. (2017). *Manual de la lengua wayuu (KARALOUTA ATÜJAAYA SAA'U WAYUUNAIKIKUWA'IPA)* (Organización Indígena de la Guajira Yanama, Ed.).

Álvarez, J. L. (2008). *Problemas de lematización verbal en una lengua aglutinante con infinitivos múltiples: el caso de guajiro/wayuunaiki*. Alicante : Biblioteca Virtual Miguel de Cervantes, 2008. https://www.cervantesvirtual.com/nd/ark:/59851/bmcrx9s2

Álvarez González, J. R. (2017). *Compendio de la gramática de la lengua wayuu*.

Captain, David y Linda Captain. (2005). Diccionario Básico Ilustrado wayuunaiki-español. Bogotá: Editorial Fundación para el Desarrollo de los Pueblos Marginados.

Casañ, G A; Castaño, M A; (2004). Generación automática de codificaciones de vocabularios para Traducción Automática mediante Redes Neuronales. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, 8() 175-186. Recuperado de http://www.redalyc.org/articulo.oa?id=92582213

Castaño Álvarez, A. (1998). Redes neuronales recurrentes para inferencia gramatical y traducción automática. Universitat Politècnica de València, España.

Castillo Romero, J. A. (2019). Big data. IFCT128PO. España: IC Editorial.

Chen, H., Zhang, Z., Huang, S., Hu, J., Ni, W., & Liu, J. (2023). TextCNN-based ensemble learning model for Japanese Text Multi-classification. *Computers and Electrical Engineering*, *109*, 108751. https://doi.org/10.1016/J.COMPELECENG.2023.108751

Chollet, F. (2018). *Deep Learning with Python*. Manning Publications.

Computadores para educar. (2021, July 8). *Llegaron computadores a La Guajira y Cesar*. https://www.computadoresparaeducar.gov.co/publicaciones/5219/llegaron-computadores-a-la-guajira-y-cesar/

Computadores para educar. (2024). Informe de gestión de Computadores para Educar 2014 (p. 33). Gobierno Nacional.

Datitos. (2021, 17 de mayo). Redes recurrentes modernas: LSTM, GRU y RNNs bidireccionales [Video]. YouTube. https://www.youtube.com/watch?v=5EyN7v-vIr4

Departamento Administrativo Nacional de Estadística (DANE). (2021). Información sociodemográfica del pueblo Wayúu (Vol. 3). En Informes de Estadística Sociodemográfica Aplicada. https://www.dane.gov.co/files/investigaciones/poblacion/informes-estadisticas-sociodemograficas/2021-09-24-Registro-Estadistico-Pueblo-Wayuu.pdf

Departamento Administrativo Nacional de Estadística (DANE). (2024). *Visor Wayuu*. https://sitios.dane.gov.co/visor-wayuu/

Departamento Administrativo Nacional de Estadística (DANE). (2025). Sistema de Información Wayuu Recuperado el 4 de enero de 2025, de https://siwayuu.dane.gov.co/es/home#tableroDeDatos

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Deng, S., Li, Q., Dai, R., Wei, S., Wu, D., He, Y., & Wu, X. (2024). A Chinese power text classification algorithm based on deep active learning. *Applied Soft Computing*, *150*, 111067. https://doi.org/10.1016/J.ASOC.2023.111067

El wayuunaiki, una lengua con 600.000 hablantes en riesgo de extinción | Noticias ONU. (n.d.). Retrieved February 23, 2024, from https://news.un.org/es/story/2019/03/1452751

Elnagar, A., Al-Debsi, R., & Einea, O. (2020). Arabic text classification using deep learning models. *Information Processing & Management*, *57*(1), 102121. https://doi.org/10.1016/J.IPM.2019.102121

Escobar, X. (2009). Traducción automatizada basada en algoritmos de redes neuronales artificiales (Te). Universitat Rovira I Virgili, España. Recuperado a partir de http://deeea.urv.cat/public/PROPOSTES/pub/pdf/1411pub.pdf

Fernández, D. I., Gamboa, O. Q., Atencia, J. M., & Bedoya, O. E. H. (2013). Design and implementation of an "Web API" for the automatic translation Colombia's language pairs: Spanish-Wayuunaiki case. *2013 IEEE Colombian Conference on Communications and Computing, COLCOM 2013 - Conference Proceedings*. https://doi.org/10.1109/COLCOMCON.2013.6564817

Firican, G. (2021). The history of machine learning. LightsOnData. Retrieved July 31, 2024, from https://www.lightsondata.com/the-history-of-machine-learning/

Gegúndez, M., & Pérez, I. (2021). Deep learning : fundamentos, teoría y aplicación. *Deep Learning*.

