



UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO

Análisis de sentimientos en español en tuits relacionados con las enfermedades infecciosas

D. Oscar Omar Apolinario Arzube

2021



UNIVERSIDAD DE MURCIA

ESCUELA INTERNACIONAL DE DOCTORADO
TESIS DOCTORAL

Análisis de sentimientos en español en tuits
relacionados con las enfermedades infecciosas

Oscar Omar Apolinario Arzube

Director
Dr. Rafael Valencia García

2021

Dedicatoria

A mis padres, Maria Arzube y Vidal Apolinario, por enseñarme que la mejor herencia es la educación. A mis hermanos y hermanas que me enseñaron la virtud de compartir lo mucho o poco que tenga. A mis hijos que aquí les dejo esta meta como si fuera suya. Y anhelando como padre que la superen en cualquier camino que escojan seguir.

A los amigos, por estar presente, en los momentos buenos y complicados. Pepe muchas gracias;!

A Rafa, director de mi tesis, por guiarme en estos cuatro años.

Contenido

1	Capítulo 1. Introducción	15
2	Capítulo 2. Estado del arte	18
2.1	Inteligencia Artificial	18
2.2	Fundamentos de la inteligencia artificial	20
2.3	Historia de la inteligencia artificial	22
2.4	El lenguaje natural	24
2.5	Procesamiento del lenguaje natural	25
2.6	Machine Learning y Deep Learning	28
2.6.1	Machine Learning.	28
2.6.2	Clasificación de algoritmos de aprendizaje.	29
2.6.3	Deep Learning.	32
2.7	Niveles del Lenguaje Natural	37
2.7.1	Recursos Lingüísticos en sistemas de procesamiento de lenguaje natural.	38
2.8	Ingeniería de características.	39
2.8.1	Características estadísticas.	39
2.8.2	Características lingüísticas.	43
2.8.3	Características contextuales.	44
2.9	Modelos de aprendizaje profundo para el procesamiento del lenguaje natural	44
2.9.1	Redes neuronales convolucionales (CNN).	44
2.9.2	Word embedding.	45
2.9.3	Redes neuronales recurrentes (RNN).	47
2.9.4	Memoria a corto plazo (LSTM).	48
2.9.5	Unidades recurrentes cerradas (GRU).	50
2.10	Técnicas de clasificación del procesamiento del lenguaje natural	51
2.10.1	Clasificación del procesamiento del lenguaje natural a nivel de documentos.	51
2.10.2	Clasificación de procesamiento del lenguaje natural a nivel de oraciones.	52
2.10.3	Clasificación del procesamiento del lenguaje natural a nivel de aspectos.	53
2.10.4	Clasificación del procesamiento del lenguaje natural multidominio.	54

2.10.5	Clasificación del procesamiento del lenguaje natural multimodal.	54
2.11	Análisis del lenguaje subjetivo	55
2.12	Análisis de sentimientos.	57
2.13	Modelos populares para la clasificación de sentimiento	58
2.14	Otras tareas de clasificación de textos: detección de la sátira	59
2.15	Resumen del estado del arte para la clasificación de textos	62
2.16	Objetivos de la tesis doctoral	63
2.16.1	Motivación.	63
2.16.2	Objetivos.	65
2.16.3	Metodología.	66
3	Capítulo 3. Materiales y métodos para la clasificación de los corpus	67
3.1	Arquitectura del sistema propuesto de Monitoreo de redes sociales.	67
3.2	Pipeline de nuestra propuesta	68
3.3	Extracción de características	69
3.3.1	Term-Counting Features.	69
3.3.2	Word Embeddings.	70
3.3.3	Clasificadores supervisados.	73
3.3.4	Clasificadores de aprendizaje automático.	73
3.3.5	Arquitecturas de aprendizaje profundo.	74
3.4	Implementación de los modelos	75
3.5	Optimización de hiperparámetros	76
3.6	Corpus desarrollados en la tesis doctoral	77
3.6.1	El corpus ZIKA.	80
3.6.1.1	Visualización del corpus ZIKA.	87
3.6.1.2	Operacionalidad de los modelos para el corpus ZIKA.	93
3.6.2	El corpus COVID-19.	95
3.6.2.1	Visualización del corpus COVID.	102
3.6.2.2	Operacionalidad de los modelos para el corpus COVID-19.	108
3.6.3	El corpus de la SATIRA	110
4	Capítulo 4. Evaluación	121
4.1	Introducción	121

