UNIVERSIDAD IBEROAMERICANA

Estudios con Reconocimiento de Validez Oficial por Decreto Presidencial
Del 3 de abril de 1981

"Análisis de Sentimientos en Redes Sociales Usando Información Multimodal"

TESIS

Que para obtener el grado de
Doctor en Ciencias de la Ingeniería

Presenta
Luis Norberto Zúñiga Morales

Bajo la dirección de
Dr. Jorge Ángel González Ordiano
Dr. José Emilio Quiroz Ibarra
Dr. Steven J. Simske

Lectores:
Dra. Katya Rodríguez Vázquez
Dr. Andrés Guillermo Molano Jiménez
Dr. Carlos Francisco Betancourt Moreno
Dr. Miguel Ángel Álvarez Carmona

Instituto de Investigación Aplicada y Tecnología

Ciudad de México, 2025

« *Our own values and desires influence our choices, from the data we choose to collect to the questions we ask. Models are opinions embedded in mathematics.* »

Cathy O'Neil

Análisis de Sentimientos en Redes Sociales Usando Información Multimodal
por Luis Norberto Zúñiga Morales

El análisis de sentimientos multimodal es una área emergente de estudio cuyo fin es determinar automáticamente la polaridad del sentimiento de documentos que contienen información multimodal como texto e imágenes. Desafortunadamente, el enfoque actual en los modelos enfocados a imágenes y texto se centran en el inglés y una única imagen, ignorando aspectos relevantes como trabajar con modelos que traten con múltiples imágenes, explorar opciones para modelar datos en idiomas distintos del inglés y estudiar el impacto que el contenido spam tiene en tales sistemas.

Para abordar las limitaciones mencionadas anteriormente, se propone un marco de trabajo para realizar análisis de sentimientos multimodal que incorpora texto, múltiples imágenes y texto en imágenes de tuits en español de dos conjuntos de datos que se construyen exclusivamente para esta labor: se anotan diferentes campos como el texto y las múltiples imágenes por separado y en conjunto, para obtener una mejor caracterización del sentimiento de cada elemento presente en una publicación.

El marco incluye un módulo de extracción de características con versiones optimizadas de la versión en español de Bidirectional Encoder Representations from Transformers y Vision Transformer para extraer características semánticas de texto e imágenes, respectivamente. Además, un módulo de detección de texto extrae el texto incrustado en imágenes y el módulo de fusión fusiona todas las modalidades mediante uno de dos enfoques propuestos: suma de vectores (fusión por suma) y mecanismos de autoatención (fusión por codificador).

Los hallazgos encontrados permiten concluir que los mejores resultados se obtienen al considerar únicamente imágenes y texto con la fusión por suma. Además, el número de imágenes, al variar en cada publicación, se vuelve un factor relevante en los modelos de clasificación. Por otro lado, el texto incrustado en imágenes no es importante para la tarea en cuestión. Con el conjunto de datos Multimodal Spanish Sentiment Analysis Impact Dataset se logra un 67.17 % de Coeficiente de Correlación de Matthews al considerar texto, la primera imagen y fusión por suma. Para el conjunto de datos Multimodal COVID19 Mexico se considera texto y hasta las primeras tres imágenes con fusión por suma, logrando un 94.26 % de Coeficiente de Correlación de Matthews. Finalmente, se observa una tendencia del spam a confundirse con la clase neutra.

Agradecimientos

A mis padres, cuyo apoyo y cariño me permitió participar y concluir este proyecto de cuatro años de forma satisfactoria. Sin ellos, nada de esto sería posible.

A mi hermana, mi cuñado y mi sobrino. Su apoyo demuestra la definición de la palabra familia de muchas formas.

A mis abuelos. Donde sea que se encuentren, espero que puedan verme.

A mis asesores de tesis, el Dr. Ángel, el Dr. Emilio y el Dr. Steve. Siempre he creído que realizar un posgrado es similar (hasta cierto punto) a contraer nupcias: debes encontrar a la persona correcta, si no, se vuelve un infierno. Ustedes permitieron el desarrollo armonioso y correcto del proyecto. Les estoy agradecido por permitirme esa paz.

A la Maestra Ana María, cuyo apoyo y constancia demuestran ser, día tras día, el pilar del posgrado en la Ibero. Sin usted, todo se desmoronaría.

Al Dr. César Villanueva y al Dr. Lázaro Bustio, quienes me permitieron participar en sus proyectos de investigación en la Ibero. Además, gracias por sus palabras de aliento y su confianza en mis habilidades.

Al Maestro Jorge Rivera, que me otorgó la confianza para formar parte de su equipo de trabajo desde el 2022. Gracias por esa oportunidad.

Finalmente, a todas las personas (incluidos mis alumnas y alumnos) con las que me he cruzado durante mi estancia en la Ibero y que, de alguna u otra forma, me han ayudado y acompañado en el proyecto. Desde comentarios hasta palabras de apoyo, siempre serán parte de este escrito.

Índice general

Resumen III

Agradecimientos V

1. Introducción 1
- 1.1. Motivación . . . 1
- 1.2. Justificación . . . 3
- 1.3. Estado del Arte . . . 5
  - 1.3.1. Conjuntos de Datos . . . 7
- 1.4. Planteamiento del Problema . . . 8
- 1.5. Objetivos . . . 9
- 1.6. Propuesta Metodológica . . . 10
- 1.7. Estructura del Documento . . . 10

2. Marco Teórico 13
- 2.1. Análisis de Sentimientos . . . 13
  - 2.1.1. Análisis de Sentimientos Tradicional . . . 14
  - 2.1.2. Análisis de Sentimientos en Imágenes . . . 17
  - 2.1.3. Análisis de Sentimientos Multimodal . . . 18
- 2.2. Transformer . . . 19
  - 2.2.1. Atención . . . 20
  - 2.2.2. Atención Multicabezal . . . 21
  - 2.2.3. El Bloque del Transformer . . . 23
  - 2.2.4. Arquitectura del Transformer . . . 23
- 2.3. Bidirectional Encoder Representations from Transformers . . . 24
- 2.4. Vision Transformer . . . 26
- 2.5. Modelos de Clasificación . . . 27
  - 2.5.1. Máquinas de Vectores de Soporte . . . 27
  - 2.5.2. Máquinas de Vectores de Soporte para el Caso Multiclase . . . 30
  - 2.5.3. Máquinas de Vectores de Soporte Sensible a Costos . . . 31
- 2.6. Métricas de Evaluación . . . 32
  - 2.6.1. Exactitud Balanceada . . . 32
  - 2.6.2. Coeficiente de Correlación de Matthews . . . 33
  - 2.6.3. Medida F_1 . . . 33

3. Metodología 35
- 3.1. Construcción de los Conjuntos de Datos . . . 35
- 3.2. Modelo Imagen y Texto . . . 37
  - 3.2.1. Extracción de Características en Textos . . . 37

3.2.2. Extracción de Características en Imágenes . . . 38
3.2.3. Detección de Texto en Imágenes . . . 38
3.2.4. Aumento de Datos en Texto . . . 39
3.2.5. Fusión de Información . . . 40
Fusión por Codificador . . . 40
Fusión por Suma . . . 41
3.2.6. Modelo de Clasificación . . . 42
3.3. Definición del Problema . . . 42
3.4. Detalles Sobre el Entrenamiento del Modelo Imagen y Texto . . . 43
3.4.1. Ajuste de los Modelos Base . . . 43
3.4.2. Ajuste de Hiperparámetros para la Fusión por Codificador . . . 44
3.4.3. Entrenamiento y Evaluación del Modelo . . . 44
3.5. Pasos Adicionales: Multimodal Spanish Sentiment Analysis Impact Dataset 45
3.5.1. Modelo Preliminar para Análisis Multimodal . . . 45
3.5.2. Métodos de Comparación . . . 46
3.5.3. Estudios de Ablación, Análisis Visual y Análisis de Error . . . 47
3.6. Pasos Adicionales: Multimodal COVID19 Mexico . . . 48
3.6.1. Pasos Adicionales . . . 48
3.6.2. Exploración de Datos . . . 48
3.7. Resumen Final . . . 48

4. Resultados y Análisis **51**
4.1. Conjuntos de Datos . . . 51
4.1.1. Multimodal Spanish Sentiment Analysis Impact Dataset . . . 51
4.1.2. Multimodal COVID19 Mexico . . . 56
4.1.3. Resumen de Resultados de la Sección . . . 57
4.2. Modelo de Detección de Texto en Imágenes . . . 61
4.3. Modelo Preliminar para Análisis Multimodal MSSAID . . . 62
4.4. Resultados Modelo Imagen y Texto: MSSAID . . . 64
4.4.1. Ajuste de los Modelos de Texto e Imágenes . . . 64
4.4.2. Ajuste de hiperparámetros para la Fusión por Codificador . . . 66
4.4.3. Resultados de Clasificación del Modelo Imagen y Texto . . . 68
4.4.4. Estudios de Ablación . . . 68
Impacto de las Modalidades . . . 68
Análisis del Impacto del Número de Imágenes en los Modelos de Clasificación Multimodal . . . 73
4.4.5. Análisis de Error . . . 76
4.4.6. Resumen de Resultados de la Sección . . . 79
4.5. Resultados Modelo Imagen y Texto: MCOVMEX . . . 79
4.5.1. Ajuste de los Modelos de Texto e Imágenes . . . 79
4.5.2. Ajuste de Hiperparámetros para la Fusión por Codificador . . . 80
4.5.3. Resultados de Clasificación del Modelo Imagen y Texto . . . 83
4.5.4. Análisis de Error . . . 83
4.5.5. Resumen de Resultados de la Sección . . . 85
4.6. Discusión Final . . . 86
4.7. Disponibilidad de Códigos . . . 88

5. Conclusiones **91**
5.1. Respuestas a las Preguntas de Investigación . . . 91
5.2. Trabajo Futuro . . . 93

A. Instrucciones de Anotación para el Multimodal COVID19 Mexico **95**

B. Resultados Completos Experimentos del Número Máximo de Imágenes en MSSAID **103**

C. Resultados Completos Experimentos del Número Máximo de Imágenes en MCOVMEX **105**

D. Proyecto Imagen de México **107**
D.1. Introducción . . . 107
D.2. Metodología . . . 107
D.2.1. Modelo Clásico . . . 108
Traducción y Aumento de Texto . . . 108
Procesamiento de Emojis . . . 109
Preprocesamiento y Procesamiento de Texto . . . 109
Representación de Texto . . . 110
Modelo de Clasificación . . . 110
D.2.2. Modelo Ajustado . . . 110
D.2.3. Entrenamiento de los Modelos . . . 110
D.3. Resultados . . . 110
D.4. Análisis de Resultados . . . 110

E. Productos Desarrollados y Participaciones Durante el Posgrado **117**
E.1. Publicaciones en Revistas JCR . . . 117
E.2. Memorias Completas en Congresos Internacionales . . . 117
E.3. Talleres . . . 117
E.4. Pósteres . . . 118

Bibliografía **119**

Índice de Figuras

1.1. Ejemplo de un tuit con información multimodal. 2

2.1. Diagrama de la idea general para realizar análisis de polaridad enfocado a textos. 15
2.2. Diagrama de la idea general para realizar análisis de polaridad multimodal. 18
2.3. Diagrama del cabezal de atención. 22
2.4. Esquema de la atención multicabezal con tres cabezas. 23
2.5. Diagrama de la estructura del bloque del Transformer. 24
2.6. Modelo de codificador y decodificador del Transformer. 25
2.7. Ejemplo de la arquitectura de BERT Base para procesar una entrada de texto. 26
2.8. Arquitectura de Vision Transformer. 27
2.9. Idea general para la MVS. 28
2.10. Ejemplos del margen suave y truco del kernel en la MVS. 29

3.1. Diagrama general del método para fusionar imágenes y texto para el análissi de sentimientos multimodal. 37
3.2. Ejemplo de una posible salida del sistema de detección de texto en imágenes. 39
3.3. Diagrama de la estrategia de fusión por codificador para tres elementos entrantes al sistema y su salida. 41
3.4. Diagrama de la estrategia de fusión por suma para tres elementos entrantes al sistema y su salida. 42
3.5. Esquema de los pasos adicionales que se requieren y se llevan a cabo para el entrenamiento del modelo propuesto de imagen y texto. 43
3.6. Diagrama general del primero proceso de clasificación multimodal. 46

4.1. Distribución de datos del MSSAID. 53
4.2. Diagrama de cajas mostrando la longitud de los tuits según el sentimiento del texto y el sentimiento general de un tuit en el MSSAID. 53
4.3. Diagrama de Sankey que muestra la transición del valor del sentimiento de los tuits para cada polaridad al considerar primero texto (izquierda), y después texto con imágenes (derecha) para el MSSAID. 54
4.4. Cantidad de tuits según el número de imágenes que contienen. 55
4.5. Distribución de datos del MCOVMEX. 57
4.6. Diagrama de cajas mostrando la longitud de los tuits según el sentimiento del texto y el sentimiento general de un tuit en el MCOVMEX. 58
4.7. Diagrama de Sankey que muestra la transición del valor del sentimiento de los tuits para cada polaridad al considerar primero texto (izquierda), y después texto con imágenes (derecha) para el MSSAID. 59

4.8. Cantidad de tuits en el MCOVMEX según el número de imágenes que contienen. 60
4.9. Ejemplo de éxitos típicos y errores comunes del sistema de extracción de texto en imágenes (detecciones incompletas). 61
4.10. Resultados del proceso de ajuste para el experimento de aumento de texto. 65
4.11. Mapas de calor generados según el CCM para los distintos modelos de clasificación multimodal al considerar diferentes combinaciones de cabezales de autoatención y número de capas. 67
4.12. Matrices de confusión para los mejores modelos de cada modalidad. . . . 71
4.13. Proyecciones 2D de los embeddings de los mejores modelos de cada modalidad. 72
4.14. Puntuaciones según el CCM para los distintos modelos de clasificación cambiando el número máximo de imágenes. 73
4.15. Proyecciones 2D de los mejores modelos para cada modalidad al considerar diferentes cantidades de imágenes. 74
4.16. Matriz de confusión y proyección 2D del mejor modelo MSSAID. 75
4.17. Mapas de calor generados según el CCM para los distintos modelos de clasificación multimodal y el MCOVMEX al considerar diferentes combinaciones de cabezales de autoatención y número de capas. 81
4.18. Resultados del sistema de clasificación multimodal al considerar diferentes cantidades de imágenes y métodos de fusión para el MCOVMEX. 84
4.19. Matriz de confusión y proyección 2D del mejor modelo MCOVMEX. . . . 84
4.20. Distribución del sentimiento general de las publicaciones de X según el mes y año de su publicación. 86

A.1. Ejemplo de una imagen con texto incrustado en ella. 96
A.2. Primer ejemplo del conjunto de datos. 97
A.3. Segundo ejemplo del conjunto de datos. 98
A.4. Tercer ejemplo del conjunto de datos. 99
A.5. Cuarto ejemplo del conjunto de datos. 100
A.6. Quinto ejemplo del conjunto de datos. 101

D.1. Frecuencia de cada tipología del conjunto de datos del MAIP ya considerando tuits y encabezados juntos. 108
D.2. Diagrama que muestra el proceso principal del modelo clásico para la clasificación de los datos. 109
D.3. Matrices de confusión normalizadas para (A) el mejor modelo ajustado y (B) el mejor modelo clásico, según la medida F_1^w. 112

Índice de Tablas

1.1. Conjuntos de datos comunes que se utilizan para la tarea de análisis de sentimientos de imágenes y texto, inspirado en [56]. 8

3.1. Campos anotados de cada conjunto de datos y su correspondiente descripción. 36
3.2. Información solicitada a la API v2 de Twitter para la construcción del conjunto de datos MCOVMEX antes de su clausura. 36
3.3. Comparación de diferencias metodológicas llevadas a cabo entre el conjunto de datos MSSAID y el MCOVMEX. 49

4.1. Resultados del modelo preliminar de clasificación multimodal. 63
4.2. Resultados del proceso de ajuste de diversos modelos de imágenes para el módulo de extracción de características en imágenes del MSSAID. 64
4.3. Resultados del proceso de ajuste de diversos modelos de texto para el módulo de extracción de características en texto. 64
4.4. Hiperparámetros seleccionados para cada combinación de los modelos ajustados en el MSSAID. 66
4.5. Métricas de desempeño de los distintos modelos de clasificación multimodal: CLIP Multilingüe, BETO y ViT base y BETO y ViT ajustados. . . . 68
4.6. Métricas de rendimiento para los distintos modelos de clasificación multimodal y la contribución de las distintas modalidades (MSSAID). 70
4.7. Tuits seleccionados para el análisis de error del MSSAID. 77
4.8. Resultados del proceso de ajuste de diversos modelos de imágenes para el módulo de extracción de características en textos del MCVOMEX. 80
4.9. Resultados del proceso de ajuste de diversos modelos de imágenes para el módulo de extracción de características en imágenes del MCVOMEX. . . 80
4.10. Hiperparámetros seleccionados para cada combinación de los modelos ajustados y número máximos de imágenes en el MCOVMEX. 82
4.11. Tuits seleccionados para el análisis de error del MCOVMEX. 85
4.12. Repositorios con los códigos utilizados en el presente trabajo del MSSAID y un demo con una herramienta desarrollada para el MCOVMEX con los resultados obtenidos. 89

B.1. Resultados completos de los experimentos del número máximo de imágenes con el conjunto de datos MSSAID. 103

C.1. Resultados completos de los experimentos del número máximo de imágenes con el conjunto de datos MCOVMEX. 105

D.1. Métricas de rendimiento de los experimentos de características adicionales del texto del modelo clásico con la MVS. 111
D.2. Métricas de rendimiento de los modelos clásicos y ajustados con el conjunto de entrenamiento según las técnicas de aumento de datos. 114
D.3. Comparación del rendimiento de clase de los mejores modelos para cada modelo de aprendizaje. 115

Lista de Abreviaturas

API	**A**pplication **P**rogramming **I**nterface
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransfomers
BP	**B**olsa de **P**alabras
CCM	**C**oeficiente de **C**orrelación de **M**atthews
CLIP	**C**ontrastive **L**anguage **I**mage **P**retraining
E	**E**mojis
HT	**H**ash**t**ags
LSTM	**L**ong-**S**hort **T**erm **M**emory
MAIP	**M**odelo **A**nalítico de **I**magen **P**aís
MCOVMEX	**M**ultimodal **C**OVID19 **M**exico
M-CLIP	**CLIP** **M**ultilingüe
MSSAID	**M**ultimodal **S**panish **S**entiment **A**nalysis **I**mpact **D**ataset
MVS	**M**áquina de **V**ectores de **S**oporte
ResNet	**R**esidual Neural **Net**work
ROC	**R**econocimiento **Ó**ptico de **C**aracteres
T	**T**exto (de una publicación)
TF-IDF	**T**erm **F**requency - **I**nverse **D**ocument **F**requency
T+I	**T**exto e **I**mágenes (de una publicación)
T+I+TI	**T**exto, **I**mágenes y **T**exto en **I**mágenes (de una publicación)
NU	**N**ombres de **U**suario
UMAP	**U**niform **M**anifold **A**pproximation and **P**rojection
ViT	**Vi**sion **T**ransfomer
YOLO	**Y**ou **O**nly **L**ook **O**nce

Lista de Símbolos

y	Constante
$\boldsymbol{x}$	Vector
M	Matriz
a_{ij}	Aspecto de u_i
h_k	Persona que alberga la opinión
s_{ijkl}	Sentimiento del aspecto a_{ij} de la entidad u_i
t_l	Tiempo cuando se expresa la opinión
u_i	Entidad de una opinión
D	Conjunto de datos D
n	Número de elementos del conjunto de datos D
n_c	Número de clases del conjunto de datos D
n_i	Número de instancias en el conjunto de datos para la clase i
d_e	Embedding del vector de entrada
$d_e^{\#}$	Embeddings de los vectores de entrada de cada palabra
K	Key
N_{d_e}	Dimensión del embedding d_e
N_{d_k}	Dimensión del embedding para Key y Query
N_{d_v}	Dimensión del embedding para Value
Q	Query
V	Value
W_K	Matriz de pesos correspondiente al Key
W_O	Matriz de pesos correspondiente a la salida
W_Q	Matriz de pesos correspondiente al Query
W_V	Matriz de pesos correspondiente al Value
$\#P$	Número de palabras que conforman el Key y Value
C	Penalización del margen suave en la MVS
γ	Parámetro del kernel de función de base radial para la MVS
k_C	Valor de $C = 2^{k_C}$ en la malla de búsqueda
k_γ	Valor de $\gamma = 2^{k_\gamma}$ en la malla de búsqueda
$\phi(x)$	Kernel aplicado a x
v_i	Vector de costos para la clase i
v_j	Vector de costos para la clase j
x_i	i-ésima característica del dato x
ξ_i	Variable de holgura del marge suave
y_i	Clase asociada al dato x_i
c	Número de muestras predichas correctamente
G	Matriz de confusión para la correlación de Matthews

F_1^w	Medida F1 ponderada
p_k	Número de veces que se predijo la clase k
t_k	Número de veces que ocurre la clase k
$\hat{w}_i$	Peso ajustado de la muestra
$\hat{y}_i$	Valor predicho de y_i
C	Número de canales de una imagen
d	Dimensión de los embeddings generados por BETO y ViT
e_j^t	Embedding de texto del j-ésimo token de entrada
e_j^{im}	Embedding de imagen del j-ésimo token de entrada
$e_j^{t_{im}}$	Embedding del texto en imagen del j-ésimo token de entrada
F	Función de fusión
F_{im}	Vector de características de imagen
$F_{im_j \setminus 1}$	Embedding de la j-ésima imagen sin el primer elemento del vector
F_{text}	Vector de características de texto
$F_{text_{im_k} \setminus 1}$	Embedding del k-ésimo texto de imagen sin el primer elemento del vector
$F_{text_{im}}$	Vector de características de texto en imágenes
H	Alto de una imagen
n_{im}	Cantidad de imágenes para la fusión por suma
n_t	Cantidad de textos para la fusión por suma
$n_{t_{im}}$	Cantidad de texto en imágenes para la fusión por suma
N_t	Número variable de tokens de la secuencia S_T
N_{im}	Número de tokens de la secuencia de parches
P	Tamaño del parche para Vision Transformer
ρ	Publicación (de una red social)
S_I	Imagen de una publicación
S_T	Secuencia de palabras
$S_{text_{im}}$	Suma de los tokens [CLS] correspondientes de F_{text}, F_{im} y $F_{text_{im}}$
w_i	Palabra i de una secuencia de palabras S_T al hablar de tokens y peso asociado a x_i al hablar de la MVS
W	Ancho de una imagen
$Z_{text_{im}}$	Concatenación de los vectores F_{text}, F_{im} y $F_{text_{im}}$

Capítulo 1

Introducción

1.1. Motivación

Las redes sociales (digitales) como Facebook, X, Instagram o TikTok, han permitido a los usuarios alrededor del mundo comunicar sus ideas, pensamientos, juicios y opiniones sobre diversos temas de interés mediante la generación e intercambio de contenido en dichas plataformas. Para el año 2025 se espera que las redes sociales reúnan a cerca de 5.42 mil millones de usuarios alrededor del mundo[1], de los cuales 94.09 millones serán mexicanos que accederán a alguna plataforma para compartir un punto de vista o mirar contenido para pasar el rato [1]. Debido a la continua evolución de estas plataformas y sus tecnologías, es necesario reconocer la importancia de las redes sociales: 1) son una gran fuente de información y 2) es prioritario crear sistemas que sean capaces de procesar tales volúmenes de referencias que se adecúen a los tipos de datos con los que se desee trabajar.

El desarrollo de la Web 2.0 promovió la generación de sitios web y herramientas que fomentaron la publicación individual y colectiva, el intercambio de imágenes, audio y vídeo, y la creación y mantenimiento de redes sociales en línea [2]. En consecuencia, los usuarios de redes sociales se encuentran constantemente interactuando con la información de su medio, ya sea consumiéndola, creándola o modificándola para darle nuevos usos. Por otro lado, el constante avance tecnológico de los teléfonos inteligentes y las redes de datos han permitido que los usuarios de redes sociales publiquen contenido en virtualmente cualquier lugar, desde la palma de su mano. En particular, el uso de teléfonos inteligentes ha sido una herramienta clave para la adopción y uso masivo de las redes sociales: por un lado se han vuelto una herramienta que ha facilitado la creación de contenido audiovisual (tomar fotos, grabar videos o editar archivos) sin preocuparse por los detalles más técnicos que esto implica. Además, se han convertido en un importante punto para diseminar el contenido creado por medio de distintas aplicaciones móviles [3].

En conjunto, estos factores permitieron a las redes sociales adoptar diversos elementos que facilitaron a los usuarios comunicarse de distintas maneras alternativas o complementarias al texto. Por ejemplo, al considerar el texto, se ha observado la adopción de nuevas formas de expresión lingüística como emojis, acrónimos, contracciones y neologismos [4] que rompen con esquemas tradicionales de escritura vistos en medios como la prensa escrita o digital. Igualmente, los elementos audiovisuales han ganado un rol más importante al ser capaces de expresar mensajes más complejos, como es el caso de las imágenes, capturas de pantalla, memes [5], animaciones cortas (GIFs) [6] y videos cortos [7]. Por lo tanto, las herramientas que los usuarios de redes sociales tienen a su disposición para

[1] https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/

FIGURA 1.1: Ejemplo de un tuit con información multimodal. Fuente: elaboración propia, tanto la imagen como el tuit, publicados en la cuenta personal del autor de X.

expresarse son multimodales, es decir, ya no están sujetos únicamente al texto para comunicarse. Sin embargo, esta valiosa información, ruidosa y sin estructura, es intrincada de analizar debido a la misma mezcla de modalidades presentes en una sola publicación. En primer lugar, es posible analizar cada elemento de una publicación por separado para determinar su intención en solitario o su contribución hacia la publicación final. En contraste, los distintos elementos multimodales interactúan entre ellos de distintas maneras, reforzando una misma idea dentro del mensaje o alterando por completo la intención final del discurso, como puede ser el caso de un meme cuyo fin es cambiar el sentido de una publicación como ejemplo de ironía, sarcasmo o ciberbullying. Lo anterior origina la necesidad de crear métodos de análisis que permitan automatizar el estudio de información que consideren dicha multimodalidad que caracteriza a las redes sociales.

Considere la imagen que se muestra en la Figura 1.1 donde se puede observar una publicación de X con información multimodal. Si se analiza únicamente el texto, podemos observar la presencia de un emoji (un elemento gráfico en el texto) que aporta cierta carga negativa al mensaje que se busca expresar. Sin embargo, éste, junto al texto, no aporta suficiente información como para descifrar la intención completa de la publicación. Al continuar con el siguiente elemento, podemos observar una imagen que resulta ser el

meme *Left Exit 12 Off Ramp*[2] cuyo fin es hacer énfasis en algo que el autor desaprueba (*hacer más experimentos*) y algo que preferiría hacer en su lugar (*terminar el capítulo de la tesis*). En este caso, se expresa con humor una situación común durante los periodos de posgrado donde los alumnos sufren al balancear tareas complejas que consumen bastante tiempo como la escritura del documento de la tesis y la realización de experimentos pendientes para su investigación, que suelen realizarse al mismo tiempo.

Al juntar todos los elementos presentes, se puede observar que la imagen complementa al texto aportando información adicional para determinar el objetivo principal que se desea plasmar en la publicación. En nuestro caso, es de particular importancia determinar la polaridad del sentimiento que se expresa en una publicación de una red social, pero este fin puede cambiar según la pregunta que se necesite responder.

Por lo tanto, el tema principal que aborda este estudio es el uso de información multimodal como lo son las imágenes y los distintos elementos que se pueden encontrar en ellas como los textos incrustados, para proponer nuevos métodos que mejoren el análisis de las publicaciones que los usuarios crean y publican en una red social. Además, los métodos propuestos deben permitir que dicho análisis se pueda hacer de forma automática: en diciembre de 2023, durante cada minuto, se enviaron aproximadamente 360,000 publicaciones en X (antes Twitter)[3], cantidad que rebasa por completo la capacidad sensorial humana para comprender y procesar de forma manual tal volumen de información.

1.2. Justificación

Del ejemplo mostrado en la Figura 1.1, surge el principal objeto de estudio de la investigación, que es analizar la polaridad del sentimiento que se expresa en una publicación de una red social. Por lo tanto, es importante empezar a introducir un concepto clave que nos permitirá construir un marco teórico y metodológico para situar el proyecto más adelante.

Definición 1 (Análisis de Sentimientos) *El análisis de sentimientos (tradicional), también llamado minería de opinión, busca construir herramientas que permitan extraer información subjetiva de fuentes de texto, tales como opiniones y sentimientos, para crear conocimiento estructurado de forma automática [8].*

La Definición 1 resalta que el análisis de sentimientos permite darle estructura a los datos no estructurados que forman una fuente de texto física o digital, tales como sitios web, entradas de blogs, redes sociales, periódicos, documentos de texto, etc. Por otro lado, se agrega el término *tradicional* a la definición ya que ésta se enfoca únicamente en fuentes de datos cuyo origen es texto.

El análisis de contenidos de redes sociales utilizando texto ha gozado de éxito en diversas aplicaciones. Por ejemplo, para la detección de depresión en redes sociales [9], analizar mensajes que refuerzan estereotipos femeninos negativos [10], estudiar el impacto de las noticias falsas en las redes sociales [11] o caracterizar la participación en las redes sociales con fines publicitarios [12]. Sin embargo, a pesar de su éxito, el análisis de sentimientos basado en texto presenta problemas que merman su efectividad. Por ejemplo,

[2] https://knowyourmeme.com/memes/left-exit-12-off-ramp

[3] https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/

la falta de gramática, vocabulario o sintaxis formales en las publicaciones complica las técnicas basadas en reglas y aprendizaje automático [13]. Por otro lado, el deseo de influir en las opiniones y decisiones conduce a contenido no deseado, como spam o reseñas falsas [14], lo que infla artificialmente los resultados. Además, el significado de las palabras y las oraciones puede variar según el contexto y el dominio en el que se utilizan, lo que genera problemas como la detección del sarcasmo [15], el manejo de la negación [16] y la desambiguación del sentido de las palabras [17]. Por último, los idiomas poco populares para su estudio suelen presentar escasez de recursos lingüísticos como lexicones, diccionarios o conjuntos de datos. Ya que los distintos enfoques para determinar el sentimiento dependen de tales recursos, se vuelve costoso acercarse a dichos idiomas debido a que se necesitan construir desde cero.

Como se mencionó anteriormente, los usuarios de redes sociales tienen acceso a herramientas para subir contenido audiovisual para complementar el texto de sus publicaciones. Aunque históricamente el desarrollo del análisis de sentimientos se ha enfocado en el análisis de textos, diversos esfuerzos se han realizado para extender estas nociones al contenido visual como imágenes y videos debido al creciente interés de utilizar tal información en diversas investigaciones [18]. Jurgenson [3] argumenta que las imágenes se han vuelto un lenguaje visual particularmente útil para expresar emociones, ideas, pensamientos y experiencias de la vida diaria que, potenciadas por las redes sociales y los teléfonos inteligentes, se pueden crear y difundir al momento. Además, es posible editarlas con software para añadirle significado o reciclarlas, dotándoles de nueva vida. El contenido visual permite una comunicación más rápida, global y directa, a diferencia de las palabras. Por ejemplo, es posible romper barreras presentes en el texto como el idioma mediante una imagen o video al alejarse del lenguaje natural e incorporar elementos que se enfoquen en el cuerpo humano, evoquen la memoria, o jueguen con la música. Banks [19] menciona que las imágenes son ubicuas en la sociedad y su interpretación va más allá de lo que muestra (el enfoque y su entorno): también incluye aquello que no se muestra, su contexto (social e histórico) y narrativa. Por lo tanto, el contenido visual resulta útil para entender el impacto que genera en la polaridad de una publicación al proveer no sólo información semántica de su contenido, sino también señales sobre el sentimiento que alberga el usuario y expresa a través de la publicación de dicha imagen o video.

El análisis de sentimientos se ha aplicado a elementos visuales como imágenes con mucho éxito [20], [21], [22], [23]. Desafortunadamente, también enfrenta ciertos retos. En primer lugar, el sentimiento presente en una imagen puede cambiar según el contexto. Por ejemplo, una imagen que muestra a dos personas gritando puede considerarse algo negativo. Sin embargo, si después añadimos que se trata de dos cantantes de ópera, el sentido de la imagen puede llegar a cambiar. Otro problema surge cuando se examinan las relaciones de intermodalidad entre imágenes y texto [24]. Algunos casos incluyen a los emojis, que son imágenes que se encuentran presentes en una fuente de texto; y el texto incrustado en imágenes, como los memes. En el caso particular de las imágenes que provienen de redes sociales, Chen et al. [18] mencionan dos problemas adicionales. En primer lugar, existe la posibilidad de presentarse un sesgo al usar este tipo de información ya que los datos no necesariamente representan a la demografía objetivo correctamente (por ejemplo, al elegir erróneamente la plataforma para la recolección), no incluyen información de aquellas personas que no emplean redes sociales o no consideran aquellas cuentas que bloquean el acceso a sus publicaciones. El segundo problema consiste en

retos éticos generados por el acceso consensuado al contenido visual de los usuarios, lo cual puede incluir la especificación de cómo se van a usar y con qué fines. Ligado a esto último, surge el problema de los derechos de autor sobre los datos.

Motivados por los desafíos que presentan los sistemas unimodales, el análisis de sentimientos multimodal busca proponer técnicas que aprendan relaciones entre las distintas modalidades presentes en los datos [24]. A pesar de la mejora que presenta el uso de este tipo de técnicas, en particular contra modelos que utilizan únicamente texto [25], se han detectado ciertas áreas de oportunidad. Por un lado, el principal obstáculo en la creación de un sistema multimodal yace en la forma de fusionar la información de cada modalidad. Además, la creación de un sistema multimodal depende de las modalidades de información consideradas y la información disponible para entrenar los modelos. Por ejemplo, para la exploración de videoblogs se integran el análisis de audio, video y declaraciones. En consecuencia, los conjuntos de datos diseñados para entrenar modelos para este contenido de redes sociales pueden no ser del todo útiles cuando se investigan escenarios que se desvían de estas condiciones, por ejemplo, publicaciones con texto e imágenes únicamente. Esto complica aún más los trabajos que abordan diversas combinaciones de información multimodal que se desvían de los esfuerzos delineados por la literatura actual. Por otro lado, los conjuntos de datos existentes para el análisis de sentimientos multimodal reflejan los tipos de problemas que se pueden resolver y los métodos que se pueden emplear dada la información que contienen y los esquemas de anotación que manejan, que pueden diferir entre ellos a pesar de centrarse en la misma aplicación. Este problema se agrava aún más por el hecho de que la mayoría de los conjuntos de datos disponibles para el análisis de sentimientos unimodal y multimodal se centran en el inglés, lo que dificulta los esfuerzos por construir sistemas en idiomas con poca o ninguna representación en el campo.

1.3. Estado del Arte

La mayoría de los esfuerzos en el análisis de sentimientos en imágenes se centran en el análisis de expresiones faciales [26] y gestos corporales [27] para determinar sentimientos y emociones. Sin embargo, se han hecho intentos para ampliar las aplicaciones a imágenes más complejas, por ejemplo, aquellas con múltiples objetos y detalles de fondo [28]. Además, el desarrollo de técnicas de aprendizaje profundo, en particular diferentes variantes basadas en Redes Neuronales Convolucionales como VGG-19, ResNet50V2, DenseNet-121 e Inception-v3, ha representado una mejora en resultados al momento de usar datos de redes sociales [20], [29]. Sin embargo, las arquitecturas basadas en Transformer [30], como Vision Transformer [31], han demostrado ser una buena alternativa para la tarea en cuestión [32].

En el caso del análisis de sentimiento enfocado a texto, han surgido distintos enfoques para trabajar con conjuntos de datos de distinto origen. Por un lado, los métodos basados en redes neuronales profundas continúan presentes en el campo de estudio. Métodos como LSTM [33], BI-LSTM [34] y CNN-LSTM [35] han expuesto una buena capacidad para resolver problemas de clasificación debido a su capacidad para capturar patrones secuenciales y jerárquicos.

Sin embargo, los métodos basados en transferencia (*Transfer Learning*) han obtenido los mejores resultados en distintas métricas, requiriendo pocos datos, mostrando alta adaptabilidad a diversos dominios de los datos, pero con la desventaja de un costo

computacional alto y baja explicabilidad [36]. La idea principal es ajustar modelos base preentrenados con una gran cantidad de datos, para después ajustar sus parámetros con otro pequeño conjunto de datos y, de esa manera, adaptar el modelo virtualmente a cualquier tema. Algunos ejemplos incluyen modelos como BioBERT, T5 y RoBERTa aplicados al campo biomédico [37] y BERT [38], [39] para casos multilingües con poca representación léxica.

La idea principal detrás del análisis de sentimientos multimodal para imágenes y texto es fusionar adecuadamente la información de ambas modalidades para obtener una buena representación del sentimiento, por ejemplo, de las publicaciones de redes sociales. En ese sentido, se han propuesto diferentes métodos utilizando distintas herramientas y modelos de representación. Los primeros esfuerzos consideraron métodos más sencillos para modelar y fusionar información multimodal como bolsas de palabras y n-gramas para texto y bolsa de palabras visuales para las imágenes [40]. No obstante, recientes avances en técnicas de aprendizaje profundo han permitido la introducción de ideas más complejas en distintas partes del proceso de tal forma que nuevos desarrollos y enfoques para realizar esta tarea han demostrado un éxito significativo. Por ejemplo, Kumar y Garg [41] construyeron un sistema que procesa distintos componentes de un tuit por separado. Las imágenes entrantes al sistema se procesan con SentiBank y Regions with Convolutional Neural Networks (R-CNN), el texto se trata con técnicas tradicionales de procesamiento del lenguaje natural y gradient boosting, el texto multimodal infográfico se extrae mediante ROC. La polaridad del sentimiento se obtiene agregando las puntuaciones de cada módulo que procesa cada modalidad.

A medida que la fama de los modelos vastos de lenguaje (*Large Language Models*) creció, los modelos preentrenados y ajustados juegan cada vez un rol más significativo en el análisis de sentimientos multimodal. Zhang et al. [42] propusieron un marco multimodal para trabajar con tuits donde las características del texto y las imágenes se extraen utilizando un modelo preentrenado de BERT y una red neuronal convolucional con atención, respectivamente, que se combinan con una red de fusión basada en tensores y un módulo de extracción para descartar información redundante. Dimitrov et al. [43] introdujeron un método para detectar el tipo de técnica de propaganda utilizada en memes. En su trabajo, exploran marcos unimodales y multimodales, utilizando versiones base y entrenadas de BERT y ResNet152, respectivamente. De igual forma, para la fusión multimodal, compararon el rendimiento de distintas técnicas: Multimodal Bitransformers (MMBT), la concatenación de características y la combinación de las predicciones de los modelos. Zhang et al. [44] utilizaron BERT y ResNet-50 para extraer características de datos multimodales para después fusionarlos utilizando una capa de codificador de Transformer para detectar sarcasmo en publicaciones de Twitter. Anshul et al. [45] propusieron un modelo que incorpora texto, imágenes, texto incrustado en imágenes, enlaces e información específica del usuario para detectar síntomas de depresión durante la pandemia de COVID-19. Zhu et al. [46] propusieron la Sentiment Knowledge Enhanced Attention Fusion Network, una red de fusión que mejora la fusión multimodal al incorporar representaciones de conocimiento de sentimientos adicionales de una base de conocimiento externa. Zhong et al. [47] propusieron un marco de mejora semántica mediante el empleo de BERT preentrenado y Vision Transformer (ViT) para codificar textos e imágenes, respectivamente, y mejorar la detección del sarcasmo.

La reciente introducción de los modelos de lenguaje para visión implica un progreso significativo en tareas multimodales que trabajan con imágenes y texto. Por ejemplo,

Rivas et al. [48] modelaron imágenes, texto, hashtags, información del perfil de los usuarios y su ubicación en un espacio de embeddings conjunto utilizando el sistema VSE++ entrenado con un perceptrón multicapa. Sin embargo, Contrastive Language Image Pre-training (CLIP) ha mostrado un reciente y mayor interés por parte de los investigadores al ser capaz de modelar relaciones entre texto e imágenes. En concreto, Lu et al. [49] propusieron un modelo de análisis de sentimientos multimodal llamado CLIP-CA-CG basado en CLIP para extraer características de texto e imágenes, un mecanismo de autoatención para fusionar información y un módulo basado en compuertas para asignar cuánta información debe transmitir cada modalidad al algoritmo de clasificación final. Chen et al. [50] introdujeron un algoritmo llamado Visual-textual Sentiment Analysis with Pre-trained Feature (VSA-PF) que consta de cuatro partes: 1) una rama visual y textual basada en el ajuste de Swin Transformer y BERTweet para la predicción de sentimientos de imágenes y texto; 2) una segunda rama que extrae un conjunto de codificadores visuales para extraer información semántica como características faciales, detección de escenas u objetos y extraer texto incrustado en imágenes; 3) una rama de CLIP para extraer características de imágenes y texto; y 4) una rama de fusión de características multimodales que utiliza un modelo BERT con cuatro cabezales de atención. Por último, An y Zainon [51] extrajeron características de pares de imágenes y texto utilizando CLIP e integraron señales de color de imágenes con mecanismos de atención para predecir etiquetas de sentimientos.

Finalmente, el análisis de sentimientos aplicado a datos en español es un área en desarrollo. Para el análisis de sentimientos en texto, Álvarez-Carmona et al. [52] trabajaron con reseñas de viajes en línea para proponer tres esquemas para combinar 14 modelos especializados construidos en el fórum Rest-Mex. Como resultado, lograron mejorar las salidas de cada modelo individual y mejora los resultados en cuatro de cinco clases de polaridad. Por otro lado, en [53], se presenta una visión general del REST-MEX 2025, evento que se centró en la tarea de clasificación automática de reseñas generadas por usuarios en tres ejes: polaridad (de 1 a 5), tipo de servicio (hotel, restaurante o atractivo), y la identificación de uno de los 40 Pueblos Mágicos predefinidos. Lo anterior funciona como un punto de referencia para las características del conjunto de datos, el protocolo de evaluación y un análisis comparativo de los resultados y modelos empleados.

En el área de análisis de sentimientos multimodal, Monsalve-Pulido et al. [54] se enfocaron en datos en español en el ámbito turístico (conjunto de datos TASS) usando algoritmos como Bosques Aleatorios y Máquinas de Vectores de Soporte. Pérez-Rosas et al. [55] presentaron un método que integra características lingüísticas, auditivas y visuales para identificar el sentimiento en vídeos de YouTube en español mediante una Máquina de Vectores de Soporte con un kernel lineal.

1.3.1. Conjuntos de Datos

Los conjuntos de datos más populares utilizados para la tarea de análisis de sentimientos de imágenes y texto se resumen en la Tabla 1.1. Como se puede apreciar, todos los conjuntos de datos que se muestran se orientan hacia el trabajo de texto e imágenes en inglés bajo distintos esquemas de anotación y la mayoría admite una sola imagen al momento de agregar el elemento multimodal al modelo de análisis de sentimientos, salvo el T4SA que puede incluir múltiples imágenes en el mismo tuit. Cabe mencionar que, para los modelos mencionados en la literatura anteriormente, la mayor parte utiliza el MVSA como modelo base para la comparación de resultados. Sin embargo, para los estudios

TABLA 1.1: Conjuntos de datos comunes que se utilizan para la tarea de análisis de sentimientos de imágenes y texto, inspirado en [56].

Nombre	Descripción
Photo Tweet Sentiment Benchmark (PTSB) [57]	Dataset de referencia que incluye 603 tweets con fotos y sus respectivas descripciones de texto que las acompañan. Se recopiló en noviembre de 2012 a través de la API PeopleBrowsr utilizando 21 hashtags. Se obtuvieron valores de sentimiento reales mediante la anotación Turk de Amazon Mechanic, lo que dio como resultado 470 etiquetas positivas y 133 negativas.
Multi-view SA Dataset (MVSA) [58]	Conjunto de datos que incluye parejas de imágenes y texto en inglés anotadas manualmente de Twitter del cual existen dos versiones: el simple (un anotador) y el múltiple (tres anotadores). Este conjunto de datos se utiliza ampliamente para la evaluación de sistemas de análisis de sentimientos de texto e imágenes. Los datos se dividen en tres categorías (positivo, neutral y negativo) con un total de 4,869 datos.
Twitter for Sentiment Analysis (T4SA) [59]	Conjunto de datos grande que accedió al 1 % del total de los tuits producidos globalmente entre julio y diciembre de 2016 para obtener un aproximado de 3.4 millones de tuits, con un total de ∼4 millones de imágenes. Cada texto e imagen del tuit se etiquetó en tres clases: positivo, negativo y neutral. Cabe mencionar que durante el proceso de construcción del conjunto de datos se descartaron todos los tuits cuyo idioma no fuese inglés.

comparativos, la etiqueta final se suele agrupar siguiendo una serie de reglas según las cuales un tuit es válido si tanto el texto como las imágenes tienen el mismo sentimiento o uno de ellos es neutral. Por otro lado, no es válido si las etiquetas son opuestas, lo que lleva a su eliminación. Como resultado, se pasan por alto las relaciones conflictivas entre pares de texto e imágenes y su causa. Otro punto de conflicto que merece la pena resaltar es la falta de una estrategia para atacar el problema del spam presente en redes sociales, cuya identificación y etiquetado es ignorado por completo en todos los conjuntos de datos. Por último, la mayoría de los conjuntos de datos suelen ignorar los casos donde las publicaciones presentan más de una imagen en ellas, que es el caso más apegado a situaciones que se observan en la realidad.