Graichen, N., van Genabith, J., & España-Bonet, C. (2023). Enriching Wayúunaiki–Spanish Neural Machine Translation with Linguistic Information. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 67–83. https://doi.org/10.18653/V1/2023.AMERICASNLP-1.9

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Google Developers. (2025, 23 de mayo). Cómo obtener embeddings. Recuperado de https://developers.google.com/machine-learning/crash-course/embeddings/obtaining-embeddings?hl=es-419

Goldberg, Y. (2016). A Primer on Neural Network Models for Natural Language Processing. J. Artif. Intell. Res. (JAIR), 57, 345-420.

Goulet, J. G., & Jusayú, M. A. (1978). El idioma guajiro: sus fonemas, su ortografía y su morfología. Universidad Católica Andrés Bello, Instituto de Investigaciones Históricas, Centro de Lenguas Indígenas. Recuperado a partir de https://books.google.com.co/books?id=1InI7GTInQ0C

Gyanendra, V., & Rajesh, D. (2023). *Deep Learning: Theory, Architectures and Applications in Speech, Image and Language Processing*. Bentham Science Publishers.

Hernández, J., Ramírez, M., & Ferri, C. (2005). *Introducción a la minería de datos*. Pearson Prentice Hall.

Hernández, M., & Gómez, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. Revista Politécnica, 32(1), 87–96.

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 328–339). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1031

IBM. (2021, August 17). *Introducción al CRISP-DM - Documentación de IBM*. https://www.ibm.com/docs/es/spss-modeler/saas?topic=guide-introduction-crisp-dm

Jatnika, D., Bijaksana, M., & Ardiyanti, A. (2019). Word2Vec Model Analysis for Semantic Similarities in English Words. Procedia Computer Science, 157, 160–167. https://doi.org/10.1016/j.procs.2019.08.153

Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746–1751).

Ley Nª 115. Congreso de la República de Colombia, Santa Fe de Bogotá, Colombia. 8 de Febrero de 1994. Recuperado de http://www.mineducacion.gov.co/1621/articles-85906_archivo_pdf.pdf

Li, L., Parra Escartín, C., Way, A., & Liu, Q. (2016). Combining translation memories and statistical machine translation using sparse features. Machine Translation, 30(3), 183–202. https://doi.org/10.1007/s10590-016-9187-6

Li, S., Deng, M., Shao, Z., Chen, X., & Zheng, Y. (2023). Automatic classification of interactive texts in online collaborative discussion based on multi-feature fusion. Computers and Electrical Engineering, 107, 108648. https://doi.org/10.1016/J.COMPELECENG.2023.108648

Llerena García, E. (2013). Software traductor de español a lengua wayuu. In Actualizaciones en Comunicación Social Centro de Lingüística Aplicada (pp. 353-356). Santiago de Cuba.

Llerena García, E. (2012). Software traductor de español a lengua kuna. Presentado en XXVII Congreso Nacional y I Internacional de Lingüística, Literatura y Semiótica, Universidad Pedagógica y Tecnológica de Colombia, Sede Tunja.

Matich, D. J. (2011). Redes Neuronales: Conceptos Básicos y Aplicaciones. Recuperado a partir de https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monografias/matich-redesneuronales.pdf

Maestre Salas, D. (2022). *Acercamiento al aprendizaje del Wayuunaiki como segunda lengua apoyado en el software lúdico "Wopü Ekirajawaa" en estudiantes del grado segundo*. https://manglar.uninorte.edu.co/handle/10584/11595

Mejía Rodríguez, P. I. (2011). *Situación sociolingüística del wayuunaiki: ranchería El Pasito* [Universidad Nacional de Colombia]. https://repositorio.unal.edu.co/handle/unal/8607

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Mintic. (2017, November 2). *Con energía solar y Tabletas para Educar, MinTIC llegó a 10 departamentos del país*. https://mintic.gov.co/portal/inicio/Sala-de-prensa/Noticias/61463:Con-energia-solar-y-Tabletas-para-Educar-MinTIC-llego-a-10-departamentos-del-pais

MinTIC. (2017). *Fotografía de entrega de tabletas en La Guajira* [Fotografía]. MinTIC. https://mintic.gov.co/portal/inicio/Sala-de-prensa/Noticias/61463:Con-energia-solar-y-Tabletas-para-Educar-MinTIC-llego-a-10-departamentos-del-pais

Muñoz Guillena, R. (2024, 19 de septiembre). Procesamiento del lenguaje natural como eje central de la inteligencia artificial generativa [Lección inaugural del curso académico 2024–2025]. Universidad de La Rioja

Pérez van Leenden, F. J. (2020). El desarrollo del léxico en wayuunaiki y en español de La Guajira colombiana -niños wayuu entre 08-15 y 15-30 meses-. *Entretextos: Revista de Estudios Interculturales Desde Latinoamérica y El Caribe, ISSN-e 2805-6159, ISSN 0123-9333, Vol. 14, Nº. 26, 2020 (Ejemplar Dedicado a: Homenaje a Francisco Justo Pérez van-Leenden (Q.E.P.D.) (Enero-Junio)), Págs. 12-32*, *14*(26), 12–32. https://dialnet.unirioja.es/servlet/articulo?codigo=8729502&info=resumen&idioma=ENG

Quero Ramones, Sandra, & Madueño Madueño, Leonel. (2006). Süchiki Walekerü: un ejemplo del uso de las TICs en escuelas indígenas. Caso Wayuu. Educere, 10(34), 435-442. Recuperado en 08 de agosto de 2024, de http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S1316-49102006000300006&lng=es&tlng=es

Radford, A., et al. (2020). *Language models are few-shot learners*. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877-1901).