4.2	Métricas de evaluación	122
4.3	Evaluación del corpus ZIKA	123
4.3.1	Evaluación para modelos tradicionales de machine learning del corpus ZIKA.	123
4.3.2	Evaluación para modelos de deep learning del corpus ZIKA.	132
4.3.3	Discusión de los resultados del corpus ZIKA.	135
4.4	Evaluación del corpus COVID	137
4.4.1	Evaluación para modelos tradicionales de machine learning del corpus COVID.	137
4.4.2	Evaluación para modelos de deep learning del corpus COVID.	145
4.4.3	Discusión de los resultados del corpus COVID.	148
4.5	Evaluación del corpus SATIRA	150
4.5.1	Evaluación para modelos tradicionales de machine learning del corpus SATIRA.	151
4.5.2	Evaluación para modelos de deep learning del corpus SATIRA.	159
4.5.3	Discusión de los resultados del corpus SATIRA.	162
5	Capítulo 5. Conclusiones y trabajos futuros.	166
5.1	Introducción	166
5.2	Conclusiones	166
5.3	Trabajos futuros	170
6	Capítulo 6. Contribuciones científicas derivadas de la tesis doctoral	174
6.1	Publicaciones en revistas	174
6.2	Publicaciones en Congresos	175

Índices de figuras

Figura 1	Mapa conceptual de la Inteligencia artificial.....	18
Figura 2	Concepto de lenguaje	25
Figura 3	Diferencias entre NLU y NLP.....	27
Figura 5	Clasificación típica de modelos de aprendizaje por computador	29
Figura 6	Diferencias entre algoritmos de clasificación y regresión	30
Figura 7	Ejemplo de algoritmo de agrupamiento	31
Figura 8	Arquitectura típica de una red neuronal	33
Figura 10	<i>One-hot encoding</i> ²	39
Figura 11	Modelo de bolsa de palabras	40
Figura 12	Representación de word embeddings en un espacio tri-dimensional	42
Figura 13	Redes neuronales CNN	45
Figura 15	Redes neuronales RNN.....	48
Figura 16	Arquitectura LSTM.....	49
Figura 17	Arquitectura combinada de LSTM y CNN	50
Figura 18	Arquitectura GRU.....	50
Figura 20	Arquitectura del módulo de monitorización de redes sociales	67
Figura 21	El pipeline de nuestra propuesta.....	69
Figura 22	Captura de pantalla de la herramienta de clasificación de tweets.....	79
Figura 23	Palabras más frecuentes del corpus del Zika (rojo).....	82
Figura 24	Histograma de longitud de textos del corpus ZIKA.....	82
Figura 25	Histograma de números de palabras en los tweets ZIKA.....	83
Figura 26	Histograma de longitud promedio de palabras en los tweets ZIKA.....	83
Figura 27	Histograma de s stop-word más utilizados en el corpus ZIKA.....	84
Figura 28	Histograma de Word más utilizados en el corpus ZIKA.....	84
Figura 29	Histograma de bigramas más utilizados en el corpus ZIKA.....	85
Figura 30	Histograma de trigramas más utilizados en el corpus ZIKA.....	85
Figura 31	Histograma de entidades más utilizados en el corpus ZIKA.....	86
Figura 32	Histograma de índice de legibilidad ARI en el corpus ZIKA.....	87
Figura 33	Numero de mensaje por zona de salud ZIKA	88
Figura 34	Mapa de calor de la información focalizada en Ecuador de ZIKA	89
Figura 35	Tendencias de sentimientos de ZIKA	90
Figura 36	Ejemplo de últimos tweets reportados ZIKA	91
Figura 37	Sentimientos agrupados por conceptos en el ZIKA	92
Figura 38	Sentimientos agrupados por conceptos en el ZIKA	92
Figura 39	Ejemplos de tweets publicado en base al corpus del ZIKA.....	94
Figura 40	Ejemplos de tweets publicado en base al corpus del ZIKA.....	95