1.4. Planteamiento del Problema

La literatura existente muestra que el enfoque actual del análisis de sentimientos multimodal se encuentra lejos de estar resuelto y presenta las siguientes preguntas abiertas:

P1 En el caso del problema de imágenes y texto, el trabajo previo apunta a un sesgo generado por los conjuntos de datos disponibles. ¿Cómo se pueden enfocar las aplicaciones para otros idiomas distintos al inglés, en particular el español?

P2 ¿Es posible crear sistemas de análisis de sentimientos multimodal para que trabajen con múltiples imágenes al mismo tiempo?

P3 ¿Cómo impacta el número de imágenes consideradas al sistema de análisis de sentimientos multimodal?

P4 ¿Cuál es el impacto del texto incrustado en imágenes a los modelos de imagen y texto de análisis de sentimientos multimodal?

P5 El contenido catalogado como spam en redes sociales no se considera un problema lo suficientemente relevante como para ser incluido dentro de los modelos de clasificación. ¿Cómo se puede incluir este tipo de datos dentro de un marco de trabajo para el análisis de sentimientos multimodal y cómo afecta su incorporación?

Dada la importancia de dichos elementos para un estudio de las redes sociales más general y comprensivo, esta investigación tiene como objetivo aprovechar estas limitantes para desarrollar nuevos métodos y recursos para el análisis de sentimientos multimodal, así como establecer nuevos resultados que sirvan como referencia para estudios futuros en el campo.

1.5. Objetivos

Dado el planteamiento del problema, los objetivos de la presente investigación se pueden resumir en los siguientes puntos:

1. (P1, Sección 3.2) Diseñar una metodología que emplee modelos de lenguaje y visión basados en la arquitectura de Transformer para construir sistemas de análisis de sentimientos multimodal para determinar la polaridad del sentimiento de publicaciones de redes sociales en español.

2. (P1, Sección 3.1) Recopilar datos en español para el análisis de sentimientos multimodal, tomando en cuenta la privacidad de los usuarios y las reglamentaciones en turno de las redes sociales pertinentes, así como principios éticos.

3. (P2, Sección 3.1) Proponer un esquema de anotación de datos multimodal que resulte útil para el trabajo de análisis de cada modalidad en conjunto y por separado.

4. (P2, Sección 3.2.5) Proponer un método de fusión de información que permita incorporar diversas modalidades de datos que considere su contexto, además de que permita el análisis de publicaciones con múltiples imágenes.

5. (P3, Secciones 4.4.4 y 4.5.3) Determinar el impacto al sistema de análisis de sentimientos multimodal al considerar diferentes números de imágenes entrantes.

6. (P4, Secciones 4.4.4 y 4.5.3) Analizar cómo afecta el texto incrustado en imágenes al sistema de análisis multimodal propuesto.

7. (P5, Secciones 4.4.4 y 4.5.3) Explorar cómo afecta el spam a los modelos de análisis de sentimientos multimodal.

1.6. Propuesta Metodológica

Para cumplir con los objetivos propuestos, se presenta una metodología para realizar análisis de sentimiento multimodal basada en arquitecturas de aprendizaje profundo y modelos vastos de lenguaje. En primer lugar, se ajustan (*fine-tuning*) los modelos preentrenados de BETO [60] y Vision Transformer (ViT) [31] con el texto y las imágenes del Multimodal Spanish Sentiment Analysis Impact Dataset (MSSAID) [61] y Multimodal COVID19 Mexico, respectivamente, bajo la idea de que la aplicación de este proceso adicional mejora su desempeño al momento de determinar la polaridad del sentimiento en el modelo multimodal. La ventaja de los conjuntos de datos que se construyen sobre otros conjuntos de datos de imágenes y texto similares se basa en el esquema de anotación que etiqueta diferentes aspectos de un tuit, lo que permite el estudio de las interacciones entre diferentes modalidades y su impacto en el sentimiento general de una publicación. Se pueden consultar detalles en extenso en la Sección 3.2.

En segundo lugar, proponemos dos métodos de fusión multimodal: el primer método se basa en la suma de vectores y el segundo utiliza bloques de codificadores del Transformer que utilizan el concepto de autoatención siguiendo la idea de que una persona interactúa con las diferentes modalidades de forma secuencial. Finalmente, para explorar el caso del texto incrustado en imágenes, construimos un conjunto de datos que contiene imágenes similares a memes donde el texto está presente dentro de las imágenes para entrenar un modelo de detección de texto. Este modelo se utiliza para detectar automáticamente las regiones donde se encuentra el texto de interés para extraerlo con la ayuda de un motor de reconocimiento óptico de caracteres e incorporarlo como una modalidad adicional de información en el marco de trabajo. El módulo de fusión combina las características del texto, múltiples imágenes y el texto detectado en ellas (si lo hay), cuya salida se utiliza como entrada para un algoritmo de clasificación. Los detalles completos se encuentran en la Sección 3.2.5.

Como resultado, nuestro método propuesto no solo aborda las limitaciones detectadas en los métodos existentes de análisis de sentimientos multimodal de imágenes y texto expuestas anteriormente, sino que también sirve como una herramienta valiosa para aplicaciones del mundo real donde los datos multimodales son abundantes.

1.7. Estructura del Documento

El documento se organiza de la siguiente manera. El Capítulo 2 presenta conceptos teóricos clave para el desarrollo de las metodologías propuestas en este trabajo. Se introducen conceptos relacionados al análisis de sentimientos en texto, imágenes y multimodal. Después, se discute sobre la arquitectura del Transformer, sus componentes y modelos derivados como BERT y Vision Transformer. Finalmente, en este capítulo se introducen los modelos de clasificación empleados en el trabajo y las métricas de evaluación utilizadas para medir el desempeño de los modelos que se construyen para la tarea de análisis de sentimientos multimodal.

En el Capítulo 3 se expone la forma de construcción de los conjuntos de datos que se emplearon en el trabajo, el modelo propuesto para trabajar con imágenes y texto, las estrategias de fusión de información y el proceso realizado para entrenar cada uno de ellos. En esta parte del trabajo se definen los experimentos preliminares, los experimentos

de ablación, la forma de analizar el error que se presenta en los resultados y la manera en la que se comparan entre sí los distintos modelos aplicados al problema.

Por otro lado, el Capítulo 4 muestra los resultados obtenidos al realizar los experimentos planteados en la Metodología para los modelos de imagen y texto. Además, se presenta a detalle los conjuntos de datos usados en el trabajo y los resultados de los experimentos de ablación sobre el impacto de las modalidades y el número de imágenes en los resultados de los modelos multimodales.

Por último, el Capítulo 5 finaliza el escrito con la conclusión de la investigación, las respuestas a las preguntas descritas en el Planteamiento del Problema, las limitaciones que se presentaron durante el estudio y la discusión sobre trabajos futuros.

Capítulo 2

Marco Teórico

En este capítulo se introducen diversos conceptos que sirven para establecer una base teórica del proyecto de investigación. En particular, se explican conceptos sobre el análisis de sentimientos aplicado a texto y análisis de sentimientos multimodal, para después cubrir nociones importantes sobre los modelos de Transformer. Además, se describen las arquitecturas de dos modelos usados en el trabajo, Bidirection Encoder Representations from Transformers y Vision Transformer. Después, se presenta brevemente el algoritmo de clasificación y las métricas de desempeño usadas en el trabajo para evaluar los resultados de los modelos.

2.1. Análisis de Sentimientos

El tipo de información con la que trabaja el análisis de sentimientos debe ser subjetiva, donde resaltan las opiniones y sentimientos que expresan las personas. Una opinión es una postura subjetiva que una persona tiene hacia cierto tema o evento, la cual suele reflejar los sentimientos, emociones o percepciones de la persona que la alberga [62]. Aunque la definición anterior es útil, el análisis de sentimientos considera otra que nos permite identificar los distintos componentes que la conforman.

Definición 2 (Opinión) *Una opinión [63] es una quíntupla*

$$(u_i, a_{ij}, h_k, t_l, s_{ijkl})$$

donde u_i es el nombre de la entidad sobre la que trata la opinión, a_{ij} es un aspecto de u_i, h_k representa quién alberga la opinión, t_l es el tiempo cuando la opinión es expresada por h_k y s_{ijkl} denota el sentimiento del aspecto a_{ij} de la entidad u_i.

El sentimiento s_{ijkl} se define como una disposición neuropsíquica importante que permite al ser humano reaccionar emocional, cognitiva y conativamente hacia un determinado objeto (o situación) de una manera estable [64]. Además, suele acompañarse por intereses y valores de cada individuo, lo cual crea una disposición a largo plazo que es adquirida y evocada cuando quien alberga el sentimiento piensa o percibe algo sobre el objeto [65]. El sentimiento se compone de dos elementos: polaridad e intensidad [66]. La polaridad representa el valor positivo, negativo o neutro contenido en cada opinión, usualmente representados numéricamente mediante 1, -1 y 0, respectivamente. Por otro lado, la intensidad del sentimiento permite medir usando una escala qué tan positiva o negativa es la opinión. Por ejemplo, se puede usar la escala $[-3, 3]$, donde -3 denota

un sentimiento muy negativo, 3 muy positivo y 0 un sentimiento neutro. Sin embargo, dependiendo de la aplicación, ambos elementos se exploran usualmente de manera separada.

Ejemplo 1 *Supongamos que* ***Ayelín*** *compra una computadora el* ***23 de marzo de 2024*** *y publica en X: «¡La nueva* ***computadora*** *tiene un problema serio con la* ***batería*** *ya que no dura más de una hora! ». Vamos a indexar de la siguiente manera: «computadora» como 1, «batería» como 2, «Ayelín» como 3 y el 23 de marzo de 2024 como 4. Entonces, Ayelín es quien alberga la opinión* h_3 *y el 23 de marzo de 2024 es el tiempo* t_4 *cuando se emite la opinión. El término «computadora» es la entidad* u_1 *sobre la que trata la opinión, «batería» corresponde al aspecto* a_{12} *de la entidad* u_1 *(«computadora») y* $s_{1234} = -1$ *es el sentimiento negativo sobre el aspecto* a_{12} *(«batería») de la entidad* u_1 *(«computadora»).*

Por otro lado, Cambria et al. [67] definen a las emociones como complejos estados sentimentales que activan reacciones físicas y fisiológicas que afectan la conducta y el pensamiento humano. Sin embargo, una definición única no se encuentra disponible ya que esta varía en la literatura: se pueden encontrar más de 90 definiciones disponibles. Por lo tanto, es difícil manejar un único marco para representar emociones, a diferencia de los sentimientos, aunque frecuentemente se eligen los modelos de Ekman [68] y Rusell [69] para dicho fin. Cuando se trabaja con sentimientos en lugar de emociones, se realiza análisis de emociones [70].

Aunque sentimiento, opinión y emoción son términos que representan subjetividad, no son lo mismo. Como indican Munezero et al. [71], los sentimientos suelen ser más duraderos y estables que las emociones. Además, su formación y dirección es hacia un objeto o tema. Por otro lado, los sentimientos son construcciones sociales que surgen de las emociones los cuales se desarrollan con el tiempo y son duraderas, mientras que las opiniones son interpretaciones personales de información que pueden o no estar cargadas de emociones o sentimientos.

2.1.1. Análisis de Sentimientos Tradicional

El análisis de sentimientos y la definición de opinión proveen un marco de trabajo para estructurar la información que puede considerarse desde diferentes niveles de detalle:

- A **nivel documento** indica que el proceso de análisis se lleva a cabo utilizando todo el documento que alberga texto bajo el supuesto que éste discute un tema en común.
- A **nivel enunciado** indica que el proceso de análisis se lleva a cabo en cada enunciado que conforma el documento de texto.
- A **nivel de entidades y aspectos** es un proceso más refinado e intensivo debido a que se enfoca en determinar las distintas opiniones que se encuentran en un documento de texto y, para cada una de ellas, realiza el análisis correspondiente.

No todos los elementos de la quíntupla se necesitan en cada situación ya que dependen de la tarea que se esté realizando. Por ejemplo, para la clasificación de polaridad a nivel documento [46] es suficiente si se determina s_{ijkl}, pero para resumir opiniones [72]

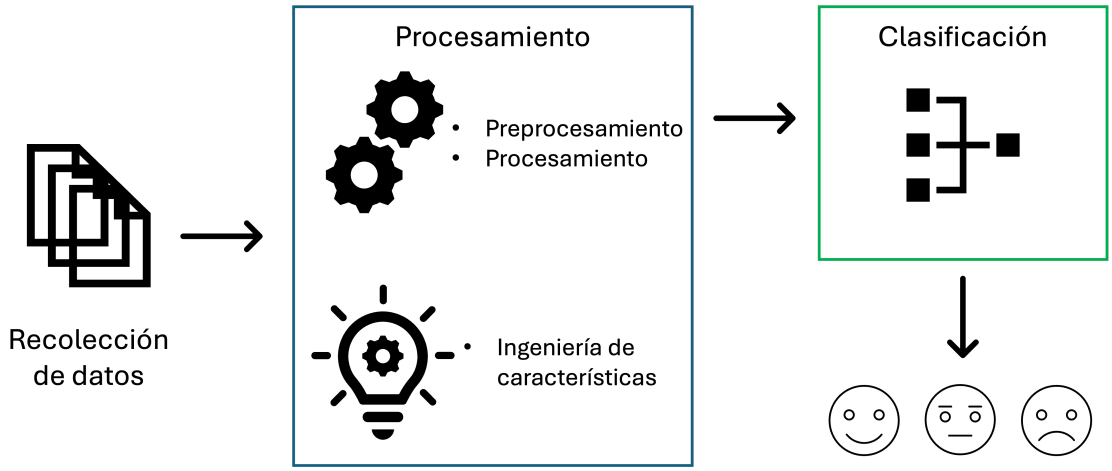


FIGURA 2.1: Diagrama que muestra la idea general que se sigue para realizar análisis de polaridad enfocado a textos. Fuente: elaboración propia.

es deseable contar con la mayor cantidad de elementos de la quíntupla posibles. Por lo tanto, al no ser un problema único sino una colección de tareas (un «gran portafolio»), el análisis de sentimientos orientado a textos requiere resolver diversas tareas de Procesamiento de Lenguaje Natural (PLN) según lo que se requiera resolver [73]. Algunas tareas comunes del análisis de sentimientos incluyen detección de sarcasmo [74], detección de spam [75], análisis de tópicos [76], generación de lexicones de sentimientos [77], resumen de opiniones, entre muchas otras. En consecuencia, a pesar de que el término «análisis de sentimientos» se usa ampliamente tanto en la academia como en la industria como sinónimo de análisis de polaridad [8], en realidad es un marco de trabajo para trabajar con información subjetiva sin estructura. De ahora en adelante, para mantener la notación que se usa en la literatura, cuando se mencione en el documento «análisis de sentimientos» en realidad se hace referencia a la tarea de análisis de polaridad.

El proceso general para realizar análisis de sentimientos tradicional consta de una fase de recolección de datos, un procesamiento de la información y un paso de clasificación, como se muestra en la Figura 2.1. Para la recolección de datos se pueden usar diversas fuentes como redes sociales, sitios web, blogs o foros mediante el uso APIs (por ejemplo, la API de X[1]), herramientas para realizar web scrapping o descargas manuales vía crowdsourcing.

La fase de procesamiento de datos busca reducir el ruido presente en la información «cruda», especialmente aquella de redes sociales, que suele presentarse en forma de errores gramaticales y de ortografía [63]. En sistemas tradicionales basados en modelos de Machine Learning, se lleva a cabo un paso adicional llamado preprocesamiento cuyo fin es reducir el tamaño final del vocabulario o dimensión del vector de entrada, por ejemplo, al eliminar palabras como artículos, preposiciones, algunos verbos, signos de puntuación, etc. Las tareas de PLN más comunes que se ejecutan en este paso son tokenización (*tokenization*), etiquetado gramatical (*part-of-speech tagging*), reconocimiento de entidades nombradas (*named entity recognition*) y lematización o stemming [78]. Dependiendo del tipo de fuente de datos, puede resultar útil agregar pasos para limpiar aún más el texto,

[1] https://developer.x.com/en/products/x-api

por ejemplo, quitando caracteres repetidos (por ejemplo, «goooooooool»), correos electrónicos o lenguaje markdown. Sin embargo, cada caso debe ser analizado de forma especial. Al trabajar con datos de redes sociales se suelen eliminar enlaces a otros sitios web, nombres de usuarios (por razones de privacidad) y etiquetas propias del sitio como hashtags y cashtags [79]. A pesar de ello, alternar entre la conservación y eliminación de tales elementos puede ayudar al rendimiento de los sistemas de análisis de sentimientos [80]. Cabe mencionar que, en el caso de modelos de lenguaje basados en la arquitectura de Transformer, el uso de una estrategia apropiada de preprocesamiento como la mencionada anteriormente puede resultar beneficiosa al mejorar los resultados de los modelos. Sin embargo, si no se manejan de manera apropiada, puede tener un efecto contrario e impactar negativamente el rendimiento de tales sistemas [81].

La extracción o ingeniería de características es un proceso que permite obtener información importante que ayuda a describir las características más representativos de la fuente de información en turno [82]. La idea clave en este paso es la forma de representar los textos en una forma numérica de dimensión fija para que los modelos de clasificación sean capaces de procesarlos. Entre las técnicas clásicas más utilizadas destacan los modelos basados en la ausencia o presencia de términos como Bolsa de Palabras y Term Frequency-Inverse Document Frequency (TF-IDF) [83]. A pesar de ser usadas ampliamente, tienen la desventaja de ignorar conceptos clave como el orden de las palabras o estructuras gramaticales. Por otra parte, los *word embeddings* son representaciones distribuidas que codifican el significado de una palabra en un espacio vectorial. Estos modelos de representación distribuida son generados usando métodos de aprendizaje profundo dentro de los que destacan Word2vec [84], Global Vectors (GloVe) [85] y FastText [86]. Desde la salida de diversos modelos de lenguaje basados en Transformer, la generación de las representaciones numéricas de las palabras se pueden hacer utilizando diferentes modelos grandes de lenguaje como Bidirectional Encoder Representations from Transformers (BERT) [87] o alguna de sus variantes en algún idioma como BETO [60], que permite trabajar con texto en español.

Los métodos para el proceso de clasificación se suelen dividir en tres clases: métodos basados en Machine Learning y Deep Learning, métodos basados en lexicones y métodos híbridos. Los métodos basados en algoritmos de Machine Learning incluyen modelos de aprendizaje supervisado como Naïve Bayes [88], Máquinas de Vectores de Soporte (MVS) [89], Máxima Entropía [90], Árboles de Decisión [91], K Vecinos Más Cercanos [92] y métodos de ensamble [93], [94]. El análisis de sentimientos también ha gozado de éxito al emplear métodos de aprendizaje profundo basados en redes neuronales con word embeddings que permiten entrenar modelos más complejos con muchos más datos [95]. Los modelos que se utilizan con más frecuencia incluyen variaciones de Redes Neuronales Recurrentes como Long Short Term Memory (LSTM) [96] o Bi-LSTM [97], y Redes Neuronales Convolucionales [98].

Los métodos basados en lexicones de opinión comprenden un conjunto de recursos léxicos donde cada palabra se asocia a su orientación semántica según la escala o polaridad del sentimiento que considere, por ejemplo, positivo, negativo o neutral [99]. Para determinar el valor del sentimiento se debe utilizar una fórmula o algoritmo. Tales lexicones se pueden generar de forma manual [100], automática [101] o derivados de grandes diccionarios de palabras como Wordnet[2] [102] bajo la idea de que los sinónimos de un

[2] https://wordnet.princeton.edu/

grupo de palabras semillas tienen la misma polaridad. A diferencia de los sistemas basados en modelos de Machine Learning, su uso es práctico ya que no requiere un conjunto de datos anotado para entrenar modelos, por lo que se pueden considerar una forma de aprendizaje no supervisado. Sin embargo, tienen la desventaja de ser dependientes del contexto en el que se usen. Por ejemplo, la palabra «explosivo» representa algo negativo en la mayoría de los casos de uso, pero si se habla de un «crecimiento *explosivo* de la empresa» se puede tratar de algo positivo. Para resolver este tipo de situaciones surgen lexicones dependientes del dominio o tema que aborde el análisis de polaridad. Asimismo, dado un conjunto lo suficiente grande de entrenamiento, los modelos de Machine Learning suelen dar mejores resultados que los métodos de clasificación basados en lexicones [99]. Finalmente, los métodos híbridos combinan las ventajas que ofrecen los métodos basados en Machine Learning y lexicones [103]. Para ilustrar esta forma de clasificación se puede citar el trabajo de Shin et al. [104] donde construyeron embeddings de palabras provenientes de diversos lexicones que fueron integrados usando una Red Neuronal Convolucional con mecanismos de atención. Como resultado, concluyen que la integración de los lexicones al proceso mejora, en lo general, el modelo de clasificación basado en Redes Neuronales Convolucionales.

2.1.2. Análisis de Sentimientos en Imágenes

La Definición 2 de opinión se puede adaptar de texto a imágenes [105]. En particular, la entidad u_i es el contenido visual que se usa y se divide en partes y atributos. Las partes pueden representarse como una sola idea plasmada en la imagen o, al contrario, se puede componer de varias subimágenes las cuales pueden comprender contenidos específicos. Los atributos corresponden a la idea en texto que describe el contenido semántico de cada elemento o a la categoría de los objetos que se encuentran en ellos. Por otro lado, es posible identificar dos titulares h_k de la opinión: quién es el dueño o quién publica el contenido (que no necesariamente son la misma persona), y quién lo consume. Por ejemplo, en el caso de campañas publicitarias, la intención del dueño de la campaña puede alinearse o no con el impacto en aquellos que la consumen [106]. Para el caso del análisis de sentimientos visual en redes sociales, se debe considerar que el sentimiento expresado por la persona que publica la imagen y el sentimiento generado por aquellos que lo consumen no son necesariamente el mismo.

Similar al caso cuando se trabaja con texto, el preprocesamiento permite modificar las imágenes para remover ruido, distorsiones o mejorar ciertas propiedades de las imágenes que permitan un mejor análisis. Algunas de las operaciones de preprocesamiento más utilizadas incluyen corrección de brillo de píxeles, transformaciones geométricas, filtrado y segmentación de imágenes, transformada de Fourier y restauración de imágenes [56].

Para la extracción de características visuales se suelen considerar tres niveles semánticos. Las características de bajo nivel son aquellas características visuales de una imagen que se extraen a nivel de los pixeles como color, textura y bordes a nivel local o global. Por otra parte, las características de mediano nivel permiten capturar e identificar información más abstracta en las imágenes como formas, partes de objetos y configuraciones de pixeles. Algunos ejemplos incluyen esquinas, uniones, contornos, patrones geométricos o agrupaciones de pixeles. Finalmente, las características de alto nivel se refieren a representaciones semánticas abstractas que capturan información significativa sobre objetos, escenas o conceptos en la imagen. A este nivel, suelen relacionarse con objetos completos

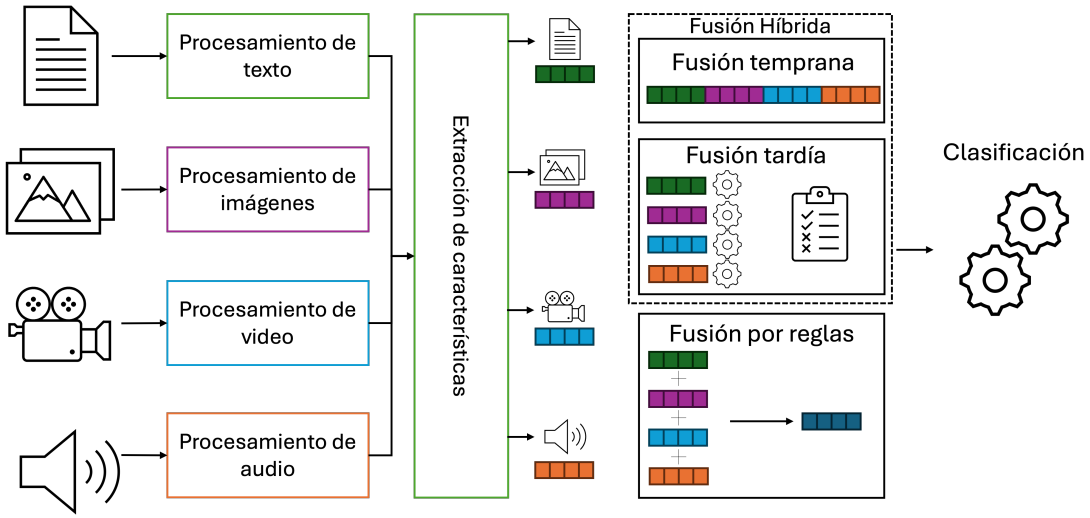


FIGURA 2.2: Diagrama que muestra la idea general que se sigue para realizar análisis de polaridad multimodal. Fuente: elaboración propia.

que se encuentren en la imagen, el contexto mediante la escena o el fondo, relaciones entre objetos o poses u orientaciones.

2.1.3. Análisis de Sentimientos Multimodal

Los sistemas de análisis de sentimientos multimodal se especializan en modelar dos tipos de dinámicas que pueden presentarse entre los distintos tipos de datos [24]. En primer lugar, por modalidades o tipos de datos se refiere a la información con la que trata el sistema. La información multimodal más común incluye señales de audio, texto, imágenes y videos, en cualquier combinación.

A partir de la interacción de las modalidades se originan dos tipos de dinámicas que describen su comportamiento [9]. Las dinámicas intramodales modelan aquellas interacciones que existen dentro de una modalidad en específico. A modo de ejemplo se puede mencionar las interacciones entre las palabras en un mismo tuit. La segunda dinámica comprende las relaciones intermodales, es decir, las interacciones entre diferentes modalidades que se dividen en síncronas y asíncronas. Una dinámica intermodal síncrona es cuando, en una presentación, alguien muestra una imagen durante la explicación de un concepto que se explica oralmente a manera de refuerzo visual. Un ejemplo de una dinámica asíncrona es una publicación de un tuit: el texto y el contenido audiovisual que lo acompaña interactúan entre ellos para dotar de información de manera mutua. Sin embargo, el usuario no interacciona con ambos al mismo tiempo, sino uno a la vez.

Por lo tanto, el análisis de sentimientos multimodal involucra la representación de características de cada modalidad considerada, a menudo empleando técnicas de redes neuronales profundas, un método de fusión para representar los datos en una forma numérica unificada que capture las relaciones entre ellos y un mecanismo de clasificación para predecir la etiqueta de polaridad de toda la publicación, como se muestra en la Figura 2.2. Como resultado, estos sistemas dependen en gran medida del enfoque para procesar cada modalidad, los medios para fusionarlos y los métodos de clasificación, los cuales cambian considerablemente dependiendo del paradigma vigente.

El principal desafío en el análisis de sentimientos multimodal involucra los métodos para fusionar características de cada modalidad. En general, Chandrasekaran et al. [107] y Abdu et al. [9] dividen los mecanismos de fusión en distintas grandes familias:

1. **Fusión temprana**: en esta categoría, todas las modalidades consideradas se concatenan en un único vector de representación que se utiliza como entrada para el modelo de predicción. A pesar de la sencillez del enfoque, este presenta problemas cuando se tienen pocos datos y no son significativos ya que en el proceso se pierden las relaciones intermodales de cada modalidad. Perez-Rosas et al. [55] combinaron las modalidades visuales, acústicas y lingüísticas en un único vector de características para predecir el sentimiento de declaraciones en videos mediante una MVS.

2. **Fusión tardía**: se construyen diferentes modelos, generalmente modelos preentrenados para cada modalidad, por lo que la etiqueta de polaridad final se determina mediante un voto mayoritario (según la predicción de sentimiento de cada modelo), suma ponderada, promedio o mecanismos de redes neuronales profundas. Estos modelos son modulares, por lo que se pueden usar versiones ajustadas de distintos modelos base unimodales, pero presentan fallas al modelar dinámicas intermodales considerando que el sentimiento final es más complejo que un voto mayoritario, además de consumir más tiempo.

3. **Fusión híbrida**: este enfoque combina las ideas principales de la fusión tardía y temprana. En [108] propusieron un sistema híbrido para predecir el sentimiento de declaraciones en videos. Primero utilizaron fusión temprana para determinar una primera predicción de las características de audio y video mediante una LSTM Bidireccional. Después, usaron fusión tardía mediante voto ponderado usando la mezcla anterior con las predicciones que realizó una MVS para las características lingüísticas para así determinar el sentimiento final de cada declaración como una suma ponderada de las puntuaciones lingüística y audiovisual.

4. **Fusión basada en reglas**: combina las distintas modalidades usando reglas como, por ejemplo, la suma ponderada, el producto o el máximo, de la representación de cada modalidad [109].

2.2. Transformer

En el 2017, el trabajo de Vaswani et al. [30] introdujo el mecanismo de atención que es la base de múltiples modelos vastos de lenguaje para texto, imágenes y multimodales que forman parte del estado del arte. Este mecanismo se encuentra en una red neuronal secuencial llamada Transfomer especializada en trabajar información que se modela como secuencia (texto, audio, series de tiempo, etc.), que además deja atrás conceptos de redes neuronales recurrentes y convolucionales.

Originalmente, la arquitectura del Transformer se basa en dos componentes:

- **Codificador**: convierte la secuencia de entrada en representaciones numéricas (*embeddings*) también llamadas estados ocultos.

- **Decodificador**: Utiliza los estados ocultos generados por el codificador para generar una secuencia de salida, un token a la vez, de forma iterativa.

Lo anterior ha originado distintas familias de modelos basados en algunos componentes del Transfomer que han permitido resolver distintos tipos de problemas según el enfoque que se requiera [110]:

- **Basados en el codificador**: su característica principal es que convierten la secuencia de entrada en representaciones numéricas que contienen información del contexto de la secuencia. Este tipo de arquitectura es útil para tareas de clasificación y cualquiera de sus variantes. Algunos ejemplos de este tipo de modelos incluyen BERT [87] (y cualquiera de sus variantes) o RoBERTa [111].
- **Basados en el decodificador**: estos modelos se caracterizan por generar de forma iterativa la secuencia de salida, prediciendo cada elemento como el más probable según el contexto dado en la misma. A esta arquitectura de modelos pertenece la familia de Generative Pretrained Transformers (GPT) [112] que se enfocan en construir secuencias de palabras según lo que ya se ha escrito anteriormente en el proceso.
- **Modelos completos**: en esta clase de modelos se utiliza tanto el codificador como el decodificador para realizar tareas que requieran mapeos complejos de una secuencia a otra, como la traducción automática. Ejemplos de este tipo de modelos incluyen a Bidirectional and Auto-Regressive Transformers (BART) [113] y Text-to-Text Transfer Transformer (T5) [114].

2.2.1. Atención

Como se mencionó anteriormente, el mecanismo de atención es lo que diferencia al Transformer de otras redes neuronales para procesar información secuencial. Si se trabaja con texto, al momento de generar la secuencia de salida, se desea elegir automáticamente la mejor palabra según aquellas que ya se encuentran en la oración.

Ejemplo 2 (Oración de ejemplo) *El carro rojo es tan ancho que no cabe en el*

Consideremos la oración del Ejemplo 2, donde se busca que el modelo de Transformer nos diga cuál es la nueva palabra más adecuada según la secuencia actual de palabras. En este caso, es cierto que el contexto que proporciona la oración nos indica que lo más adecuado resulten ser palabras como «garaje». En ese sentido, algunas palabras son importantes para determinar qué palabra es la mejor para continuar con la secuencia, por lo que términos como «refrigerador» o «lavamanos» quedan totalmente descartados por hablarse del carro como objeto principal de la misma. Por otro lado, otros términos contribuyen menos para determinar la nueva palabra: ciertamente el color del carro (rojo) no afecta su tamaño. En resumen, se presta atención a ciertas partes de la oración y a otras no para extraer información de manera eficiente.

Para lograr esto, el mecanismo de atención (también llamado cabezal de atención), realiza un mapeo de tres vectores llamados *query* (Q), *value* (V) y *key* (K), en una única salida que contiene la información más relevante del contexto de la oración, como se muestra en la Figura 2.3. El query Q se puede interpretar como la búsqueda de la siguiente palabra de la oración, en este caso, ¿qué sigue después de «el»? Matemáticamente, como se aprecia en la Ecuación 2.1, Q es el producto de multiplicar el embedding del vector de entrada de la palabra «el», d_e, por una matriz de pesos W_Q. En esta ecuación, $d_e \in \mathbb{R}^{N_{d_e}}$,

donde N_{d_e} es su dimensión. Por otro lado, $W_Q \in \mathbb{R}^{N_{d_e} \times N_{d_k}}$, donde N_{d_k} es la dimensión de los queries y keys que adoptará el vector d_e. Es decir, su dimensión pasa de ser N_{d_e} a N_{d_k}.

$$Q = d_e W_Q \tag{2.1}$$

Los vectores *key* K son representaciones de cada palabra de la secuencia hasta el momento («El», «carro», «rojo», etc.). De forma similar al vector Q, los embeddings de cada palabra (ya no es una, sino varias) $d_e^{\#} \in \mathbb{R}^{\#P \times N_{d_e}}$, donde $\#P$ es el número de palabras que conforman el key y value, se multiplican por una matriz de pesos $W_K \in \mathbb{R}^{N_{d_e} \times N_{d_k}}$ para establecer la dimensión de cada embedding $d_e^{\#}$ a $\#P \times N_{d_k}$, como se muestra en la Ecuación 2.2.

$$K = d_e^{\#} W_K \tag{2.2}$$

El vector V son los embeddings $d_e^{\#} \in \mathbb{R}^{\#P \times N_{d_e}}$ de representación de cada palabra en la oración sin modificación alguna, multiplicadas por la matriz de pesos $W_V \in \mathbb{R}^{N_{d_e} \times N_{d_v}}$. El fin de esta acción es cambiar la dimensionalidad de $d_e^{\#}$ a $\#P \times N_{d_v}$, como se muestra en la Ecuación 2.3.

$$V = d_e^{\#} W_V \tag{2.3}$$

Después, cada key se compara con el query para determinar cuales son los términos más relevantes usando un producto punto, ya que ambos vectores son del mismo tamaño. Entre más grande sea el valor del producto punto, más relacionados están el query con el key en turno. El vector resultante de esta operación se escala por $\sqrt{N_{d_k}}$ para mantener la varianza de la suma del vector estable, y se le aplica una transformación softmax para que los resultados sumen 1, obteniendo los pesos de atención. Estos pesos después se multiplican por V para determinar la importancia de cada palabra. La Definición 3 representa la salida matricial de la atención resultante para una secuencia dada.

Definición 3 (Ecuación de atención) *Dados vectores de representación K, Q y V, la atención se define como:*

$$Atención(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{N_{d_k}}}\right) V \tag{2.4}$$

Finalmente, para obtener un único vector de salida del cabezal de atención, la atención se suma para obtener un vector de tamaño $d_v \in \mathbb{R}^{N_{d_v}}$, llamado vector de contexto, que captura las opiniones de las palabras para determinar el nuevo término. Cabe mencionar que, como toda red neuronal, lo que se tiene que ajustar son las matrices de pesos W_Q, W_K y W_V.

2.2.2. Atención Multicabezal

En lugar de utilizar el mecanismo de atención con un solo cabezal, resulta útil realizar la proyección de Q, K y V h veces, cada una con diferentes proyecciones aprendidas. El resultado de cada cabezal de atención se concatena y la forma final se somete a una matriz de pesos $W_O \in \mathbb{R}^{hd_v \times d_e}$ para proyectar el vector a la dimensión de salida final deseada,

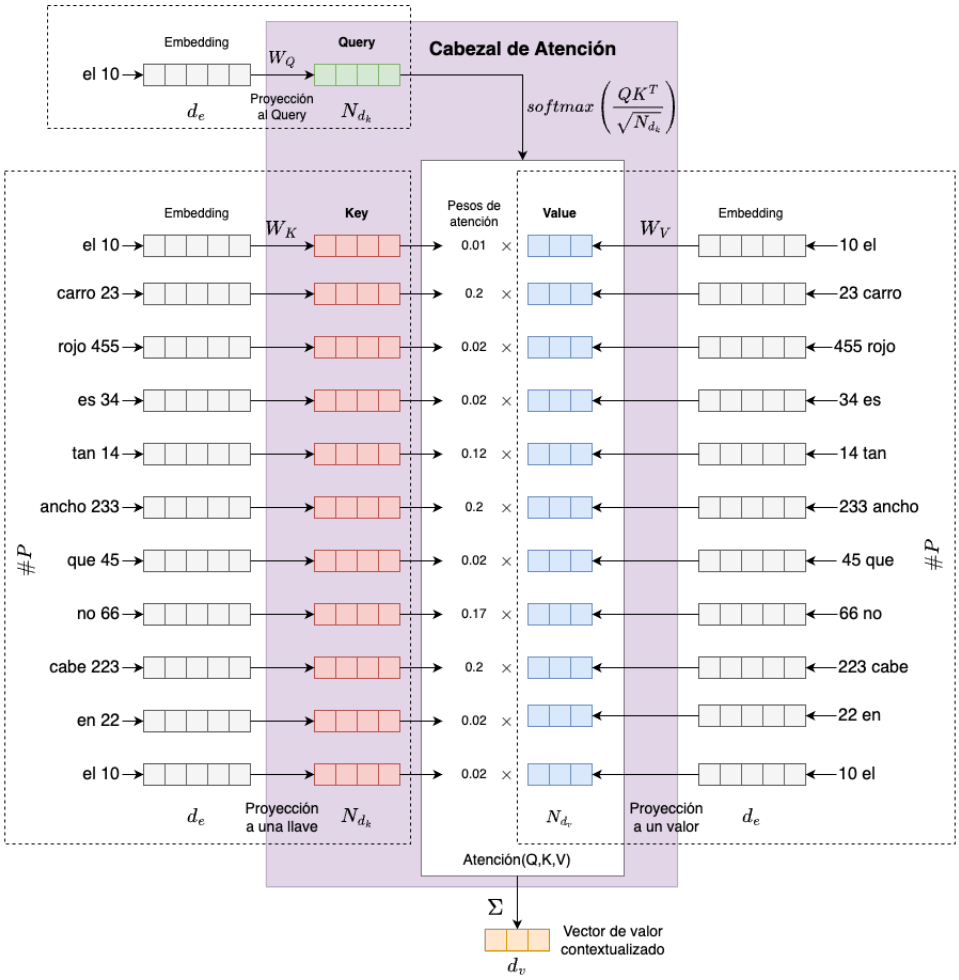


FIGURA 2.3: Diagrama del cabezal de atención utilizado en el Ejemplo 2.
Fuente: elaboración propia, inspirado en [115].

que en este caso es d_e. La Figura 2.4 ejemplifica lo anterior. Al realizar lo anterior se obtiene la atención multicabezal, como se explica en la Definición 4.

Definición 4 (Ecuación de Atención Multicabezal) *Dados vectores de representación K, Q y V, la atención multicabezal se define como:*

$$MultiCabezal(Q,K,V) = \Big\|_{i=1}^{h} (cabezal_i) W_O \tag{2.5}$$

donde

$$cabezal_i = softmax\left(\frac{d_e W_Q^i \cdot d_e^{\#} W_K^{i\,T}}{\sqrt{N_{d_k}}}\right) d_e^{\#} W_V^i \tag{2.6}$$

y $\|$ representa la operación de concatenación de vectores.

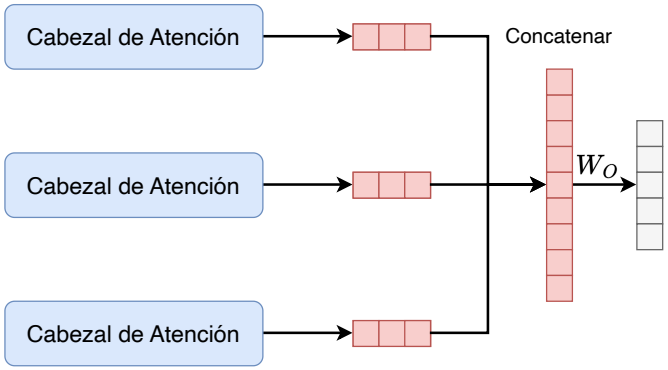


FIGURA 2.4: Esquema de la atención multicabezal con tres cabezas. Fuente: elaboración propia, inspirado en [115].

2.2.3. El Bloque del Transformer

El bloque del Transformer, ilustrado en la Figura 2.5, es un componente que se encuentra dentro de toda la arquitectura del Transformer. En primer lugar, como entradas se tienen los vectores Q, K y V que pasan a la capa de atención multicabezal. Sin embargo, el query se añade a la salida de la capa. A esto se le conoce como *skip connection* y sirve para evitar problemas de gradiente descendiente. Después, la salida se somete a una capa de normalización para proveer estabilidad durante el proceso de entrenamiento y mejorar el rendimiento de la red neuronal [116]. Finalmente, un perceptrón multicapa se incluye con el fin de extraer características de alto nivel conforme se avanza en el proceso.

2.2.4. Arquitectura del Transformer

La Figura 2.6 muestra la arquitectura propuesta para el Transformer en su totalidad. Del lado izquierdo se tiene el codificador, en total $N = 6$ bloques, que incluyen el mecanismo de atención multicabezal y el perceptrón multicapa. En este punto, se codifica la secuencia para aprender una buena representación de la misma para pasarla al decodificador. Del lado derecho, se tienen bloques de Transformer, también $N = 6$ bloques, que actúan como el decodificador. Se introducen los estados ocultos, que son los vectores K y V, permaneciendo únicamente el query Q de la secuencia de entrada del decodificador, lo que resulta en atención referencial cruzada que permite influenciar el proceso de generación de secuencias. En la obra original, Vaswani et al. [30] usan al Transformer para una tarea secuencia a secuencia, como la traducción automática, donde entra una cadena de texto y se debe generar la traducción correspondiente en el idioma objetivo. En este caso, la salida del decodificador se utiliza como entrada del mismo, un token generado a la vez, para generar la secuencia completa de salida. Usando el ejemplo anterior, se genera una palabra a la vez y cada palabra generada se usa como entrada del decodificador para completar el proceso. Por otro lado, en el trabajo original, la atención multicabezal original consta de $h = 8$ cabezales con una dimensión de $N_{d_k} = N_{d_v} = 64$, aunque ésto puede cambiar según la arquitectura con la que se trabaje.

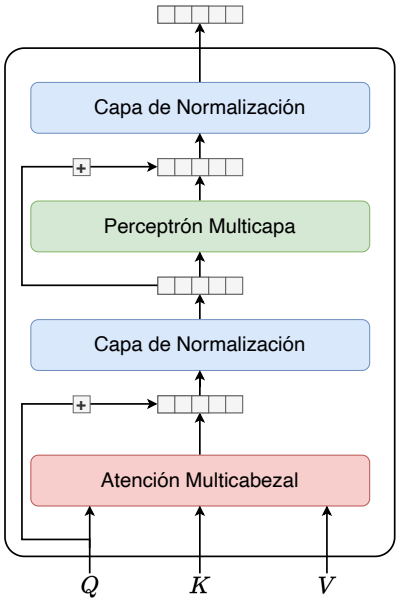


FIGURA 2.5: Diagrama de la estructura del bloque del Transformer. Fuente: elaboración propia.

2.3. Bidirectional Encoder Representations from Transformers

Bidirectional Encoder Representations from Transformers (BERT) [87] es un modelo de lenguaje desarrollado por Google en 2018 que revolucionó el campo del procesamiento del lenguaje natural. Su diseño se basa en la arquitectura Transformer, en particular su codificador. A diferencia de otras arquitecturas como la LSTM [117] que procesan secuencias de forma unidireccional, destaca por su capacidad para procesar el contexto de las palabras de manera bidireccional, lo que le permite capturar relaciones semánticas más profundas. Su arquitectura se puede apreciar en la Figura 2.7.

BERT tiene dos versiones originales que fueron entrenadas con grandes corpus de datos, específicamente Wikipedia en inglés y BooksCorpus de Google:

- **BERT Base** tiene 12 capas de Transformer, vectores de tamaño 768 y alrededor de 110 millones de parámetros.
- **BERT Large** cuenta con 24 capas de Transformer, vectores de tamaño 1024 y cerca de 340 millones de parámetros, lo que le permite manejar tareas más complejas.