Ramírez González, R., & Dieck Novial, M. E. (2019). Los pronombres personales en Wayuunaiki. Lenguaje, 47(2S). https://doi.org/10.25100/lenguaje.v47i3.6824

Rodrawangpai, B., & Daungjaiboon, W. (2022). Improving text classification with transformers and layer normalization. *Machine Learning with Applications*, *10*, 100403. https://doi.org/10.1016/J.MLWA.2022.100403

Saint-Exupéry, A. de. (2016). *Pürinsipechonkai*. The Jean-Marc Probst Foundation. (J. Álvarez, A. Dorado, J. A. Fernández & E. Montiel, Trads.).

Sanango, E. U., Hallo, M. A. (2024). Aplicación de la Metodología CRISP-DM en la Detección de Necesidades Ciudadanas en base al Análisis de Tweets. En N. Callaos, J. Horne, E. F. Ruiz-Ledesma, B. Sánchez, A. Tremante (Eds.), *Memorias de la Décima Cuarta Conferencia Iberoamericana de Complejidad, Informática y Cibernética: CICIC 2024,* pp. 150-156. International Institute of Informatics and Cybernetics. https://doi.org/10.54808/CICIC2024.01.150

Salazar, D., & Rodriguez, J. (2024, 21-25 de octubre). Análisis exploratorio de frases en idioma indígena Wayuunaiki para el adecuado Procesamiento de Lenguaje Natural.

Suarez Suarez, H. M., & López Uriana, R. S. (2022). *Articulación de la oralidad Wayuu como eje transversal para la propuesta de transformación curricular del proyecto educativo comunitario para mejorar la calidad educativa en la institución etnoeducativa integral rural internado indígena de Puerto Estrella y Nuestra Señora de Fátima de Nazareth*. https://intellectum.unisabana.edu.co/handle/10818/50637

Sutskever, I., Vinyals, O., & Le, Q. (2014). Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, *4*.

The MathWorks Inc. (2021). *Classify Text Data Using Convolutional Neural Network*. https://www.mathworks.com/help//releases/R2021a/deeplearning/ug/classify-text-data-using-convolutional-neural-network.html

The mathworks inc. (2024-a). *Introducción a la memoria a corto-largo plazo (LSTM) - MATLAB & Simulink*. https://la.mathworks.com/discovery/lstm.html.

The MathWorks Inc. (2024-c). *Classify Text Data Using Convolutional Neural Network, Natick, Massachusetts: The MathWorks Inc.* https://la.mathworks.com/help/deeplearning/ug/classify-text-data-using-convolutional-neural-network.html

The MathWorks Inc. (2024-d). *Classify Text Data Using Deep Learning*. https://la.mathworks.com/help/deeplearning/ug/classify-text-data-using-deep-learning.html

The MathWorks Inc. (2024 -b). Exploratory Analysis of Data. https://la.mathworks.com/help/stats/exploratory-analysis-of-data.html

Vásquez, A. C., Vásquez, A. C., huerta, H. V., Quispe, J. P., & Huayna, A. M. (2009). Procesamiento de lenguaje natural. Revista de Investigación de Sistemas e Informática, 6(2), 45–54. https://revistasinvestigacion.unmsm.edu.pe/index.php/sistem/article/view/5923

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gómez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

Vázquez, J., & Constable, L. (2019). Redes Neurales Artificiales con Aprendizaje Profundo. *Wicc*.

Viloria Rodríguez, Leonel, De La Rosa Yacomelo, Johan, Ramírez González, Rudecindo, & García Segura, Daniela. (2022). Pronombres en wayuunaiki y español: una mirada contrastive. Íkala, Revista de Lenguaje y Cultura, 27(1), 153-173. Epub March 11, 2022.https://doi.org/10.17533/udea.ikala.v27n1a08

Wayús pueden leer en wayuunaiki sentencia que pide frenar desnutrición en La Guajira - Otras Ciudades - Colombia - ELTIEMPO.COM. (n.d.). Retrieved March 1, 2024, from https://www.eltiempo.com/colombia/otras-ciudades/wayus-pueden-leer-en-wayuunaiki-sentencia-que-pide-frenar-desnutricion-en-la-guajira-837019