Figura 41	Palabras más frecuentes del corpus de COVID (verde).	97
Figura 42	Histograma de longitud de textos del corpus COVID.	98
Figura 43	Histograma de números de palabras en los tweets COVID.	98
Figura 44	Histograma de longitud promedio de palabras en los tweets COVID.	99
Figura 45	Histograma de stop-word más utilizados en el corpus COVID.	99
Figura 46	Histograma de Word más utilizados en el corpus COVID.	100
Figura 47	Histograma de bigramas más utilizados en el corpus COVID.	100
Figura 48	Histograma de trigramas más utilizados en el corpus COVID.	101
Figura 49	Histograma de entidades más utilizados en el corpus COVID.	101
Figura 50	Histograma de índice de legibilidad ARI en el corpus COVID.	102
Figura 51	Numero de mensaje por zona de salud COVID-19	103
Figura 52	Mapa de calor de la información focalizada en Ecuador de COVID-19	104
Figura 53	Tendencias de sentimientos de COVID-19	105
Figura 54	Ejemplo de últimos tweets reportados COVID-19	106
Figura 55	Sentimientos agrupados por conceptos en el COVID-19	107
Figura 56	Sentimientos agrupados por conceptos en el COVID-19	108
Figura 57	Ejemplos de tweets publicado en base al corpus del COVID-19.	109
Figura 58	Ejemplos de tweets publicado en base al corpus del COVID-19.	110
Figura 59	Histograma de cantidad de textos del corpus SATIRA.	111
Figura 60	Histograma de números de palabras en los tweets SATIRA.	112
Figura 61	Histograma de longitud promedio de palabras en los tweets SATIRA.	112
Figura 62	Histograma de stop Word más utilizados en el corpus SATIRA.	113
Figura 63	Histograma de Word más utilizados en el corpus SATIRA.	113
Figura 64	Histograma de bigramas más utilizados en el corpus SATIRA.	114
Figura 65	Histograma de trigramas más utilizados en el corpus SATIRA	114
Figura 66	Histograma de entidades más utilizados en el corpus SATIRA.	115
Figura 67	Histograma de entidades como PERSO más utilizados en el corpus SATIRA	115
Figura 68	Histograma de índice de legibilidad ARI en el corpus SATIRA.	116
Figura 69	Distribución de corpus por cuentas para el español europeo (izquierda) y el español mexicano derecho) conjuntos de datos compilados en (Salas-Zárate et al., 2017). En rojo, relatos satíricos. En verde, relatos no satíricos	Figura 3. Ejemplos de español europeo no satírico (izquierda) y satírico español europeo (derecho) tweets. 118
Figura 70	Ejemplos de español europeo no satírico (izquierda) y satírico español europeo (derecho) tweets.	119
Figura 71	Ejemplos de español mexicano no satírico (izquierda) y satírico español mexicano derecho) tweets.	119
Figura 72	Matriz de confusión de los dos mejores modelos de machine learning en el corpus ZIKA	130
Figura 73	Matriz de confusión del mejor modelo para n-gramas y el mejor modelo para char-gramas para el corpus ZIKA	131
Figura 74	Matriz de confusión de los dos mejores modelos de deep learning en el corpus ZIKA	135
Figura 75	Matriz de confusión de los dos mejores modelos de machine learning en el corpus COVID	143
Figura 76	Matriz de confusión del mejor modelo para n-gramas y el mejor modelo para char-gramas para el corpus COVID	145

Figura 77	Matriz de confusión de los dos mejores modelos de deep learning en el corpus COVID	148
Figura 78	Matriz de confusión de los dos mejores modelos de machine learning en el corpus SATIRA	157
Figura 79	Matriz de confusión del mejor modelo para n-gramas y el mejor modelo para char-gramas para el corpus SATIRA.....	159
Figura 80	Matriz de confusión de los dos mejores modelos de deep learning en el corpus SATIRA	162