Otro punto importante sobre BERT son los tokens especiales que ocupan un rol importante en el modelo durante su entrenamiento y para darle estructura a la secuencia de entrada. Por ejemplo, el token [CLS], que se ubica al inicio de la secuencia, sirve como embedding para ser usado como entrada en otros modelos de clasificación. Otros tokens especiales incluyen [SEP] para le separación de las oraciones, [EOS] para marcar el fin de una oración, [MASK] para indicar las palabras que se enmascaran durante el

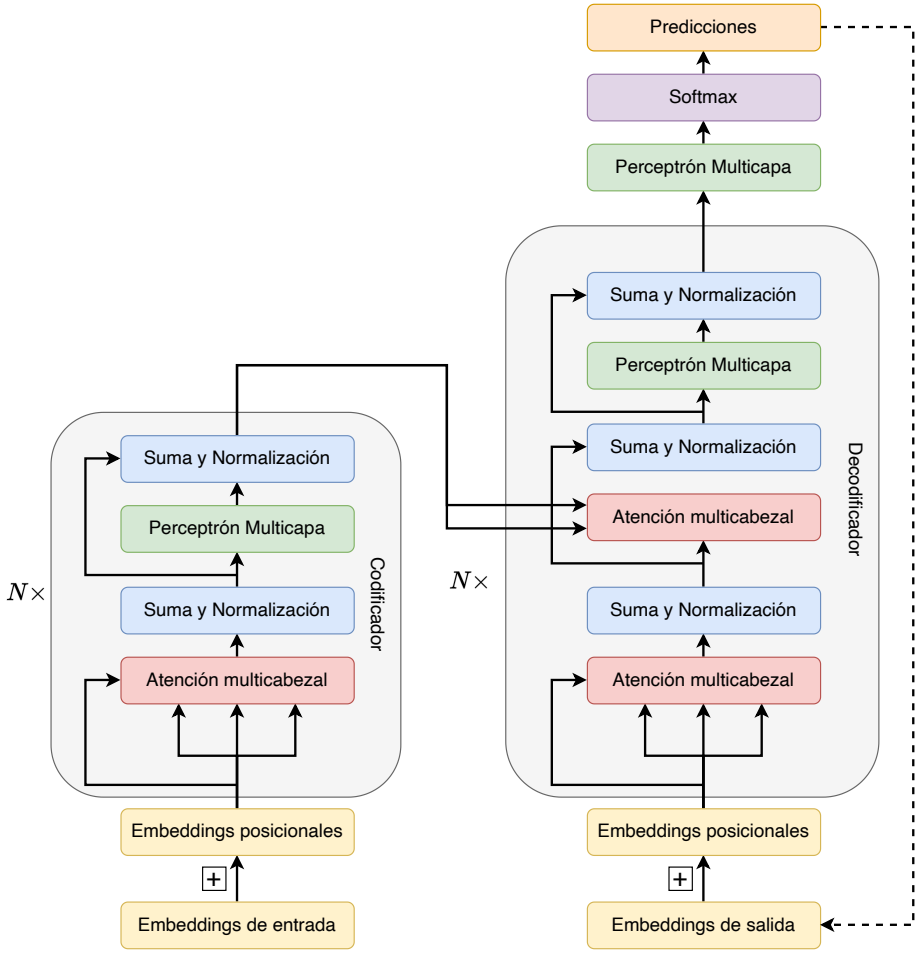


FIGURA 2.6: Modelo de codificador y decodificador del Transformer. Como entrada se toman los embeddings de la cadena de texto a codificar. Fuente: elaboración propia, basado en [30].

entrenamiento de la red y [UNK] para denotar a los tokens desconocidos que no forman parte del vocabulario del modelo.

Por otro lado, modelos como BERT utilizan embeddings posicionales para incorporar información sobre el orden o la posición de los elementos dentro de una secuencia. La codificación posicional es un componente esencial en los Transformers que introduce información sobre el orden de los tokens en una secuencia, compensando la falta de estructura secuencial propia del mecanismo de atención. Al representar la posición mediante funciones seno y coseno, el modelo puede inferir relaciones relativas y absolutas entre tokens [118]. En el contexto de la recuperación contextual (por ejemplo, la generación de texto), la codificación posicional permite que la autoatención relacione tokens no solo por su contenido semántico, sino también por su ubicación en el texto. Esto es importante para mantener la coherencia en relaciones de dependencia larga (es decir, que se encuentren separadas en el texto) y la correcta interpretación de estructuras gramaticales o narrativas.

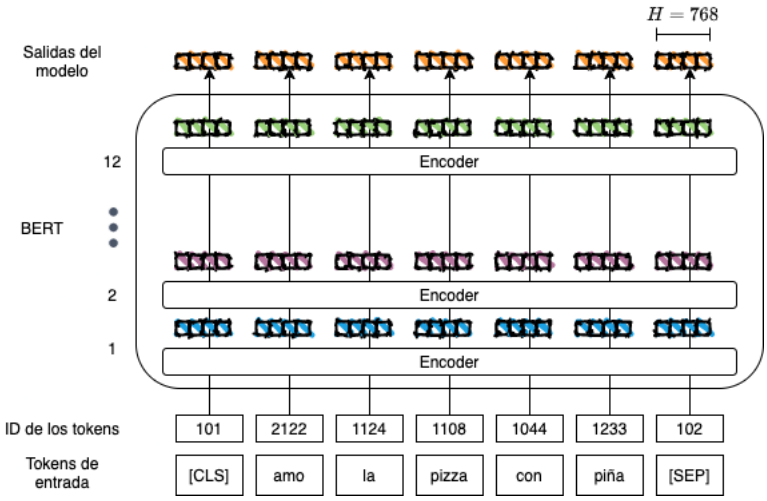


FIGURA 2.7: Ejemplo de la arquitectura de BERT Base para procesar una entrada de texto. Fuente: elaboración propia.

Esta codificación permite que la autoatención relacione elementos no solo por su significado, sino también por su ubicación contextual, lo cual es fundamental para tareas como la traducción, la comprensión lectora o la generación de texto. En la recuperación contextual, el PE actúa como un ancla que facilita que el modelo reconozca dependencias largas y relaciones sintácticas o semánticas precisas.

Dentro de las ventajas de usar BERT se encuentra la gran variedad de modelos derivados de esta arquitectura, desde modelos con menos parámetros pero más ligeros, hasta versiones grandes en distintos idiomas. Por otro lado, es posible usar y adaptar las distintas versiones de BERT mediante fine-tuning, lo que permite trabajar con pocos datos y no construir un modelo de aprendizaje desde cero. Algunas desventajas incluyen el costo computacional que implica usar las versiones grandes del modelo y sensibilidad a errores en el ajuste de hiperparámetros

2.4. Vision Transformer

Vision Transformer (ViT) es un modelo de aprendizaje profundo propuesto por Dosovitskiy et al. [31], que adapta la arquitectura Transformer para tareas de visión por computadora, reemplazando el uso de redes convolucionales, lo que permite modelar relaciones globales entre regiones de una imagen. Lo anterior demuestra que el Transformer puede ser útil en tareas de clasificación de imágenes cuando se entrena con suficientes datos.

Las imágenes en dos dimensiones primero deben adaptarse para que la arquitectura del Transformer las pueda manejar, como se observa en la Figura 2.8. Cada imagen S_I en dos dimensiones del conjunto de datos D, se divide en parches cuadrados S_{I_p}. $S_I \in \mathbb{R}^{H\times W\times C}$, donde H, W, C representan la altura, el ancho y los canales de la imagen, respectivamente. Además, $S_{I_p} \in \mathbb{R}^{N\times P\times P\times C}$, donde P es la resolución de cada parche de la imagen y $N = HW/P^2$ es la cantidad de parches resultantes que también funciona como la longitud de la secuencia de entrada al Transformer. Después, los parches generados se aplanan y se asignan a un vector de tamaño D mediante una proyección lineal que aprende

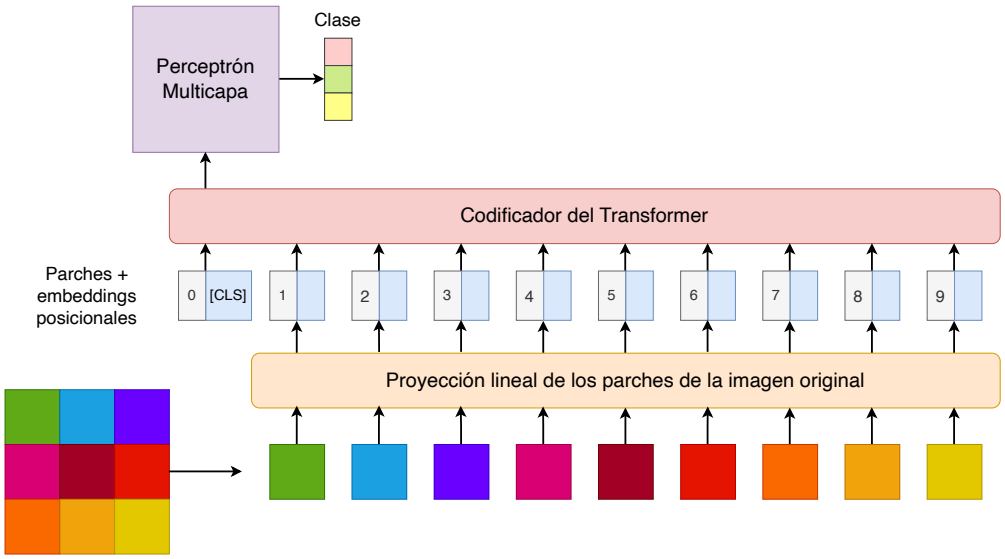


FIGURA 2.8: Arquitectura de Vision Transformer. Fuente: elaboración propia, inspirado en [31].

el modelo, es decir, los embeddings resultantes de cada parche. ViT utiliza embeddings posicionales, los cuales son sumados a las proyección lineales, para retener información posicional.

De forma similar a BERT, ViT considera el token [CLS] que se añade al inicio de la secuencia de parches para que sirva como embedding para la tarea de clasificación al salir de la última capa del modelo. En la arquitectura original, el modelo de clasificación es un perceptrón multicapa.

ViT cuenta con las siguientes versiones [31]:

- **ViT-Base**: Modelo base con 12 capas, vectores de tamaño 768, 12 cabezales de atención, y parches de tamaño 16×16. En total, cuenta con 86 millones de parámetros.
- **ViT-Large**: Modelo grande con 24 capas de Transformer, vectores de tamaño 1024, 16 cabezales de atención y parches de tamaño 16×16. En total, cuenta con 307 millones de parámetros.
- **ViT-Huge**: Modelo más grande con 32 capas de Transformer, vectores de tamaño 1280, 16 cabezales de atención y parches de tamaño 16×16. En total, cuenta con 632 millones de parámetros.

2.5. Modelos de Clasificación

2.5.1. Máquinas de Vectores de Soporte

La Máquina de Vectores de Soporte (MVS) [119] es un modelo de aprendizaje supervisado utilizado para clasificación binaria. Dado un conjunto de entrenamiento $D =$

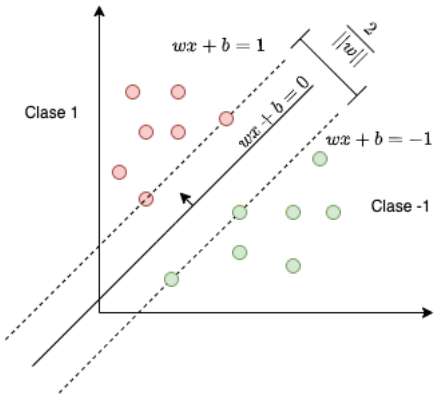


FIGURA 2.9: Idea general para la MVS. El hiperplano óptimo (línea sólida) para separar los datos se encuentra entre el margen generado por los hiperplanos de apoyo (línea puntuada). Fuente: elaboración propia.

$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ compuesto de n datos, donde $\boldsymbol{x}_i$ es el vector de representación de los atributos y $y_i \in \{-1, 1\}$ es la clase a la que pertenece el vector $\boldsymbol{x}_i$, se busca construir un hiperplano que separe óptimamente los puntos de una clase de la otra. Ese hiperplano óptimo se define como aquel donde la distancia entre cualquiera de los puntos de ambas clases tenga distancia máxima. La Figura 2.9 muestra gráficamente la idea general de la construcción del clasificador.

Un hiperplano es el conjunto de puntos que satisfacen la Ecuación 2.7

$$\boldsymbol{w}^T \boldsymbol{x} + b = 0 \tag{2.7}$$

donde $\boldsymbol{w}$ representa un vector normal (perpendicular) al hiperplano no necesariamente normalizado y b representa una constante. Suponiendo que los datos del conjunto D sean linealmente separables, se puede seleccionar dos hiperplanos de apoyo cuya distancia entre ellos sea lo más grande posible y, en el margen que generan, ubicar al hiperplano óptimo. Dichos hiperplanos de apoyo se describen como

$$\boldsymbol{w}^T \boldsymbol{x} + b = 1 \tag{2.8}$$

para el hiperplano que separa los puntos cuya clase es 1, y como

$$\boldsymbol{w}^T \boldsymbol{x} + b = -1 \tag{2.9}$$

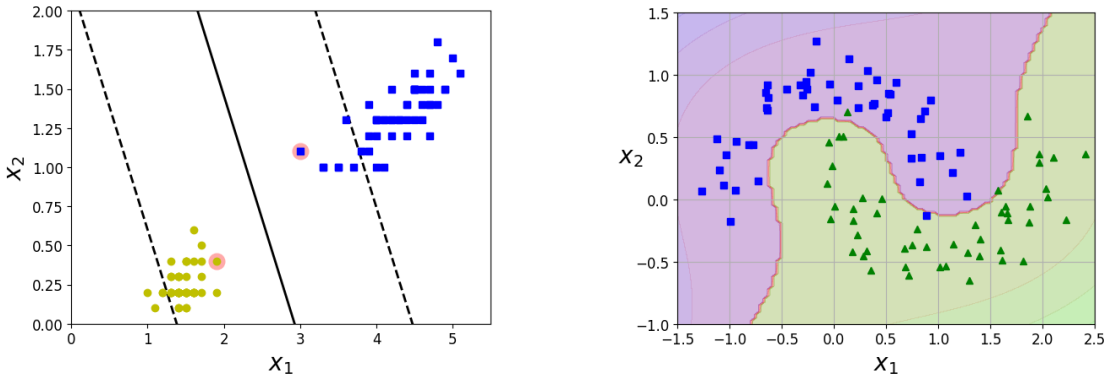
para el hiperplano que separa los puntos cuya clase es -1.

La distancia entre los dos hiperplanos de apoyo es igual a $\frac{2}{||\boldsymbol{w}||}$, donde $||\boldsymbol{w}|| = \boldsymbol{w}^T \boldsymbol{w}$, por lo que para maximizar la distancia entre ellos se debe minimizar $\boldsymbol{w}$. Además, se debe evitar que los puntos $\boldsymbol{x_i}$ entren al margen de separación:

$$\boldsymbol{w}^T \boldsymbol{x} + b \geq 1 \qquad \text{si } y_i = 1 \tag{2.10}$$

$$\boldsymbol{w}^T \boldsymbol{x} + b \leq -1 \qquad \text{si } y_i = -1 \tag{2.11}$$

Las expresiones anteriores se pueden juntar de la siguiente forma:



(A) (B)

FIGURA 2.10: (A): Caso de clasificación para el margen suave donde se admite la entrada al margen de separación. Sombreados en rojo se encuentran los vectores de soporte encontrados por el margen duro comparados con los encontrados por el margen suave. (B) Ejemplo de un problema linealmente no separable que se intenta separar mediante un kernel polinomial. Fuente: elaboración propia.

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \ . \tag{2.12}$$

Por lo tanto, el problema de optimización resultante es:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimizar}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} \\ \text{sujeto a} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 \ . \end{aligned} \tag{2.13}$$

La Ecuación (2.13) representa la forma genérica del clasificador, la cual se puede modificar para que se adapte a instancias particulares que dependen de la naturaleza de los datos.

Existen casos en que una separación perfecta no siempre es posible. Para manejar este caso se introduce un nuevo parámetro C creando un margen blando que permite errores en la clasificación al mismo tiempo que se penalizan. Esto se refleja en la Ecuación (2.13) como [120]:

$$\begin{aligned} \underset{\mathbf{w},b,\xi_i}{\text{minimizar}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_i \xi_i \\ \text{sujeto a} \quad & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{2.14}$$

Por otro lado, al considerar casos en que no es posible separar los datos mediante un hiperplano, es decir, no es linealmente separable, se utiliza el truco del kernel. La Figura 2.10 muestra un ejemplo de problemas linealmente separables y no separables. Un kernel es una función $K(\boldsymbol{v}, \boldsymbol{w})$ tal que $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, es decir, define un producto punto entre $\boldsymbol{v}$ y $\boldsymbol{w}$ que cumple ciertas propiedades [121].

La idea general es que, si un problema en cierta dimensión no es separable linealmente, puede que al aplicar cierta transformación no lineal lleve al problema a otra dimensión

donde sí lo sea. Una vez separados los datos podemos regresar a la dimensión original aplicando una transformación inversa.

Sean $\boldsymbol{x}_i$, $\boldsymbol{x}_j \in \mathbb{R}^n$ filas del conjunto de datos D y $\langle \cdot, \cdot \rangle$ denota el producto punto entre dos vectores. Los kernels más utilizados son los siguientes:

- Kernel Lineal: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$
- Kernel Polinomial: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + r)^d$
- Kernel Función de Base Radial: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma \cdot ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right)$
- Kernel Sigmoide: $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh\left(\gamma \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + r\right)$

donde $\gamma > 0$ y $r, d \in \mathbb{R}$. Lo anterior se puede adaptar de la siguiente forma:

$$\begin{aligned} \underset{\mathbf{w}, b, \xi_i}{\text{minimizar}} \quad & \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + c\sum_i \xi_i \\ \text{sujeto a} \quad & y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{2.15}$$

donde $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ es una función que mapea un vector a otra dimensión y $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)$ denota al kernel.

2.5.2. Máquinas de Vectores de Soporte para el Caso Multiclase

Originalmente, las MVS se plantearon considerando un modelo de clasificación binaria, es decir, solo separan dos clases. Múltiples enfoques se han considerado para extender el algoritmo para el caso de clasificación multiclase. Se consideran dos formas de atacar el problema: uno contra todos y uno contra uno [122].

El enfoque uno contra todos [122] construye k modelos de MVS, uno para cada clase considerada. El i-ésimo modelo considera la i-ésima clase y sus elementos como la clase positiva y las $k-1$ clases restantes se consideran como las negativas. Dado el conjunto de datos $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n)$ donde $y_j \in \{1, ..., k\}$ es la clase del vector $\boldsymbol{x}_j$, la i-ésima MVS resuelve el siguiente problema:

$$\begin{aligned} \underset{\mathbf{w}, b, \xi_i^j}{\text{minimizar}} \quad & \frac{1}{2}||\boldsymbol{w}^j||^2 + C\sum_{i=1}^{L} \xi_i^j \\ \text{sujeto a} \quad & (\boldsymbol{w}^j)^T \cdot \phi(\boldsymbol{x}_i) + b^j \geq 1 - \xi_i^j \qquad \text{si} \quad y_j = i \\ & (\boldsymbol{w}^j)^T \cdot \phi(\boldsymbol{x}_i) + b^j \leq -1 + \xi_i^j \qquad \text{si} \quad y_j \neq i \\ & \xi_i^j \geq 0 \end{aligned} \tag{2.16}$$

Al resolver la Eq.(2.16) se tienen k funciones de decisión:

$$(\boldsymbol{w}^1)^T\phi(\boldsymbol{x}) + b^1$$

$$\vdots$$

$$(\boldsymbol{w}^k)^T\phi(\boldsymbol{x}) + b^k$$

Se dice que un vector desconocido $\boldsymbol{x}$ pertenece a la clase con el mayor valor en la función de decisión:

$$y = \underset{j=\in\{1,\ldots,k\}}{\operatorname{argmax}} \ (\boldsymbol{w}^j)^T \boldsymbol{\phi}(\boldsymbol{x}) + b^j. \tag{2.17}$$

El segundo método es llamado el enfoque de clasificación uno contra uno. Dicho enfoque construye $k(k-1)/2$ clasificadores donde cada uno se entrena con información de dos clases. Considerando los datos de la clase i y la clase j, se resuelve el siguiente problema [122]:

$$\begin{aligned} \underset{\mathbf{w^{jk}}, b^{jk}, \xi_i^{jk}}{\text{minimizar}} \quad & \frac{1}{2}||\boldsymbol{w}^{jk}||^2 + C\sum_{i=1}^{L} \xi_i^{jk} \\ \text{sujeto a} \quad & (\boldsymbol{w}^{jk})^T \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b^{jk} \geq 1 - \xi_i^{jk} \qquad \text{si} \quad y_i = j \\ & (\boldsymbol{w}^{jk})^T \cdot \boldsymbol{\phi}(\boldsymbol{x}_i) + b^{jk} \leq -1 + \xi_i^{jk} \qquad \text{si} \quad y_j = k \\ & \xi_i^{jk} \geq 0 \end{aligned} \tag{2.18}$$

Hsu y Lin [122] deciden usar la siguiente estrategia basada en votos: si $\text{sign}((\boldsymbol{w}^{ij})^T \boldsymbol{\phi}(\boldsymbol{x}) + b^{ij})$ dice que $\boldsymbol{x}$ pertenece a la i-ésima clase, se suma un voto a esa clase. Si dice que pertenece a la clase j entonces se da el voto a la clase j. La clase que se elige para $\boldsymbol{x}$ es aquella que tenga más votos.

2.5.3. Máquinas de Vectores de Soporte Sensible a Costos

Cuando se trabaja con conjuntos de datos no balanceados, es decir, aquellos donde una clase se encuentra con mayor representación que otra [123], existen diversas maneras para mitigar el efecto negativo que estos pueden causar en los modelos de aprendizaje. Una forma es utilizar aprendizaje sensible a costos (*cost-senstive learning*), una modificación a nivel algorítmico que supone costos de clasificación errónea asimétricos entre clases, que generalmente se realiza definiendo una matriz de costos [124] o un vector [125] que afecta la función de pérdida. En particular, una variación del vector de costos como se especifica en el paquete Scikit-learn[3] se adopta, donde cada componente del vector de costos $\boldsymbol{v}$ se define como:

$$v_i = \frac{n}{n_c \cdot n_i} \tag{2.19}$$

donde v_i es el costo ponderado asociado a i-ésima clase, n representa el número de muestras en el conjunto de datos D, n_c representa el número total de clases y n_i indica el número de instancias en el conjunto de datos para la clase i.

Después, cada modelo se modifica considerando un enfoque de Costo de Error Diferente (*Different Error Cost*) [126], adaptando el enfoque uno contra uno para el problema de clasificación multiclase de la MVS con margen suave de la siguiente manera:

[3] https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

$$\begin{aligned} \underset{\mathbf{w^{jk}},b^{jk},\xi_i^{jk}}{\text{minimizar}} \quad & \frac{1}{2}||\boldsymbol{w}^{jk}||^2 + v_j C \sum_{i|y_i=j}^{|D|} \xi_i^{jk} + v_k C \sum_{i|y_i=k}^{|D|} \xi_i^{jk} \\ \text{sujeto a} \quad & y_i \left((\boldsymbol{w}^{jk})^T \phi(\boldsymbol{x}_i) + b^{jk} \right) \geq 1 - \xi_i^{jk} \\ & \xi_i^{jk} \geq 0 \end{aligned} \tag{2.20}$$

Como resultado, se crean un total de $n(n-1)/2$ modelos de MVS, cada uno de los cuales entrena dos clases j y k, $j \neq k$. En la Ecuación (2.20), $\phi(\boldsymbol{x}_i)$ representa el truco del kernel y ξ_i^{jk} es la distancia del punto $\boldsymbol{x}_i$ a su límite de margen correcto definido por $\left((\boldsymbol{w}^{jk})^T \phi(\boldsymbol{x}_i) + b^{jk}\right)$, siendo $\boldsymbol{w}^{jk}$ un vector normal, b^{jk} la intersección y y_i la clase correspondiente de $\boldsymbol{x}_i$.

2.6. Métricas de Evaluación

Para evaluar el desempeño del modelo de clasificación, se seleccionó un grupo de métricas para evaluar los datos haciendo énfasis en tratar el problema del desequilibrio entre las clases. Un conjunto de datos se dice no balanceado cuando existen una desproporción significativa entre el número de ejemplos de cada clase presente en el problema [127]. Es decir, existen más elementos de una clase que de otras. La confiabilidad de la métrica de exactitud (*accuracy*) se ve socavada por su tendencia a ofrecer una evaluación demasiado optimista de la clase mayoritaria [128]. Por el contrario, se recomiendan métricas como la precisión equilibrada (*balanced accuracy*), la medida F_1 y el Coeficiente de Correlación de Matthews (CCM) debido a su capacidad para abordar este problema [129]. A pesar de la limitación mencionada anteriormente, la métrica de exactitud se incluye con el fin de proporcionar un punto de comparación.

2.6.1. Exactitud Balanceada

La métrica de exactitud balanceada evita estimaciones de rendimiento infladas en conjuntos de datos no balanceados. La definición adoptada para calcular esta métrica es la proporcionada por el paquete Scikit-learn[4]:

$$\text{exactitud balanceada}(y, \hat{y}, w) = \frac{1}{\sum \hat{w}_i} \sum_i 1(\hat{y}_i = y_i)\hat{w}_i \tag{2.21}$$

donde y_i es el valor verdadero de la i-ésima clase, $\hat{y}_i$ es el valor predicho y

$$\hat{w}_i = \frac{w_i}{\sum_j 1(y_j = y_i) w_j} \tag{2.22}$$

es el peso de la muestra ajustado para el peso correspondiente $1(y_j = y_i)$ es la función indicadora y, si n_i es el número de elementos de la clase i en la muestra de tamaño n:

$$w_i = \frac{n_i}{n} \tag{2.23}$$

[4] https://scikit-learn.org/stable/modules/model_evaluation.html#balanced-accuracy-score

2.6.2. Coeficiente de Correlación de Matthews

El CCM [130], una métrica de rendimiento originada en el campo de la bioinformática, es una adaptación del coeficiente de correlación de Pearson para evaluar la correlación en matrices de confusión [131]. Para el caso multiclase, si $t_k = \sum_i^K G_{ik}$ es el número de veces que la clase k ocurrió verdaderamente, $p_k = \sum_i^K G_{ki}$ es el número de veces que se predijo la clase k, $c = \sum_k^K G_{kk}$ es el número de muestras predichas correctamente y $s = \sum_i^K \sum_j^K G_{ij}$ es el número total de muestras para una matriz de confusión G con K clases, la fórmula es [132]:

$$\text{CCM} = \frac{c \cdot s - \sum_k^K p_k \cdot t_k}{\sqrt{(s^2 - \sum_k^K p_k^2)(s^2 - \sum_k^K t_k^2)}} \tag{2.24}$$

A diferencia del caso binario, al aplicar el CCM para problemas multiclase, el mínimo se encuentra en el rango de 0 a -1, y el máximo siempre es 1.

2.6.3. Medida F_1

La medida F_1 puede considerarse como una media armónica ponderada de las puntuaciones de precisión (*precision*) y exhaustividad (*recall*) [133]. Para el caso binario, la medida F_1 se define como:

$$\mathrm{F}_1 = 2 \cdot \frac{\text{precisión} \cdot \text{exhaustividad}}{\cdot\text{precisión} + \text{exhaustividad}} \tag{2.25}$$

Sin embargo, para el problema multiclase la definición cambia y existen distintas versiones. La puntuación F_1 se puede determinar mediante promedios micro, macro y ponderados [134]. En este trabajo, se considera el promedio ponderado F_1^w, que utiliza la media de la puntuación F_1 de cada clase mientras considera el número de ocurrencias reales de la clase en el conjunto de dato, expresado de la siguiente manera:

$$\mathrm{F}_1^w = \frac{1}{\sum_{l \in L} |y_l|} \sum_{l \in L} |y_l| \mathrm{F}_1(y_l, \hat{y}_l) \tag{2.26}$$

donde y_l es el conjunto de datos de la clase l, L es el conjunto de clases posibles y $\mathrm{F}_1(y_l, \hat{y}_l)$ es la puntuación binaria F_1 para la clase l.

Para comparar los resultados del modelo de clasificación, en este trabajo se elige la puntuación CCM debido a sus ventajas como métrica de evaluación entre las tareas de clasificación cuando se trabaja con conjuntos de datos no balanceados [135].

Capítulo 3

Metodología

En este capítulo se presenta el marco metodológico diseñado para contestar las preguntas que se plantean en la Sección 1.4. Primero, se presenta la forma de construir los conjuntos de datos usados durante los experimentos. Después, se exponen los detalles del modelo imagen y texto, las formas de fusión de información propuestas y se especifica la forma de entrenamiento de los modelos. Finalmente, se muestran las consideraciones adicionales que se toman en cuenta para cada experimento con cada conjunto de datos, tales como modelos para análisis preliminares, estudios de ablación y exploración de resultados.

3.1. Construcción de los Conjuntos de Datos

Los modelos de clasificación de imagen y texto necesitan datos para realizar su labor. Considerando las necesidades detectadas en la literatura actual, mencionadas en la Sección 1.4, se decidió construir dos conjuntos de datos mediante la API v2 de Twitter [1] (antes de que se convirtiera en X) para explorar el problema de clasificación multimodal: el Multimodal Spanish Sentiment Analysis Impact Dataset (MSSAID) que trata sobre eventos deportivos que involucran al boxeador Saúl "Canelo"Álvarezz el Multimodal COVID19 Mexico (MCOVMEX) sobre el COVID-19 en México.

En primer lugar, cada conjunto de datos debe conformarse de publicaciones que contengan texto e imágenes que, por la plataforma de X, éstas últimas pueden ser de uno a cuatro elementos visuales. Cabe mencionar que las imágenes pueden ser miniaturas de videos en lugar del video mismo, si es que existe. Por otro lado, el esquema de etiquetado considera cuatro categorías diferentes que reflejan el análisis de polaridad y detección de spam deseado para el análisis de información multimodal:

- +1 para un sentimiento positivo presente en un tuit, como lo son halagos, felicitaciones o análisis positivos.
- -1 para un sentimiento negativo presente en el tuit, incluyendo burlas, insultos y cualquier otra opinión o crítica negativa.
- 0 para un sentimiento neutral presente en un tuit. Por ejemplo, publicaciones cuyo fin principal es informar los resultados del evento de manera objetiva o tuits que redirigen a los usuarios a otras plataformas digitales que hablan sobre los eventos.

[1] https://developer.x.com/en/docs/x-api

Tabla 3.1: Campos anotados de cada conjunto de datos y su correspondiente descripción.

Etiqueta	**Descripción**
Sentimiento del texto	Sentimiento del texto en el tuit.
Sentimiento de Texto en Imagen	Sentimiento del texto que se considera relevante dentro de una imagen.
Sentimiento de Imagen	Sentimiento plasmado en cada imagen, de forma individual.
Sentimiento General de las Imágenes	Sentimiento expresado en conjunto por todas las imágenes de un tuit.
Sentimiento General del Tuit	Sentimiento de un tuit considerando el texto y las imágenes que contiene.

Tabla 3.2: Información solicitada a la API v2 de Twitter para la construcción del conjunto de datos MCOVMEX antes de su clausura.

Campo	**Descripción**
text	Texto del tuit.
has:media	Obtiene los elementos multimedia (imágenes). Entre 1 y 4 imágenes diferentes o una miniatura de video.
lang	Idioma del tuit, en este caso, español.
place_country	Especifica la ubicación para la recolección de los tuits. Se solicita que sean de México.
tweet.fields	Id del autor, fecha de creación y métricas públicas.
media.fields	Enlace de cada imagen para su recuperación.
date	Fechas para recuperar los datos

- +2 para spam, contenido presente en redes sociales digitales que no habla sobre el tema principal del conjunto de datos.

El esquema de anotación de cada conjunto de datos, resumido en la Tabla 3.1, etiqueta diferentes aspectos de un tweet, permitiendo un análisis de los componentes de forma individual y en conjunto, sus interacciones y el efecto de que generan las distintas modalidades en el sentimiento general de un tweet. De igual manera, los campos solicitados a la API v2 de Twitter y su descripción se pueden consultar en la Tabla 3.2. Las etiquetas de los datos consideradas en este trabajo son *Text Sentiment* (sentimiento del texto), *Image Sentiment* (sentimiento de la imagen) y *Overall Tweet Sentiment* (sentimiento general del tweet), donde este último campo es el objetivo principal del proceso de inferencia de los modelos de clasificación multimodal.

Para la anotación del conjunto de datos, se usaron todos los tuits anotados utilizando una estrategia de anotación cruzada con tres anotadores [136]. Dos anotadores trabajaron con los tuits en cuestión para después discutir con un tercero cualquier divergencia sobre la etiqueta asignada a cualquiera de los campos. De esta manera, se llega a un acuerdo para conservar toda la información del conjunto de datos y evitar discrepancias en los resultados que se presentan en el Capítulo 4.

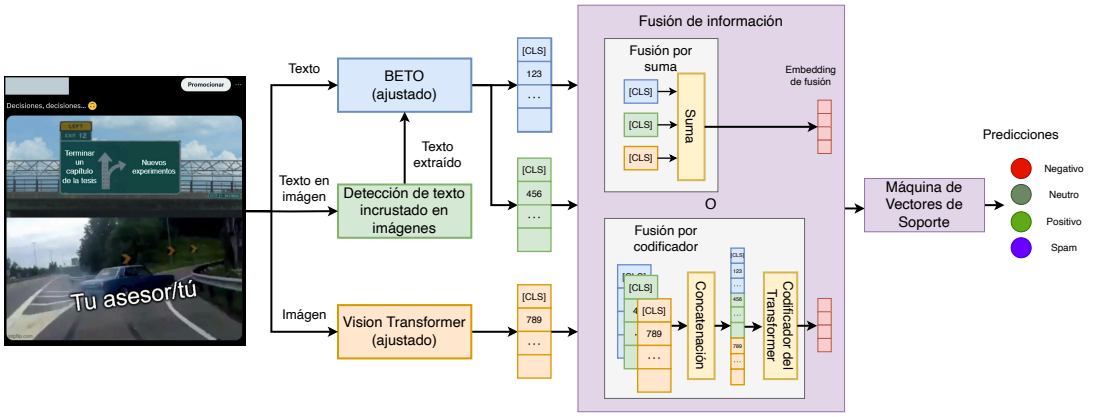


FIGURA 3.1: Diagrama general del método para trabajar con imágenes y texto propuesto para la tarea de análisis de sentimientos multimodal. Fuente: elaboración propia.

3.2. Modelo Imagen y Texto

Para realizar análisis de sentimientos multimodal, el estudio propone un marco de trabajo que incorpora y fusiona información en forma de texto, texto incrustado en imágenes y múltiples imágenes para determinar el sentimiento final de una publicación de X. El enfoque que se adopta para lograr este objetivo es procesar cada información por separado y fusionarlas mediante un método propuesto, como se puede observar en la Figura 3.1. En primer lugar, se extraen características de las distintas modalidades por separado mediante el ajuste de modelos vastos de lenguaje preentrenados basados en Transformer usando los conjuntos de datos disponibles, en particular las etiquetas de sentimientos relevantes, para entrenar los módulos de texto e imágenes. Después, el módulo de fusión se encarga de mejorar las representaciones individuales de cada modalidad al juntarlas y agregarlas por medio de uno de dos métodos de fusión de información: fusión por suma y fusión por codificador. Finalmente, el sentimiento final de cada publicación se obtiene al usar la representación multimodal unificada como entrada en una Máquina de Vectores de Soporte sensible al costo.

3.2.1. Extracción de Características en Textos

La extracción de características para la modalidad de texto se efectúa en un proceso que involucra dos pasos. En primer lugar, se ajusta (*fine-tune*) el modelo base de BETO, la versión en español de BERT, para después representar el texto de las publicaciones y generar las características correspondientes. En específico se puede ajustar la versión capitalizada[2] o sin capitalizar[3] de BETO, usando únicamente los textos de los tuits.

Sea $S_T = \{w_1, w_2, \ldots, w_m\}$ una secuencia de palabras que pertenecen a un elemento del conjunto de datos D y m es la longitud máxima admitida por BETO. Cada palabra se convierte en tokens y se transforma en un vector denso en función de su tipo de token. Luego, las características del texto F_{text} se extraen mediante el mapeo de cada token en un embedding de tamaño 768 mediante el modelo ajustado.

[2] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[3] https://huggingface.co/dccuchile/bert-base-spanish-wwm-uncased

$$F_{text} = [e_0^t, e_1^t, e_2^t, \ldots, e_{N_t}^t]^T \in \mathbb{R}^{N_t+1\times d} \tag{3.1}$$

En la Ecuación (3.1), e_j^t, $j \geq 1$, representa al embedding del j-ésimo token del texto de entrada (indicado por el superíndice t) tokenizado, d es la dimensión de salida de los embeddings generados por BETO, que en este caso es 768, y N_t es el número variable de tokens de la secuencia S_T. Cabe recordar que los modelos basados en BERT como BETO utilizan tokens especiales para indicar tareas especiales como el token [CLS] en e_0^t, el cual resulta ser útil para la tarea de clasificación.

3.2.2. Extracción de Características en Imágenes

Para la extracción de características en imágenes se sigue una idea similar que para el caso de texto al ajustar modelos preentrenados para clasificación de imágenes. La selección del modelo se centra en Vision Transformer, un modelo con una arquitectura similar a BERT que utiliza únicamente el codificador del Transformer para realizar la tarea de clasificación de imágenes. Cada imagen S_I del conjunto de datos D, $S_I \in \mathbb{R}^{H\times W\times C}$ entra al modelo de ViT para ser ajustado y generar embeddings F_{im} para las imágenes.

$$F_{im} = [e_0^{im}, e_1^{im}, e_2^{im}, \ldots, e_{N_{im}}^{im}]^T \in \mathbb{R}^{N_{im}+1\times d} \tag{3.2}$$

En la Ecuación (3.2), e_j^{im}, $j \geq 1$, representa el embedding contextual del j-ésimo parche de la imagen (indicado por el superíndice im) de la secuencia de imágenes aplanadas, d es la dimensión de salida (también 768) y N_{im} es el número de tokens de la secuencia de parches. De manera similar a BETO, ViT considera el token especial [CLS] en la posición e_0^{im}.

La principal ventaja de usar ViT sobre otros modelos para la clasificación de imágenes es el uso del token [CLS] en la arquitectura, el cual se usa posteriormente en el marco de trabajo. Además, ViT emplea la misma idea para procesar la secuencia de vectores utilizando un codificador Transformer como BERT, generando embeddings del mismo tamaño. El modelo base de ViT empleado para el ajuste es el modelo google/vit-base-patch16-224-in21k[4]. Sin embargo, se contrasta además con otros modelos. Es particular, se eligen para esta tarea ResNet [137] (resnet-50[5]) y vit-base-patch16-224[6].

3.2.3. Detección de Texto en Imágenes

Como se muestra en el diagrama de la Figura 3.1, es posible que algunas imágenes contengan texto incrustado en ellas. En este trabajo, se explora el efecto de incluir el análisis de tales dinámicas intermodales entre el texto de las imágenes y los demás elementos de una publicación ya que puede proveer información adicional y contribuir a la tarea de determinar de manera más certera el sentimiento de una publicación.

Con este fin, se entrena un modelo de detección de texto en imágenes. En primer lugar, se entrena un primer modelo para delimitar regiones de interés que contienen texto en imágenes. En este caso, se utiliza la red neuronal profunda para detectar objetos You Only Look Once (YOLO) v8, una arquitectura de detección de objetos rápida y precisa

[4] https://huggingface.co/google/vit-base-patch16-224-in21k
[5] https://huggingface.co/microsoft/resnet-50
[6] https://huggingface.co/google/vit-base-patch16-224

FIGURA 3.2: Ejemplo de una posible salida del sistema de detección de texto en imágenes para una imagen prototipo. Fuente: realización propia.

propuesta originalmente por Redmon et al. [138]. Dado que la tarea de identificar texto en una imagen es similar a la de detección de objetos realizada por YOLO, podemos aplicar la técnica de transfer learning para ajustar el modelo a nuestros datos.

El modelo de detección de regiones de texto se ajusta con un conjunto de 471 imágenes recolectadas manualmente de redes sociales que contienen texto en español, similar al formato de los memes[7]. Primero, las regiones de texto en la imagen se etiquetan manualmente para determinar las coordenadas del cuadro que delimita cada región de interés para cada imagen usando la herramienta de etiquetado de imágenes de Roboflow[8]. Luego, las coordenadas del cuadro delimitador se envían al modelo de YOLO v8 para ajustar sus parámetros utilizando el módulo de entrenamiento de Ultralytics[9] en Python 3. Cabe mencionar que se realiza un aumento de datos sobre las imágenes aplicando diferentes transformaciones para alcanzar un total de 1027 imágenes, entre ellas, traslación, rotación, voltear, recortar, cambio de colores y ruido sal y pimienta. La Figura 4.9 muestra un ejemplo de una posible salida del módulo de detección de texto.

Una vez que se detecta una región que contiene texto, las coordenadas del cuadro de esta se utilizan para crear una subimagen que se envía al motor de reconocimiento óptico de caracteres (ROC) para extraer el texto identificado y guardarlo como texto plano. En este trabajo, se emplea el motor de easyOCR[10] sin entrenamiento adicional dado su buen rendimiento durante experimentos previos con el idioma español. Luego, cualquier texto extraído se maneja como si fuera texto de una publicación, donde sus características se extraen utilizando el modelo BETO ajustado, como se explica en la Sección 3.2.1, generando el vector $F_{text_{im}}$.

3.2.4. Aumento de Datos en Texto

Trabajar con conjuntos de datos desequilibrados puede llevar a que el modelo se sesgue y únicamente prediga la clase más representada. Por lo tanto, surge la necesidad de mitigar este problema ya sea a nivel algoritmo, como se hace con la MVS con penalización, o con los datos [125]. Aumentar datos permite atacar el problema del desequilibrio

[7] https://universe.roboflow.com/canelo-q9gl8/textfinder-qhjhx
[8] https://roboflow.com/
[9] https://github.com/ultralytics/ultralytics
[10] https://github.com/JaidedAI/EasyOCR

entre clases y también aliviar, hasta cierto punto, el problema de los conjuntos de datos pequeños. En particular, cuando se aplica a texto, el aumento de datos ha permitido mejorar el rendimiento de modelos de aprendizaje [139]. Algunas técnicas comunes para aumentar datos incluyen sustituir palabras por sinónimos, cambiar el orden de palabras al azar e inclusive insertar palabras nuevas [140].

Las técnicas que se adoptan en este trabajo incluyen sustituir adjetivos y sustantivos por sinónimos mediante la versión en español de Wordnet[11], disponible en NLTK[12]; además, se elige como opción adicional la técnica de *back translation*: se parte del idioma objetivo (en este caso, español) y se traduce n veces a un idioma aleatorio para que, al final, se vuelva a traducir de vuelta al idioma original con el fin de modificar el documento original. El método propuesto incluye elegir una forma de aumento de datos al azar para cada documento (sinónimo o back translation) y se específica cuantas veces se busca aumentar el conjunto de datos.

3.2.5. Fusión de Información

La idea de fusionar la información de distintas modalidades previo al paso de clasificación sigue el concepto de mejorar las representaciones finales de la información en conjunto para así mejorar el rendimiento de los modelos de aprendizaje [107]. Para este proceso se proponen dos técnicas distintas y exclusivas para capturar dos paradigmas contrastantes de fusión: la fusión por suma y la fusión por codificador.

Fusión por Codificador

La primera técnica propuesta, denominada fusión por codificador, fusiona la información de las diferentes modalidades utilizando un bloque codificador del Transformer, como se observa en la Figura 3.3. La idea principal detrás de esta técnica es que, al mirar una publicación como un tuit, analizamos cada modalidad presente de manera secuencial. Es decir, primero se obtiene información de, por ejemplo, el texto, se procede a observar la imagen y cualquier posible elemento en ella para finalmente generar una opinión o determinar un sentimiento a partir de las diversas piezas de información disponibles. Dado que este proceso se modela como una secuencia de pasos, justificamos el uso de la estructura del codificador del Transformer ya que modela principalmente entradas secuenciales. Para realizar esta técnica de fusión y obtener el vector de representación fusionado F^{enc}, utilizamos los vectores generados F_{text}, F_{im} y $F_{text_{im}}$ a partir de los modelos previamente ajustados.

Para empezar, se concatenan los vectores F_{text}, F_{im} y $F_{text_{im}}$ en un nuevo vector $Z_{text_{im}}$ siguiendo un conjunto de reglas. Para permitir escenarios más generales, se adopta la idea de que las imágenes complementan la modalidad principal de información, que es el texto en este trabajo. Por lo tanto, solo se mantiene el token [CLS] presente en F_{text}, descartando los de las imágenes o el texto en imágenes. La decisión se toma porque no todos los elementos del conjunto de datos tienen imágenes disponibles en el momento de su uso. Por ejemplo, si un punto de datos contiene texto, dos imágenes y el módulo de detección de texto extrae dos regiones de texto, la operación resultante es la siguiente:

[11] https://wordnet.princeton.edu/
[12] https://www.nltk.org/howto/wordnet.html

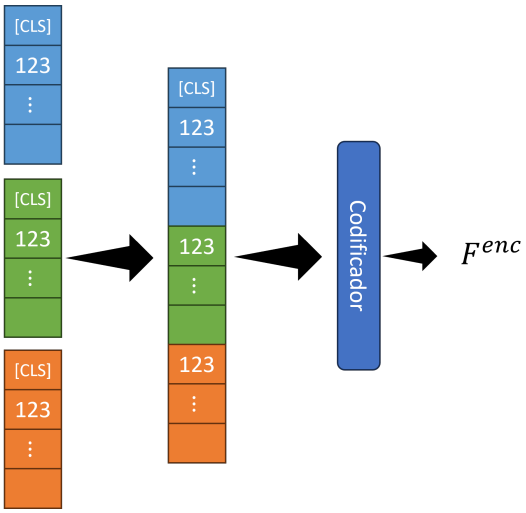


FIGURA 3.3: Diagrama que muestra la estrategia de fusión por codificador para tres elementos entrantes al sistema y su salida correspondiente. Fuente: elaboración propia.

$$Z_{text_{im}} = [e_0^t, e_1^t, \ldots, e_{N_t}^t, e_1^{im_1}, \ldots, e_{N_{im_1}}^{im_1}, e_1^{im_2}, \ldots, e_{N_{im_2}}^{im_2}, e_1^{t_{im_1}}, \ldots, e_{N_{t_{im_1}}}^{t_{im_1}}, e_1^{t_{im_2}}, \ldots, e_{N_{t_{im_2}}}^{t_{im_2}}] \tag{3.3}$$

En la Ecuación (3.3), los superíndices t corresponden a los embeddings del texto (F_{text}), im_1 e im_2 indican los embeddings de la primera (F_{im_1}) y la segunda (F_{im_2}) imagen, respectivamente, sin el primer elemento del vector. Además, los superíndices t_{im_1} y t_{im_2} indican los embeddings del primer ($F_{text_{im_1}}$) y el segundo ($F_{text_{im_2}}$) texto en imagen detectados. De forma general para cualquier cantidad de elementos n_t de textos, n_{im} imágenes y $n_{t_{im}}$ elementos de texto extraídos de imágenes:

$$Z_{text_{im}} = \Big\|_{i=1}^{n_t} \Big\|_{j=1}^{n_{im}} \Big\|_{k=1}^{n_{t_{im}}} F_{text_i} F_{im_j \setminus 1} F_{text_{im_k} \setminus 1} \tag{3.4}$$

donde $\|$ es el operador de concatenación, $F_{im_j \setminus 1}$ representa el embedding de la j-ésima imagen sin el primer elemento del vector y $F_{text_{im_k} \setminus 1}$ representa el embedding del k-ésimo texto de imagen sin el primer elemento del vector.

El nuevo vector $Z_{text_{im}}$ se emplea como elemento de entrada para el bloque del codificador del Transformer, para obtener la representación fusionada de las modalidades F^{enc}. Por último, el nuevo token [CLS] generado por este bloque, ubicado al inicio de la secuencia de salida, se utiliza como entrada para el método de clasificación.

Fusión por Suma

La segunda técnica para fusionar información, denominada fusión por suma, crea un único vector de características F^{sum} sumando cada token [CLS] de cada modalidad, ilustrado en la Figura 3.5. Dado que todas las características se extraen utilizando arquitecturas basadas en Transformer, el tamaño de los vectores de embeddings es el mismo

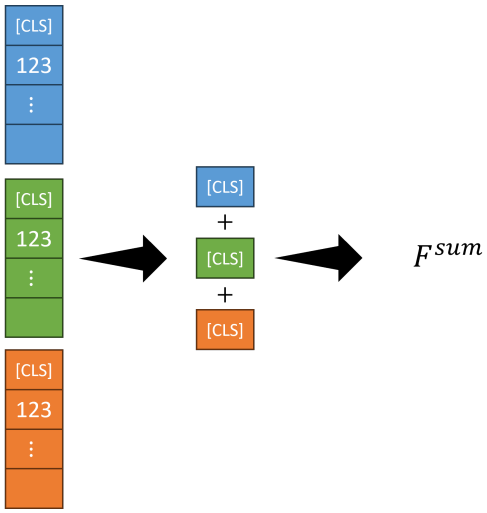


FIGURA 3.4: Diagrama que muestra la estrategia de fusión por suma para tres elementos entrantes al sistema y su salida correspondiente. Fuente: elaboración propia.

para todos esos tokens. Por lo tanto, es posible realizar una suma usando los vectores disponibles F_{text}, F_{im} y $F_{text_{im}}$. Por ejemplo, si un tuit tiene texto y dos imágenes, donde una de ellas contiene texto incrustado, el marco de trabajo extraerá un total de cuatro embeddings que se suman para formar el vector de características final, como se explica en la Ecuación (3.5):

$$S_{text_{im}} = e_0^t + e_0^{im_1} + e_0^{im_2} + e_0^{t_{im_1}} \tag{3.5}$$

y de forma general:

$$S_{text_{im}} = \sum_{i=1}^{n_t} e_0^{t_i} + \sum_{j=1}^{n_{im}} e_0^{im_j} + \sum_{k=1}^{n_{t_{im}}} e_0^{t_{im_k}} \tag{3.6}$$

El nuevo vector resultante del proceso se alimenta al algoritmo de clasificación para inferir el sentimiento general de un tuit.

3.2.6. Modelo de Clasificación

El modelo de clasificación que se considera en el Modelo Imagen y Texto es una Máquina de Vectores de Soporte con aprendizaje sensible al costo, como se especifica en la Sección 2.5.3.

3.3. Definición del Problema

En la Sección 1.4 se plantearon las preguntas de investigación que motivan la presente investigación. A continuación, se presenta una definición formal del problema que aborda el trabajo.

Definición 5 (Planteamiento del Problema) *Dada una publicación ρ, se busca sus características textuales F_{text}, sus características F_{im} de las imágenes S_I, donde $I =$*

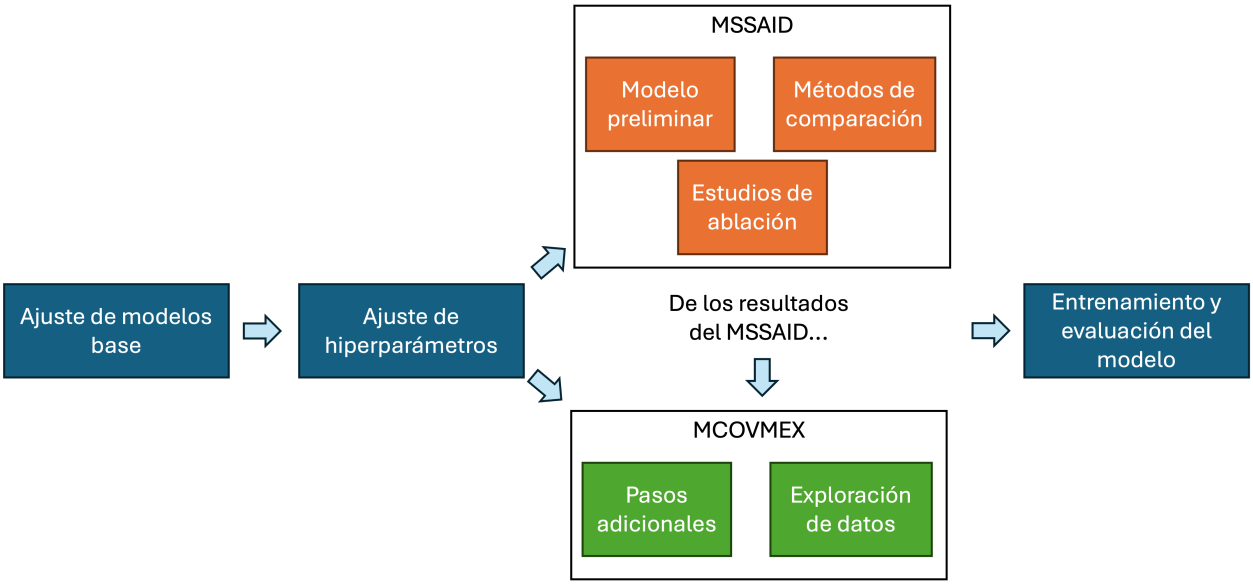


FIGURA 3.5: Esquema de los pasos adicionales que se requieren y se llevan a cabo para el entrenamiento del modelo propuesto de imagen y texto. Fuente: elaboración propia.

$\{1, 2, \ldots, k\}$ *y las características de texto en imágenes* $F_{text_{im}}$*, para definir una función de predicción:*

$$f(\rho) = f(F(F_{text}, F_{im}, F_{text_{im}})) = sentimiento(\rho) \tag{3.7}$$

En la Ecuación 3.7, F *representa el mecanismo de fusión multimodal (*F^{enc} *o* F^{sum}*) y* $sentimiento(\rho) \in \{-1, 0, +1, +2\}$ *designa la clase asignada a cada publicación.*

3.4. Detalles Sobre el Entrenamiento del Modelo Imagen y Texto

El flujo de trabajo que consolida la idea del modelo de imagen texto ilustrado en la Figura 3.1 consta de los pasos descritos anteriormente en la Sección 3.2. Sin embargo, existen una serie de pasos técnicos que deben llevarse a cabo para que el modelo de imagen y texto propuesto se pueda desarrollar de manera satisfactoria. Tales pasos se muestran en la Figura 3.5 y se describen con más detalle a continuación.

3.4.1. Ajuste de los Modelos Base

Los modelos base de BETO y ViT se ajustan con la biblioteca Transformers[13] de Hugging Face[14] [141] con Python 3 utilizando el modelo Sequence Classification con cuatro clases (positivo, negativo, neutro y spam). La etiqueta de clase objetivo que se usa para ajustar el modelo base de texto e imágenes es el sentimiento general del tuit. La métrica de rendimiento monitoreada durante el entrenamiento es el CCM. Cada modelo

[13] https://huggingface.co/docs/transformers/index
[14] https://huggingface.co/

se entrena 100 epochs con detención temprana y 10 epochs de tolerancia para evitar el sobreajuste, una tasa de aprendizaje de 2^{-5} para evitar modificar demasiado los parámetros originales de las redes originales y un tamaño de batch de 4. Durante el proceso de entrenamiento, un 10 % adicional de los datos, tomados del conjunto de entrenamiento, se reserva para el conjunto de validación necesario en el proceso.

3.4.2. Ajuste de Hiperparámetros para la Fusión por Codificador

Es importante recalcar que, para el método de fusión por codificador, se deben de ajustar una serie de hiperparámetros adicionales que se originan al usar los bloques de codificador del Transformer para fusionar los datos. A saber, se necesita encontrar el mejor número de capas de codificadores y el número de cabezales por capa. Para el hiperparámetro de número de cabezales, este debe dividir a la dimensión del embedding entrante. Aunque para el número de capas no existe limitación alguna, el número máximo se limita a cinco dado que la carga computacional de los experimentos se vuelve un componente crítico y se vuelve imposible obtener resultados de ciertas combinaciones de hiperparámetros. Para obtener estos resultados, se realiza una malla de búsqueda para las distintas combinaciones de cabezales y capas posibles considerando el tamaño del embedding entrante.

3.4.3. Entrenamiento y Evaluación del Modelo

El conjunto de datos se divide en dos partes: un conjunto de entrenamiento y un conjunto de prueba en una proporción de 90/10, respectivamente. Esta proporción se elige para permitir que los modelos de clasificación aprendan con la mayor cantidad de datos posible dada la limitante del tamaño de los mismos. Cada modelo se entrena con el conjunto de entrenamiento. Por otro lado, el conjunto de prueba se emplea para evaluar su rendimiento, comparar los resultados entre modelos y elegir el mejor de ellos. Para entrenar los modelos de clasificación, se tuvo acceso a una computadora con las siguientes especificaciones: Windows 10 de 64 bits, procesador Intel Xeon W-2295 3.00 GHz, 64 GB de memoria RAM y una tarjeta gráfica RTX A4000 de 16 GB.

Se entrenan los modelos de clasificación mediante validación cruzada con 10 pliegues usando el conjunto de entrenamiento. Para determinar el mejor conjunto de parámetros, se utiliza una malla de búsqueda en la que se consideran secuencias de crecimiento exponencial de C y γ con un kernel de función de base radial (FBR) [142]. Específicamente, $C = 2^{k_C}$ y $\gamma = 2^{k_\gamma}$, donde $k_C \in [-5, 16]$ y $k_\gamma \in [-15, 4]$. El rango de parámetros se selecciona empíricamente debido a la tendencia de γ de adoptar valores pequeños, mientras que C puede seguir valores más grandes.

El proceso de búsqueda de los mejores parámetros considera una primera iteración para acercar los valores iniciales de C y γ, para después refinar la exploración en repeticiones subsecuentes examinando la vecindad de los parámetros encontrados en la última salida considerando una ventana de $0.25/2^n$ tanto para k_γ como para k_C, donde $n \in \mathbb{N}$ corresponde a la iteración adicional en curso. Este proceso se repite hasta que se alcanza un umbral de tolerancia en el CCM de 1×10^{-4}. El valor de tolerancia se selecciona en función de su impacto en el tiempo de entrenamiento de los modelos, ya que el aumento de iteraciones requeridas afecta su duración. Finalmente, las métricas de evaluación para todo modelo se obtienen utilizando el mejor modelo entrenado con validación cruzada con los parámetros optimizados.

Para evaluar los modelos de clasificación se consideran las siguientes métricas de evaluación para contrastar los resultados: coeficientes de correlación de Matthews, exactitud balanceada, medida F_1 y exactitud, donde el CCM es el que se utiliza para vigilar los entrenamientos de los modelos, como se definen en la Sección 2.6.

Cabe resaltar que, durante el ajuste que se realiza a los modelos base BETO y ViT con el texto e imágenes, respectivamente, cada modelo genera su propio error y se optimiza para reducirlo. En cambio, el codificador del Transformer empleado para fusionar las características del texto e imágenes de las publicaciones, ilustrado en la Figura 3.1, no se ajusta con el error generado después de fusionar los datos y alimentar dicha representación a la MVS. La única parte que se entrena y ajusta al error de salida en el modelo multimodal propuesto es el algoritmo de clasificación.

3.5. Pasos Adicionales: Multimodal Spanish Sentiment Analysis Impact Dataset

3.5.1. Modelo Preliminar para Análisis Multimodal

Aunque el marco del trabajo descrito en la Sección 3.2 explica el proceso principal al que se someten los datos de imágenes y texto de publicaciones de redes sociales en esta investigación, durante el trabajo con el MSSAID se realizaron trabajos adicionales con métodos alternativos con la finalidad de construir distintas técnicas de clasificación y llegar a la base del modelo propuesto.

En primer lugar, se construyó un primer esquema de clasificación multimodal bajo la premisa de considerar el mejor rendimiento posible del módulo de etiquetado de sentimientos de imágenes. Con este fin, se aprovechan los campos de sentimientos etiquetados manualmente de las imágenes como si fueran los valores de salida del módulo y evitar instancias mal clasificadas, como se observa en la Figura 3.6. Los valores que se manejan son la polaridad general del sentimiento de las imágenes.

Para el caso del texto se consideran métodos clásicos como Bolsa de Palabras y Term Frequency - Inverse Document Frequency (TF-IDF) con combinaciones de unigramas (1-gramas) y bigramas (1-2-gramas); y unigramas, bigramas y trigamas (1-3-gramas). La razón para seleccionar estos modelos es que son un buen punto de partida para la investigación principalmente por su simplicidad y facilidad de uso. El paso de preprocesamiento y procesamiento de texto se implementa con NLTK 3.7 en Python 3.8.9, mientras que para la extracción de embeddings con Bolsa de Palabras y TF-IDF, se usa Scikit-learn.

También se incorporan el procesamiento de emojis ya que son una forma de contenido ampliamente adoptada y utilizada por los usuarios en varios sitios de redes sociales digitales [143]. Para incorporarlos al flujo de trabajo, se tiene en cuenta un enfoque que consiste en traducir cada emoji a su equivalente en español dentro del texto con la ayuda de la biblioteca de Python, Emoji[15]. De esta manera, se extraen características usando n-gramas con los emoticones, como se explicó anteriormente.

Similar al modelo principal, se extrae texto de las imágenes mediante una variación de YOLO v3 [144] entrenada usando la idea de transfer learning y los mismos principios que se explicaron en la Sección 3.2.3. Para realizar transfer learning, YOLO v3 se entrenó con

[15]https://github.com/carpedm20/emoji/

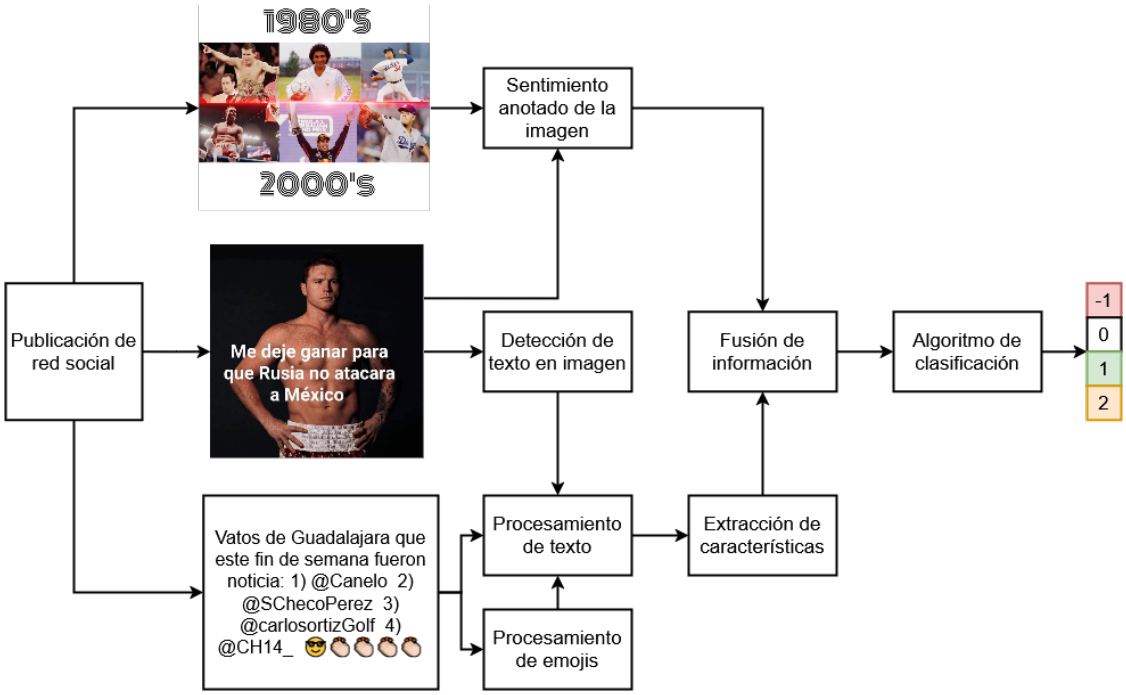


FIGURA 3.6: Diagrama general que muestra el primer proceso de clasificación multimodal que considera directamente las etiquetas de los sentimientos de las distintas modalidades de la información disponible. Fuente: elaboración propia.

Keras 2.8.0 y Tensorflow 2.8.0. El procesamiento de imágenes se realiza con scikit-image 0.19.2 [145].

Para fusionar la información obtenida de las distintas modalidades se concantenan las características del texto y la etiqueta del sentimiento general de la imagen. El texto obtenido de los emojis, así como el texto obtenido de las imágenes, se agrega al texto del tweet, de modo que toda esa información se procesa en conjunto durante el paso de procesamiento del texto. Finalmente, el algoritmo de clasificación utilizado es una MVS sensible a costos con kernel de función de base radial, una optimización de parámetros similar a la explicada en la Sección 3.2.6 y la métrica que se vigila es la exactitud balanceada.

3.5.2. Métodos de Comparación

Para comparar los resultados del Modelo de Imagen y Texto propuesto con los distintos métodos de fusión, se decidió usar un modelo representativo en la literatura para la clasificación multimodal que involucra imágenes y texto. Contrastive Language Image Pretraining [146] (CLIP) es un modelo de red neuronal que asocia imágenes y descripciones de texto en inglés mediante aprendizaje contrastivo. La particularidad de CLIP es que permite comprensión intermodal, permitiendo tareas como la clasificación de imágenes en función de descripciones textuales y viceversa, aprovechando un único modelo para tareas multimodales. A pesar de su popularidad, tiene una representación limitada en español. Por este motivo, se emplea una versión multilingüe de CLIP [147] con

capacidad para trabajar con español, entre otros idiomas. El modelo seleccionado es sentence-transformers/clip-ViT-B-32-multilingual-v1[16].

Dado que CLIP solo puede manejar una imagen a la vez, se adapta su uso para comparar sus resultados contra aquellos obtenidos por los métodos propuestos. En primer lugar, se usa el mismo codificador de texto e imágenes que sugiere el modelo de CLIP mencionado anteriormente para generar los embeddings correspondientes para textos e imágenes. Para permitir una comparación justa, los embeddings generados por CLIP multilingüe se utilizan de la misma manera que los embeddings generados por BETO y ViT en la fusión por codificador y suma. En particular, para el enfoque de fusión por codificador, el token [CLS] se inicializa aleatoriamente antes de agregarlo al comienzo de la incrustación, siguiendo la misma idea que en los modelos de la familia BERT. Cabe mencionar que el tamaño del embedding generado por CLIP multilingüe es de 512.

Finalmente, el modelo de CLIP multilingüe se compara con los modelos base de BETO y ViT (sin ajustar) y los modelos ajustados de BETO y ViT.

3.5.3. Estudios de Ablación, Análisis Visual y Análisis de Error

Los estudios de ablación son experimentos diseñados para evaluar la importancia de diferentes componentes de un modelo eliminándolos y observando el impacto en el rendimiento [148] de, en este caso, los modelos de clasificación multimodal. Se utilizan para comprender mejor qué partes del modelo son esenciales para su funcionamiento y cómo contribuyen al desempeño general del mismo. Algunos componentes que suelen estudiarse son las características del modelo, arquitecturas de algoritmos y parámetros específicos para determinar su impacto.

Los estudios de ablación comprendidos en este experimento incluyen el estudio de la contribución de distintas modalidades de la información considerada en el modelo de clasificación multimodal. Es decir, ¿cómo afecta la incorporación del texto en imágenes y las imágenes al rendimiento de los modelos considerados y los esquemas de fusión de información? En segundo lugar, se efectúa un análisis del impacto del número de imágenes que se incorporan al modelo de clasificación multimodal. En otras palabras, ¿qué le sucede al rendimiento del clasificador cuando se usa una, las primeras dos, las primeras tres o todas las imágenes disponibles?

Además de los estudios de ablación se procede a visualizar la calidad de los embeddings generados por los métodos de fusión (suma y codificador). Con este fin, se realiza la proyección en dos dimensiones de los embeddings originales generados por CLIP, BETO + ViT base y BETO + ViT ajustados mediante Uniform Manifold Approximation and Projection (UMAP) [149]. La visualización de los clústeres generados por las proyecciones ofrece información valiosa sobre los diferentes modelos, especialmente al ser capaz de observar la facilidad para separar las clases y las posibles debilidades de los marcos para la tarea de análisis de sentimientos multimodales.

El análisis de error considera análisis cualitativo de los errores generados por el mejor clasificador multimodal obtenido para detectar debilidades y posibles oportunidades.

[16] https://huggingface.co/sentence-transformers/clip-ViT-B-32-multilingual-v1

3.6. Pasos Adicionales: Multimodal COVID19 Mexico

Con este conjunto de datos se siguió el marco de trabajo descrito en la Sección 3.2 con algunos pasos adicionales que se eligieron al tomar en cuenta los resultados de los experimentos con el conjunto de datos MSSAID, además de agregar una tarea de exploración de datos usando los datos anotados para todo el conjunto de datos MCOVMEX.

3.6.1. Pasos Adicionales

Gracias a los resultados de los estudios de ablación llevados a cabo con el conjunto de datos MSSAID, se optó por evaluar el impacto de agregar determinado número de imágenes a los modelos de imagen y texto como un hiperparámetro adicional del modelo. En otras palabras, para ambas modalidades de fusión, se evaluó el desempeño de los clasificadores al considerar la primera, las primeras dos, las primeras tres y todas las imágenes existentes en las publicaciones, si es que las hay, desde un inicio. Además, para el paso de ajuste del modelo base para representación de las imágenes, se consideran modelos adicionales para su contraste. En específico, se descarta ResNet y se sustituye por Swin Transformer [150]. Los modelos considerados incluyen swin-tiny-patch4-window7[17], swin-base-patch4-window7-224[18] y swinv2-tiny-patch4-window16-256[19]. Adicionalmente, no se contrastan los modelos contra CLIP multilingüe y no se aumentaron los datos (texto ni imágenes).

3.6.2. Exploración de Datos

La exploración de datos se realiza con el fin de descubrir información latente en el conjunto de datos una vez que se logró estructurar y agregar parte de su información. En este caso, el proceso de agregación se lleva a cabo al etiquetar todos los tuits (no sólo los que conforman el conjunto de entrenamiento) usando el mejor clasificador encontrado al entrenar el modelo de aprendizaje usando el método de análisis multimodal de imagen y texto propuesto. Dicha separación de mensajes según su polaridad permite inspeccionar el discurso que se lleva a cabo en las publicaciones en cada categoría.

3.7. Resumen Final

Para concluir esta sección se incluye un pequeño resumen, expuesto en la Tabla 3.3, sobre qué es lo que se realiza con cada conjunto de datos para resaltar sus diferencias y lo personalizable que resulta ser el marco de trabajo propuesto para trabajar con distintos datos según se necesite. Cabe resaltar la cantidad de caminos disponibles al aplicar la metodología presentada en este capítulo, especialmente al decidir si se realiza o no un paso en particular, o los modelos usados para representar cada modalidad. Lo anterior permite personalizar y adaptar cada proyecto según se necesite. Además, no dependen de un conjunto de datos en particular: se pueden usar para elementos que tengan texto y cualquier número de imágenes.

[17] https://huggingface.co/microsoft/swin-tiny-patch4-window7-224
[18] https://huggingface.co/microsoft/swin-base-patch4-window7-224-in22k
[19] https://huggingface.co/microsoft/swinv2-tiny-patch4-window16-256

Tabla 3.3: Comparación de diferencias metodológicas llevadas a cabo entre el conjunto de datos MSSAID y el MCOVMEX.

	MSSAID	MCOVMEX
Modelos de imagen	vit-base-patch16-224-in21k vit-base-patch16-224 resnet-50	vit-base-patch16-224-in21k vit-base-patch16-224 vit-tiny-patch16-224 swin-tiny-patch4-window7-224 swin-base-patch4-window7-224 swinv2-tiny-patch4-window16-256
Modelos de texto	bert-base-spanish-wwm-cased bert-base-spanish-wwm-uncased	bert-base-spanish-wwm-cased bert-base-spanish-wwm-uncased
¿Aumento de texto?	Sí	No
¿Aumento de imágenes?	No	No
Métodos de comparación	CLIP multilingüe, modelos base y ajustados	No
Hiperparámetros fusión por codificador	CLIP multilingüe, modelos base y ajustados	Modelos ajustados + número de imágenes
Estudios de ablación	Impacto de modalidades y efecto de número de imágenes en modelos	No
Análisis visual	CLIP multilingüe, modelos base y ajustados	Mejor modelo
Análisis de error	Sí	Sí

Finalmente, muchos de los módulos que componen el modelo de imagen y texto se originaron como productos del trabajo realizado en el proyecto Imagen de México vinculado al proyecto de posgrado. Para su consulta, se puede dirigir al Apéndice D.

Capítulo 4

Resultados y Análisis

En este capítulo se presentan y discuten los resultados obtenidos en los diversos experimentos llevados a cabo con el modelo de imagen y texto propuesto en la Sección 3.2. En particular, se presentan los conjuntos de datos que se utilizan para los experimentos principales y una breve discusión de ellos. Además, se comparten los resultados obtenidos por los modelos de clasificación del modelo de imagen y texto propuestos usando los conjuntos de datos MSSAID y MCOVMEX. Aún más, se presentan los experimentos de ablación del impacto de las modalidades y el número de imágenes en los sistemas de clasificación.

4.1. Conjuntos de Datos

La estrategia propuesta en la Sección 3.1 para etiquetar las publicaciones de X supone una gran ventaja al mostrar aspectos multimodales de los datos que se manejan. A continuación, se muestran los conjuntos de datos que se utilizan para alimentar los modelos de imagen y texto y un breve análisis de datos de cada uno de ellos.

4.1.1. Multimodal Spanish Sentiment Analysis Impact Dataset

El Multimodal Spanish Sentiment Analysis Impact Dataset [61] es un conjunto de datos que contiene un total de 674 tuits sobre dos eventos deportivos diferentes que involucran al boxeador mexicano Saúl "Canelo" Álvarez, de los cuales se extrajeron el texto, imágenes y miniaturas de videos (en forma de imágenes) correspondientes. Los eventos contemplados son la pelea contra Caleb Plant (6 de noviembre de 2021) y la pelea contra Dmitri Bivol (7 de mayo de 2022). Todos los tuits se recolectaron en una ventana de tiempo de una semana antes, durante y una semana después del evento.

El esquema de etiquetado considera cuatro categorías diferentes:

- +1 para un sentimiento positivo presente en un tuit, por ejemplo, cualquier muestra de apoyo o cualquier otra opinión y crítica positiva hacia el boxeador.
- -1 para un sentimiento negativo presente en el tuit, incluyendo burlas, insultos y cualquier otra opinión o crítica negativa hacia el boxeador.
- 0 para un sentimiento neutral presente en un tuit. Por ejemplo, publicaciones cuyo fin principal es informar los resultados de una pelea de manera objetiva o tuits que redirigen a los usuarios a otras plataformas digitales que hablan sobre los eventos.

- +2 para spam, contenido presente en redes sociales digitales que no habla sobre el tema principal del conjunto de datos.

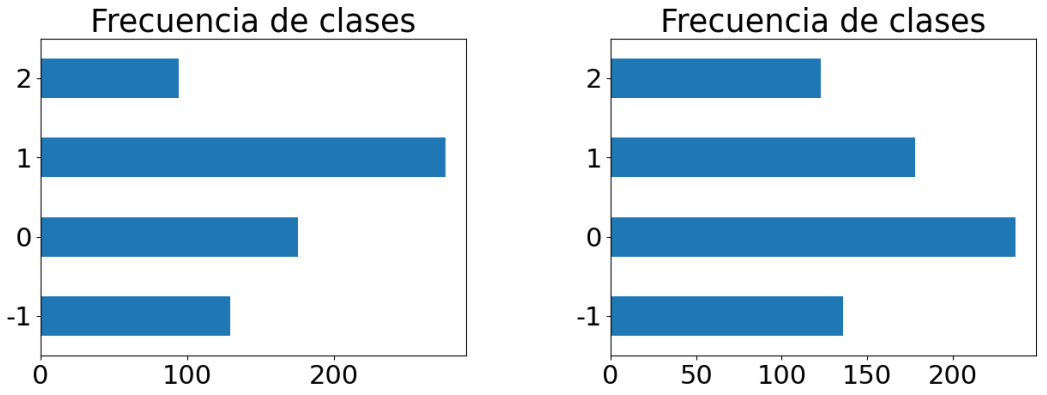
Dado que un tweet puede contener múltiples imágenes, el conjunto de datos involucra un total de 804 imágenes, aunque solo se pudieron recuperar 767 de ellas. Para este conjunto de datos se anotaron todas las publicaciones.

Un breve análisis del conjunto de datos, que se muestra en la Figura 4.1, arroja la frecuencia para cada clase de sentimiento considerada, donde se puede apreciar que los datos presentan desequilibrio entre ellos, lo cual justifica la elección de la MVS penalizada (véase la Sección 2.5.3) para aliviar este problema de forma algorítmica. La Figura 4.2, por otro lado, representa la dispersión de la longitud del texto, medida en palabras, de los tweets dadas las etiquetas de sentimiento del texto y sentimiento general. La dispersión de la longitud observada no supera las 512 palabras de longitud, lo que comprueba el uso de modelos basados en Transformer, específicamente la familia de BERT. Aún más, para el caso del MSSAID se puede observar que los mensajes de spam son ligeramente más largos. Sin embargo, esta característica (la longitud de los mensajes) no es suficiente para agregarla al modelo como discriminante dado que para las otras clases se asemeja mucho esta métrica.

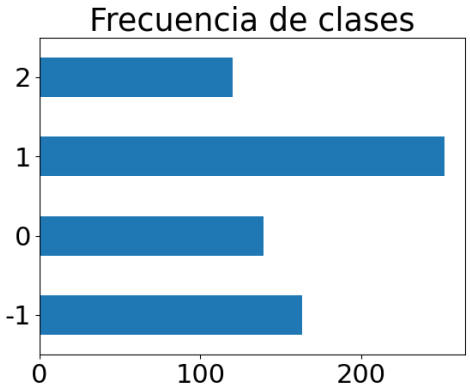
La Figura 4.3 muestra un diagrama de Sankey ilustrando la transición del sentimiento de los tuits al considerar únicamente el texto, a la izquierda, en contraste con el sentimiento considerado al incorporar las imágenes, a la derecha. El supuesto principal del trabajo es que el texto se complementa de las imágenes para determinar la polaridad final o general de cada publicación. En el caso del MSSAID, en el diagrama de Sankey, la mayor parte de los elementos que conforman cada polaridad al considerar texto e imágenes se compone de la misma polaridad. Sin embargo, se pueden observar cambios de polaridad para cada una de ellas. Por ejemplo, se puede apreciar que, para los elementos neutros, se cambia la polaridad de 55 elementos (31.43 % del total original), al considerar ahora texto con imágenes. Para el caso de positivos, se observa un cambio en 43 publicaciones (15.58 %); para los negativos, se presenta un cambio en 14 elementos (10.85 %); y para spam, 7 elementos (4.26 %). Se puede concluir que la clase neutra es la que mayores componentes sufren cambio de polaridad al agregar las imágenes como información adicional, mientras que el spam es la que menor transiciones reporta.

Todo lo anterior apunta a que, de manera efectiva, las imágenes ayudan a darle forma a la polaridad del sentimiento general del tuit y existen instancias (en este caso, 119 o el 17.66 % del total de datos) donde el texto no es suficiente para determinar de forma adecuada la polaridad de una publicación cuando se consideran elementos multimodales. Existen clases que son más sensibles que otras ante este tipo de comportamiento, pero en general, todas presentan transiciones.

Finalmente, en la Figura 4.4 se encuentra la cantidad de tuits según el número de imágenes que contienen y cuántas tienen texto incrustado en ellas. Es posible apreciar que las personas prefieren agregar una única imagen en sus publicaciones. Dado que el sistema de X permite subir hasta cuatro imágenes, conforme aumenta el número de imágenes cargadas, es más difícil observar a usuarios con tales preferencias al subir dicho número de imágenes, lo que hace que surja la pregunta sobre el impacto del número de las imágenes en los modelos de aprendizaje. Además, el texto en imágenes no es tan dominante en los elementos multimedia. A pesar de ello, no es razón suficiente para evitar el estudio de su impacto en los modelos de clasificación multimodal propuestos.

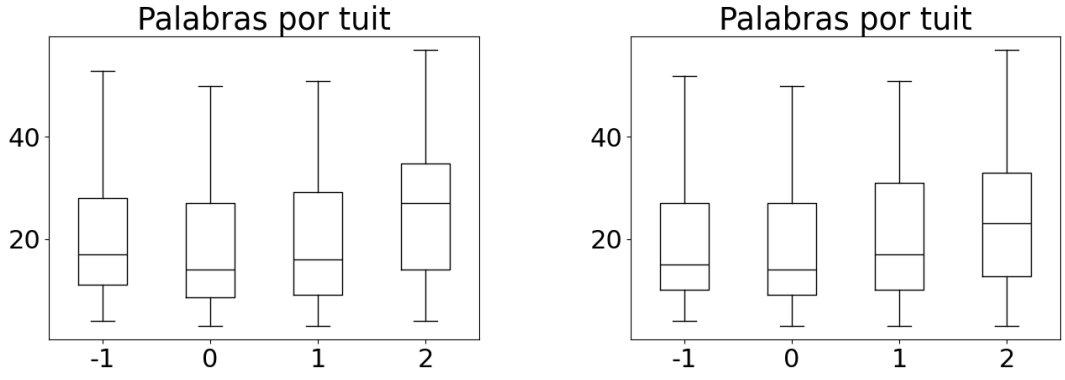


(A) (B)



(C)

FIGURA 4.1: Distribución de datos del MSSAID considerando: (A) sentimiento del texto, (B) sentimiento general de las imágenes y (C) sentimiento general de un tuit. Fuente: elaboración propia.



(A) (B)

FIGURA 4.2: Diagrama de cajas mostrando la longitud de los tuits según (A) el sentimiento del texto y (B) el sentimiento general de un tuit medido en palabras para las clases consideradas en el MSSAID. Fuente: elaboración propia.

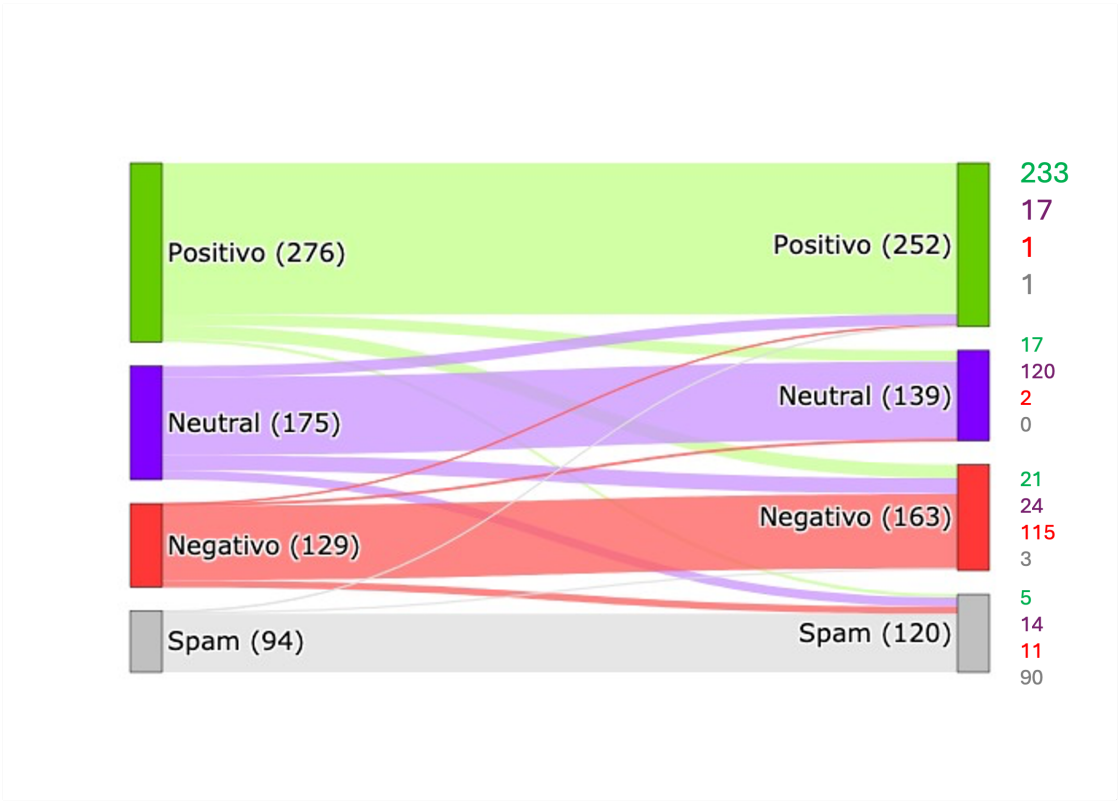


FIGURA 4.3: Diagrama de Sankey que muestra la transición del valor del sentimiento de los tuits para cada polaridad al considerar primero texto (izquierda), y después texto con imágenes (derecha) para el MSSAID. A la derecha se indica el número de elementos que transicionan de cada polaridad en la izquierda después de considerar el texto con imágenes para formar el total indicado en la derecha. En verde: positivo, morado: neutro, rojo: negativo, gris: spam. Fuente: elaboración propia.

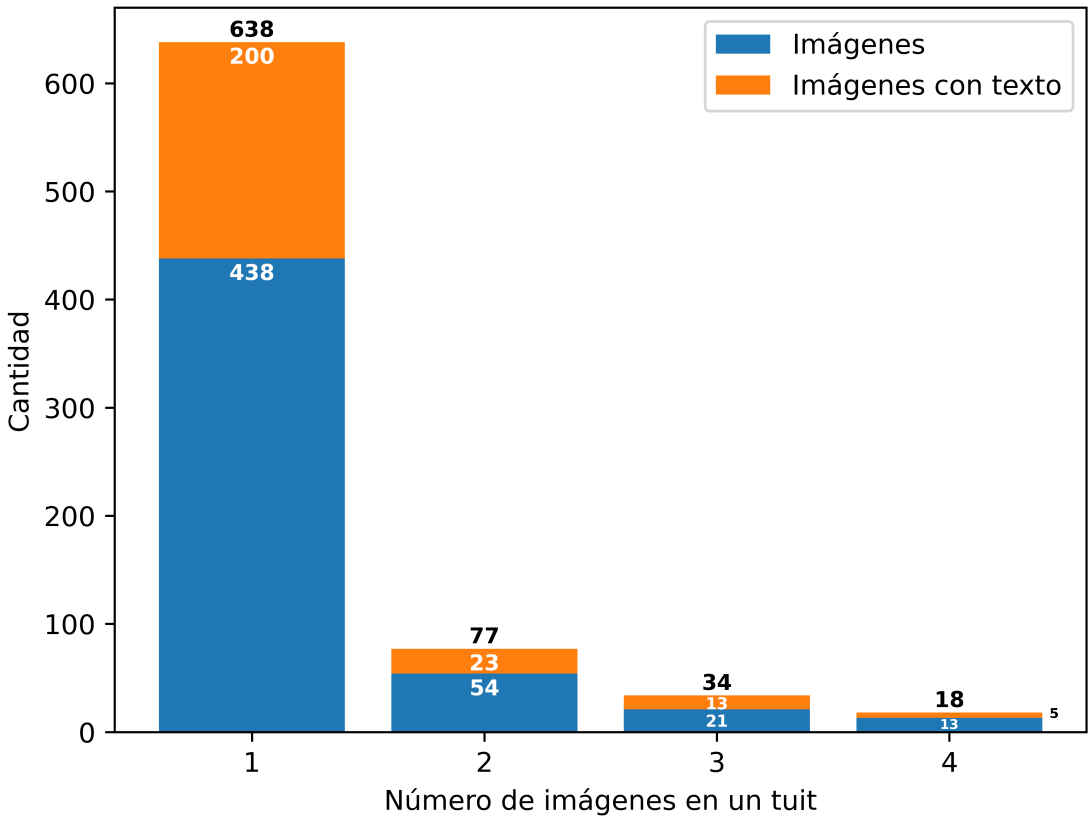


FIGURA 4.4: Cantidad de tuits según el número de imágenes que contienen. Fuente: elaboración propia.

4.1.2. Multimodal COVID19 Mexico

El conjunto de datos Multimodal COVID19 Mexico es un conjunto de datos que contiene 94,866 tuits compuesto por texto, imágenes y sus correspondientes métricas básicas (likes, retuits, comentarios y quotes), recolectados entre abril de 2020 y abril de 2023.

Se anotó manualmente una muestra aleatoria de 1000 tuits, de los cuáles, únicamente en 982 se logró recuperar sus respectivas imágenes. Las instrucciones de anotación se pueden consultar en el Apéndice A. El esquema de anotación considera las siguientes etiquetas:

- -1 para tuits negativos. Es decir, aquellas publicaciones que muestran claramente un sentimiento negativo predominante en su mensaje. Esto incluye mensajes sobre defunciones, problemas del personal médico, críticas hacia figuras públicas o decisiones tomadas por ellos, etc.
- 0 para tuits neutros. Esta categoría incluye publicaciones que no demuestran sentimiento preponderante alguno, pero cuyo tema se relaciona con el COVID. Esta categoría reúne principalmente a las publicaciones objetivas de medios noticiosos que dirigen tráfico a un sitio web (página web o canal de Youtube, por ejemplo).
- +1 para tuits positivos. Es decir, aquellas publicaciones que muestran claramente un sentimiento positivo predominante en su mensaje. Esto incluye mensajes de apoyo emocional, declaraciones positivas de salud, crítica positiva hacia algún desarrollo de la pandemia como las vacunas, etc.
- +2 para spam. Es posible encontrar mensajes que no hablan sobre el COVID. Por ejemplo, publicaciones que se cuelguen de las tendencias para vender productos o dirigir tráfico hacia otros sitios con fines de lucro o captar la atención de los usuarios con otros fines.
- -2 para indicar que la imagen no se encuentra disponible al momento de hacer la petición de búsqueda o descarga.