Índices de tablas

Tabla 1 Comparación de frameworks de aprendizaje profundo	36
Tabla 2 Características de los Niveles del lenguaje natural.....	38
Tabla 3 Opciones de hiperparámetros para los clasificadores tradicionales de aprendizaje automático.	76
Tabla 4 Opciones de hiperparámetros para las arquitecturas de aprendizaje profundo.	77
Tabla 5 Estadísticas de los corpus compilados y etiquetados para el ZIKA, COVID y SATIRA.	78
Tabla 6 Estadísticas generales del corpus ZIKA.	81
Tabla 7 Estadísticas generales del corpus COVID.	96
Tabla 8 Estadísticas generales del corpus SATIRA.....	111
Tabla 9 Estadísticas de corpus de la Sátira.	118
Tabla 10 Resultados de las evaluaciones en modelos de machine learning para el corpus ZIKA.....	125
Tabla 11 Comparación de los dos mejores modelos de machine learning en el corpus ZIKA.....	129
Tabla 12 Comparación por modelo de los dos mejores n-gramas y los dos mejores char-gramas en el corpus ZIKA ...	131
Tabla 13 Resultados de las evaluaciones en modelos de deep learning para el corpus ZIKA.....	133
Tabla 14 Comparación de los dos mejores modelos de deep learning en el corpus ZIKA.....	134
Tabla 15 Mejores accuracy de los modelos en machine learning y Deep learning en el corpus ZIKA.....	137
Tabla 16 Resultados de las evaluaciones en modelos de machine learning para el corpus COVID.....	139
Tabla 17 Comparación de los dos mejores modelos de machine learning en el corpus COVID.	142
Tabla 18 Comparación por modelo de los dos mejores n-gramas y los dos mejores char-gramas en el corpus COVID	144
Tabla 19 Resultados de las evaluaciones en modelos de deep learning para el corpus COVID.....	146
Tabla 20 Comparación de los dos mejores modelos de deep learning en el corpus COVID.....	147
Tabla 21 Mejores accuracy de los modelos en machine learning y Deep learning en el corpus COVID.....	149
Tabla 22 Resultados de las evaluaciones en modelos de machine learning para el corpus SATIRA.....	152
Tabla 23 Comparación de los dos mejores modelos de machine learning en el corpus SATIRA.....	156
Tabla 24 Comparación por modelo de los dos mejores n-gramas y los dos mejores char-gramas en el corpus SATIRA	158
Tabla 25 Resultados de las evaluaciones en modelos de deep learning para el corpus SATIRA.....	160
Tabla 26 Comparación de los dos mejores modelos de deep learning en el corpus SATIRA.....	161
Tabla 27 Mejores accuracy de los modelos en machine learning y Deep learning en el corpus SATIRA.....	163
Tabla 28 Comparación de los resultados en la predicción para los corpus ZIKA, COVID y SATIRA.....	164
Tabla 29 Comparación de los mejores modelos de machine y deep learning para los corpus ZIKA, COVID y SATIRA.....	165

Lista de acrónimos

Termino	Significado
IA	Inteligencia artificial
OMS	Organización mundial de la salud
PLN	Procesamiento del lenguaje natural
ML	Machine learning
DL	Deep learning
SVM	Support Vector Machine
TF-IDF	Term frequency – Inverse document frequency
URL	Uniform resource locator
API	Interfaz de programación de aplicaciones
CNN	Convolutional neural network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
DENSE	Called fully-connected layer
LR	Logistic Regression
RF	Random Forest
K	The k-nearest neighbors (KNN)
J48	Decision Tree

Resumen

Desde que Tim Berners-lee creara la web en el año de 1990 permitió que la humanidad compartiera información como su principal objetivo. Esta herramienta ha seguido creciendo con mayores funcionalidades que permiten a los usuarios de la red de internet maximizar su uso. Así nace la Web 2.0 que permite a las personas colaborar y compartir información en línea. Una mejora inmensamente considerable para las necesidades humanas. Este hito permitió la creación de blog, wiki, twitter y otras plataformas de uso libre que potencializan la expresión más pura del ser humano en comunicarse.

Lo que al inicio empezó por escribirse en piedra como jeroglíficos; luego paso a ser expresado en papel y ahora ver a esta comunicación ser plasmada en un mundo digital, hace que la información se potencialice en cantidad y forma. Esto implica un gran beneficio, pero a su vez un gran problema. Tener una cantidad grande de información permite saber que una noticia que hace décadas generada hoy, al otro lado del mundo terminaba llegando en meses. Pero ahora una noticia que se genera es casi conocida al instante por el uso cotidiano de las diferentes redes sociales. Información que expresan los usuarios en diferentes dominios como memes, noticias, deportes, política, comercio por nombras unos cuantos temas. Pero es el dominio de la salud donde se centra nuestra atención.

La salud es un factor de ser humano que debe ser cuidado desde diferentes aristas. Estas van de la mano con la evolución en la sociedad. Creamos y conceptualizamos conocimientos a través de lo que denominamos ciencias, específicamente la medicina para tratar enfermedades que aquejan a la humanidad. Ahora si tomamos a la salud como una fuente de generación de información y que se trasmite o se escribe en las redes sociales; pues tenemos una formula multiplicadora de datos. Y si de cantidad hablamos son las ciencias exactas una parte de la solución a este problema. Pero que no se terminaría de resolver sin las tecnologías información o las ciencias computacionales.

Para resolver el problema de la existencia de enorme información que se genera sobre la salud, ha nacido una nueva ciencia denominada la infodemiología. Que conlleva tomar la información que se genera en sitios webs o en las redes sociales para crear una vigilancia comunitaria. Pero que es de poco valor si la información obtenida no es transformada para que se visualice o tenga rostro que se consiguen con gráficos, indicadores, índices, tendencias y todo lo que podamos bosquejar. Para pasar de una información a una descriptiva.