Un breve análisis del conjunto de datos, que se muestra en la Figura 4.5, arroja la frecuencia para cada clase de sentimiento. También se puede observar un desequilibrio entre las clases presentes en el conjunto de datos. La Figura 4.6 representa la dispersión de la longitud del texto, medida en palabras, de los tweets dadas sus etiquetas de sentimiento del texto y sentimiento general. Igual que con el MSSAID, las medidas de dispersión y las longitudes máximas justifican el uso de modelos basados en BERT.

El diagrama de Sankey del MCOVMEX de la Figura 4.7 muestra un menor número de transiciones: para la clase positiva se observa una (1) transición (0.53 % del total); en la clase neutral se presentan 4 transiciones (0.90 %); en la clase negativa existen 6 transiciones (2.28 %); y para la clase spam, ninguna transición.

Todo apunta a que el texto parece ser suficiente para determinar la mayor parte de la polaridad general del sentimiento en éste conjunto de datos. Sin embargo, la presencia de las transiciones, aunque mucho menor que en el caso pasado, específicamente 11 (1.12 % del total de datos), no debe ser ignorada: su incorporación presenta un potencial para mejorar los sistemas de análisis de sentimientos.

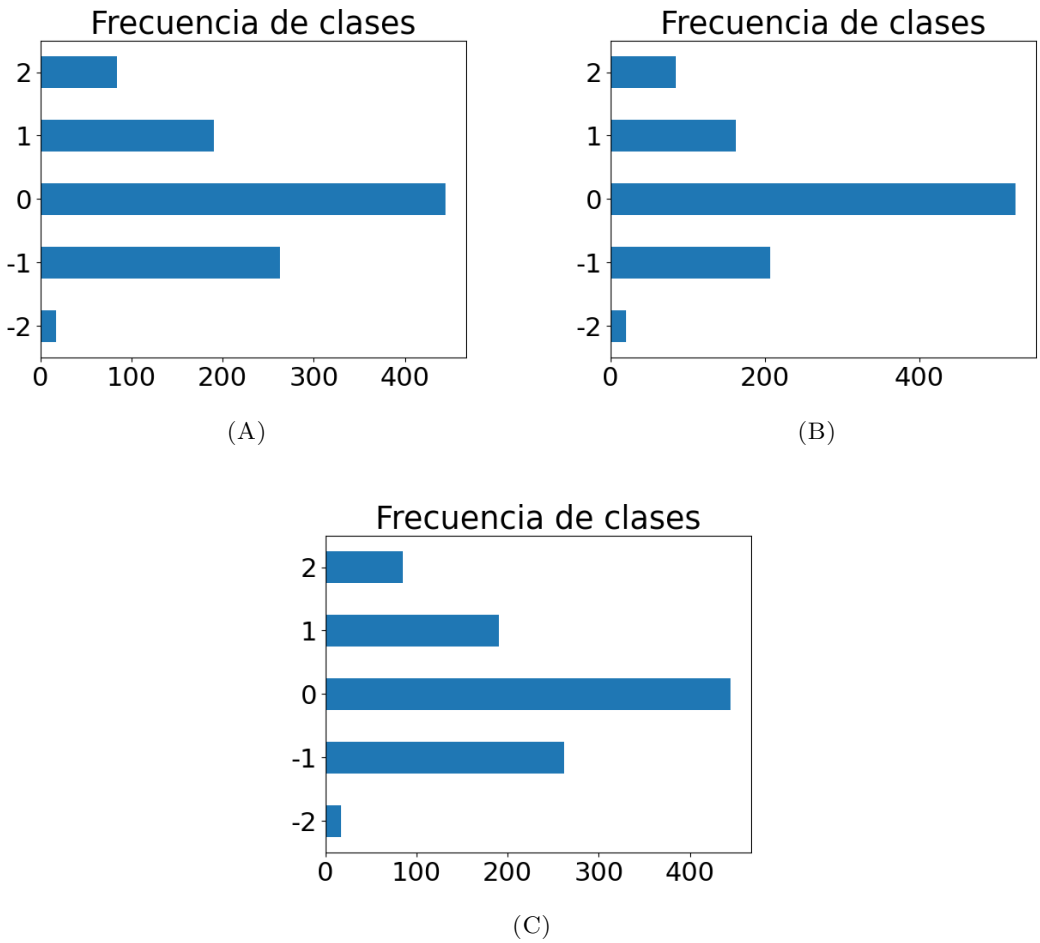


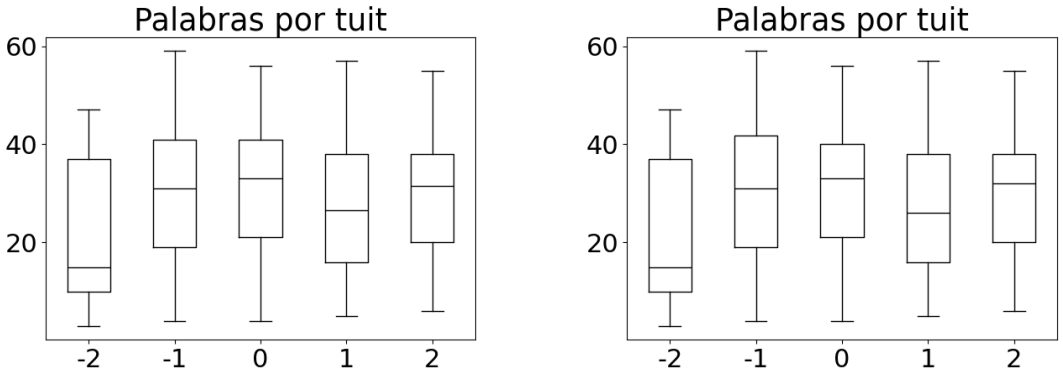
FIGURA 4.5: Distribución de datos del MCOVMEX considerando: (A) sentimiento del texto, (B) sentimiento general de las imágenes y (C) sentimiento general de un tuit. Fuente: elaboración propia.

Juntando los análisis del MSSAID y el MCOVMEX se puede concluir que incorporar el estudio de las imágenes como modalidad adicional presenta la oportunidad de mejorar los sistemas para la tarea de análisis de sentimientos. Existen mensajes donde, utilizando únicamente el texto, resulta complejo descifrar la polaridad general de una publicación. Tal es el caso de la clase neutra.

Finalmente, en la Figura 4.8 se encuentra la cantidad de tuits según el número de imágenes que contienen y cuántas tienen texto incrustado en ellas. Igual que en el caso anterior, las personas prefieren incorporar una única imagen a sus publicaciones y el texto en imágenes no es preponderante entre ellas.

4.1.3. Resumen de Resultados de la Sección

1. El esquema de anotación propuesto en la Sección 3.1 permite explorar de una mejor manera los conjuntos de datos al mejorar la determinación de la polaridad del sentimiento de una publicación cuando se consideran múltiples modalidades como el texto y las imágenes.



(A) (B)

FIGURA 4.6: Diagrama de cajas mostrando la longitud de los tuits según (A) el sentimiento del texto y (B) el sentimiento general de un tuit medido en palabras para las clases consideradas en el MCOVMEX. Fuente: elaboración propia.

2. Lo anterior se refleja en los diagramas de Sankey y, al analizar las transiciones, se puede observar el nivel de cambio que genera introducir una modalidad adicional al texto para determinar la polaridad general de una publicación. En este caso, el conjunto de datos MSSAID es más sensible ante esta situación que el MCOVMEX.

3. Los conjuntos de datos se encuentra desequilibrados, por lo que surge la necesidad de tratar dicho problema con un enfoque algorítmico (la MVS penalizada).

4. El análisis de frecuencias de las imágenes mostró que los usuarios prefieren usar, en su mayoría, una única imagen para acompañar el texto de una publicación. Por esta razón, ¿vale la pena mantener todas las imágenes en los modelos? o ¿sólo es necesario mantener algunas y no todas? Preguntas que se toman en cuenta en los experimentos de ablación (véase la Sección 3.5.3).

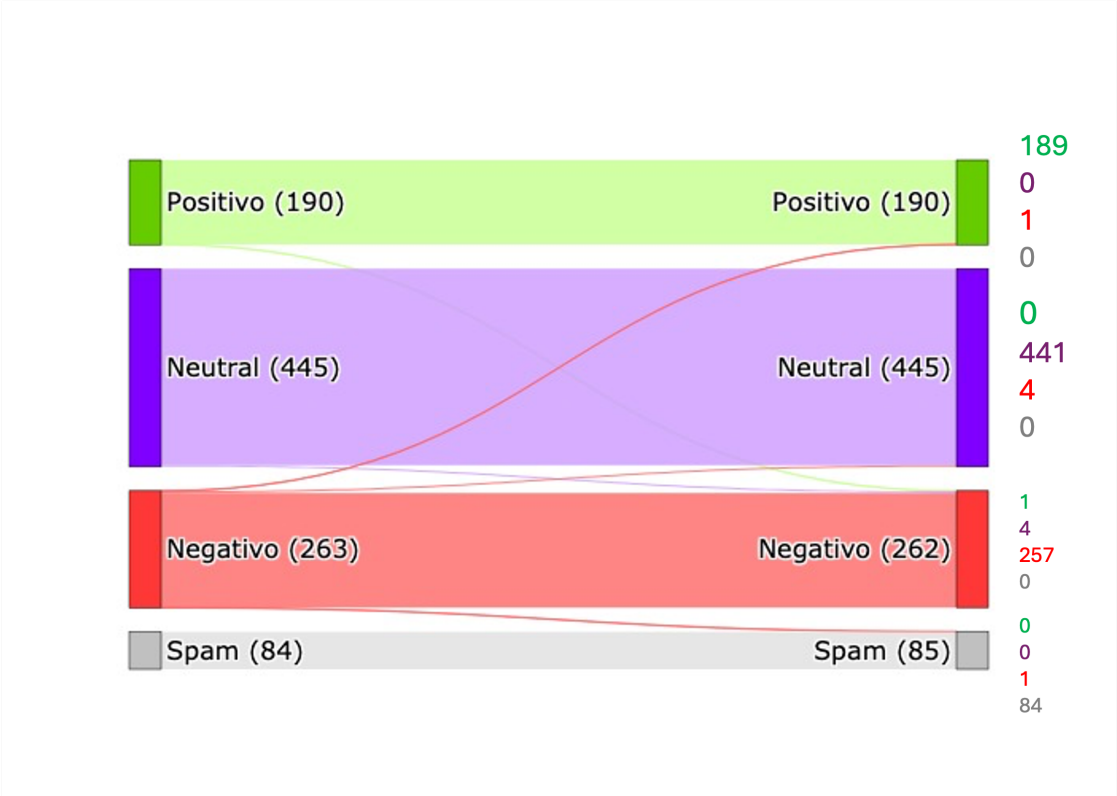


FIGURA 4.7: Diagrama de Sankey que muestra la transición del valor del sentimiento de los tuits para cada polaridad al considerar primero texto (izquierda), y después texto con imágenes (derecha) para el MCOVMEX. A la derecha se indica el número de elementos que transicionan de cada polaridad en la izquierda después de considerar el texto con imágenes para formar el total indicado en la derecha. En verde: positivo, morado: neutro, rojo: negativo, gris: spam. Fuente: elaboración propia.

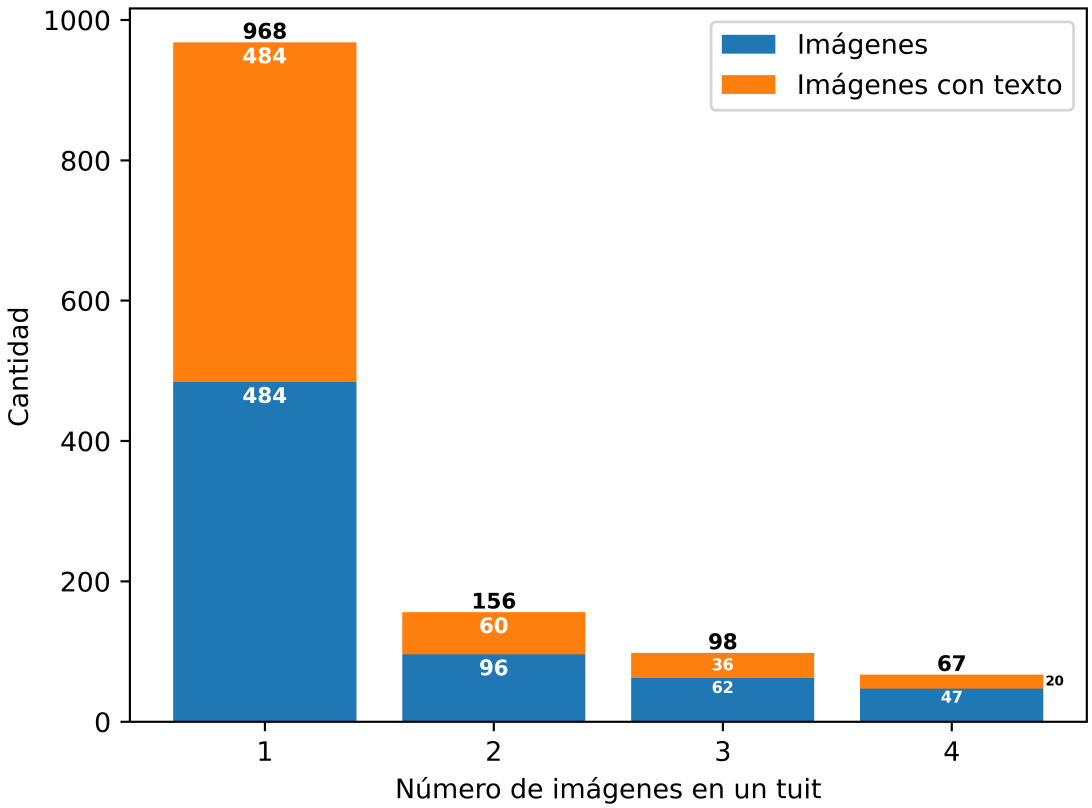


FIGURA 4.8: Cantidad de tuits en el MCOVMEX según el número de imágenes que contienen. Fuente: elaboración propia.

FIGURA 4.9: Ejemplo de éxitos típicos y errores comunes del sistema de extracción de texto en imágenes (detecciones incompletas).

4.2. Modelo de Detección de Texto en Imágenes

Se evaluó el rendimiento en cuanto a la detección de cuadros delimitadores. El sistema logró detectar texto contra el fondo con las siguientes puntuaciones: precision del 82.18 %, recall del 86.67 % y medida F1 del 84.37 %. En la Figura 4.9 se presentan ejemplos cualitativos de detecciones correctas y errores típicos, que ilustran tanto las ventajas del proceso de captura de texto superpuesto legible como sus limitaciones con fuentes estilizadas o de bajo contraste.

4.3. Modelo Preliminar para Análisis Multimodal MSSAID

La finalidad de proponer un modelo preliminar multimodal es explorar de una forma rápida los efectos de agregar modos adicionales al texto para determinar su sentimiento general y determinar el impacto de diversas técnicas para tratar ciertos tipos de datos. Como se explicó en la Sección 3.5.1, se utilizaron modelos tradicionales para representar texto: Bolsa de Palabras y TF-IDF con combinaciones de n-gramas, además de determinar el impacto de los emojis en el sistema propuesto y la construcción de una primera versión del sistema de detección de texto en imágenes.

La Tabla 4.1 muestra los resultados obtenidos al considerar diferentes combinaciones de modelos de representación de texto y modalidades de información disponibles. El mejor clasificador obtenido, con un 74.74 % de exactitud balanceada, considera un modelo de Bolsa de Palabras con combinaciones de 1, 2 y 3-gramas, incorporando únicamente texto e imágenes. También se puede apreciar que los modelos que sólo incorporan texto o texto con emojis no logran superar el umbral del 50 % de exactitud balanceada. Además, incorporar emojis no ayuda a los modelos de clasificación y el texto en imágenes impacta negativamente en el rendimiento de los mismos, independientemente del modelo de representación de texto.

Como conclusión principal de este modelo preliminar, una primera y rápida exploración apunta a que el texto y las imágenes es suficiente para construir el mejor modelo de clasificación multimodal: en promedio, agregar imágenes mejora un 25.5 % el rendimiento de los modelos en contra de aquellos que usan únicamente texto. Sin embargo, en este caso se utiliza de manera directa el sentimiento de las imágenes, lo cual no puede estar disponible en todos los casos. Por lo tanto, se necesita construir modelos de extracción de características tanto para texto como para imágenes y determinar sus efectos al momento de realizar la tarea de análisis de polaridad.

TABLA 4.1: Resultados del modelo preliminar de clasificación multimodal, disponible también en [61]. Los valores de k_C y k_γ son el valor de $C = 2^{k_C}$ y $\gamma = 2^{k_\gamma}$, respectivamente. T indica texto, I son imágenes, E apunta a la presencia de emojis y TI significa texto en imágenes. BP significa Bolsa de Palabras y TF-IDF, Term Frequency Inverse Document Frequency. El mejor resultado se resalta en negritas. Obtenida de [61].

Exactitud Balanceada	Desviación Estándar	K_C	k_γ	Modelo de Lenguaje	Modalidades Añadidas
0.4631	0.0679	12.75	-15.75	BP, 1-2 *n*-gramas	T
0.4532	0.0663	15.5	-13.75	BP, 1-3 *n*-gramas	T
0.4935	0.0598	13.75	-15.75	TF-IDF, 1-2 *n*-gramas	T
0.4942	0.0678	9.5	-11.25	TF-IDF, 1-3 *n*-gramas	T
0.7437	0.0604	5.75	-10.5	BP, 1-2 *n*-gramas	T+I
0.7474	**0.0621**	**6**	**-11**	**BP, 1-3 *n*-gramas**	**T+I**
0.7359	0.0566	6.25	-7.5	TF-IDF, 1-2 *n*-gramas	T+I
0.7312	0.0560	1.75	-2	TF-IDF, 1-3 *n*-gramas	T+I
0.4630	0.0679	12.75	-15.75	BP, 1-2 *n*-gramas	T+E
0.4532	0.0663	15.5	-13.75	BP, 1-3 *n*-gramas	T+E
0.4935	0.0598	13.75	-15.75	TF-IDF, 1-2 *n*-gramas	T+E
0.4942	0.0678	9.5	-11.25	TF-IDF, 1-3 *n*-gramas	T+E
0.7260	0.0613	6.25	-11	BP, 1-2 *n*-gramas	T+E+I
0.7279	0.0637	6.5	-11.5	BP, 1-3 *n*-gramas	T+E+I
0.7359	0.0566	6.25	-7.5	TF-IDF, 1-2 *n*-gramas	T+E+I
0.7349	0.0498	1.5	0	TF-IDF, 1-3 *n*-gramas	T+E+I
0.7270	0.0543	2.5	-7	BP, 1-2 *n*-gramas	T+I+TI
0.7258	0.0565	6.5	-11.5	BP, 1-3 *n*-gramas	T+I+TI
0.7367	0.0508	2	-2.5	TF-IDF, 1-2 *n*-gramas	T+I+TI
0.7342	0.0491	1.5	0	TF-IDF, 1-3 *n*-gramas	T+I+TI
0.7270	0.0543	2.5	-7	BP, 1-2 *n*-gramas	T+I+TI+E
0.7258	0.0565	6.5	-11.5	BP, 1-3 *n*-gramas	T+I+TI+E
0.7367	0.0508	2	-2.5	TF-IDF, 1-2 *n*-gramas	T+I+TI+E
0.7342	0.0491	1.5	0	TF-IDF, 1-3 *n*-gramas	T+I+TI+E

TABLA 4.2: Resultados del proceso de ajuste de diversos modelos de imágenes para el módulo de extracción de características en imágenes del MSSAID. El mejor resultado se resalta en negritas.

Modelo	Exactitud	Exactitud Balanceada	F_1^w	CCM
google/vit-base-patch16-224-in21k	**0.7317**	**0.5625**	**0.7017**	**0.5892**
google/vit-base-patch16-224	0.7073	0.5932	0.7019	0.5634
microsoft/resnet-50	0.3902	0.2544	0.3351	-0.0406

TABLA 4.3: Resultados del proceso de ajuste de diversos modelos de texto para el módulo de extracción de características en texto. El mejor resultado se resalta en negritas.

Modelo	Exactitud	Exactitud Balanceada	F_1^w	CCM
dccuchile/bert-base-spanish-wwm-cased	**0.6111**	**0.5927**	**0.6121**	**0.4495**
dccuchile/bert-base-spanish-wwm-uncased	0.5926	0.5215	0.5895	0.3993

4.4. Resultados Modelo Imagen y Texto: MSSAID

En esta parte se presentan los resultados del primer experimento con el modelo de imagen y texto con el conjunto de datos MSSAID. Por un lado, como se expuso en las Secciones 3.4 y 3.5, se llevan a cabo ciertas tareas relacionadas con el entrenamiento del modelo de imagen y texto como el ajuste de hiperparámetros para la fusión por codificador. Por otro lado, se presentan los primeros resultados de clasificación del modelo imagen y texto, para después continuar con los estudios de ablación y sus implicaciones en el modelo.

4.4.1. Ajuste de los Modelos de Texto e Imágenes

Para el caso del MSSAID, se usaron 607 elementos en total para entrenar los modelos de clasificación para texto. Es decir, el 90 % de los 674 elementos que conforman el total del conjunto de datos MSSAID. En el caso de las imágenes, se usaron un total de 767 imágenes, de las cuales 690 formaron parte del conjunto de entrenamiento.

En el caso de las imágenes, se ajustaron diversos modelos, de los cuales vit-base-patch16-224-in21k resultó ser el mejor al obtener un 58.92 % de CCM después de su ajuste como se muestra en la Tabla 4.2. En la misma línea, la Tabla 4.3 arroja que bert-base-spanish-wwm-cased es el mejor modelo al obtener un 44.95 % de CCM. Cabe mencionar que el modelo final para el texto contempla un aumento de datos un promedio de 17 veces, valor que se eligió al observar los resultados en la Figura 4.10. Los modelos ajustados finales de texto[1] e imágenes[2] se encuentran disponibles para su uso y descarga en repositorios públicos de Hugging Face.

[1] https://huggingface.co/lzun/spanish-social-media-boxing-text
[2] https://huggingface.co/lzun/spanish-social-media-boxing-images

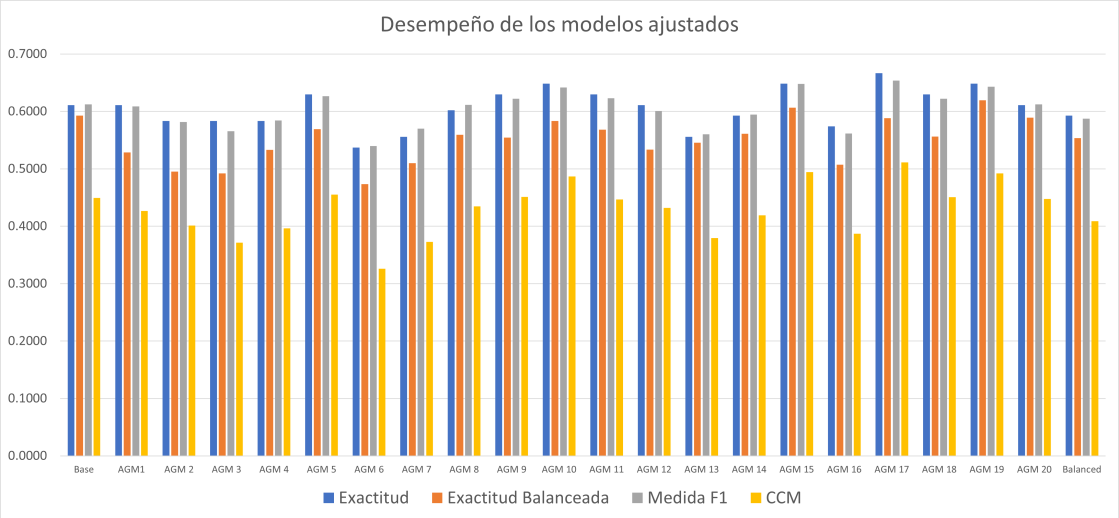


FIGURA 4.10: Resultados del proceso de ajuste para el experimento de aumento de texto. Fuente: elaboración propia.

TABLA 4.4: Hiperparámetros finales seleccionados para entrenar los modelos de clasificación multimodal para la fusión por codificador y las distintas modalidades de información consideradas en el MSSAID. T: texto, T+I: texto e imágenes, T+I+IT: texto, imágenes y texto en imágenes. M-CLIP: CLIP multilingüe.

Modelo	T		T+I		T+I+TI	
	Cabezales	Capas	Cabezales	Capas	Cabezales	Capas
M-CLIP	128	3	256	2	256	1
BETO y ViT Base	8	1	48	1	256	1
BETO y ViT Ajustados	4	1	128	1	6	1

4.4.2. Ajuste de hiperparámetros para la Fusión por Codificador

Uno de los detalles que surge con la fusión por codificador es la necesidad de ajustar los hiperparámetros que rigen el codificador del Transformer. En específico, se necesita encontrar el número óptimo de capas y número de cabezales por capa de la arquitectura utilizada. Además, es indispensable que el número de cabezales divida al tamaño del embedding entrante, el cual es 768 para los modelos basados en BERT y 512 para CLIP. La figura 4.11 muestra los mapas de calor resultantes de las mallas de búsqueda usando el CCM como métrica de vigilancia. Con estos resultados gráficos es posible elegir la mejor pareja de parámetros para cada modelo para la fusión por codificador. Con fines comparativos, se incluyen los resultados para los modelos base y ajustados de BETO y ViT, además de los modelos multilingües de CLIP.

Es posible observar que para CLIP multilingüe, las Subfiguras 4.11A, 4.11B y 4.11C muestran mapas de calor con resultados que no logran superar el 10 % de CCM al emplear este esquema de fusión de información, por lo que se anticipa un mal rendimiento. Por otro lado, los mejores modelos (Subfiguras 4.11G, 4.11H y 4.11I) apuntan a ser aquellos donde se trabajan con los modelos ajustados. De forma temprana, el mejor modelo apunta a ser el que considera los modelos ajustados considerando texto e imágenes únicamente. Al final, las mejores combinaciones de números de capas de codificador y número de cabezales por capa para cada posible modelo y modalidades se presentan en la Tabla 4.4. Dichas combinaciones se utilizan para entrenar la MVS con fusión por codificador final.

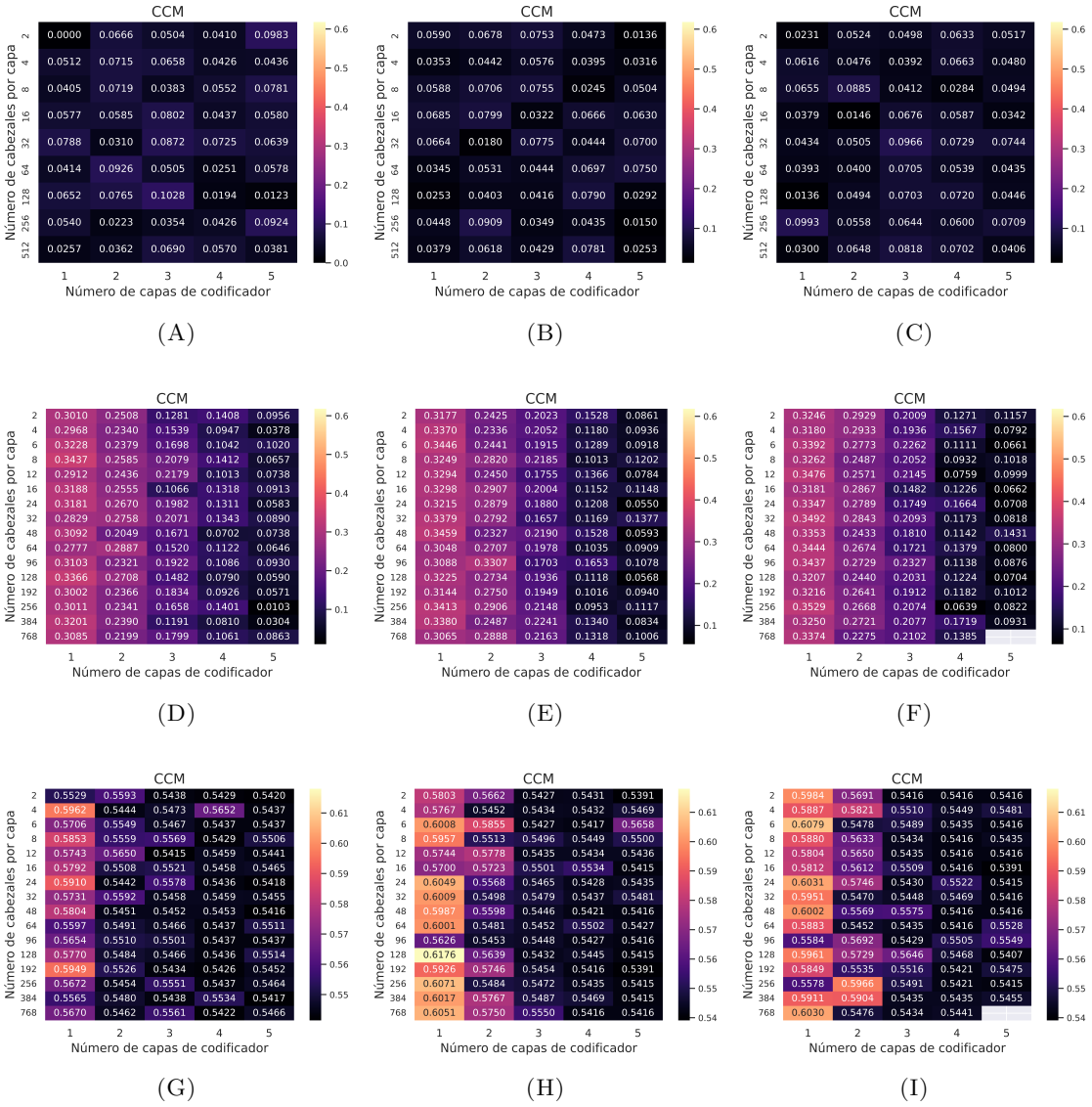


FIGURA 4.11: Mapas de calor generados según el CCM para los distintos modelos de clasificación multimodal al considerar diferentes combinaciones de cabezales de autoatención y número de capas. Arriba se encuentran los modelos multilingües basados en CLIP que consideran (A) solo texto; (B) texto e imágenes; y (C) texto, imágenes y texto en imágenes. En el medio, los modelos base de BETO y ViT que utilizan (D) solo texto; (E) texto e imágenes; (F) texto, imágenes y texto en imágenes. En la parte inferior, los modelos BETO y ViT ajustados con (G) solo texto; (H) texto e imágenes; e (I) texto, imágenes y texto en imágenes. Debido a limitaciones por RAM, no fue posible determinar los resultados para la combinación de 5 capas de codificador con 768 cabezas de autoatención para los mapas de calor (F) e (I). Fuente: elaboración propia.

4.4.3. Resultados de Clasificación del Modelo Imagen y Texto

Los resultados de los modelos de análisis de sentimientos multimodal propuestos para las publicaciones de X y el MSSAID se encuentran en la Tabla 4.5. Cabe mencionar que estos contemplan los modelos de CLIP multilingüe, BETO y ViT base y BETO y ViT ajustados, además de considerar toda la información disponible reflejada en las modalidades presentes: texto, imágenes y texto en imágenes. Dado que el sistema propuesto presenta dos enfoques para fusionar las distintas modalidades, se contrastan los resultados obtenidos por la fusión por suma y la fusión por codificador para su mejor comparación. En resumen, el mejor modelo es aquel que considera fusión por suma, con los modelos ajustados y todas las modalidades presentes, obteniendo 60.52 % de CCM. Además, de forma general, la fusión por suma presenta mejores resultados que la fusión por codificador.

Tabla 4.5: Métricas de desempeño de los distintos modelos de clasificación multimodal: CLIP Multilingüe, BETO y ViT base y BETO y ViT ajustados. Los modelos consideran la información de todas las modalidades disponibles (texto, imágenes y texto en imágenes) que se pueden encontrar en un tuit, fusionadas por la fusión por suma o por codificador. Los valores de k_C y k_γ apuntan al valor de $C = 2^{k_C}$ y $\gamma = 2^{k_\gamma}$, respectivamente, de la MVS. El mejor resultado obtenido para estas combinaciones se marca en negritas.

	Modelo	Exactitud	Exactitud Balanceada	F_1^w	CCM	k_C	k_γ
Fusión Suma	M-CLIP	0.4776	0.4950	0.4939	0.3008	9.9219	-17.0156
	Base	0.5075	0.5136	0.5221	0.3352	1.8750	-10.8750
	Ajustado	**0.7313**	**0.6679**	**0.7370**	**0.6052**	**11.1875**	**-17.3750**
Fusión Codific.	M-CLIP	0.4925	0.2500	0.3251	0.0000	-2.0000	-13.5000
	Base	0.5672	0.4728	0.5405	0.3285	1.1250	-7.2813
	Ajustado	0.6567	0.5603	0.6543	0.4883	3.0000	-11.0000

4.4.4. Estudios de Ablación

Impacto de las Modalidades

Del análisis de datos surgen preguntas sobre diversos aspectos naturales de los datos y cómo pueden llegar a afectar los resultados finales. En específico, se plantean las siguientes preguntas:

1. ¿Cómo contribuye cada modalidad al sistema de clasificación?
2. ¿Cómo afecta el número de imágenes que se incorporan a los modelos de análisis de polaridad?

Para ambos métodos de fusión se prueba el sistema de clasificación propuesto con CLIP multilingüe, BETO y ViT base y BETO y ViT ajustado utilizando solo el texto de los tweets, luego considerando solo texto e imágenes y finalmente incorporando texto, imágenes y texto en imágenes.

Para responder la primera pregunta, se presenta la Tabla 4.6 donde se observan las diferentes contribuciones de las modalidades (texto, texto e imágenes y texto, imágenes y texto en imágenes). Una vez más, el mejor modelo se presenta en la fusión por suma al incorporar todas las modalidades posibles. Sin embargo, se puede observar que para los modelos ajustados con fusión por suma, la diferencia en CCM para el modelo con texto e imágenes es apenas del 0.66 %. CLIP multilingüe continúa con su bajo desempeño al aplicarle la fusión por codificador, pero en general presenta un mejor resultado para la fusión por suma. Desafortunadamente, estos resultados revelan lo complicado que es usar los modelos multilingües cuando se tiene uno dedicado para un idioma en específico, ya que los modelos base resultan mejores que CLIP multilingüe. La fusión por codificador, a pesar del mecanismo complejo detrás de ella, no logra superar a la fusión por suma.

Finalmente, una selecta colección de matrices de confusión de los mejores modelos de aprendizaje para cada alternativa (base, ajustado y CLIP multilingüe) se puede apreciar en la Figura 4.12. En general, se evidencia que las matrices de confusión generadas por los modelos con fusión por suma (4.12A, 4.12B y 4.12C) presentan mejores resultados que aquellas generadas por la fusión por codificador al presentar un menor error de clasificación reflejado en los valores altos que se presentan en la diagonal de las matrices. En particular, al considerar la fusión por codificador, la matriz de confusión para el mejor modelo de CLIP multilingüe, que se muestra en la Subfigura 4.12D, presenta un atractor en la clase con mayor representación para reducir el error del modelo. Sin embargo, es el peor modelo ya que no predice elementos de otras clases. En el caso de las matrices de confusión para el modelo BETO y ViT, base y ajustados (Subfigura 4.12E y 4.12F, respectivamente), presenta mayor confusión en los resultados que su contraparte generada por la fusión por suma.

Lo anterior también se refleja en los embeddings en dos dimensiones de los puntos generados para cada polaridad, disponibles en la Figura 4.13. Aquí se puede distinguir que las representaciones vectoriales fusionadas de las publicaciones se enciman y es difícil distinguir una de otra para la mayoría de los modelos, particularmente en las Subfiguras 4.13A, 4.13B, 4.13D y 4.13E. A pesar de ello, se aprecia una mejor separación y agrupación en la Subfigura 4.13F que en la Subfigura 4.13C, la correspondiente al mejor modelo obtenido.

TABLA 4.6: Métricas de rendimiento para los distintos modelos de clasificación multimodal y la contribución de las distintas modalidades para ambos métodos de fusión de información. T indica texto, I denota la presencia de imágenes e IT expresa texto en imágenes. Los valores de k_C y k_γ son el valor de $C = 2^{k_C}$ y $\gamma = 2^{k_\gamma}$, respectivamente. El mejor resultado se resalta en negritas.

	Modelo	Exactitud	Exactitud Balanceada	F_1^w	CCM	k_C	k_γ	Datos
Fusión Suma	M-CLIP	0.4627	0.5025	0.4747	0.3081	7.5000	-12.5000	T
		0.4776	0.4660	0.4994	0.2931	7.6250	-14.6250	T+I
		0.4776	0.4950	0.4939	0.3008	9.9219	-17.0156	T+I+TI
	Modelos Base	0.5224	0.4654	0.5562	0.3456	4.5313	-10.8438	T
		0.6119	0.5713	0.6109	0.4362	1.3125	-7.5625	T+I
		0.5075	0.5136	0.5221	0.3352	1.8750	-10.8750	T+I+TI
	Modelos Ajustados	0.7164	0.6372	0.7179	0.5771	2.0000	-9.0000	T
		0.7313	0.6657	0.7310	0.5986	9.0000	-15.2500	T+I
		0.7313	**0.6679**	**0.7370**	**0.6052**	**11.1875**	**-17.3750**	**T+I+TI**
Fusión Codificador	M-CLIP	0.3284	0.2435	0.3388	0.0363	1.4219	-10.4531	T
		0.4925	0.2500	0.3251	0.0000	-0.6250	-11.1250	T+I
		0.4925	0.2500	0.3251	0.0000	-2.0000	-13.5000	T+I+TI
	Modelos Base	0.5672	0.4671	0.5477	0.3278	1.4375	-8.1172	T
		0.4776	0.4345	0.4893	0.2560	2.2500	-10.7500	T+I
		0.5672	0.4728	0.5405	0.3285	1.1250	-7.2813	T+I+TI
	Modelos Ajustados	0.6866	0.5987	0.6898	0.5343	2.1875	-10.7500	T
		0.7164	0.6442	0.7181	0.5821	1.0000	-9.2500	T+I
		0.6567	0.5603	0.6543	0.4883	3.0000	-11.0000	T+I+TI

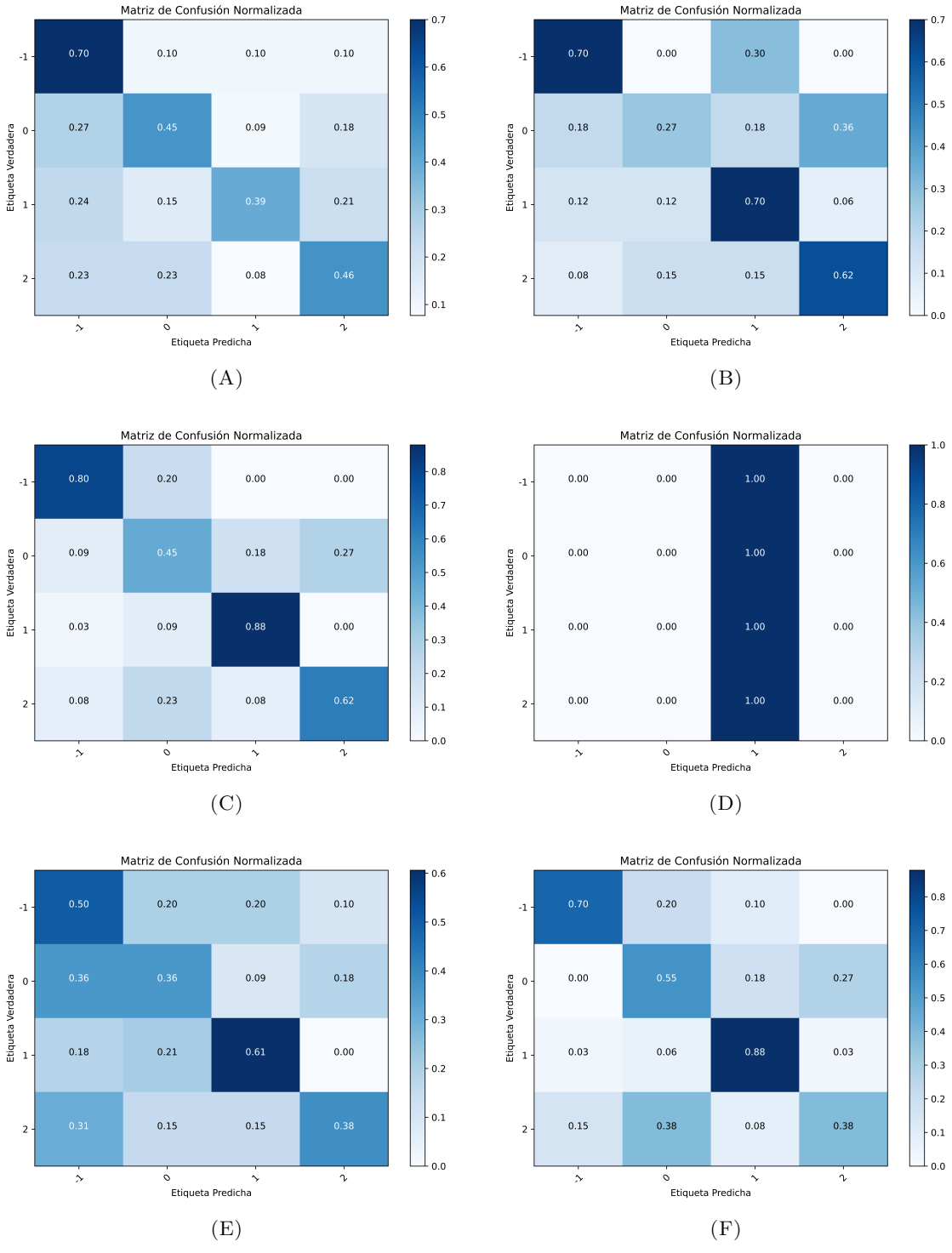


FIGURA 4.12: Matrices de confusión para los mejores modelos de cada modalidad. Arriba, con el método de fusión por suma: (A) CLIP multilingüe con texto; (B) BETO y ViT base con texto e imágenes; y (C) BETO y ViT ajustados con texto, imágenes y texto en imágenes. En la parte inferior, con el método de fusión por codificador: (D) CLIP multilingüe con texto; (E) BETO y ViT base con texto, imágenes y texto en imágenes; y (F) BETO y ViT ajustados con texto e imágenes. Fuente: elaboración propia.

FIGURA 4.13: Proyecciones 2D de los embeddings de los mejores modelos de cada modalidad. Arriba, con el método de fusión por suma: (A) CLIP multilingüe con texto; (B) BETO y ViT base con texto e imágenes; y (C) BETO y ViT ajustados con texto, imágenes y texto en imágenes. En la parte inferior, con el método de fusión por codificador: (D) CLIP multilingüe con texto; (E) BETO y ViT base con texto, imágenes y texto en imágenes; y (F) BETO y ViT ajustados con texto e imágenes. Fuente: elaboración propia.