El gran reto está en pasar de las técnicas descriptivas a pasar a las técnicas de predicción. Esto nos lleva a buscar dentro de las ciencias una; que nos permita proponer un marco de estudio, metodología

he investigación. Así las ciencias computacionales con su rama de la inteligencia artificial y su enfoque en el procesamiento del lenguaje natural para obtener el texto y poder describirlo. Pero esta tesis doctoral va más allá y propone utilizar algoritmos de aprendizaje automático y aprendizaje profundo. Esto permitirá aportar con hallazgos importantes al mundo científico, producto de aplicar algoritmos sobre los datos que nos arroja la infodemiología.

Este trabajo doctoral termina por realizar comparaciones de los mejores modelos de predicción sobre tres dominios. Dos de ellos enfocados en la salud y uno enfocado en textos lingüísticos o psicológico. Esto permitirá evaluar nuestra propuesta de estudio en el uso de n-gramas o char-gramas de los dominios en enfermedades infectocontagiosas como el zika, dengue o chikungunya. Ampliamos nuestro estudio sobre el dominio de covid-19 y se vuelve más interesante; sobre un hecho histórico como lo es una pandemia.

1 Capítulo 1. Introducción

Existe una nueva ciencia denominada infodemiología que indica como gestionar la información sobre la salud que se genera en el internet. Al igual que los patógenos en las epidemias, la información sesgada se expande más y agrega complejidad a las respuestas de emergencia en la salud. Por lo cual, el mundo avanza con apariciones de nuevas epidemias, nuevas formas de comunicarse a través de las tecnologías de información y el uso constante de redes sociales que dejan una huella en la comunicación digital. Esto se materializa a través de medios tecnológicos y la aparición de los teléfonos celulares ha permitido que la información se expanda de forma exponencial. Los problemas de as salud siempre requerirán una acción coordinada a tiempo para hacer frente a los problemas que aquejan a la humanidad y el entorno que lo rodea.

Al abarcar gran cantidad de datos que se genera sobre la salud con la infodemiología se presenta un segundo concepto que es la abundancia de información o también llamada infodemia; y de cual, implica que existe información precisa como alguna otra que requiere de alguna verificación. A las personas les resulta complicado encontrar fuentes confiables y más cuando se trata de factores que correlacionan a la salud. Incluso el acceso a la información de alta calidad sigue siendo un factor clave entre las necesidades que tenemos como sociedad.

Tomar decisiones sobre un conjunto de información de poca calidad conlleva poner en riesgo las actividades que enmarcan la sociedad como la salud, el comercio, la educación, las actividades de recreación y hasta la vida misma se ponen en entredicho. Provocando dudas o incertidumbres sobre la sociedad. Existe también información que van a desprestigiar acciones o instituciones; teniendo como resultado que se obstaculice las políticas públicas encaminadas a combatir el impacto de enfermedades en la salud global.

La sociedad humana ha tomado como costumbre ir resolviendo los problemas de la vida por medio de las ciencias computacionales y más específicamente con la inteligencia artificial (IA) cuya informática es la inteligencia humana expresada por maquinas. Esta rama ha demostrado ser capaz de ayudar al ser humano en actividades como reconocimiento de voz, reconocimiento de imágenes, entre otras tareas que resaltan su uso. Es por eso, que hemos tomado esta ciencia

como herramienta para ayudar a resolver los numerosos problemas que se le presentan a la infodemiología.

En este trabajo doctoral, creo un entorno de trabajo que permitió diseñar una arquitectura integral compuesta de metodología, topología computacional e investigación científica; para así, tomar a la infodemiología como fuente, procesarla con las técnicas que provee la inteligencia artificial y lograr predecir el sentimiento sobre lenguaje natural. El trabajo se focalizó en tres grandes dominios para la investigación y comprobación; dos de ellos enfocados en la salud; específicamente en las enfermedades infectocontagiosas como zika, dengue o chikungunya y la pandemia que se ha generado por el covid-19. A este trabajo también incorporamos otro tipo de fuente expresado en texto lingüísticos o psicológicos que nos da el género literario de la sátira; para tener una segunda comprobación de la eficiencia aquí presentada.

Por lo antes expuesto, en los párrafos anteriores la presente tesis doctoral describe la investigación que se ha llevado a cabo en el análisis de sentimientos en español sobre tuits relacionados con las enfermedades infecciosas. El documento se divide en seis capítulos que detallan el trabajo que se realizó.