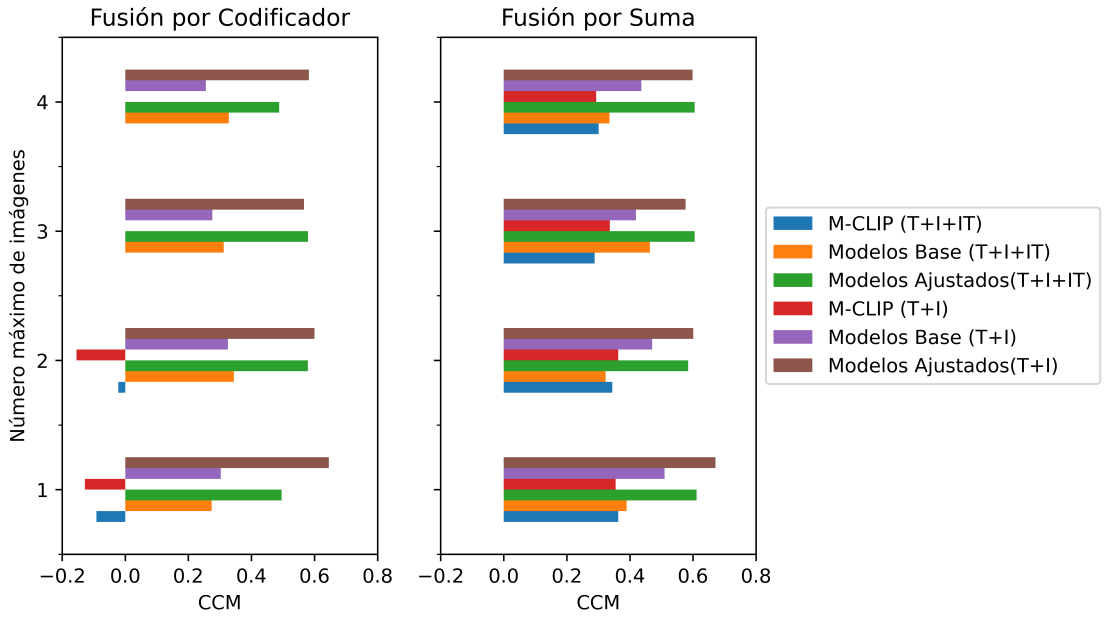


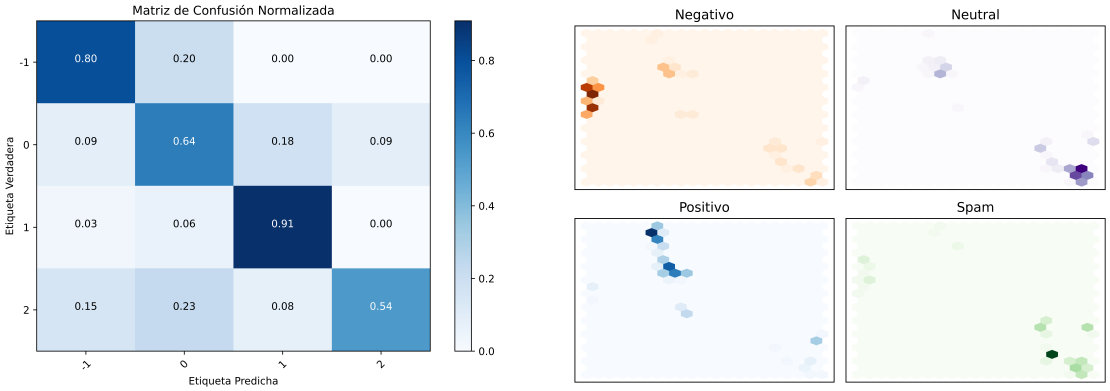
FIGURA 4.14: Puntuaciones según el CCM para los distintos modelos de clasificación, los dos modos de fusión de información, y diferentes tipos de información, cambiando el número máximo de imágenes que se agregan a ellos. Fuente: elaboración propia.

Análisis del Impacto del Número de Imágenes en los Modelos de Clasificación Multimodal

Para la segunda pregunta, la Figura 4.14 y la Tabla B.1 en el Anexo B indican las distintas puntuaciones del CCM para los modelos de clasificación multimodal contemplando ambos modos de fusión, la contribución de cada modalidad de información a los modelos y el efecto de alterar el número máximo de imágenes que se agregan para determinar su impacto en los modelos de clasificación. De esta figura se puede concluir que el mejor modelo solamente considera, a lo más, la primera imagen de cada publicación, además de su correspondiente texto e imágenes, logrando una puntuación del 67.17 % en CMM y 72.10 % en exactitud balanceada. A diferencia de los experimentos anteriores, donde el mejor modelo consideraba texto, imágenes y texto en imágenes con un 60.52 % de CCM, ahora no es necesario agregar el texto de las imágenes ni todas las imágenes y se logra un mejor resultado, aliviando al mismo tiempo la carga computacional del sistema. Las proyecciones en dos dimensiones de los embeddings de modelos selectos, disponibles en la Figura 4.15, muestran resultados similares al primer experimento de ablación.

Finalmente, la matriz de confusión y la proyección de los embeddings del mejor modelo (imágenes y texto fusionados por fusión por suma, únicamente la primera imagen) se encuentran en la Figura 4.16. La matriz de confusión refleja una buena capacidad del modelo para discriminar entre publicaciones positivas y es aceptable para publicaciones negativas. Sin embargo, muestra problemas para diferenciar entre publicaciones neutras y spam. Esto último también se evidencia en las proyecciones de los embeddings, donde existen zonas en las clases neutral y spam que se traslapan considerablemente.

FIGURA 4.15: Proyecciones 2D de los embeddings de los mejores modelos de cada modalidad al considerar diferentes cantidades de imágenes. Arriba, con el método de fusión por suma: (A) CLIP multilingüe con texto, la primera imagen; (B) BETO y ViT base con texto e imágenes, la primera imagen; y (C) BETO y ViT ajustados con texto, imágenes y texto en imágenes, la primera imagen. En la parte inferior, con el método de fusión por codificador: (D) CLIP multilingüe con texto, las primeras tres imágenes; (E) BETO y ViT base con texto, imágenes y texto en imágenes, las primeras dos imágenes; y (F) BETO y ViT ajustados con texto e imágenes, la primera imagen. Fuente: elaboración propia.



(A) (B)

FIGURA 4.16: (A) Matriz de confusión y (B) proyecciones 2D de los embeddings del mejor modelo obtenido: 67.16 % CCM para tuits fusionados por fusión por suma, BETO y ViT ajustados, texto e imágenes, considerando únicamente la primera imagen de cada publicación. Fuente: elaboración propia.

4.4.5. Análisis de Error

A continuación, en la Tabla 4.7 se presenta una colección de 10 tuits seleccionados de forma aleatoria cuya clasificación por el mejor modelo multimodal fue errónea para realizar un breve análisis de error. En general, los textos que se presentan son relativamente largos (dada la longitud de un tuit, como se exploró en la Figura 4.2). Sin embargo, se puede apreciar algunos mensajes cortos donde el modelo no puede extraer mucho contexto del texto, como lo son el ejemplo 7 y 9. En el caso en particular del ejemplo 7, todo apunta a un error de etiqueta mal asignada por un humano, ya que se predice como positivo cuando en realidad es neutro. Dado el mensaje y la imagen, es posible que lo que prediga el modelo sea correcto. En contraste, el ejemplo 9, la imagen carga cierta connotación negativa con un tono burlón similar a un meme. El principal problema en este caso es que el texto no provee mayor información para darle un mejor significado a las imágenes.

También se puede apreciar una de fotos de la pelea (por ejemplo, tomadas al televisor) que se anexan en los mensajes, como lo son los ejemplos 2 y 10. Estas imágenes cargan un mensaje neutro ya que no aportan mucho mensaje alguno más que el de «*estuve aquí, en la pelea*». En el caso del ejemplo 2, el mensaje se predice como negativo cuando se anota como neutro dado el tono informativo de la publicación. Sin embargo, existe cierta ambigüedad dado el lenguaje que explora la derrota del boxeador, aunque se expresa de manera objetiva y no como objeto de burla. Por otro lado, el ejemplo 10 menciona elementos ajenos al evento deportivo, por lo que se etiqueta como spam. Desafortunadamente, el sistema no es capaz de comprender tales elementos como ajenos a la pelea. De forma similar, el ejemplo 6, aunque menciona al Canelo, no se encuentra relacionado con el evento deportivo que se discute y se etiqueta como neutro en lugar de spam. El ejemplo 1 puede ser un error de etiquetado ya que la publicación trata sobre el Canelo, pero falta contexto, elemento que la imagen no es capaz de proveer y se considera neutra.

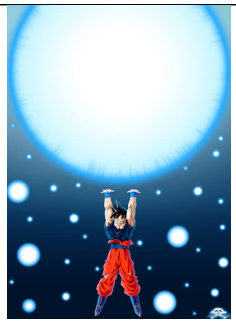
Un caso particular es el del ejemplo 4, donde se aprecia una imagen de una apuesta que se hizo en torno al evento deportivo. En este caso, el sistema de detección de texto puede que no haya detectado todo el texto y, aunque si lo hiciese, resulta complicado entender el significado de la apuesta (¿apostó a favor o en contra?) para el sistema de clasificación salvo que se entrene para dicha tarea. Dado que el tema del Canelo se mezcla con otros eventos deportivos y parece mezclarse el sentimiento de cada uno de ellos, el sistema predice un valor negativo cuando es positivo.

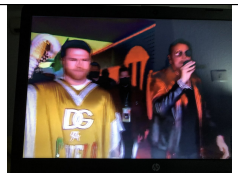
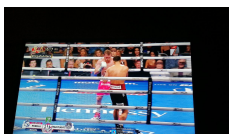
Por otro lado, en el ejemplo 8, tomado antes de la pelea, la imagen muestra a los boxeadores cara a cara de forma amenazante. A pesar de que el texto del mensaje es a favor del boxeador mexicano, parece que la imagen complementó de más al texto e hizo cambiar la polaridad del mensaje hacia algo negativo cuando en realidad es positivo. En el ejemplo 3, donde se expresa de forma humorística un milagro para que gane el canelo, se predice como positivo cuando es neutro dado que el mensaje se publicó durante la pelea. En general, el sistema no sabe sobre situaciones exteriores (como es de esperar) ya que en ese punto el boxeador mexicano se encontraba perdiendo el encuentro. Al final, a pesar del toque humorístico para disfrazar la desgracia, el sistema no fue capaz de determinar la verdadera intención dada la falta de contexto.

Finalmente, el ejemplo 5 muestra la captura de pantalla de un video corto de una entrevista. En este caso, el video proporciona más información que complementa al texto. Desafortunadamente, como el sistema no modela video, no es capaz de acceder a esa

información adicional y no puede capturar la verdadera intención del usuario.

Tabla 4.7: Tuits seleccionados para el análisis de error del MSSAID.

	Tuit	Imagen	Valor Verdadero	Valor Predicho
1	EXCLUSIVA. Rigo, el Español Álvarez espera una pelea larga entre @Canelo y Vibol. Su hermano, sigue con el hambre de superarse y asumir retos. @TVAztecaJalisco https://t.co/pnF0Wtkq1w		2	0
2	🥊 Pierde Saúl @Canelo Álvarez por decisión unánime ante el boxeador ruso #DmitriBivol, se mantiene invicto en peso semipesado de la AMB. Es la segunda derrota para #SaulAlvarez, Los jueces le dieron la victoria a Dimitry Bivol por decisión unánime 115-113 (x3). https://t.co/ziJhNewQMS	 	0	-1
3	Yo pidiendo una Genkidama para que gane el Canelo... @AztecaDeportes @ESPNmx https://t.co/viV1pcb80o		0	1
4	🚨 Pick de último minuto, si hoy el Pachuca no saca mínimo el empate ante el equipo más débil de la liga, seria un insulto hacia ellos y afición, necesitan ganar si o si para poder pelear por algo, y lo del Canelo pues, es Canelo, el resto es historia, métanle buena lana 🔥💰🍀 https://t.co/RflfOjdywB		2	-1

	Tuit	Imagen	Valor Verdadero	Valor Predicho
5	Coincido totalmente con @JMMarquezOf // La neta @Canelo https://t.co/B0PDAtxC6f		1	0
6	Mira @Canelo este cinturón es el que te hace falta Nadamas. Si compras a mis @CHIVAS Puede ser tuyo junto con @peladoalmeyda paquete completo. #CaneloPlant #caneloVsPlant https://t.co/HbXjjHNwRH		0	2
7	Con todo @Canelo 👊🏼 https://t.co/OtsK09JKnL		0	1
8	Llegó el día!! Voy @Canelo ¿Alguien piensa lo contrario? #canelovsbivol #caneloalvarez 👊🏼 🥊 🇲🇽 🇷🇺 https://t.co/CYfXdCNyyo		1	-1
9	Y diay vos [redacted] #Canelo https://t.co/yHbMVfJamo		-1	0
10	Las morras que se ven atrás del cuadrilátero (que son como 6) oh lala! Les vale [redacted] la pelea, están en el celular todo el tiempo, están platicando entre ellas y de seguro están en lugares que valen lo que tú y yo ganamos en dos años pero .. oh Lala 🥵🥴 #Canelo https://t.co/1MlJMbxcJR		2	0

4.4.6. Resumen de Resultados de la Sección

Al emplear el uso del modelo de clasificación multimodal con el conjunto de datos MSSAID se llegó a los siguientes resultados clave sobre su aplicación que sirven para mejorar el uso del modelo de imagen y texto propuesto:

1. Los modelos ajustados resultan ser mejores que sus respectivas versiones base y CLIP multilingüe. Por lo tanto, es recomendable agregar al marco de trabajo, desde un principio, el ajuste de los modelos vastos de lenguaje base con el conjunto de datos.

2. El modelo es sensible al número de imágenes que se agregan, por lo que es indispensable realizar un análisis preliminar para determinar dicho hiperparámetro de antemano.

3. Ligado al punto anterior, los modelos de clasificación multimodal son sensibles a las modalidades que se incluyen. De la misma forma, se debe realizar un análisis preliminar para determinar cuál o cuáles de ellas se deben incorporar para obtener el mejor modelo posible.

4. En este caso de uso, el texto de las imágenes no es relevante para mejorar el proceso de clasificación.

5. Existe confusión entre contenido neutro y spam.

4.5. Resultados Modelo Imagen y Texto: MCOVMEX

En esta sección se presentan los resultados obtenidos con el modelo de imagen y texto aplicados al conjunto de datos MCOVMEX. A diferencia del caso de uso anterior (MSSAID), donde se realizaron diversos experimentos adicionales para determinar el impacto de diversos factores en los modelos de clasificación multimodal, en esta aplicación se implementan las conclusiones obtenidas para obtener un mejor modelo desde el principio. En particular, se usan únicamente modelos ajustados a los datos, se determina el número óptimo de imágenes para los modelos y también las contribuciones de cada modalidad para determinar las mejores.

4.5.1. Ajuste de los Modelos de Texto e Imágenes

El ajuste de los modelos base para los extractores de texto e imágenes se llevaron a cabo con 1000 elementos de texto y 1289 imágenes en total. Los resultados del ajuste de los distintos modelos de texto base para el MCOVMEX se pueden seguir en la Tabla 4.8 y para las imágenes, en la Tabla 4.9. Como consecuencia, el mejor modelo para texto es la versión sin capitalizar de BETO (bert-base-spanish-wwm-uncased) y para las imágenes, ViT (vit-base-patch16-224-in21k). En este último caso, a pesar de que la versión vit-base-patch16-224 obtuvo un mejor rendimiento en el CCM, se optó por el primero debido a que superó al último en las primeras tres métricas de contraste. Los modelos ajustados finales de texto[3] e imágenes[4] finales se encuentran disponibles para su uso y descarga en repositorios públicos de Hugging Face.

[3] https://huggingface.co/lzun/mcovmex-text
[4] https://huggingface.co/lzun/mcovmex-image

Tabla 4.8: Resultados del proceso de ajuste de diversos modelos de imágenes para el módulo de extracción de características en textos del MCOVMEX.

Modelo	Exactitud	Exactitud Balanceada	F_1^w	CCM
dccuchile/bert-base-spanish-wwm-cased	0.6960	0.5441	0.6717	0.5577
dccuchile/bert-base-spanish-wwm-uncased	**0.7760**	**0.6532**	**0.7548**	**0.6780**

Tabla 4.9: Resultados del proceso de ajuste de diversos modelos de imágenes para el módulo de extracción de características en imágenes del MCVOMEX.

Modelo	Exactitud	Exactitud Balanceada	F_1^w	CCM
google/vit-base-patch16-224-in21k	**0.6436**	**0.4084**	**0.5924**	**0.3428**
google/vit-base-patch16-224	0.5691	0.3991	0.3991	0.3991
WinKawaks/vit-tiny-patch16-224	0.6117	0.4155	0.5851	0.3066
microsoft/swin-tiny-patch4-window7-224	0.5638	0.3453	0.5131	0.1715
microsoft/swin-base-patch4-window7-224	0.6383	0.4408	0.6039	0.3428
microsoft/swinv2-tiny-patch4-window16-256	0.5585	0.3872	0.5401	0.2324

4.5.2. Ajuste de Hiperparámetros para la Fusión por Codificador

Del mismo modo que en el caso del MSSAID, se debe empezar el trabajo con la fusión por codificador y la selección óptima de hiperparámetros: cantidad de capas y número de cabezales por capa. En este caso, de las lecciones aprendidas de los experimentos con el MSSAID, no se exploran los casos de los modelos base y CLIP multilingüe. Sin embargo, ahora se determina el número óptimo de los hiperparámetros mencionados anteriormente para diferentes cantidades de imágenes que se agregan a los modelos. Es decir, se exploran los modelos con la primera imagen, las primeras dos, las primeras tres y, finalmente, todas las imágenes disponibles. Lo anterior se realiza con la finalidad de explorar mejor el impacto, desde un inicio, de las imágenes en diferentes cantidades en los modelos de clasificación multimodal. Debido al mayor número de elementos con los que se debe trabajar, se encontró que en varios de los experimentos no fue posible determinar valores por encima de 256 cabezales, por lo que se limitó este hiperparámetro a dicho valor.

Los mapas de calor que se obtienen como resultado de los experimentos se pueden consultar en la Figura 4.17 y los hiperparámetros seleccionados para cada combinación en la Tabla 4.10. A diferencia de los resultados del MSSAID, los mapas de calor del MCVOMEX plasman resultados más altos según el CCM. Sin embargo, la mayor diferencia es que ahora se necesitan entre una y dos capas de codificador para obtener los mejores resultados, reflejando que se necesita una arquitectura más compleja para extraer información de más datos.

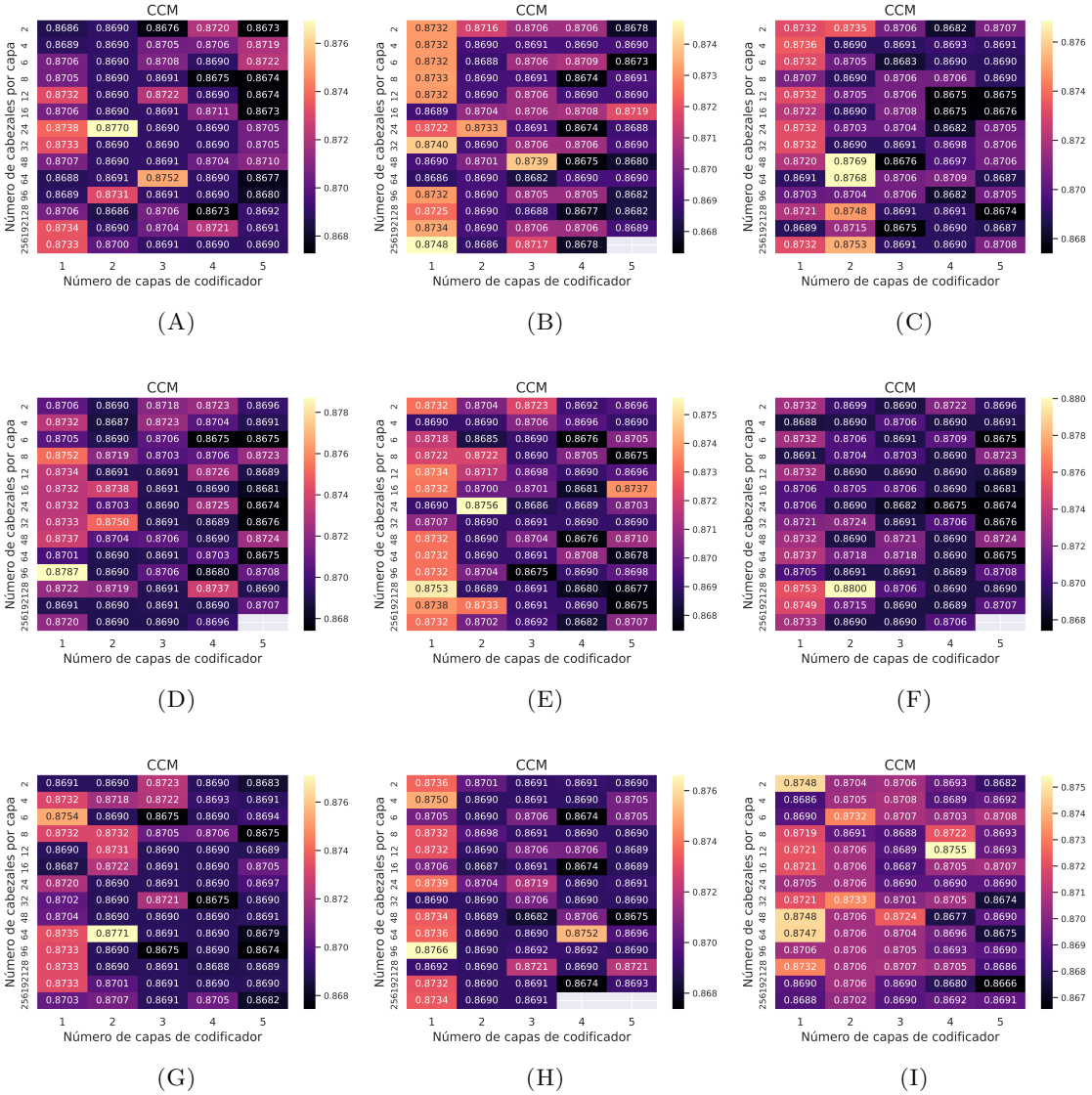


FIGURA 4.17: Mapas de calor generados según el CCM para los distintos modelos de clasificación multimodal al considerar diferentes combinaciones de cabezales de autoatención y número de capas. (A) La primera imagen, sólo imágenes; (B) la primera imagen, con imágenes y texto en imágenes; (C) las primeras dos imágenes, sólo imágenes; (D) las primeras dos imágenes, con imágenes y texto en imágenes; (E) las primeras tres imágenes, sólo imágenes; (F) las primeras tres imágenes, con imágenes y texto en imágenes; (G) todas las imágenes disponibles, sólo imágenes; (H) todas las imágenes disponibles, con imágenes y texto en imágenes; e (I) únicamente texto. Fuente: elaboración propia.

TABLA 4.10: Hiperparámetros seleccionados para cada combinación de los modelos ajustados y número máximos de imágenes. Para las modalidades T: texto, I: imágenes, TI: texto en imágenes.

Modalidades	Número de Imágenes	Capas	Cabezales
T+I+TI	1	1	256
	2	1	96
	3	2	128
	4	1	96
T+I	1	2	24
	2	2	48
	3	2	24
	4	2	64
T		4	12

4.5.3. Resultados de Clasificación del Modelo Imagen y Texto

Los resultados del modelo de imagen y texto para el conjunto de datos MCOVMEX se pueden apreciar en la Figura 4.18 y la tabla con los datos completos se puede consultar en el Apéndice C. Se puede observar que el mejor modelo obtenido utiliza fusión por suma logrando un CCM de 94.26 % al incorporar únicamente imágenes, considerando hasta las primeras tres de ellas en el sistema. Una vez más, la fusión por suma es superior a la fusión por codificador y agregar texto en imágenes no ayuda al proceso de clasificación. Por otro lado, la matriz de confusión y los embeddings en dos dimensiones generados por este mejor modelo de clasificación se siguen en la Figura 4.19. Se puede apreciar que la matriz de confusión exhibe una diagonal muy marcada, indicando una gran capacidad para predecir todas las clases de forma satisfactoria. Similar al caso del MSSAID, se logra apreciar cierta confusión entre los mensajes que son spam que se etiquetan como neutros y entre positivos y negativos. Las proyecciones de los embeddings expresan regiones claras para cada polaridad de sentimientos, con ligeros empalmes ente negativo y positivo, y para spam y neutral.

Con el mejor modelo ya entrenado, se procedió a utilizarlo para anotar todo el conjunto de datos restante para explorar la polaridad del mismo. El resultado final, filtrado por mes y año, se puede observar en la Figura 4.20. En la imagen se muestra la tendencia del COVID como tema a través del tiempo, perdiendo relevancia conforme sale del 2021. En general, la mayor cantidad de los mensajes son neutros.

4.5.4. Análisis de Error

A continuación, en la Tabla 4.7 se presenta una colección de cuatro tuits cuya clasificación por el mejor modelo multimodal fue errónea para realizar un breve análisis de error. El primer error es un claro ejemplo de sarcasmo, elemento que cambia el valor de la polaridad del mensaje, al describir algo positivo pero que en realidad es negativo. En el segundo ejemplo, se tiene una mezcla de sentimientos donde los anotadores concluyen que predomina el sentimiento positivo. En este caso, el sistema no logra determinar cual sentimiento es el más fuerte ya que se presentan ambos en la publicación y la imagen ciertamente no aporta información adicional para lograr diferenciarlo.

Por otro lado, el tercer ejemplo es otro caso donde se necesita una comprensión de lectura en la imagen, aunque en esta ocasión no todo el texto es relevante. El sistema en general no logra determinar que la donación de sangre es algo urgente y negativo. Finalmente, el último caso se cataloga como spam ya que las vacunas son veterinarias y no del COVID. El sistema no logra determinar a partir de la imagen esta información y, por lo tanto, no liga ambos pedazos de información correctamente y lo cataloga como neutro. Este caso eleva la importancia de sistemas que comprendan las imágenes en un nivel más alto, en este caso, al agregar compresión de lectura.

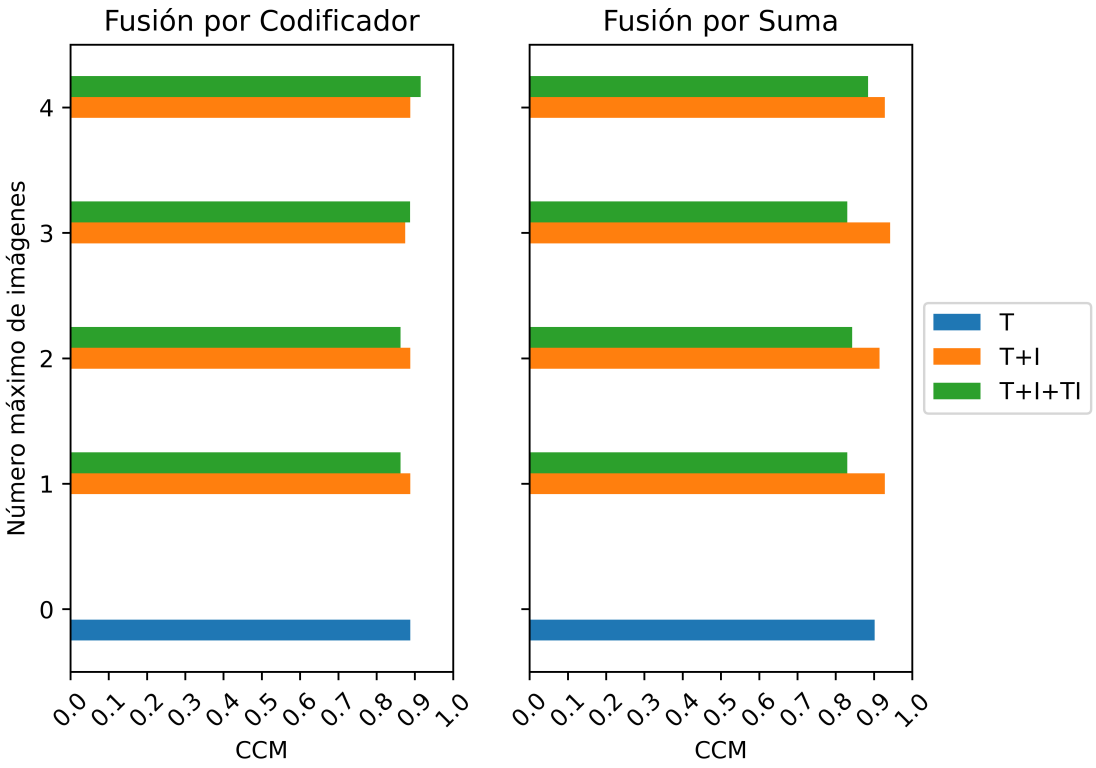
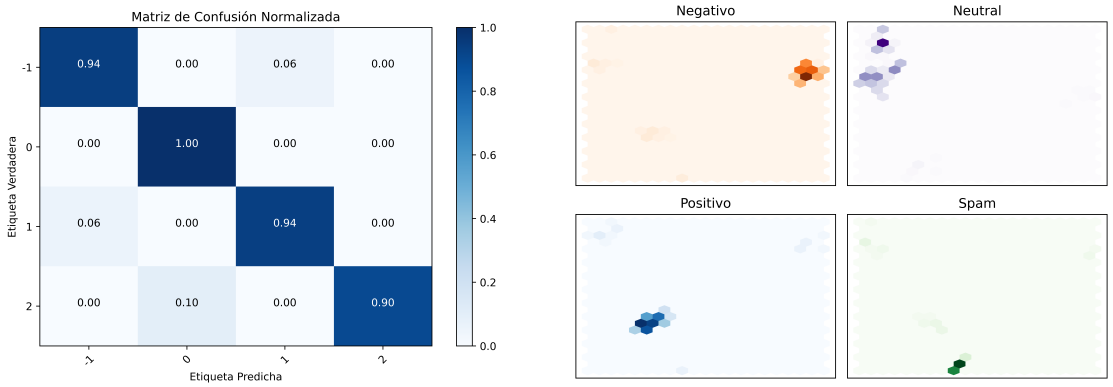


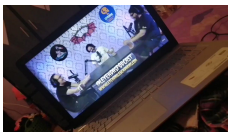
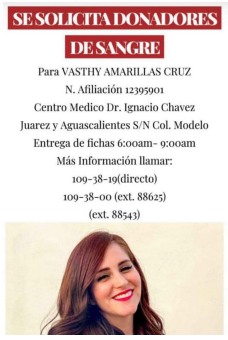
FIGURA 4.18: Resultados del sistema de clasificación multimodal al considerar diferentes cantidades de imágenes y métodos de fusión para el MCOVMEX. T: texto, I: imágenes, TI: texto en imágenes. Fuente: elaboración propia.



(A) (B)

FIGURA 4.19: (A) Matriz de confusión y (B) proyecciones 2D de los embeddings del mejor modelo obtenido: 96.00 % CCM para tuits fusionados por fusión por suma, BETO y ViT ajustados, texto e imágenes, considerando únicamente las primeras tres imágenes de cada publicación. Fuente: elaboración propia.

Tabla 4.11: Tuits seleccionados para el análisis de error del MCOVMEX.

	Tuit	Imagen	Valor Verdadero	Valor Predicho
1	Éste niño nació en un México con gobernantes que proveían de salud y vacunas para los recién nacidos. Nació y tuvo la suerte de tener acceso a una operación que le pudo haber costado la vida sí sus padres no hubieran tenido la fortuna de acceder a servicios médicos de calidad >2 https://t.co/UqLsWC-TizP		-1	1
2	Me ando retorciendo del dolor por la reacción de la vacuna. Pero nada que no pueda arreglar @LeyendasPodcast @ElBadiablo @NingunEduardo @MarioLopez-Capi https://t.co/VMnh7Uw8cq		1	-1
3	¡Ella lucha no solo por su vida, también por la de su bebé de 6 meses 🤰! Se solicitan DONADORES DE SANGRE para Vasthy 🙏😔 #Covid19 Ella es parte del personal de salud que ha estado activo durante la contingencia (Química del laboratorio del Isssteson)👩‍🔬. Favor de compartir. https://t.co/AAXxFG3aDX		-1	1
4	Llegó la vacuna para continuar la promoción de 3 x 2 hasta el 30 de septiembre Aprovecha https://t.co/krkQSLVLkR		2	0

4.5.5. Resumen de Resultados de la Sección

De los experimentos llevados a cabo con el conjunto de datos MCOVMEX se llegó a los siguientes resultados principales:

1. En el caso del MCOVMEX, la modalidad del texto se alinea de mejor manera con el sentimiento general de las publicaciones. Sin embargo, agregar las imágenes

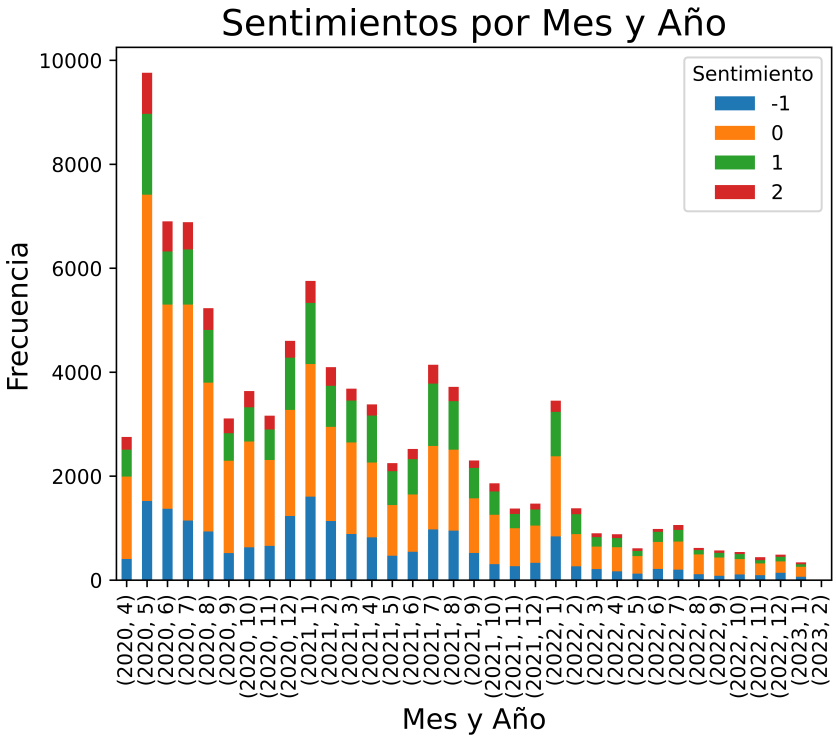


FIGURA 4.20: Distribución del sentimiento general de las publicaciones de X según el mes y año de su publicación. Fuente: elaboración propia.

contribuye a la mejora de la correcta detección de la polaridad del sentimiento general de una publicación.

2. Al contrario que con el caso del MSSAID, las transiciones mostradas por el diagrama de Sankey en la Figura 4.7 no son tan notorias.

3. Se vuelve a confirmar que el mejor modelo necesita únicamente del texto y las imágenes, es decir, el texto en las imágenes no es relevante. Sin embargo, el hiperparámetro del número de imágenes que se deben considerar es diferente: ahora es 3, con el MSSAID fue 1.

4. Permanece la confusión entre contenido que se cataloga como spam y neutro.

4.6. Discusión Final

Los resultados obtenidos en los experimentos permiten determinar que el modelo de imagen y texto propuesto mejora la clasificación de las publicaciones para determinar la polaridad general de las mismas, en especial si se compara con aquellas que usan únicamente texto. También, se manifiesta una dominancia de la fusión por suma sobre la fusión por codificador. A pesar del mecanismo más complejo que sustenta la idea de la fusión por codificador, la regla de sumar los embeddings correspondientes al token [CLS] de cada modelo ajustado es la mejor, tanto en resultados obtenidos como en tiempo computacional ya que es más sencilla de obtener. En la misma línea, todo apunta a que

agregar texto en imágenes resulta ser poco útil y únicamente es necesario considerar texto e imágenes. Sin embargo, esta conclusión puede deberse a que el proceso necesita realizarse de una manera más compleja. Por ejemplo, Sidorov et al. [151] introducen el problema de subtitulado de imágenes con comprensión de lectura donde se recalca la necesidad de una mejor compresión de las imágenes para entender lo que se expresa a través de ellas, especialmente cuando existe texto incrustado y como se relaciona con su entorno. En el análisis de error del MSSAID y el MCOVMEX es posible apreciar que este caso originó algunos errores de clasificación, por lo que tal adición puede presentar una mejora significativa.

Ligado al punto anterior, es relevante discutir acerca de las imágenes con texto, un formato popular en redes sociales donde destacan los memes. En particular, se puede argumentar que analizar el texto de la imagen por separado no supone ninguna ventaja adicional a los sistemas. Sin embargo, agregar las imágenes ayuda considerablemente al rendimiento general de los modelos. Esto apunta a que el meme es un caso especial de multimodalidad donde el texto forma parte intrínseca de la imagen y aislar ambos elementos es poco útil: se deben proponer mejores métodos para analizar la fusión de modalidades al mismo tiempo y no por separado, lo cual agrega una capa adicional de complejidad al proceso de clasificación multimodal.

Los experimentos del número de imágenes llevados a cabo con el conjunto de datos MSSAID arrojó resultados importantes que permitieron mejorar el modelo para el MCOVMEX. A diferencia de los conjuntos de datos usuales que se manejan para la tarea de análisis multimodal de imagen y texto (Tabla 1.1), no es del todo evidente cómo la incorporación de múltiples (más de una) imágenes afecta los modelos de aprendizaje ya que todos consideran una sola. En particular, los resultados muestran que, en efecto, se debe realizar este análisis de forma previa para determinar el número adecuado de imágenes ya que no es el mismo para todos los casos y presenta una mejora significativa. Aún más, no se sabe, cuando existen múltiples imágenes, cuál de ellas o qué combinación de ellas es la más importante (¿la primera, la segunda, la tercera con la cuarta?) o la que puede proporcionar más información al texto en turno. Una posible opción es realizar un modelo preliminar con las imágenes como lo hacen Vempala and Preoţiuc-Pietro [152] para determinar cuáles son las más relevantes y cuales no aportan significado adicional.

Finalmente, las matrices de confusión de los mejores modelos (Figuras 4.16 y 4.19) demuestran buenas capacidades en general para discriminar clases positivas y negativas. Sin embargo, existe una confusión entre contenido catalogado como spam y neutro. Se puede debatir sobre la necesidad de crear la cuarta clase de spam y juntarla con el contenido neutro. Por un lado el spam no es un sentimiento y su incorporación podría aumentar las métricas de rendimiento considerablemente. Sin embargo, el spam, al ser abundante en redes sociales, presenta la oportunidad de crear sistemas de detección temprana que pueden beneficiar a los sistemas de análisis de sentimientos, evitando el ruido que aportan a los modelos, mejorando el análisis de datos y optimizando su rendimiento general.

Los conjuntos de datos analizados presentan resultados contrastantes. Por un lado, el mejor modelo del MSSAID presenta un CCM del 67.17 % y una exactitud balanceada del 72.10 %, que no logra superar al modelo preliminar con un resultado del 74.74 % en esta última métrica. Lo anterior se puede deber al modelo de la MVS que se entrena con pocos datos (607) que no rebasan la dimensión de los modelos de BETO (768). Además, se presenta una mayor cantidad de transiciones de sentimiento entre las parte negativa

y positiva, apuntando a la presencia de situaciones como sarcasmo e ironía. Se puede concluir, además, que la estrategia adoptada de ajustar los modelos base es correcta y que CLIP multilingüe no es un modelo adecuado debido a que no existe una versión dedicada para el español. En contraste, el MCVOMEX es un conjunto de datos donde el texto se alinea mejor con el sentimiento general del texto. Lo anterior se refleja al presentar un resultado base (al considerar únicamente texto) de 88.84 % de CCM. A pesar de esto, incorporar hasta las tres primeras imágenes ayuda a mejorar los resultados del modelo de imagen y texto, elevando el rendimiento del sistema a un 94.26 % de CCM.

Cabe resaltar el hecho de que, a pesar de los buenos resultados obtenidos por BETO, estos modelos presentan ciertas limitantes. Para el español, existen opciones limitadas para elegir sobre otros modelos vastos de lenguaje ya que la mayoría de los esfuerzos se orientan hacia el inglés y, de los ejercicios con CLIP multilingüe, se puede concluir que no son una buena opción cuando se tienen versiones dedicadas para cada idioma. En el hub de Hugging Face, al buscar «spanish» en la categoría de modelos para clasificación de texto, se obtienen (para abril de 2025, fecha en la que se realizó la consulta) apenas 291 modelos, en contraste de los 92,168 que existen en total: apenas el 0.31 % del total de los modelos para dicha tarea. Muchos de esos modelos no son útiles ya que se encuentran ajustados para tareas específicas, por lo que la cantidad de modelos viables se vuelve aún más pequeña. Por otro lado, BERT (y sus variantes), aunque provechoso, se está haciendo viejo: la versión grande de BERT cuenta con 340 millones de parámetros, mientras que la versión de Llamma 4 cuenta con 109 mil millones de parámetros activos. Además, muchos de los modelos vastos de lenguaje que forman parte del estado del arte se encuentran fuera del libre acceso, lo cual complica el entrenamiento de versiones especializadas como las que se ajustaron en este trabajo que se pueden hacer de uso libre para el público.

Otra limitante es el cierre de las APIs para la recolección de información. Afortunadamente, se logró recolectar información antes de la compra de Twitter y su posterior transformación en X. Esto significó para la API su completa monetización y la pérdida de los accesos académicos a los datos. Lo anterior ocasiona una grave privación para trabajos académicos que complica el estudio de diversos fenómenos en la red social, especialmente cuando la información se encuentra detrás de una barrera de pago.

Por último, es importante mencionar que el modelo propuesto permite trabajar con datos de diversos temas, haciendo énfasis en su versatilidad. Aunque en el presente trabajo se delimitó el análisis a peleas de un boxeador mexicano y un enfoque general del COVID-19 en México, se puede recolectar datos y anotarlos usando el esquema propuesto en la Sección 3.1. En consecuencia, se puede entrenar un modelo base para el texto y las imágenes, aplicar el método de fusión por suma (dado que es el mejor), elegir un algoritmo de clasificación y obtener un modelo de clasificación multimodal general para trabajar con texto y múltiples imágenes de redes sociales.

4.7. Disponibilidad de Códigos

Para facilitar la accesibilidad y reproducibilidad de los resultados expuestos en esta investigación, se facilita el acceso a los códigos en Python utilizados para procesar y analizar los datos con los modelos de imagen y texto propuestos para el conjunto de datos MSSAID y MCOVMEX. Los enlaces correspondientes se pueden consultar en la Tabla 4.12.

Repositorio	Enlace
Análisis del MSSAID	https://github.com/lzun/mssaid-msa
Demo del MCOVMEX	https://huggingface.co/spaces/lzun/multimodal-covid-19-spanish

Tabla 4.12: Repositorios con los códigos utilizados en el presente trabajo del MSSAID y un demo con una herramienta desarrollada para el MCOVMEX con los resultados obtenidos.

Capítulo 5

Conclusiones

A lo largo del trabajo de investigación se presentó un modelo de clasificación multimodal para trabajar con múltiples imágenes y texto en español de publicaciones de redes sociales. Con este fin, se construyeron dos conjuntos de datos con un esquema de anotación especial. Además, se propuso un marco de trabajo para extraer características de cada modalidad (texto, imagen y texto en imágenes) mediante el entrenamiento de modelos base de BETO y ViT, para después fusionar las representaciones vectoriales usando dos métodos de fusión: fusión por suma y fusión por codificador. A continuación se presentan las respuestas a las preguntas de investigación planteadas en la Sección 1.4 y el trabajo futuro.

5.1. Respuestas a las Preguntas de Investigación

P1. En el caso del problema de imágenes y texto, el trabajo previo apunta a un sesgo generado por los conjuntos de datos disponibles. ¿Cómo se pueden enfocar las aplicaciones para otros idiomas distintos al inglés, en particular el español?

Se crearon dos conjuntos de datos usando información de X sobre eventos relevantes en México enfocados en el idioma español y el dialecto mexicano: el MSSAID y el MCOVMEX. A diferencia de conjuntos de datos en el campo, resalta la inclusión de varias imágenes que acompañen una publicación. Se propone un esquema de anotación que incluye la anotación por separado y en conjunto de texto e imágenes en cuatro clases: negativo (-1), neutral (0), positivo (+1) y spam (+2). Dicho esquema permitió obtener una mejor configuración y entendimiento necesario para entrenar los modelos propuestos y el análisis de diferentes componentes, tanto de forma individual como en conjunto.

A pesar del tamaño y la poca diversidad de temas presentes en los conjuntos de datos, los elementos que los caracterizan permiten crear un ambiente más controlado para evaluar futuros avances metodológicos antes de aplicarse a conjuntos de datos más grandes.

P2. ¿Es posible crear sistemas de análisis de sentimientos multimodal para que trabajen con múltiples imágenes al mismo tiempo?

En este trabajo se propuso una metodología que emplea modelos de lenguaje (BETO) y visión (ViT) basados en la arquitectura de Transformer para crear un sistema de análisis de sentimientos multimodal aplicado a dos conjuntos de datos en español. La metodología

incluye la extracción de características de texto, imágenes y texto en imágenes de las publicaciones.

Para fusionar los datos, se expusieron dos estrategias: la fusión por suma y la fusión por codificador. La primera aprovecha la presencia del token especial para clasificación ubicado al inicio de las secuencias de embeddings [CLS] y crea una regla donde los suma para obtener la representación final de las modalidades. Por otro lado, la fusión por suma aprovecha el mecanismo de autoatención del Transformer para analizar el contexto de las secuencias de los datos de entrada. Se concatentan los embeddings de cada modalidad y se analiza con capas de codificador del Transformer para obtener la representación fusionada de las modalidades presentes en una publicación. Ambas modalidades permiten incorporar múltiples imágenes de forma nativa.

Finalmente, la representación final de las modalidades se alimenta a una MVS con entrenamiento penalizado para obtener la clase de cada publicación.

P3. ¿Cómo impacta el número de imágenes consideradas al sistema de análisis de sentimientos multimodal?

El modelo de clasificación de imagen y texto propuesto presenta buenos beneficios al determinar las clases positiva y negativa, tanto en el MSSAID como con el MCOVMEX. El mejor modelo del MSSAID considera texto y la primera imagen con fusión por suma, logrando un 67.17 % de CCM. Para el MCVOMEX, se considera texto y hasta las primeras tres imágenes con fusión por suma, logrando un 94.26 % de CCM.

El método de fusión por suma obtuvo los mejores resultados con un bajo costo computacional. Además, el número variable de imágenes debe ser considerado de antemano como un hiperparámetro de los modelos, ya que al contener un número entre una y cuatro de ellas, no se sabe el valor indicado para obtener el mejor modelo de clasificación. Este resultado se obtuvo de los experimentos llevados a cabo en el conjunto de datos MSSAID (Sección 4.4.4) para aplicarse después al MCVOMEX.

P4 ¿Cuál es el impacto del texto incrustado en imágenes a los modelos de imagen y texto de análisis de sentimientos multimodal?

Gracias a los resultados del experimento sobre el impacto de las modalidades en los datos (Sección 4.4.4) en el MSSAID, se concluye que agregar el texto incrustado en las imágenes no es relevante para el proceso y solo se necesita incluir las imágenes.

P5 El contenido catalogado como spam en redes sociales no se considera un problema lo suficientemente relevante como para ser incluido dentro de los modelos de clasificación. ¿Cómo se puede incluir este tipo de datos dentro de un marco de trabajo para el análisis de sentimientos multimodal y cómo afecta su incorporación?

Gracias a los experimentos de proyección de los embeddings en la Sección 4.4.4 se detectó que las publicaciones neutrales se confunden con aquellas que se catalogan como spam. Lo anterior se confirma al analizar la matriz de confusión del mejor modelo (Figuras 4.16 y 4.19), tanto para el MSSAID como el MCOVMEX, donde se detectan errores de clasificación entre dichas clases.

5.2. Trabajo Futuro

El trabajo futuro consiste en superar varias limitaciones presentes en el marco propuesto. En primer lugar, con respecto al conjunto de datos, se busca ampliar la evaluación de los modelos con conjuntos de datos más amplios y diversos. Por otro lado, existe la oportunidad de incorporar nuevos modelos de fusión, como la fusión de tensores o arquitecturas más complejas basadas en Transfomer, particularmente en el contexto de conjuntos de datos más amplios y con mayor diversidad temática, para explorar a fondo la generalización del análisis multimodal de sentimientos. Ligado a este punto, surge la oportunidad de usar una red neuronal densa como algoritmo de clasificación en lugar de la MVS para explorar si su inclusión mejora el proceso de fusión por codificador.

Existen nuevos sistemas de ROC basados en Transformer y modelos de lenguaje de visión para explorar y mejorar la detección en diversas imágenes de redes sociales. Por lo tanto, existen nuevas maneras de mejorar el sistema de ROC utilizado en el marco propuesto y determinar si dichas herramientas justifican la inclusión de texto en imágenes al marco multimodal.

Si bien los modelos basados en BERT han mostrado resultados prometedores en diferentes aplicaciones, modelos vastos de lenguaje más recientes con más parámetros podrían mejorar el rendimiento general del modelo multimodal. Además, existe la oportunidad de abordar el problema de determinar qué imágenes de una colección aportan significado al texto de una publicación en redes sociales y complementan aún más su significado.

Finalmente, se enfatiza la necesidad de mejorar la detección de contenido spam en redes sociales digitales. Una opción es probar un sistema de dos pasos que fusiona las clases neutrales y spam en una sola superclase para trabajar con un problema de clasificación de tres clases y luego construir un segundo clasificador para separar la superclase fusionada.

En conclusión, el marco de Análisis de Sentimiento Multimodal propuesto proporciona un punto de partida para investigaciones adicionales y futuras aplicaciones.

Apéndice A

Instrucciones de Anotación para el Multimodal COVID19 Mexico

El conjunto de datos que se busca anotar consta de 1000 tuits (en ese entonces todavía Twitter) que se extrajeron usando como términos de consulta aquellos relacionados con el COVID en México entre 2020 y 2023.

El conjunto de datos recopila una gran cantidad de datos de múltiples etapas del fenómeno, por lo que se espera cubrir distintos aspectos relacionados con este. Por ejemplo, se espera que aparezcan mensajes sobre COVID y política pública, COVID y educación, COVID y deportes, COVID y salud pública, COVID y relaciones internacionales, por mencionar algunos. Por lo tanto, es necesario, al momento de considerar toda la información disponible, ubicarse dentro de ese aspecto particular del COVID para determinar el sentimiento de cada aspecto que se pide a continuación. Por ejemplo, dentro del deporte, se espera encontrar que el COVID paró muchas actividades (algo negativo), pero una vez superada la emergencia, poco a poco las competencias se fueron reactivando con éxito (algo positivo).

La tarea principal es anotar el sentimiento de texto e imágenes del conjunto de datos que se proporciona usando el siguiente esquema de anotación, donde se explica lo que indica cada clase:

- -1 para tuits negativos. Es decir, aquellas publicaciones que muestran claramente un sentimiento negativo predominante en su mensaje. Esto incluye mensajes sobre defunciones, problemas del personal médico, críticas hacia figuras públicas o decisiones tomadas por ellos, etc.

- 0 para tuits neutros. Esta categoría incluye publicaciones que no demuestran sentimiento preponderante alguno, pero cuyo tema se relaciona con el COVID. Esta categoría reúne principalmente a las publicaciones objetivas de medios noticiosos que dirigen tráfico a un sitio web (página web o canal de YouTube, por ejemplo).

- +1 para tuits positivos. Es decir, aquellas publicaciones que muestran claramente un sentimiento positivo predominante en su mensaje. Esto incluye mensajes de apoyo emocional, declaraciones positivas de salud, crítica positiva hacia algún desarrollo de la pandemia como las vacunas, etc.

- +2 para spam. Es posible encontrar mensajes que no hablan sobre el COVID. Por ejemplo, publicaciones que se cuelguen de las tendencias para vender productos o dirigir tráfico hacia otros sitios con fines de lucro o captar la atención de los usuarios con otros fines.

En el caso de las imágenes, se debe acceder a cada enlace para acceder a ellas. Es posible que una o varias imágenes no se encuentren disponibles ya que el autor las borró o bien, la cuenta que las publicó ya no existe en el sistema. En dado caso, usar el valor -2 para indicar que la imagen no se encuentra disponible. Para anotar el sentimiento de cada imagen, se debe usar la misma guía que en el caso del texto. Sin embargo, se debe anotar el sentimiento que causa la imagen al momento de analizar todos sus componentes.

- -1 para imágenes que evoquen sentimientos negativos. Por ejemplo, ira, tristeza, confusión, confusión, etc.
- 0 para imágenes que no evoquen ningún sentimiento. Esto incluye imágenes informativas como miniaturas de videos o infografías.
- +1 para imágenes que evoquen sentimientos positivos. Por ejemplo, felicidad, esperanza, alivio, etc.
- +2 para imágenes que no tengan nada que ver con el COVID. Por ejemplo, contenido para adultos, imágenes con fines de lucro, información falsa, menús de restaurantes, fotos de comida, etc.
- -2 si la imagen no se encuentra disponible al momento de realizar la consulta correspondiente.

Se debe tener particular atención con aquellas imágenes que tengan texto incrustado en ellas, como lo son memes o similares al caso de la Figura A.1:

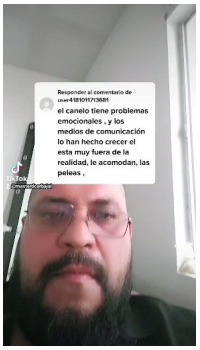


FIGURA A.1: Ejemplo de una imagen con texto incrustado en ella. Fuente: X.

En esta situación, se debe vigilar cómo interactúa el texto con las imágenes y tenerlo en cuenta al momento de anotar el sentimiento correspondiente de la imagen.

En total, se deben anotar los siguientes campos de un tuit, que aparecen en las última columna del archivo de Google Sheets:

1. El sentimiento presente en el texto en la columna text_sentiment.
2. El sentimiento presente en cada una de las imágenes presentes en el tuit, que pueden ser de una a cuatro. En este caso, el sentimiento es el que genera cada imagen de forma individual, en la columnas image_sent_1, image_sent_2, image_sent_3, image_sent_4.

3. El sentimiento que generan las imágenes en conjunto en la columna img_overall_sent.
4. El sentimiento que genera la combinación de texto e imágenes en conjunto en la columna overall_sent.
5. El sentimiento presente en el texto incrustado en las imágenes, si es que lo hay en alguna de ellas. Si hay más de una imagen con texto en el tuit, anotar el sentimiento como una lista separada por comas, e.g, 1,2 para dos imágenes, o 1,0,0 para tres, etc.

Algunos ejemplos:

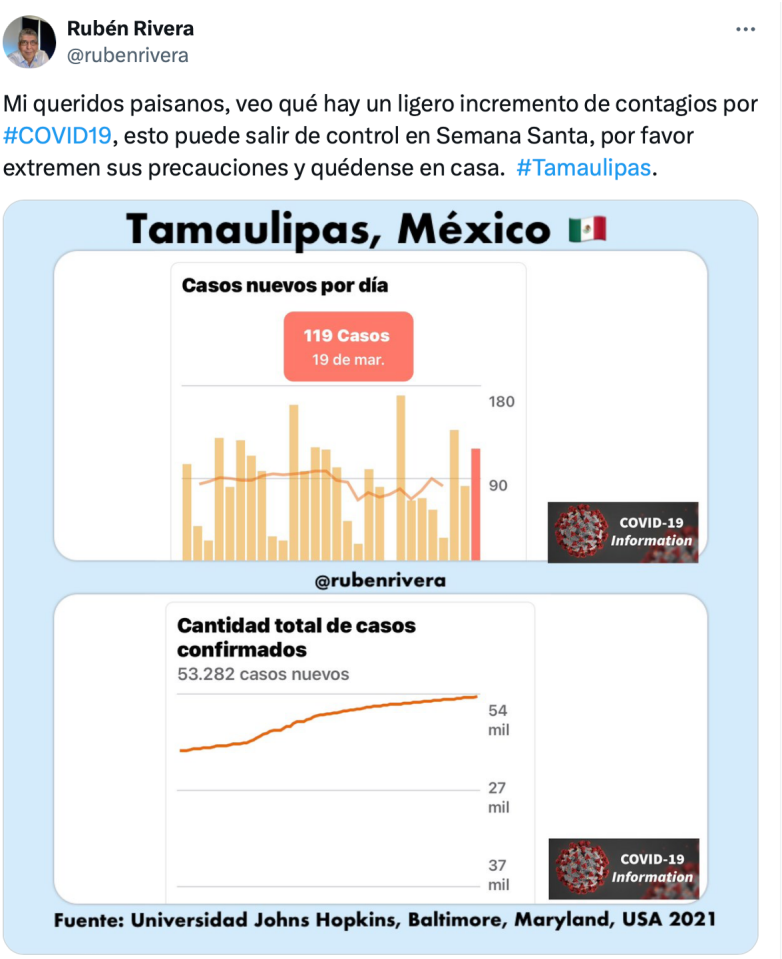


FIGURA A.2: Primer ejemplo del conjunto de datos. Fuente: X.

Para el ejemplo de la Figura A.2:

- El texto expresa preocupación, por lo que se puede considerar negativo.
- La imagen en general, puede considerarse neutra al ser informativa, en lo general.
- Al considerar ambos elementos en conjunto, el tuit tiene una carga en su mayoría negativa.

- Aunque hay texto en la imagen, no lo vamos a considerar ya que necesitamos que tenga una estructura similar a un meme o un video corto donde se emita una opinión. No hay sentimiento de texto en imágenes.

Figura A.3: Segundo ejemplo del conjunto de datos. Fuente: X.

Para el ejemplo de la Figura A.3:

- Sentimiento negativo en el texto ya que es una crítica sobre el manejo del COVID en México por parte de las autoridades.
- La imagen tiene texto incrustado que nos interesa, el cual es negativo por la referencia a Díaz Ordaz.
- La imagen, analizándola toda, es negativa ya que es una crítica hacia el gobierno en turno.
- Juntando texto e imagen, el sentimiento del tuit es negativo.

FIGURA A.4: Tercer ejemplo del conjunto de datos. Fuente: X.

Para el ejemplo de la Figura A.4:

- El texto del tuit si se relaciona con el COVID, aunque no con la enfermedad en sí misma. Estos casos pueden etiquetarse como positivos si se considera que el mensaje implica un mensaje o situación de apoyo, o que ayuda a aliviar la carga que causó la enfermedad. Por otro lado, se puede considerar neutro ya que el mensaje cuenta con una estructura más cercana a una noticia o reportaje.
- Se tienen cuatro imágenes, por lo que cada una de ellas debe tener su propia etiqueta de sentimiento.
- En general, el tuit tiene una carga neutra. Sin embargo, también se puede considerar positiva según el criterio del anotador por el mensaje de "salir adelante en esta etapa de crisis".

FIGURA A.5: Cuarto ejemplo del conjunto de datos. Fuente: X.

Para el ejemplo de la Figura A.5:

- Otro elemento común es que el COVID se encuentre como un tema acompañado de otros, aunque no sea el principal, similar al caso del boxeador anterior. No se debe considerar spam.
- En este caso, tanto el texto, la imagen y el sentimiento general del tuit son negativos.

FIGURA A.6: Quinto ejemplo del conjunto de datos. Fuente: X.

Para el ejemplo de la Figura A.6:

- Cuenta oficial cuyo fin es informar. Por lo tanto todo se considera neutral (texto, imagen y sentimiento general del tuit).
- Aunque la imagen tiene texto, este no nos interesa al ser una infografía.
- La imagen al ser infografía, se considera neutra.

Apéndice B

Resultados Completos Experimentos del Número Máximo de Imágenes en MSSAID

Tabla B.1: Resultados completos de los experimentos del número máximo de imágenes con el conjunto de datos MSSAID. Los valores de k_C y k_γ apuntan al valor de $C = 2^{k_C}$ y $\gamma = 2^{k_\gamma}$, respectivamente, de la MVS. En negritas se resalta el mejor resultado obtenido

Modalidades	Método de Fusión	Modelo	Imágenes	Exactitud	Exactitud Balanceada	F_1^w	CCM	k_C	k_γ
T+I	Fusión Suma	M-CLIP	1	0.5522	0.5270	0.5564	0.3547	1.1250	-6.0000
			2	0.5224	0.5213	0.5452	0.3631	3.0000	-10.0000
			3	0.5075	0.4985	0.5317	0.3363	3.5000	-10.5000
			4	0.4776	0.4660	0.4994	0.2931	7.6250	-14.6250
		Modelos Base	1	0.6716	0.6319	0.6739	0.5103	1.6250	-7.6250
			2	0.6418	0.5899	0.6409	0.4705	1.7656	-7.7031
			3	0.5970	0.5637	0.5985	0.4189	1.2500	-7.7500
			4	0.6119	0.5713	0.6109	0.4362	1.3125	-7.5625
		Modelos Ajustados	**1**	**0.7761**	**0.7210**	**0.7776**	**0.6717**	**2.6875**	**-6.3750**
			2	0.7313	0.6634	0.7325	0.6008	9.9063	-15.7188
			3	0.7164	0.6407	0.7176	0.5766	7.1250	-13.2500
			4	0.7313	0.6657	0.7310	0.5986	9.0000	-15.2500
	Fusión Codificador	M-CLIP	1	0.2239	0.1427	0.2320	-0.1280	-0.5000	-9.2500
			2	0.1642	0.1311	0.1819	-0.1545	4.6250	-13.1250
			3	0.4925	0.2500	0.3251	0.0000	0.1250	-8.1250
			4	0.4925	0.2500	0.3251	0.0000	-0.6250	-11.1250
		Modelos Base	1	0.5075	0.4542	0.5175	0.3031	0.6875	-7.9688
			2	0.5522	0.4653	0.5347	0.3255	1.5625	-7.6875
			3	0.5075	0.4460	0.4949	0.2762	1.0313	-8.1250
			4	0.4776	0.4345	0.4893	0.2560	2.2500	-10.7500
		Modelos Ajustados	1	0.7612	0.7054	0.7615	0.6457	0.7656	-9.6250
			2	0.7313	0.6599	0.7334	0.6001	1.2031	-9.8438

Modalidades	Método de Fusión	Modelo	Imágenes	Exactitud	Exactitud Balanceada	F_1^w	CCM	k_C	k_γ
			3	0.7015	0.6540	0.7060	0.5668	2.8750	-8.5000
			4	0.7164	0.6442	0.7181	0.5821	1.0000	-9.2500
T+I+TI	Fusión Suma	M-CLIP	1	0.5224	0.5364	0.5384	0.3631	7.0000	-14.0000
			2	0.5075	0.5311	0.5237	0.3441	6.1875	-13.4375
			3	0.4776	0.4834	0.4947	0.2884	6.8125	-14.0625
			4	0.4776	0.4950	0.4939	0.3008	9.9219	-17.0156
		Modelos Base	1	0.5522	0.5654	0.5614	0.3898	1.7500	-9.7500
			2	0.5075	0.5136	0.5203	0.3231	1.8672	-10.5078
			3	0.6119	0.6108	0.6235	0.4637	5.0000	-12.0000
			4	0.5075	0.5136	0.5221	0.3352	1.8750	-10.8750
		Modelos Ajustados	1	0.7313	0.6925	0.7428	0.6115	5.0000	-10.7500
			2	0.7164	0.6487	0.7222	0.5847	6.7500	-13.0000
			3	0.7313	0.6679	0.7370	0.6052	10.3750	-16.6250
			4	0.7313	0.6679	0.7370	0.6052	11.1875	-17.3750
	Fusión Codificador	M-CLIP	1	0.2537	0.1811	0.2379	-0.0913	0.2500	-12.0000
			2	0.1940	0.2182	0.1899	-0.0219	0.0625	-11.4219
			3	0.4925	0.2500	0.3251	0.0000	-1.0000	-9.0000
			4	0.4925	0.2500	0.3251	0.0000	-2.0000	-13.5000
		Modelos Base	1	0.4776	0.4519	0.4923	0.2739	2.0000	-10.0000
			2	0.5522	0.4863	0.5600	0.3445	0.8945	-7.5703
			3	0.5224	0.4712	0.5334	0.3125	2.0000	-9.0000
			4	0.5672	0.4728	0.5405	0.3285	1.1250	-7.2813
		Modelos Ajustados	1	0.6567	0.5719	0.6596	0.4959	3.5625	-9.3750
			2	0.7164	0.6407	0.7191	0.5788	0.2500	-8.7500
			3	0.7164	0.6464	0.7174	0.5796	0.1250	-6.8438
			4	0.6567	0.5603	0.6543	0.4883	3.0000	-11.0000

Apéndice C

Resultados Completos Experimentos del Número Máximo de Imágenes en MCOVMEX

Tabla C.1: Resultados completos de los experimentos del número máximo de imágenes con el conjunto de datos MCOVMEX. T: texto, I: imágenes, TI: texto en imágenes. Los valores de k_C y k_γ apuntan al valor de $C = 2^{k_C}$ y $\gamma = 2^{k_\gamma}$, respectivamente, de la MVS. En negritas se resalta el mejor resultado obtenido.

Fusión	Modalidades	Imágenes	Exactitud	Exactitud Balanceada	F_1^w	CCM	k_C	k_γ
Fusión Suma	T+I+TI	1	0.8800	0.8651	0.8821	0.8303	8.3438	-16.0313
		2	0.8900	0.8723	0.8916	0.8431	1.3380	-14.5000
		3	0.8800	0.8657	0.8835	0.8304	8.2969	-16.6250
		4	0.9200	0.8931	0.9202	0.8848	7.0000	-14.0000
	T+I	1	0.9500	0.9389	0.9499	0.9285	8.2500	-15.2500
		2	0.9400	0.9317	0.9396	0.9146	5.8750	-12.3750
		3	**0.9600**	**0.9460**	**0.9600**	**0.9426**	**7.5000**	**-13.2500**
		4	0.9500	0.9389	0.9499	0.9285	5.0000	-11.7500
	T		0.9300	0.9246	0.9311	0.9020	5.0000	-13.7500
Fusión Codificador	T+I+TI	1	0.9000	0.9032	0.9004	0.8627	-6.2500	-12.5000
		2	0.9000	0.9032	0.9004	0.8627	-4.2500	-14.7500
		3	0.9200	0.9174	0.9192	0.8877	6.0000	-14.0000
		4	0.9400	0.9317	0.9407	0.9152	-6.2500	-12.5000
	T+I	1	0.9200	0.9174	0.9196	0.8880	5.7500	-14.0000
		2	0.9200	0.9174	0.9205	0.8884	-6.2500	-10.5000
		3	0.9100	0.9103	0.9097	0.8749	-4.2500	-13.5000
		4	0.9200	0.9174	0.9205	0.8884	-6.2500	-10.5000
	T		0.9200	0.9174	0.9205	0.8884	2.5000	-11.7500

Apéndice D

Proyecto Imagen de México

Adicional al proyecto de clasificación de polaridad multimodal, durante el programa de posgrado se participó en diversas colaboraciones que sirvieron de proyectos preliminares para crear diversos módulos del marco de clasificación multimodal propuesto en el Capítulo 3. El proyecto Imagen de México en el Mundo [153] tiene la finalidad de analizar la imagen país de México vista desde el exterior a partir de publicaciones de redes sociales y noticias de medios y personalidades reconocidas en ámbitos como el político, económico, deportivo, entretenimiento, entre otros. El trabajo que se llevó a cabo consistió en la creación de un sistema de recolección de datos y la creación de un sistema de clasificación automática en 18 clases de diversos tuits y encabezados de noticias relacionados con México. A continuación se presentan brevemente los resultados obtenidos de la colaboración y, en particular, del sistema de clasificación. Los resultados completos se puede consultar en [80].

D.1. Introducción

El Modelo Analítico de Imagen País (MAIP) introduce un marco de trabajo que consta de nueve tipologías mayores y nueve tipologías menores, que de manera colectiva se conocen como imagen pública de un país. Las tipologías de imagen de país del MAIP abarcan una amplia gama de categorías, entre ellas: cosmopolita, amigo, moderno, emergente, cooperador, aliado, rival, adversario menor, neutral, independiente, exótico, degenerado, dependiente, marginal, intervencionista, colonizador, enemigo y bárbaro.

El MAIP se acompaña de un conjunto de datos compuesto por tuits y encabezados de noticias del New York Times recolectados entre 2012 y 2018. En total, se cuenta con 7,569 tuits y 204 encabezados de noticias para un total de 7,773 elementos que conforman el conjunto de datos. De la Figura D.1 se puede concluir que el conjunto de datos es desequilibrado, donde las dos tipologías menos comunes no rebasan los cinco elementos.

D.2. Metodología

Para la creación del sistema de clasificación del conjunto de datos del MAIP, se propusieron dos modelos: uno clásico basado en la perspectiva del Machine Learning tradicional, y otro basado en el ajuste de modelos vastos de lenguaje. Además, se trataron diversos componentes que aparecen en los tuits como emojis, etiquetas como hashtags y nombres de usuarios, elementos en diversos idiomas y el tratamiento de un conjunto de datos pequeño (para el número de clases que se maneja) y no balanceado.

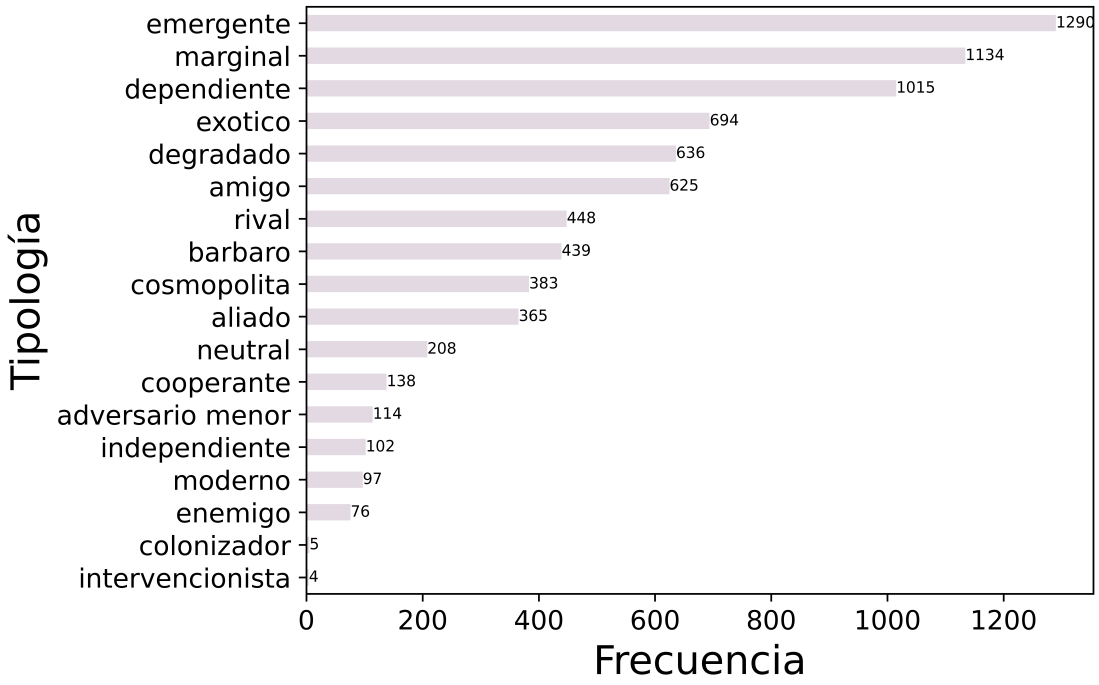


FIGURA D.1: Frecuencia de cada tipología del conjunto de datos del MAIP ya considerando tuits y encabezados juntos. Realización propia, inspirada en [80].

D.2.1. Modelo Clásico

El modelo clásico incorpora elementos tradicionales del Machine Learning, en particular en la extracción de características para representar el texto de los tuits y los encabezados de noticias. La idea general de este sistema y sus diversos componentes se puede consultar en la Figura D.2.

Traducción y Aumento de Texto

Como se discutió en la Sección 3.2.4, el desequilibrio entre datos puede causar diversos problemas durante la fase de aprendizaje de los modelos de clasificación. Por tal razón se creó y se utilizó el sistema de aumento de datos utilizada en dicha sección en este sistema con un paso adicional: dado que los tuits se encuentran en diversos idiomas, se tomó la decisión de traducirlos todos al inglés. Lo anterior se logró mediante el uso de la API de DeepL[1] en Python. Dado que en este caso el idioma es inglés, se utilizó Wordnet en ese idioma para encontrar sinónimos de sustantivos y adjetivos.

Se abordan dos estrategias para realizar el aumento de datos. La primera consiste en experimentar con incrementos graduales a las clases menos representadas, en específico de uno a cuatro aumentos de su tamaño original: AGM1 para un incremento, AGM2 para dos, AGM3 para tres y AGM4 para cuatro. Las tipologías incrementadas corresponden a intervencionista, colonizador, enemigo, moderno, independiente, adversario menor,

[1] https://www.deepl.com/translator

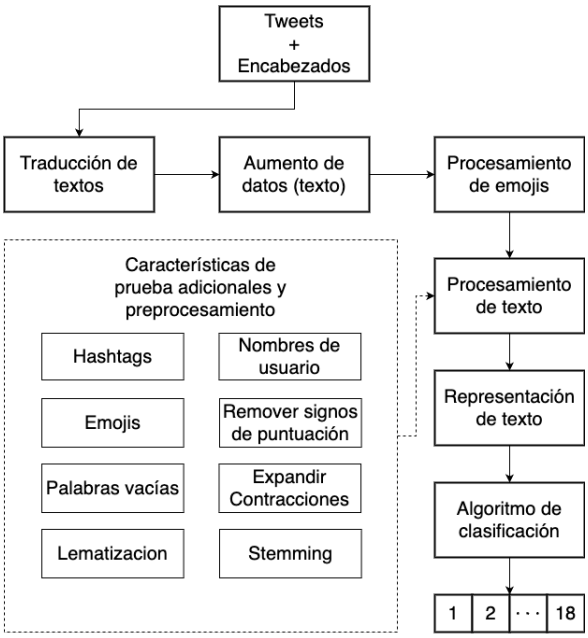


FIGURA D.2: Diagrama que muestra el proceso principal del modelo clásico para la clasificación de los datos. Realización propia, inspirada en [80].

cooperativo y neutral. Sin embargo, con fines comparativos se añade una segunda estrategia adicional que comprende el aumento de todos los datos cinco veces.

Procesamiento de Emojis

Para trabajar con emojis se adopta la misma estrategia que en el marco preliminar para análisis de sentimientos multimodal propuesta en la Sección 3.5.1, solo que en este caso se convierten los emojis a su correspondiente interpretación en inglés.

Preprocesamiento y Procesamiento de Texto

Como se trabajan con tuits y encabezados de noticias, se debe proponer un marco para tratar este tipo de datos previo al paso de extracción de características. En específico para información de redes sociales como X, en el paso de preprocesamiento, la literatura recomienda eliminar elementos como nombres de usuario, hashtags y cashtags [79]. Sin embargo, dado que los tuits y encabezados se seleccionaron manualmente, no queda totalmente claro si los elementos mencionados anteriormente deben ser incluidos o descartados. Por lo tanto, se prueba su inclusión y exclusión en los modelos de aprendizaje para explorar su impacto, en solitario o en conjunto, y determinar su permanencia. Por otro lado, elementos como enlaces y números se eliminan. El paso de procesamiento de datos consiste en tareas de normalización del texto y reducción del vocabulario. Por ello, se tokeniza el texto, se pasa a letra minúscula, se expanden contracciones y se lematiza cada documento.

Representación de Texto

En este sistema de clasificación se eligió usar fastText [86], una biblioteca que permite el representar palabras y clasificación de oraciones en distintos idiomas que se han utilizado con éxito con datos de Twitter [154]. En lugar de considerar la representación vectorial de cada palabra, se elige la representación de cada oración que se genera al promediar los vectores de cada palabra usando los modelos preetrenados en inglés con una dimensión igual a 300.

Modelo de Clasificación

El modelo de clasificación que se considera es una Máquina de Vectores de Soporte con aprendizaje sensible al costo, como se especifica en la Sección 2.5.3.

D.2.2. Modelo Ajustado

Los modelos preentrenados que se utilizaron para el ajuste (*fine-tunning*) son el modelo base de BERT, el modelo de base de Distilbert[2] [155] y el modelo base de Twitter-roBERTa-Base[3] para análisis de sentimientos [156]. El proceso de ajuste reemplaza al algoritmo de clasificación por una red neuronal densa para minimizar el error generado. Finalmente, se aprovecha el marco de trabajo propuesto para aumentar y tratar los datos con los que se operan.

D.2.3. Entrenamiento de los Modelos

El entrenamiento de los modelos, tanto para la MVS como para el ajuste de los modelos vastos de lenguaje, se hace de la misma forma como se explica en la Sección 3.4.3. Las métricas de evaluación consideradas son las mismas que las expuestas en la Sección 2.6, sin embargo, la métrica que se vigila en la medida F_1.

D.3. Resultados

Los resultados del modelo clásico considerando las diferentes características adicionales se pueden consultar en la Tabla D.1. Por otro lado, en la Tabla D.2 se muestran los resultados obtenidos en los experimentos de aumento de datos propuestos en un inicio tanto para el mejor modelo clásico encontrado en el primer experimento, y los distintos modelos ajustados. Finalmente, la matriz de confusión de los mejores modelos encontrados, tanto el clásico como el ajustado, se pueden consultar en la Figura D.3. Para mejorar la comparación entre modelos, en la Tabla D.3 se indica qué tipología se predice mejor en cada modelo utilizado.

D.4. Análisis de Resultados

En el trabajo, se comparó el rendimiento de un sistema de aprendizaje automático basado en una MVS (el modelo clásico) con embeddings de fastText contra otro sistema basado en el ajuste de modelos vastos de lenguaje como BERT, DistilBERT y Twitter

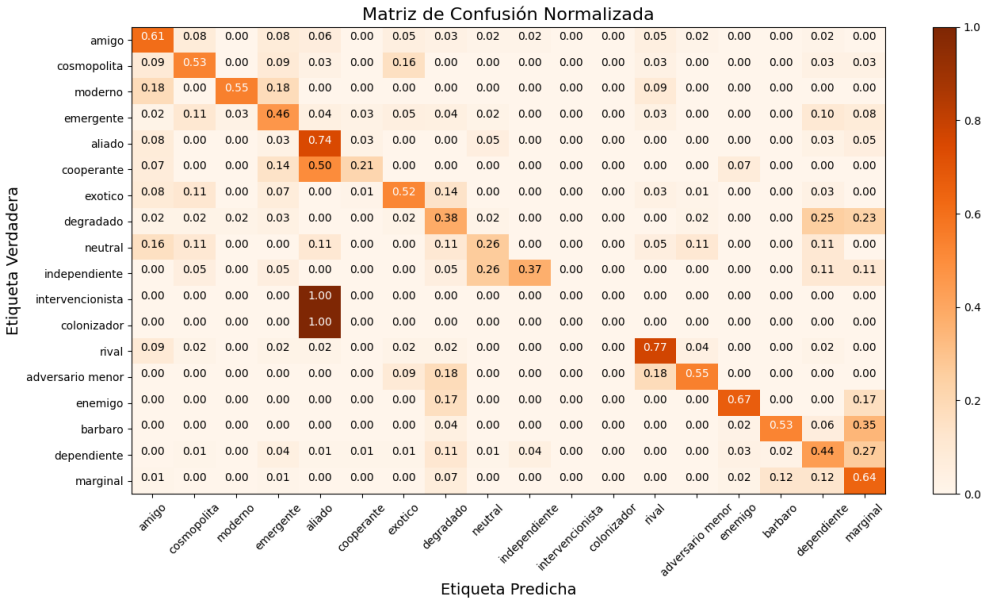
[2] https://huggingface.co/distilbert/distilbert-base-cased
[3] https://huggingface.co/cardiffnlp/twitter-roberta-base

TABLA D.1: Métricas de rendimiento de los experimentos de características adicionales del texto del modelo clásico con la MVS. Los valores mostrados en las columnas k_C y k_γ indican el valor de $C = 2^{k_C}$ y $\gamma = 2^{k_\gamma}$, respectivamente. HT significa hashtag, NU indica nombres de usuario, emoji es E, y «Ninguno» implica que se eliminan todas las características adicionales del texto. Las mejores puntuaciones de cada experimento y sus características asociadas se resaltan en negritas. Obtenida de [80]

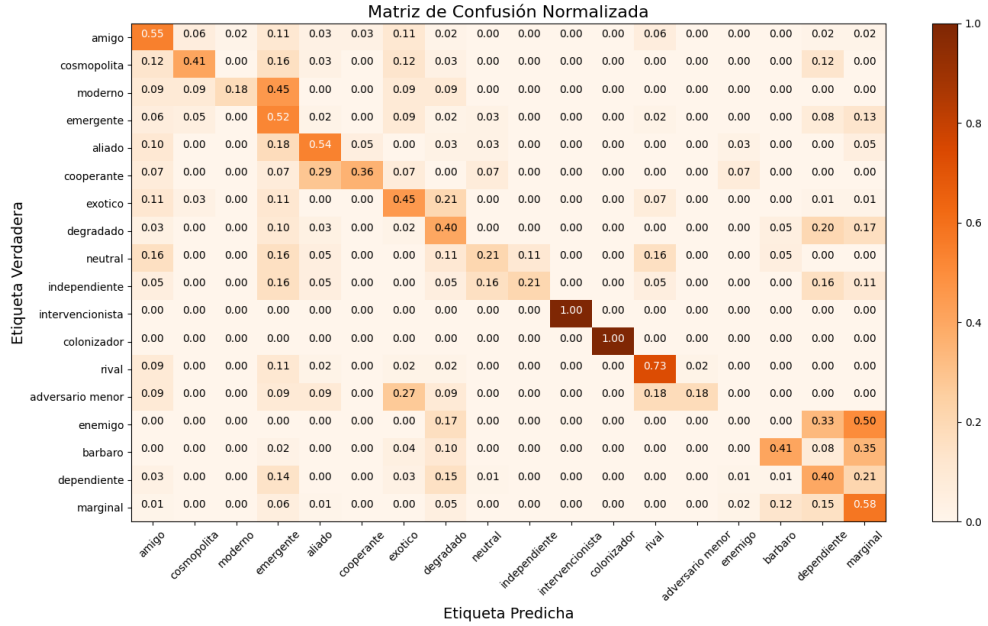
Con Palabras Vacías						
Características	Exactitud	Exactitud Balanceada	F_1^w	CCM	k_C	k_γ
Ninguna	**0.4692**	**0.4523**	**0.4654**	**0.4108**	**2.0195**	**3.5234**
HT	**0.4692**	**0.4523**	**0.4654**	**0.4108**	**2.0195**	**3.5234**
NU	0.4704	0.4514	0.4650	0.4113	2.7500	3.3750
E	0.4704	0.4514	0.4650	0.4113	2.7500	3.3750
HT + NU	0.4704	0.4514	0.4650	0.4113	2.7500	3.3750
HT + E	**0.4692**	**0.4523**	**0.4654**	**0.4108**	**2.0195**	**3.5234**
NU + E	0.4704	0.4514	0.4650	0.4113	2.7500	3.3750
HT + NU + E	0.4704	0.4514	0.4650	0.4113	2.7500	3.3750
Sin Palabras Vacías						
Características	Exactitud	Exactitud Balanceada	F_1^w	CCM	k_C	k_γ
Ninguna	0.4794	0.4544	0.4758	0.4228	2.2500	2.8125
HT	**0.4820**	**0.4669**	**0.4788**	**0.4262**	**2.0938**	**2.7188**
NU	0.4794	0.4544	0.4758	0.4228	2.2500	2.8125
E	**0.4820**	**0.4669**	**0.4788**	**0.4262**	**2.0938**	**2.7188**
HT + NU	0.4794	0.4544	0.4758	0.4228	2.2500	2.8125
HT + E	**0.4820**	**0.4669**	**0.4788**	**0.4262**	**2.0938**	**2.7188**
NU + E	0.4794	0.4544	0.4758	0.4228	2.2500	2.8125
HT + NU + E	0.4794	0.4544	0.4758	0.4228	2.2500	2.8125

roBERTa Base con el objetivo de automatizar la clasificación de datos de texto para analizar la imagen de México utilizando 18 clases basadas en teoría de las Relaciones Internacionales.

Sobre el modelo clásico, en las pruebas de características adicionales (hashtags, nombres de usuario, emojis y palabras vacías), se descubrió que eliminar las palabras vacías y conservar los hashtags, emojis, o hashtags y emojis arrojaron un mejor rendimiento. Se seleccionó el modelo con hashtags y sin palabras vacías debido a su menor dimensionalidad. Por otro lado, las propuestas de aumento de datos no mejoraron el rendimiento del modelo basado en la MVS, observando incluso una disminución en los puntajes durante el aumento parcial y el peor resultado con el aumento uniforme. Esto sugiere un posible sobreajuste durante el entrenamiento con datos aumentados. Además, el mejor rendimiento del modelo clásico, usando la medida F_1^w, fue de 47.88 %, obtenido con la eliminación de palabras vacías y la conservación de hashtags, sin aplicar aumento de datos. Finalmente,



(A)



(B)

FIGURA D.3: Matrices de confusión normalizadas para (A) el mejor modelo ajustado y (B) el mejor modelo clásico, según la medida F_1^w. Realización propia, inspirada en [80]

el modelo clásico mostró una mayor tasa de predicción en algunas tipologías, en específico la emergente, cooperante, degenerado, intervencionista y colonizador, en comparación con el mejor modelo ajustado.

En el caso del sistema basado en el ajuste de modelos vastos de lenguaje, el mejor resultado general se obtuvo al ajustar el modelo Twitter roBERTa Base sin aplicar técnicas de aumento de datos, alcanzando una medida F_1^w de 52.92 %. Los modelos de aprendizaje profundo mostraron pequeños beneficios de las técnicas de aumento de datos, una diferencia importante sobre los modelos clásicos. En esa misma línea, DistilBERT Base Cased tuvo una ligera mejora al aumentar todas las clases, mientras que BERT Base Cased alcanzó su punto más alto al aumentar tres veces las clases menos representadas. Sin embargo, el aumento de rendimiento fue marginal en la mayoría de los modelos y no justificó la inclusión de este paso en el mejor sistema desarrollado. El modelo Twitter roBERTa Base no pudo predecir correctamente las dos tipologías menos representadas, Intervencionista y Colonizador (aunque solo tenían un elemento en el conjunto de prueba). Finalmente se observó una eficacia significativa en las etiquetas Rival y Aliado, con un rendimiento alrededor del 77 % y 74 % respectivamente. Las tipologías Amigo, Enemigo y Marginal mostraron un rendimiento superior al 60 %, mientras que el resto estuvo por debajo de dicho umbral.

En conclusión, los resultados sugieren que el ajuste fino de modelos vastos de lenguaje, específicamente Twitter roBERTa Base, es más efectivo para la tarea de clasificación de la Imagen de México en este conjunto de datos. Sin embargo, el modelo clásico demostró ser mejor en la predicción de algunas tipologías, aunque no lo suficiente como para proponer una estrategia de dos modelos para predecir etiquetas en conjunto. La presentación de desafíos relacionados con el tamaño y el desbalance del conjunto de datos, así como la necesidad de mejorar la clasificación para tipologías específicas obligó a explorar diversas alternativas para aliviar sus efectos en los modelos de aprendizaje. Sin embargo, resultaron ser poco efectivas, por lo que sigue el problema abierto de cómo se puede mejorar el rendimiento de estos modelos con el conjunto de datos base.

TABLA D.2: Métricas de rendimiento de los modelos clásicos (MVS) y ajustados con el conjunto de entrenamiento según las técnicas de aumento de datos. El número en la columna AGM indica el número de veces que se incrementan las ocho clases menos pobladas, mientras que Base y Todo especifican si el modelo de entrenamiento es el original o si el aumento se aplica a todas las clases, respectivamente. Las mejores puntuaciones de cada modelo de aprendizaje se resaltan en negritas. Obtenida de [80]

Modelo	AGM	Exactitud	Exactitud Balanceada	F_1^w	CCM
MVS + Hashtags	**Base**	**0.4820**	**0.4669**	**0.4788**	**0.4262**
	1	0.4781	0.4471	0.4702	0.4200
	2	0.4781	0.4258	0.4652	0.4185
	3	0.4859	0.4395	0.4729	0.4272
	4	0.4833	0.4290	0.4701	0.4239
	Todos	0.4692	0.4083	0.4068	0.4523
DistilBERT Base Cased	Base	0.5141	0.4048	0.5090	0.4610
	1	0.4936	0.4540	0.4873	0.4379
	2	0.4961	0.4961	0.4994	0.4438
	3	0.4807	0.5053	0.4826	0.4260
	4	0.5129	0.4497	0.5021	0.4611
	Todos	**0.5129**	**0.4280**	**0.5094**	**0.4609**
BERT Base Cased	Base	0.5192	0.4156	0.5145	0.4683
	1	0.5013	0.4738	0.5008	0.4515
	2	0.5141	0.5298	0.5097	0.4649
	3	**0.5129**	**0.5350**	**0.5156**	**0.4633**
	4	0.5103	0.4484	0.5022	0.4581
	Todos	0.5116	0.4218	0.5047	0.4600
Twitter roBERTa Base	**Base**	**0.5296**	**0.4573**	**0.5292**	**0.4832**
	1	0.5154	0.4173	0.5115	0.4627
	2	0.5103	0.4723	0.5073	0.4586
	3	0.5231	0.4307	0.5214	0.4722
	4	0.5116	0.4823	0.5099	0.4605
	All	0.4884	0.5062	0.4908	0.4348

TABLA D.3: Comparación del rendimiento de clase de los mejores modelos para cada modelo de aprendizaje, extraído de las matrices de confusión normalizadas. La mejor puntuación de cada tipología se resalta en negritas. Obtenida de [80]

Tipología	Mejor Modelo Clásico	Mejor Modelo Ajustado
Amigo	0.59	**0.61**
Cosmopolita	0.44	**0.53**
Moderno	0.18	**0.55**
Emergente	**0.54**	0.46
Aliado	0.51	**0.74**
Cooperante	**0.36**	0.21
Exótico	0.45	**0.52**
Degenerado	**0.42**	0.38
Neutral	0.21	**0.26**
Independiente	0.21	**0.37**
Intervencionista	**1**	0
Colonizador	**1**	0
Rival	0.73	**0.77**
Adversario Menor	0.18	**0.55**
Enemigo	0.17	**0.67**
Bárbaro	0.47	**0.53**
Dependiente	0.41	**0.44**
Marginal	0.54	**0.64**

Apéndice E

Productos Desarrollados y Participaciones Durante el Posgrado

E.1. Publicaciones en Revistas JCR

1. **L. N. Zúñiga-Morales**, J. Á. González-Ordiano, J. E. Quiroz-Ibarra, and C. Villanueva Rivas, "Machine learning framework for country image analysis", Journal of Computational Social Science, vol. 7, no. 1, pp. 523–547, 2024, doi: 10.1007/s42001-023-00246-3.