El capítulo 2, implica un detallado estudio del arte sobre las tecnologías para el desarrollo del trabajo propuesto. El estudio inicia con una visión general de lo que indica la inteligencia artificial y sus campos de acción; luego pasamos a detallar lo que implica la comunicación humana expresada en el lenguaje natural y descritas con la ingeniería de características; Para luego, profundizar en el campo de inteligencia artificial, así como el procesamiento del lenguaje natural (PLN) que me da las directrices para procesar texto. Incursionamos en los algoritmos de aprendizaje automático y aprendizaje profundo, que me permitirán realizar las predicciones y saber la eficiencia obtenida sobre cada dominio. También existe un apartado de cómo se realizan la clasificación del procesamiento del lenguaje natural a nivel de documentos, oraciones, aspectos, multidominio y multimodal. Para terminar con los conocimientos del análisis de sentimiento.

El capítulo 3, describe en detalle la arquitectura topológica que se diseñó para procesar los tres dominios del estudio. Cada dominio fue detallado en su composición de forma resumida y visual con indicadores que dan una vista global a la calidad de los tuis recopilados.

El capítulo 4, se centra en la ejecución de los experimentos y el análisis de los resultados obtenidos. En cada dominio describimos los resultados sobre a técnicas tradicionales de aprendizaje automático o también conocido como machine learning; así como los resultados obtenidos aplicando las técnicas de aprendizaje profundo o deep learning. Para terminar con una comparación entre los mejores dos resultados de ambas técnicas y una comparación de los resultados entre los dominios.

En el capítulo 5, se presentan las conclusiones que se realizó en este trabajo doctoral y posibles nuevas vías para investigaciones futuras.

El capítulo 6, presenta el aporte científico sobre publicaciones de alto impacto realizadas a partir de este trabajo.

2 Capítulo 2. Estado del arte

En esta sección contiene información focalizada sobre las tecnologías involucradas en este trabajo tenemos lo siguiente: (1) Inteligencia artificial, (2) fundamentos de la inteligencia artificial, (3) historia de la inteligencia artificial, (4) el lenguaje natural, (5) procesamiento del lenguaje natural, (6) machine learning y deep learning, (7) niveles del lenguaje natural, (8) ingeniería de características, (9) modelos de aprendizaje profundo, (10) técnicas de clasificación del procesamiento del lenguaje natural, (11) análisis del lenguaje subjetivo, (12) análisis de sentimiento, (13) modelos populares para la clasificación de sentimiento, (14) otras tareas de clasificación, (15) resumen del estado del arte y (16) objetivo de la tesis doctoral.

2.1 Inteligencia Artificial

El principio universal de la inteligencia artificial (IA) es imitar el pensamiento y capacidad de aprender como el ser humano, crear seres artificiales capaces de realizar tareas inteligentes con base en unos principios(Serna, Acevedo, & Serna, 2017).

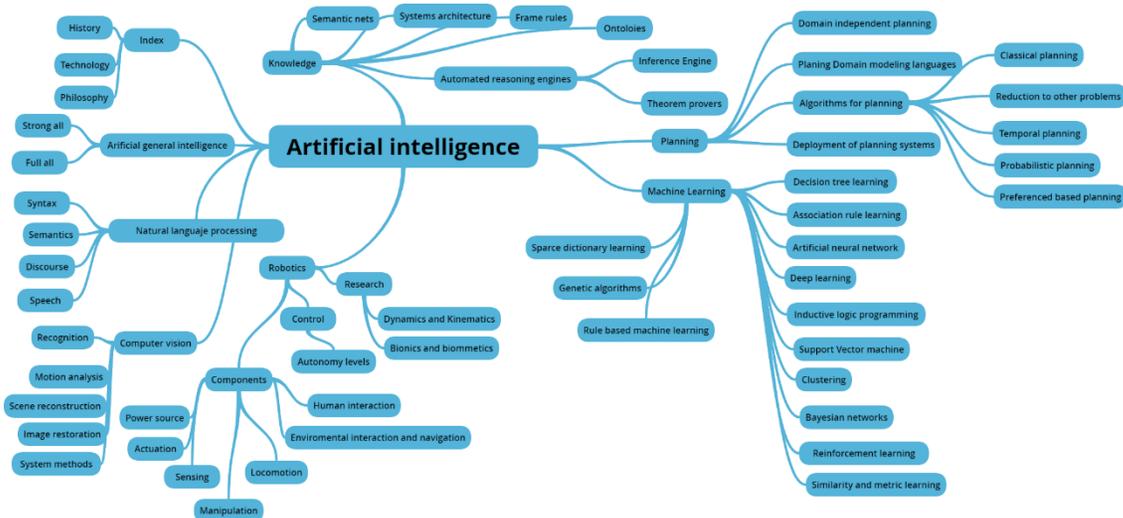


Figura 1 Mapa conceptual de la Inteligencia artificial

La Figura 20 muestra las ramas de la inteligencia artificial (IA) como una parte de las ciencias computacionales que se encarga del diseño y construcción de sistemas capaces de realizar tareas asociadas con la inteligencia humana. Está claro que existen muchas variantes para la aplicación de dicha tecnología. Sus campos de actuación pueden atravesar desde el marketing

digital, la ciencia, medicina, medio ambiente, hasta en disciplinas sociales como la antropología y sociología. A medida que avance y se genere mayor conocimiento científico, más grandes serán las oportunidades de uso que brindará la Inteligencia Artificial.

En la historia se han seguido distintos enfoques que se refieren a sistemas que piensan como humanos, sistemas que actúan como humanos, sistemas que piensan racionalmente y sistemas que actúan racionalmente(Andrew Burgess.Palgrave Macmillan, 2018, pp 56).

En el comportamiento, el enfoque de la prueba de Turing propuesta por Alan Turing (1950), se construyó para presentar el concepto operacional y de inteligencia. Él sugirió una acción basada en la incapacidad de diferenciar entre entidades inteligentes indiscutibles y seres humanos. El computador superaba la prueba si un evaluador humano no es capaz de distinguir si las respuestas a una serie de interrogantes planteados son de una persona o no. Para superar la prueba de Turing el computador debe está dotado de las siguientes características(Cabanelas Omil, 2019):

- Procesamiento del lenguaje natural, que le permita comunicarse satisfactoriamente.
- Presentación del conocimiento para almacenar lo que se conoce o siente.
- Razonamiento automático para utilizar la información almacenada para responder a preguntas y extraer nuevas conclusiones.
- Visión computacional para percibir objetos.
- Robótica para manipular y mover objetos.

El enfoque del modelo cognitivo permite decir que un programa dado piensa como un ser humano. El mecanismo para indicar como piensan los seres humanos puede ser mediante introspección (intentando atrapar nuestros propios pensamientos) y a través de experimentos psicológicos. En el campo de la ciencia cognitiva convergen modelos computacionales de IA y técnicas experimentales de psicología intentando elaborar precisas y variables sobre el procesamiento de la mente humana(Gregersen, 2014, pp. 147-150).

El pensamiento racional nos permite estructurar la argumentación de la mente con leyes para llegar a conclusiones. Las leyes del pensamiento gobiernan la manera de operar de la mente lo que se conoce como el campo llamado lógica(Nakamura & Kawaguchi, 2016, pp. 70-95). Actuar de forma racional permite definir un enfoque donde el agente racional es aquel que

actúa con la intención de alcanzar el mejor resultado esperado. En el caso de la IA, las leyes de pensamiento ponen énfasis en hacer inferencias correctas. Sin embargo, el efectuar una inferencia correcta no depende siempre de la racionalidad. Las inferencias es un mecanismo existente para garantizar la racionalidad y es más afín a la forma en que se ha producido el avance científico(Bunker, Magarkar, & Viitala, 2016, pp. 32-35).

2.2 Fundamentos de la inteligencia artificial

Las disciplinas que han contribuido con ideas y técnicas en el desarrollo de la IA son diversas. En la filosofía (desde el año 428 a.C. hasta el presente), Aristóteles (384-322 a.C.) fue uno de los primero en formular un conjunto preciso de leyes que gobernaban la parte racional de la inteligencia. Él desarrollo un sistema informal para razonar adecuadamente que en principio permitía extraer conclusiones mecánicamente a partir de premisas iniciales (Gotthelf, 2013).