2. L. Bustio-Martínez et al., "Uncovering phishing attacks using principles of persuasion analysis", Journal of Network and Computer Applications, p. 103964, 2024, doi: 10.1016/j.jnca.2024.103964.

E.2. Memorias Completas en Congresos Internacionales

1. L. Bustio-Martínez et al., "Towards Automatic Principles of Persuasion Detection Using Machine Learning Approach", in Progress in Artificial Intelligence and Pattern Recognition, V. and R. S. J. Hernández Heredia Yanio and Milián Núñez, Ed., Cham: Springer Nature Switzerland, 2024, pp. 155–166.

2. **L. N. Zúñiga-Morales**, J. Á. González-Ordiano, J. E. Quiroz-Ibarra, and S. J. Simske, "Impact Evaluation of Multimodal Information on Sentiment Analysis", in Advances in Computational Intelligence, O. Pichardo Lagunas, J. Martínez-Miranda, and B. Martínez Seis, Eds., Cham: Springer Nature Switzerland, 2022, pp. 18–29. doi: 10.1007/978-3-031-19496-2_2

E.3. Talleres

1. **Inteligencia artificial para Relaciones Internacionales con un modelo de clasificación automatizada.** Conmemoración de los 40 años del programa de Licenciatura en Relaciones Internacionales y 25 años de la fundación del DEI. Universidad Iberoamericana Ciudad de México. Octubre 2023.

2. **Taller de recopilación y clasificación de datos para el modelo Analítico de Imagen País.** Construyendo el futuro de la inteligencia artificial. Universidad Iberoamericana Ciudad de México. Agosto 2023.

E.4. Pósteres

1. **Análisis de redes sociales usando información multimodal.** Seminario de Avances de Investigación. Universidad Iberoamericana Ciudad de México. Noviembre 2024.

2. **Análisis de redes sociales usando información multimodal.** Seminario de Avances de Investigación. Universidad Iberoamericana Ciudad de México. Diciembre 2023.

3. **Análisis de redes sociales usando información multimodal.** Seminario de Avances de Investigación. Universidad Iberoamericana Ciudad de México. Diciembre 2022

Bibliografía

[1] «Social media usage in Mexico,» Statista, inf. téc., 2023.

[2] S. Bennett, A. Bishop, B. Dalgarno, J. Waycott y G. Kennedy, «Implementing Web 2.0 technologies in higher education: A collective case study,» *Computers & Education*, vol. 59, págs. 524-534, 2 2012, ISSN: 0360-1315. DOI: https://doi.org/10.1016/j.compedu.2011.12.022. dirección: https://www.sciencedirect.com/science/article/pii/S0360131511003381.

[3] N. Jurgenson, *The Social Photo: On Photography and Social Media*. Verso, 2020.

[4] M. A. Nwala e I. Tamunobelema, «The Social Media and Language Use: The Case of Facebook,» *Advances in Language and Literary Studies*, vol. 10, n.º 4, 2019, ISSN: 2203-4714.

[5] U. Russmann y J. Svensson, «Introduction to Visual Communication in the Age of Social Media: Conceptual, Theoretical and Methodological Challenges,» *Media and Communication*, vol. 5, n.º 4, págs. 1-5, 2017, ISSN: 2183-2439. DOI: 10.17645/mac.v5i4.1263. dirección: https://www.cogitatiopress.com/mediaandcommunication/article/view/1263.

[6] Q. Zhang, R. Qing-Dao-Er-Ji y N. Li, «Research on Animated GIFs Emotion Recognition Based on ResNet-ConvGRU,» *Mathematical Problems in Engineering*, vol. 2022, F. Lolli, ed., pág. 3 143 748, 2022, ISSN: 1024-123X. DOI: 10.1155/2022/3143748. dirección: https://doi.org/10.1155/2022/3143748.

[7] L. Chen, «Exploring the Impact of Short Videos on Society and Culture: An Analysis of Social Dynamics and Cultural Expression,» *Pacific International Journal*, vol. 6, págs. 115-118, 3 sep. de 2023. DOI: 10.55014/pij.v6i3.420. dirección: https://rclss.com/pij/article/view/420.

[8] F. A. Pozzi, E. Fersini, E. Messina y B. Liu, «Chapter 1 - Challenges of Sentiment Analysis in Social Networks: An Overview,» en F. A. Pozzi, E. Fersini, E. Messina y B. Liu, eds. Morgan Kaufmann, 2017, págs. 1-11, ISBN: 978-0-12-804412-4. DOI: https://doi.org/10.1016/B978-0-12-804412-4.00001-2. dirección: https://www.sciencedirect.com/science/article/pii/B9780128044124000012.

[9] S. A. Abdu, A. H. Yousef y A. Salem, «Multimodal video sentiment analysis using deep learning approaches, a survey,» *Information Fusion*, vol. 76, págs. 204-226, 2021.

[10] D. Felmlee, P. I. Rodis y A. Zhang, «Sexist Slurs: Reinforcing Feminine Stereotypes Online,» *Sex Roles*, vol. 83, págs. 16-28, 1 2020, ISSN: 1573-2762. DOI: 10.1007/s11199-019-01095-z. dirección: https://doi.org/10.1007/s11199-019-01095-z.

[11] F. Olan, U. Jayawickrama, E. O. Arakpogun, J. Suklan y S. Liu, «Fake news on Social Media: the Impact on Society,» *Information Systems Frontiers*, 2022, ISSN: 1572-9419. DOI: 10.1007/s10796-022-10242-z. dirección: https://doi.org/10.1007/s10796-022-10242-z.

[12] H. A. M. Voorveld, G. van Noort, D. G. Muntinga y F. Bronner, «Engagement with Social Media and Social Media Advertising: The Differentiating Role of Platform Type,» *Journal of Advertising*, vol. 47, págs. 38-54, 1 2018. DOI: 10.1080/00913367.2017.1405754. dirección: https://doi.org/10.1080/00913367.2017.1405754.

[13] Y. K. Lal, V. Kumar, M. Dhar, M. Shrivastava y P. Koehn, «De-Mixing Sentiment from Code-Mixed Text,» F. Alva-Manchego, E. Choi y D. Khashabi, eds., Association for Computational Linguistics, jul. de 2019, págs. 371-377. DOI: 10.18653/v1/P19-2052. dirección: https://aclanthology.org/P19-2052.

[14] İ. Yurtseven, S. Bagriyanik y S. Ayvaz, «A Review of Spam Detection in Social Media,» 2021, págs. 383-388. DOI: 10.1109/UBMK52708.2021.9558993.

[15] L. Ren, B. Xu, H. Lin, X. Liu y L. Yang, «Sarcasm Detection with Sentiment Semantics Enhanced Multi-level Memory Network,» *Neurocomputing*, vol. 401, págs. 320-326, 2020, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.03.081. dirección: https://www.sciencedirect.com/science/article/pii/S0925231220304689.

[16] P. K. Singh y S. Paul, «Deep Learning Approach for Negation Handling in Sentiment Analysis,» *IEEE Access*, vol. 9, págs. 102 579-102 592, 2021. DOI: 10.1109/ACCESS.2021.3095412.

[17] M. Bevilacqua, T. Pasini, A. Raganato y R. Navigli, «Recent Trends in Word Sense Disambiguation: A Survey,» Z.-H. Zhou, ed., Survey Track, International Joint Conferences on Artificial Intelligence Organization, jun. de 2021, págs. 4330-4338. DOI: 10.24963/ijcai.2021/593. dirección: https://doi.org/10.24963/ijcai.2021/593.

[18] Y. Chen, K. Sherren, M. Smit y K. Y. Lee, «Using social media images as data in social science research,» *New Media & Society*, vol. 25, págs. 849-871, 4 2023. DOI: 10.1177/14614448211038761. dirección: https://doi.org/10.1177/14614448211038761.

[19] M. Banks, *Using Visual Data in Qualitative Research*. 2007. DOI: 10.4135/9780857020260. dirección: https://methods.sagepub.com/book/using-visual-data-in-qualitative-research.

[20] G. Chandrasekaran, N. Antoanela, G. Andrei, C. Monica y J. Hemanth, «Visual Sentiment Analysis Using Deep Learning Models with Social Media Data,» *Applied Sciences*, vol. 12, 3 2022, ISSN: 2076-3417. DOI: 10.3390/app12031030. dirección: https://www.mdpi.com/2076-3417/12/3/1030.

[21] J. Chen, Q. Mao y L. Xue, «Visual Sentiment Analysis With Active Learning,» *IEEE Access*, vol. 8, págs. 185 899-185 908, 2020. DOI: 10.1109/ACCESS.2020.3024948.

[22] N. Desai, S. Venkatramana y B. Sekhar, «Automatic Visual Sentiment Analysis with Convolution Neural network,» *international journal of industrial Engineering & Production Research*, vol. 31, 3 2020. DOI: 10.22068/ijiepr.31.3.351. dirección: http://ijiepr.iust.ac.ir/article-1-1070-en.html.

[23] Y. Wang y B. Li, «Sentiment Analysis for Social Media Images,» 2015, págs. 1584-1591. DOI: 10.1109/ICDMW.2015.142.

[24] S. S. Rajagopalan, L.-P. Morency, T. Baltrusăaitis y R. Goecke, «Extending Long Short-Term Memory for Multi-View Structured Learning,» en *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe y M. Welling, eds., Cham: Springer International Publishing, 2016, págs. 338-353, ISBN: 978-3-319-46478-7.

[25] B. Liu et al., «Context-aware social media user sentiment analysis,» *Tsinghua Science and Technology*, vol. 25, n.º 4, págs. 528-541, 2020.

[26] G. Meena, K. K. Mohbey, A. Indian, M. Z. Khan y S. Kumar, «Identifying emotions from facial expressions using a deep convolutional neural network-based approach,» *Multimedia Tools and Applications*, vol. 83, págs. 15 711-15 732, 6 2024, ISSN: 1573-7721. DOI: 10.1007/s11042-023-16174-3. dirección: https://doi.org/10.1007/s11042-023-16174-3.

[27] M. de Meijer, «The contribution of general features of body movement to the attribution of emotions,» *Journal of Nonverbal Behavior*, vol. 13, págs. 247-268, 4 1989, ISSN: 1573-3653. DOI: 10.1007/BF00990296. dirección: https://doi.org/10.1007/BF00990296.

[28] S. Z. Hassan et al., «Visual Sentiment Analysis from Disaster Images in Social Media,» *Sensors*, vol. 22, 10 mayo de 2022, ISSN: 14248220. DOI: 10.3390/s22103628.

[29] S. Liang, D. Wu y C. Zhang, «Enhancing image sentiment analysis: A user-centered approach through user emotions and visual features,» *Information Processing & Management*, vol. 61, pág. 103 749, 4 2024, ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2024.103749. dirección: https://www.sciencedirect.com/science/article/pii/S0306457324001092.

[30] A. Vaswani et al., «Attention Is All You Need,» jun. de 2017.

[31] A. Dosovitskiy et al., «An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,» oct. de 2020.

[32] X. Wang, J. Yang, M. Hu y F. Ren, «EERCA-ViT: Enhanced Effective Region and Context-Aware Vision Transformers for image sentiment analysis,» *Journal of Visual Communication and Image Representation*, vol. 97, pág. 103 968, 2023, ISSN: 1047-3203. DOI: https://doi.org/10.1016/j.jvcir.2023.103968. dirección: https://www.sciencedirect.com/science/article/pii/S1047320323002183.

[33] P. Bharti, V. Sagar y B. Wadhwa, «An Analysis on Sentiments Using Deep Learning Approaches,» en *2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, 2022, págs. 355-360. DOI: 10.1109/CISES54857.2022.9844318.

[34] Y. Dong, Y. Fu, L. Wang, Y. Chen, Y. Dong y J. Li, «A Sentiment Analysis Method of Capsule Network Based on BiLSTM,» *IEEE Access*, vol. 8, págs. 37 014-37 020, 2020. DOI: 10.1109/ACCESS.2020.2973711.

[35] B. Chen, Q. Huang, Y. Chen, L. Cheng y R. Chen, «Deep Neural Networks for Multi-class Sentiment Classification,» en *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2018, págs. 854-859. DOI: 10.1109/HPCC/SmartCity/DSS.2018.00142.

[36] K. Alahmadi, S. Alharbi, J. Chen y X. Wang, «Generalizing sentiment analysis: a review of progress, challenges, and emerging directions,» *Social Network Analysis and Mining*, vol. 15, pág. 45, 1 2025, ISSN: 1869-5469. DOI: 10.1007/s13278-025-01461-8. dirección: https://doi.org/10.1007/s13278-025-01461-8.

[37] X. Zhu, Y. Chen, Y. Gu y Z. Xiao, «SentiMedQAer: A Transfer Learning-Based Sentiment-Aware Model for Biomedical Question Answering,» *Frontiers in Neurorobotics*, vol. Volume 16 - 2022, 2022, ISSN: 1662-5218. DOI: 10.3389/fnbot.2022.773329. dirección: https://www.frontiersin.org/journals/neurorobotics/articles/10.3389/fnbot.2022.773329.

[38] N. J. Prottasha et al., «Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning,» *Sensors*, vol. 22, 11 2022, ISSN: 1424-8220. DOI: 10.3390/s22114157. dirección: https://www.mdpi.com/1424-8220/22/11/4157.

[39] S. Rastogi, «Weak Supervision and Transformed-based Sentiment Analysis on Multi-lingual Data,» en *2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, 2023, págs. 706-712. DOI: 10.1109/COMSNETS56262.2023.10041286.

[40] R. Safa, P. Bayat y L. Moghtader, «Automatic detection of depression symptoms in twitter using multimodal analysis,» *The Journal of Supercomputing*, vol. 78, págs. 4709-4744, 4 2022, ISSN: 1573-0484. DOI: 10.1007/s11227-021-04040-8. dirección: https://doi.org/10.1007/s11227-021-04040-8.

[41] A. Kumar y G. Garg, «Sentiment analysis of multimodal twitter data,» *Multimedia Tools and Applications*, vol. 78, n.º 17, págs. 24 103-24 119, sep. de 2019, ISSN: 1573-7721. DOI: 10.1007/s11042-019-7390-1. dirección: https://doi.org/10.1007/s11042-019-7390-1.

[42] K. Zhang, Y. Geng, J.-X. Zhao, W. Li y J. Liu, «Multimodal Sentiment Analysis Based on Attention Mechanism and Tensor Fusion Network,» *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2021. DOI: 10.1109/SMC52423.2021.9658940.

[43] D. Dimitrov et al., «Detecting Propaganda Techniques in Memes,» C. Zong, F. Xia, W. Li y R. Navigli, eds., Association for Computational Linguistics, ago. de 2021, págs. 6603-6617. DOI: 10.18653/v1/2021.acl-long.516. dirección: https://aclanthology.org/2021.acl-long.516.

[44] S. Zhang, Y. He, L. Li e Y. Dou, «Multimodal sentiment analysis with BERT-ResNet50,» K. Subramaniam y A. P. Muthuramalingam, eds., vol. 12635, SPIE, 2023, pág. 1 263 510. DOI: 10.1117/12.2679113. dirección: https://doi.org/10.1117/12.2679113.

[45] A. Anshul, G. S. Pranav, M. Z. U. Rehman y N. Kumar, «A Multimodal Framework for Depression Detection During COVID-19 via Harvesting Social Media,» *IEEE Transactions on Computational Social Systems*, págs. 1-17, 2023. DOI: 10.1109/TCSS.2023.3309229.

[46] C. Zhu et al., «SKEAFN: Sentiment Knowledge Enhanced Attention Fusion Network for multimodal sentiment analysis,» *Information Fusion*, vol. 100, pág. 101 958, 2023, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2023.101958. dirección: https://www.sciencedirect.com/science/article/pii/S1566253523002749.

[47] W. Zhong, Z. Zhang, Q. Wu, Y. Xue y Q. Cai, «A Semantic Enhancement Framework for Multimodal Sarcasm Detection,» *Mathematics*, vol. 12, 2 2024, ISSN: 2227-7390. DOI: 10.3390/math12020317. dirección: https://www.mdpi.com/2227-7390/12/2/317.

[48] R. Rivas, S. Paul, V. Hristidis, E. E. Papalexakis y A. K. Roy-Chowdhury, «Task-agnostic representation learning of multimodal twitter data for downstream applications,» *Journal of Big Data*, vol. 9, 18 2022. DOI: https://doi.org/10.1186/s40537-022-00570-x.

[49] X. Lu, Y. Ni y Z. Ding, «Cross-Modal Sentiment Analysis Based on CLIP Image-Text Attention Interaction,» *International Journal of Advanced Computer Science and Applications*, vol. 15, 2 2024. DOI: 10.14569/IJACSA.2024.0150290. dirección: http://dx.doi.org/10.14569/IJACSA.2024.0150290.

[50] J. Chen, J. An, H. Lyu, C. Kanan y J. Luo, «Holistic Visual-Textual Sentiment Analysis with Prior Models,» *arXiv preprint arXiv:2211.12981*, jun. de 2024.

[51] J. An y W. M. N. W. Zainon, «Integrating color cues to improve multimodal sentiment analysis in social media,» *Engineering Applications of Artificial Intelligence*, vol. 126, pág. 106 874, 2023, ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2023.106874. dirección: https://www.sciencedirect.com/science/article/pii/S0952197623010588.

[52] M. Álvarez-Carmona, R. Aranda, R. Guerrero-Rodríguez, A. Rodríguez-González y A. López-Monroy, «A Combination of Sentiment Analysis Systems for the Study of Online Travel Reviews: Many Heads are Better than One,» *Computación y Sistemas*, vol. 26, n.º 2, págs. 977-987, 30 de jun. de 2022. DOI: 10.13053/cys-26-2-4055.

[53] M. Álvarez-Carmona, Á. Díaz-Pacheco, R. Aranda, A. Rodríguez-González, L. Bustio-Martínez y V. Herrera-Semenets, *Overview of REST-MEX at IberLEF 2025: Researching Sentiment Evaluation in Text for Mexican Magical Towns*, 2025.

[54] J. Monsalve-Pulido, C. A. Parra y J. Aguilar, «Multimodal model for the Spanish sentiment analysis in a tourism domain,» English, *Social Network Analysis and Mining*, vol. 14, n.º 1, dic. de 2024, Publisher Copyright: © The Author(s), under exclusive licence to Springer-Verlag GmbH Austria, part of Springer Nature 2024., ISSN: 1869-5450. DOI: 10.1007/s13278-024-01202-3.

[55] V. P. Rosas, R. Mihalcea y L.-P. Morency, «Multimodal Sentiment Analysis of Spanish Online Videos,» *IEEE Intelligent Systems*, vol. 28, n.º 3, págs. 38-45, 2013. DOI: 10.1109/MIS.2013.9.

[56] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh y M. Asadpour, «A comprehensive review of visual–textual sentiment analysis from social media networks,» *Journal of Computational Social Science*, 2024, ISSN: 2432-2725. DOI: 10.1007/s42001-024-00326-y. dirección: https://doi.org/10.1007/s42001-024-00326-y.

[57] D. Borth, R. Ji, T. Chen, T. Breuel y S.-F. Chang, «Large-scale visual sentiment ontology and detectors using adjective noun pairs,» en *Proceedings of the 21st ACM International Conference on Multimedia*, Association for Computing Machinery, 2013, págs. 223-232, ISBN: 9781450324045. DOI: 10.1145/2502081.2502282. dirección: https://doi.org/10.1145/2502081.2502282.

[58] Shiai, P. Lei, E. S. A. N. Teng y Zhu, «Sentiment Analysis on Multi-View Social Data,» Nicu, Q. Guo-Jun, H. Benoit, H. Richang, L. X. T. Qi y Sebe, eds., Springer International Publishing, 2016, págs. 15-27, ISBN: 978-3-319-27674-8.

[59] L. Vadicamo et al., «Cross-Media Learning for Image Sentiment Analysis in the Wild,» en *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, págs. 308-317. DOI: 10.1109/ICCVW.2017.45.

[60] J. Cañete, G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang y J. Pérez, «Spanish Pre-Trained BERT Model and Evaluation Data,» en *PML4DC at ICLR 2020*, 2020.

[61] L. N. Zúñiga-Morales, J. Á. González-Ordiano, J. Quiroz-Ibarra y S. J. Simske, «Impact Evaluation of Multimodal Information on Sentiment Analysis,» en *Advances in Computational Intelligence*, O. Pichardo Lagunas, J. Martínez-Miranda y B. Martínez Seis, eds., Cham: Springer Nature Switzerland, 2022, págs. 18-29, ISBN: 978-3-031-19496-2.

[62] V. Singh y S. K. Dubey, «Opinion mining and analysis: A literature review,» en *2014 5th International Conference - Confluence The Next Generation Information Technology Summit (Confluence)*, 2014, págs. 232-239. DOI: 10.1109/CONFLUENCE.2014.6949318.

[63] B. Liu, *Sentiment Analysis and Opinion Mining*. Springer International Publishing, 2012, ISBN: 978-3-031-01017-0. DOI: 10.1007/978-3-031-02145-9.

[64] R. B. Catell, «Sentiment Or Attitude? The Core Of A Terminology Problem In Personality Research,» *Journal of Personality*, vol. 9, págs. 6-17, 1 1940. DOI: https://doi.org/10.1111/j.1467-6494.1940.tb02192.x. dirección: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-6494.1940.tb02192.x.

[65] C. D. Broad, «Emotion and sentiment,» *Journal of Aesthetics and Art Criticism*, vol. 13, 2 1954.

[66] L. Tian, C. Lai y J. D. Moore, «Polarity and Intensity: the Two Aspects of Sentiment Analysis,» jul. de 2018. arXiv: 1807.01466 [cs.CL]. dirección: https://arxiv.org/abs/1807.01466.

[67] Devamanyu, P. Soujanya, H. Amir, S. R. B. V. C. Erik y Hazarika, «Benchmarking Multimodal Sentiment Analysis,» A. Gelbukh, ed., Springer International Publishing, 2018, págs. 166-179, ISBN: 978-3-319-77116-8.

[68] P. Ekman, «Are there basic emotions?» *Psychological review*, vol. 99, págs. 550-3, 3 jul. de 1992, ISSN: 0033-295X. DOI: 10.1037/0033-295x.99.3.550.

[69] J. A. Russell y A. Mehrabian, «Evidence for a three-factor theory of emotions,» *Journal of Research in Personality*, vol. 11, págs. 273-294, 3 1977, ISSN: 0092-6566. DOI: https://doi.org/10.1016/0092-6566(77)90037-X. dirección: https://www.sciencedirect.com/science/article/pii/009265667790037X.

[70] P. Nandwani y R. Verma, «A review on sentiment analysis and emotion detection from text,» *Social Network Analysis and Mining*, vol. 11, pág. 81, 1 2021, ISSN: 1869-5469. DOI: 10.1007/s13278-021-00776-6. dirección: https://doi.org/10.1007/s13278-021-00776-6.

[71] M. Munezero, C. S. Montero, E. Sutinen y J. Pajunen, «Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text,» *IEEE Transactions on Affective Computing*, vol. 5, págs. 101-111, 2 2014. DOI: 10.1109/TAFFC.2014.2317187.

[72] Y. Zhao, J. Ma y T. W. S. Chow, «Extractive Negative Opinion Summarization of Consumer Electronics Reviews,» *IEEE Transactions on Consumer Electronics*, vol. 70, págs. 3521-3528, 1 2024. DOI: 10.1109/TCE.2023.3302851.

[73] E. Cambria, S. Poria, A. Gelbukh y M. Thelwall, «Sentiment Analysis Is a Big Suitcase,» *IEEE Intelligent Systems*, vol. 32, págs. 74-80, 6 2017, ISSN: 1941-1294. DOI: 10.1109/MIS.2017.4531228.

[74] S. Kannangara, «Mining Twitter for Fine-Grained Political Opinion Polarity Classification, Ideology Detection and Sarcasm Detection,» en *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, Association for Computing Machinery, 2018, págs. 751-752, ISBN: 9781450355810. DOI: 10.1145/3159652.3170461. dirección: https://doi.org/10.1145/3159652.3170461.

[75] R. Singh, M. Bansal, S. Gupta, A. Singh, G. Bhardwaj y A. D. Dhariwal, «Detection of Social Network Spam Based on Improved Machine Learning,» en *2022 5th International Conference on Contemporary Computing and Informatics (IC3I)*, 2022, págs. 2257-2261. DOI: 10.1109/IC3I56241.2022.10073448.

[76] S. Rani y M. Kumar, «Topic modeling and its applications in materials science and engineering,» *Materials Today: Proceedings*, vol. 45, págs. 5591-5596, 2021, Second International Conference on Aspects of Materials Science and Engineering (ICAMSE 2021), ISSN: 2214-7853. DOI: https://doi.org/10.1016/j.matpr.2021.02.313. dirección: https://www.sciencedirect.com/science/article/pii/S2214785321014127.

[77] K. E. N. Kumar y V. Uma, «Intelligent sentinet-based lexicon for context-aware sentiment analysis: optimized neural network for sentiment classification on social media,» *The Journal of Supercomputing*, vol. 77, págs. 12 801-12 825, 11 2021, ISSN: 1573-0484. DOI: 10.1007/s11227-021-03709-4. dirección: https://doi.org/10.1007/s11227-021-03709-4.

[78] A. Pinto, H. Gonçalo Oliveira y A. Oliveira Alves, «Comparing the Performance of Different NLP Toolkits in Formal and Social Media Text,» en *5th Symposium on Languages, Applications and Technologies (SLATE'16)*, M. Mernik, J. P. Leal y H. Gonçalo Oliveira, eds., ép. Open Access Series in Informatics (OASIcs), vol. 51, Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2016, 3:1-3:16, ISBN: 978-3-95977-006-4. DOI: 10.4230/OASIcs.SLATE.2016.3. dirección: https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SLATE.2016.3.

[79] N. Oliveira, P. Cortez y N. Areal, «Stock market sentiment lexicon acquisition using microblogging data and statistical measures,» *Decision Support Systems*, vol. 85, págs. 62-73, 2016.

[80] L. N. Zúñiga-Morales, J. Á. González-Ordiano, J. E. Quiroz-Ibarra y C. Villanueva Rivas, «Machine learning framework for country image analysis,» *Journal of Computational Social Science*, 2024, ISSN: 2432-2725. DOI: 10.1007/s42001-023-00246-3. dirección: https://doi.org/10.1007/s42001-023-00246-3.

[81] M. Siino, I. Tinnirello y M. L. Cascia, «Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers,» *Information Systems*, vol. 121, pág. 102 342, 2024, ISSN: 0306-4379. DOI: https://doi.org/10.1016/j.is.2023.102342. dirección: https://www.sciencedirect.com/science/article/pii/S0306437923001783.

[82] M. Birjali, M. Kasri y A. Beni-Hssane, «A comprehensive survey on sentiment analysis: Approaches, challenges and trends,» *Knowledge-Based Systems*, vol. 226, pág. 107 134, 2021, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2021.107134. dirección: https://www.sciencedirect.com/science/article/pii/S095070512100397X.

[83] S. Akuma, T. Lubem e I. T. Adom, «Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets,» *International Journal of Information Technology*, vol. 14, págs. 3629-3635, 7 2022, ISSN: 2511-2112. DOI: 10.1007/s41870-022-01096-4. dirección: https://doi.org/10.1007/s41870-022-01096-4.

[84] T. Mikolov, K. Chen, G. Corrado y J. Dean, «Efficient Estimation of Word Representations in Vector Space,» ene. de 2013.

[85] J. Pennington, R. Socher y C. Manning, «GloVe: Global Vectors for Word Representation,» en *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang y W. Daelemans, eds., Doha, Qatar: Association for Computational Linguistics, oct. de 2014, págs. 1532-1543. DOI: 10.3115/v1/D14-1162. dirección: https://aclanthology.org/D14-1162.

[86] A. Joulin, E. Grave, P. Bojanowski y T. Mikolov, «Bag of Tricks for Efficient Text Classification,» en *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Association for Computational Linguistics, abr. de 2017, págs. 427-431.

[87] J. Devlin, M.-W. Chang, K. Lee y K. Toutanova, «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,» oct. de 2018.

[88] A. Goel, J. Gautam y S. Kumar, «Real time sentiment analysis of tweets using Naive Bayes,» en *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*, 2016, págs. 257-261. DOI: 10.1109/NGCT.2016.7877424.

[89] V. Kumar y B. Subba, «A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus,» en *2020 National Conference on Communications (NCC)*, 2020, págs. 1-6. DOI: 10.1109/NCC48643.2020.9056085.

[90] Radhia, A. J. C. Yasmine y Toujani, «Sentiment Analysis Method for Tracking Touristics Reviews in Social Media Network,» en *Intelligent Interactive Multimedia Systems and Services 2017*, Luigi, H. R. J., J. L. C. D. P. Giuseppe y Gallo, eds., Springer International Publishing, 2018, págs. 299-310, ISBN: 978-3-319-59480-4.

[91] Jitendra, P. D. Kumar, P. S. P. S. Shanker y Kumar, «Sentiment Classification: An Approach for Indian Language Tweets Using Decision Tree,» en *Mining Intelligence and Knowledge Exploration*, A. Kumar, K. T. P. Rajendra y Vuppala, eds., Springer International Publishing, 2015, págs. 656-663, ISBN: 978-3-319-26832-3.

[92] A. R. Lubis, S. Prayudani, M. Lubis y O. Nugroho, «Sentiment Analysis on Online Learning During the Covid-19 Pandemic Based on Opinions on Twitter using KNN Method,» en *2022 1st International Conference on Information System & Information Technology (ICISIT)*, 2022, págs. 106-111. DOI: 10.1109/ICISIT54091.2022.9872926.

[93] Anikó, B. C. D., F. D. R. B. J. J. y Ekárt, «High Resolution Sentiment Analysis by Ensemble Classification,» en *Intelligent Computing*, Rahul, K. S. A. Kohei y Bhatia, eds., Springer International Publishing, 2019, págs. 593-606, ISBN: 978-3-030-22871-2.

[94] S. Neelakandan y D. Paulraj, «A gradient boosted decision tree-based sentiment classification of twitter data,» *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 18, pág. 2 050 027, 04 2020. DOI: 10.1142/S0219691320500277. dirección: https://doi.org/10.1142/S0219691320500277.

[95] N. C. Dang, M. N. Moreno-García y F. la Prieta, «Sentiment Analysis Based on Deep Learning: A Comparative Study,» *Electronics*, vol. 9, 3 2020, ISSN: 2079-9292. DOI: 10.3390/electronics9030483. dirección: https://www.mdpi.com/2079-9292/9/3/483.

[96] C.-R. Ko y H.-T. Chang, «LSTM-based sentiment analysis for stock price forecast,» *PeerJ Computer Science*, vol. 7, e408, mar. de 2021, ISSN: 2376-5992. DOI: 10.7717/peerj-cs.408. dirección: https://doi.org/10.7717/peerj-cs.408.

[97] Olarik, L. S. C. K. Sanya y Surinta, «Sentiment Analysis of Local Tourism in Thailand from YouTube Comments Using BiLSTM,» en *Multi-disciplinary Trends in Artificial Intelligence*, K. S. Olarik y K. F. Yuen, eds., Springer International Publishing, 2022, págs. 169-177, ISBN: 978-3-031-20992-5.

[98] M. Mhamed, R. Sutcliffe, X. Sun, J. Feng, E. Almekhlafi y E. A. Retta, «Improving Arabic Sentiment Analysis Using CNN-Based Architectures and Text Preprocessing,» *Computational Intelligence and Neuroscience*, vol. 2021, pág. 5 538 791, 1 2021. DOI: https://doi.org/10.1155/2021/5538791. dirección: https://onlinelibrary.wiley.com/doi/abs/10.1155/2021/5538791.

[99] C. S. G. Khoo y S. B. Johnkhan, «Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons,» *Journal of Information Science*, vol. 44, págs. 491-511, 4 2018. DOI: 10.1177/0165551517703514. dirección: https://doi.org/10.1177/0165551517703514.

[100] M. Taboada, J. Brooke, M. Tofiloski, K. Voll y M. Stede, «Lexicon-Based Methods for Sentiment Analysis,» *Computational Linguistics*, vol. 37, págs. 267-307, 2 jun. de 2011, ISSN: 0891-2017. DOI: 10.1162/COLI_a_00049. dirección: https://doi.org/10.1162/COLI_a_00049.

[101] S. Wu, F. Wu, Y. Chang, C. Wu e Y. Huang, «Automatic construction of target-specific sentiment lexicon,» *Expert Systems with Applications*, vol. 116, págs. 285-298, 2019, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2018.09.024. dirección: https://www.sciencedirect.com/science/article/pii/S0957417418306018.

[102] G. A. Miller, «WordNet: A Lexical Database for English,» en *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. dirección: https://aclanthology.org/H94-1111.

[103] «A Hybrid Model for Social Media Sentiment Analysis for Indonesian Text,» en *Proceedings of the 20th International Conference on Information Integration and Web-Based Applications & Services*, Association for Computing Machinery, 2018, págs. 297-301, ISBN: 9781450364799. DOI: 10.1145/3282373.3282850. dirección: https://doi.org/10.1145/3282373.3282850.

[104] B. Shin, T. Lee y J. D. Choi, «Lexicon Integrated CNN Models with Attention for Sentiment Analysis,» en *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, A. Balahur, S. M. Mohammad y E. der Goot, eds., Association for Computational Linguistics, sep. de 2017, págs. 149-158. DOI: 10.18653/v1/W17-5220. dirección: https://aclanthology.org/W17-5220.

[105] A. Ortis, G. M. Farinella y S. Battiato, «Survey on visual sentiment analysis,» *IET Image Processing*, vol. 14, págs. 1440-1456, 8 jun. de 2020, ISSN: 1751-9659. DOI: https://doi.org/10.1049/iet-ipr.2019.1270. dirección: https://doi.org/10.1049/iet-ipr.2019.1270.

[106] Y.-Y. Chen, T. Chen, W. H. Hsu, H.-Y. M. Liao y S.-F. Chang, «Predicting Viewer Affective Comments Based on Image Content in Social Media,» en *Proceedings of International Conference on Multimedia Retrieval*, Association for Computing Machinery, 2014, págs. 233-240, ISBN: 9781450327824. DOI: 10.1145/2578726.2578756. dirección: https://doi.org/10.1145/2578726.2578756.

[107] G. Chandrasekaran, T. N. Nguyen y J. H. D., «Multimodal sentimental analysis for social media applications: A comprehensive review,» *WIREs Data Mining and Knowledge Discovery*, vol. 11, e1415, 5 sep. de 2021, https://doi.org/10.1002/widm.1415, ISSN: 1942-4787. DOI: https://doi.org/10.1002/widm.1415. dirección: https://doi.org/10.1002/widm.1415.

[108] M. Wöllmer et al., «Youtube movie reviews: sentiment analysis in an audio-visual context,» *IEEE Intelligent Systems*, vol. 28, n.º 3, págs. 46-53, 2013. DOI: 10.1109/MIS.2013.34.

[109] S. Al-Azani y E.-S. M. El-Alfy, «Enhanced Video Analytics for Sentiment Analysis Based on Fusing Textual, Auditory and Visual Information,» *IEEE Access*, vol. 8, págs. 136 843-136 857, 2020. DOI: 10.1109/ACCESS.2020.3011977.

[110] L. Tunstall, L. von Werra y T. Wolf, *Natural Language Processing with Transformers, Revised Edition*, 1st. USA: O'Reilly Media, Incorporated, 2022, ISBN: 9781098136796.

[111] Y. Liu et al., «RoBERTa: A Robustly Optimized BERT Pretraining Approach,» jul. de 2019.

[112] A. Radford, K. Narasimhan, T. Salimans e I. Sutskever, *Improving Language Understanding by Generative Pre-Training*, jun. de 2018.

[113] M. Lewis et al., «BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,» oct. de 2019.

[114] C. Raffel et al., «Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,» oct. de 2019.

[115] D. Foster, *Generative Deep Learning*, Second Edition. O'Reilly Media, Inc., 2023.

[116] J. L. Ba, J. R. Kiros y G. E. Hinton, «Layer Normalization,» jul. de 2016.

[117] S. Hochreiter y J. Schmidhuber, «Long Short-term Memory,» *Neural computation*, vol. 9, págs. 1735-1780, jun. de 1997. DOI: 10.1162/neco.1997.9.8.1735.

[118] P. Shaw, J. Uszkoreit y A. Vaswani, «Self-Attention with Relative Position Representations,» en *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, jun. de 2018, págs. 464-468. DOI: 10.18653/v1/N18-2074. dirección: https://aclanthology.org/N18-2074/.

[119] V. Vapnik y C. Cortes, «Support-Vector Networks,» *Machine Learning*, vol. 20, págs. 273-297, 1995.

[120] W. Chao, «A Tutorial for Support Vector Machine,» Disponible en: http://disp.ee.ntu.edu.tw/ pujols/Support %20Vector %20Machine.pdf.

[121] M. Hoffmann, *Support Vector Machines – Kernels and the Kernel Trick*, An elaboration for the Hauptseminar "Reading Club: Support VectorMachines", jun. de 2006.

[122] C.-W. Hsu y C.-J. Lin, «A comparison of methods for multiclass support vector machines,» *IEEE Transactions on Neural Networks*, vol. 13, n.º 2, págs. 415-425, 2002. DOI: 10.1109/72.991427.

[123] W. C. S. IV y B. Krawczyk, «Multi-class imbalanced big data classification on Spark,» *Knowledge-Based Systems*, vol. 212, pág. 106 598, 2021, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2020.106598. dirección: https://www.sciencedirect.com/science/article/pii/S0950705120307279.

[124] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk y F. Herrera, «Cost-Sensitive Learning,» en *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018, cap. 4, págs. 63-78, ISBN: 978-3-319-98074-4. DOI: 10.1007/978-3-319-98074-4_4.

[125] V. Ganganwar, «An overview of classification algorithms for imbalanced datasets,» *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, n.º 4, págs. 42-47, 2012.

[126] K. Veropoulos, C. Campbell y N. Cristianini, «Controlling the Sensitivity of Support Vector Machines,» *Proceedings of International Joint Conference Artificial Intelligence*, jun. de 1999.

[127] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk y F. Herrera, «Foundations on Imbalanced Classification,» en *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018, cap. 2, págs. 19-46, ISBN: 978-3-319-98074-4. DOI: 10.1007/978-3-319-98074-4_4.

[128] Q. Gu, L. Zhu y Z. Cai, «Evaluation Measures of the Classification Performance of Imbalanced Data Sets,» en *Computational Intelligence and Intelligent Systems*, Z. Cai, Z. Li, Z. Kang e Y. Liu, eds., 2009, págs. 461-471, ISBN: 978-3-642-04962-0.

[129] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk y F. Herrera, «Cost-Sensitive Learning,» en *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, 2018, cap. 4, págs. 47-61, ISBN: 978-3-319-98074-4. DOI: 10.1007/978-3-319-98074-4_4. dirección: https://doi.org/10.1007/978-3-319-98074-4%5C_4.

[130] B. W. Matthews, «Comparison of the predicted and observed secondary structure of T4 phage lysozyme,» *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, págs. 442-451, 2 1975, ISSN: 0005-2795. DOI: https://doi.org/10.1016/0005-2795(75)90109-9. dirección: https://www.sciencedirect.com/science/article/pii/0005279575901099.

[131] D. Powers, «Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,» *Journal of Machine Learning Technologies*, vol. 2, n.º 1, págs. 37-63, ene. de 2008.

[132] J. Gorodkin, «Comparing two K-category assignments by a K-category correlation coefficient,» *Computational Biology and Chemistry*, vol. 28, n.º 5, págs. 367-374, dic. de 2004, ISSN: 14769271. DOI: 10.1016/j.compbiolchem.2004.09.006.

[133] Z. C. Lipton, C. Elkan y B. Naryanaswamy, «Optimal Thresholding of Classifiers to Maximize F1 Measure,» en *Machine Learning and Knowledge Discovery in Databases*, T. Calders, F. Esposito, E. Hüllermeier y R. Meo, eds., 2014, págs. 225-239, ISBN: 978-3-662-44851-9.

[134] M. Sokolova y G. Lapalme, «A systematic analysis of performance measures for classification tasks,» *Information Processing & Management*, vol. 45, n.º 4, págs. 427-437, 2009, ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2009.03.002. dirección: https://www.sciencedirect.com/science/article/pii/S0306457309000259.

[135] D. Chicco y G. Jurman, «The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,» *BMC Genomics*, vol. 21, 1 ene. de 2020, ISSN: 14712164. DOI: 10.1186/s12864-019-6413-7.

[136] F. Stollenwerk et al., «Text Annotation Handbook: A Practical Guide for Machine Learning Projects,» oct. de 2023.

[137] K. He, X. Zhang, S. Ren y J. Sun, «Deep Residual Learning for Image Recognition,» dic. de 2015.

[138] J. Redmon, S. Divvala, R. Girshick y A. Farhadi, «You Only Look Once: Unified, Real-Time Object Detection,» jun. de 2015. arXiv: 1506.02640 [cs.CV].

[139] M. Fadaee, A. Bisazza y C. Monz, «Data Augmentation for Low-Resource Neural Machine Translation,» mayo de 2017. DOI: 10.18653/v1/P17-2090. arXiv: 1705.00440 [cs.CL].

[140] X. Zhang, J. Zhao e Y. LeCun, «Character-Level Convolutional Networks for Text Classification,» en *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, ép. NIPS'15, Montreal, Canada: MIT Press, 2015, págs. 649-657.

[141] T. Wolf et al., «Transformers: State-of-the-Art Natural Language Processing,» en *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, oct. de 2020, págs. 38-45. dirección: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[142] C.-W. Hsu, C.-C. Chang y C.-J. Lin, «A Practical Guide to Support Vector Classification,» National Taiwan University, inf. téc., 2016.

[143] G. Guibon, M. Ochs y P. Bellot, «From emojis to sentiment analysis,» en *WACAI 2016*, Lab-STICC and ENIB and LITIS, Brest, France, jun. de 2016. dirección: https://hal-amu.archives-ouvertes.fr/hal-01529708.

[144] J. Redmon y A. Farhadi, «YOLOv3: An Incremental Improvement,» abr. de 2018. arXiv: 1804.02767 [cs.CV].

[145] S. Van der Walt et al., «Scikit-image: image processing in Python,» *PeerJ*, vol. 2, e453, jun. de 2014, ISSN: 2167-8359. dirección: https://doi.org/10.7717/peerj.453.

[146] A. Radford et al., «Learning Transferable Visual Models From Natural Language Supervision,» 2021. dirección: https://api.semanticscholar.org/CorpusID:231591445.

[147] N. Reimers e I. Gurevych, «Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,» en *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, nov. de 2019. dirección: http://arxiv.org/abs/1908.10084.

[148] R. Meyes, M. Lu, C. W. de Puiseau y T. Meisen, «Ablation Studies in Artificial Neural Networks,» ene. de 2019.

[149] L. McInnes, J. Healy y J. Melville, «UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,» feb. de 2018. arXiv: 1802.03426 [stat.ML].

[150] Z. Liu et al., «Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,» mar. de 2021.

[151] O. Sidorov, R. Hu, M. Rohrbach y A. Singh, «TextCaps: a Dataset for Image Captioning with Reading Comprehension,» inf. téc.

[152] A. Vempala y P.-P. Daniel, *Categorizing and Inferring the Relationship between the Text and Image of Twitter Posts*. dirección: https://github.com/.

[153] C. V. Rivas, *La imagen de México en el Mundo 2006-2015*. Fernández Editores, 2016.

[154] A. Alessa, M. Faezipour y Z. Alhassan, «Text Classification of Flu-Related Tweets Using FastText with Sentiment and Keyword Features,» en *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, 2018, págs. 366-367. DOI: 10.1109/ICHI.2018.00058.

[155] V. Sanh, L. Debut, J. Chaumond y T. Wolf, «DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,» *ArXiv*, vol. abs/1910.01108, 2019.

[156] J. Camacho-collados et al., «TweetNLP: Cutting-Edge Natural Language Processing for Social Media,» en *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Abu Dhabi, UAE: Association for Computational Linguistics, dic. de 2022, págs. 38-49. dirección: https://aclanthology.org/2022.emnlp-demos.5.