Dentro de las matemáticas (aproximadamente desde el año 800 al presente) la filosofía delimito las ideas más importantes de la IA, pero para una ciencia formal es necesario contar con una formulación matemática en tres áreas fundamentales: la probabilidad, la computación y la lógica. El concepto de lógica formal se remonta a los filósofos de la antigua Grecia, pero para su desarrollo matemático comenzó realmente con el trabajo de George Boole (1815-1864) que definió la lógica proposicional o booleana (Barkalov, Titarenko, & Mazurkiewicz, 2019). El siguiente paso fue la definición de lo que se puede hacer con la lógica y la informática. El primer algoritmo no trivial fue el algoritmo Euclídeo para el cálculo del máximo común divisor (Eisenbrand, 2011). La tesis de Church-Turing, que afirma que la máquina de Turing (Turing 1936) es capaz de calcular cualquier función computable (Graves, 2014). Además de la lógica y el cálculo, la tercera contribución de las matemáticas a la IA es la teoría de la probabilidad. El italiano Gerolamo Cardano (1501-1576) fue el primero en proponer la idea de probabilidad, presentándola en términos de resultados de juegos de apuesta (Evren, Evren, Dalbudak, Topcu, & Kutlu, 2018, pp. 7-10).

Por otro lado, en la economía (desde el año 1776 hasta el presente), empezó cuando el filósofo escoces Adam Smith (1723-1790) fue el primero en tratarlo como una ciencia, utilizando la idea de que las economías pueden concebirse como un conjunto de elementos individuales que maximizan su propio estado en el bienestar de la economía o también llamada utilidad

(Sandmo, 2019, pp. 10-15). Los economistas indican que ellos estudian cómo la gente toma decisiones por lo que nace la teoría de la toma de decisiones (Stigler, Irwin, Stigler, & Irwin, 2018). La teoría de la decisión que combina la teoría de la probabilidad con la teoría de la utilidad proporciona un entorno global y formal para la toma de decisiones.

La neurociencia (desde el año 1861 hasta el presente) es el estudio del sistema neurológico y fundamentalmente del cerebro. La forma exacta en la que en un cerebro se genera el pensamiento es uno de los grandes misterios de la ciencia. El estudio de Paul Broca (1824-1880) sobre la afasia (dificultad para hablar) en paciente con el cerebro dañado, en 1861, le dio fuerza a este campo con la existencia de áreas localizadas en el cerebro responsable de funciones cognitivas específicas (Esteves, Ganz, Sousa, & Leite-Almeida, 2020, pp. 3-14). La ley de Moore indica que el número de puertas lógicas de Unidad de Procesamiento Central de un computador estándar se igualará con el de neuronas del cerebro (Waldrop, 2016, pp. 1-4).

La psicología científica (desde el año 1879 hasta el presente) se inició con los trabajos del físico alemán Hermann von Helmholtz quien aplicó el método científico al estudio de la vista humana (Cahan, 2015). La conceptualización del cerebro como un dispositivo de procesamiento de información, característica principal de la psicología cognitiva, se remonta según lo indicado por William James (1842-1910) que pone énfasis en que la percepción indica un tipo de inferencia lógica inconsciente. La ciencia cognitiva se puede encontrar en el siguiente trabajo (Cowan, 2015). Este artículo influyente mostró cómo se puede utilizar los modelos informáticos para modelar la psicología de la memoria, el lenguaje y el pensamiento lógico.

La ingeniería computacional (desde el año 1940 hasta el presente) ayuda a la inteligencia artificial para que pueda llegar a ser una realidad. Para ello se necesitan dos cosas: inteligencia y un artefacto. El computador ha sido el artefacto elegido para cumplir ese rol. El equipo de Alan Turing construyó el primer computador en 1940 llamada Heath Robinson, con un único propósito descifrar mensajes alemanes (Wright, 2017, pp. 295-328). La investigación en IA ha generado numerosas ideas de las que se ha beneficiado la informática: así como el tiempo compartido, los intérpretes, los computadores personales con interfaces gráficas y ratones, entornos integrados de desarrollo, listas de punteros, administración de la memoria y conceptos

de programación simbólica como la orientada a objetos (Sung, Ahn, & Black, 2017, pp. 443–463).

En la teoría de control y cibernética (desde el año 1948 hasta el presente); Ktesibios de Alejandría (250 a.C.) construyó la primera máquina auto controlada: un reloj de agua con un regulador que mantenía el flujo de agua circulando por él, con un ritmo constante y predecible(Fragaki, 2012). La figura central del desarrollo de lo que ahora se llama teoría de control fue Norbert Wiener (1894-1964) quien fue un matemático que trabajo en sistemas de control biológicos y mecánicos en sus vínculos con la cognición(Franchi & Bianchini, 2018). La IA se fundamentó para superar las limitaciones matemáticas de la teoría de control en los años 50. Las herramientas de inferencia lógica y computaciones permitieron a los investigadores de IA afrontar problemas relacionados con el lenguaje, visión y planificación(Van Long, 2015).

La lingüística (desde el año 1957 hasta el presente) también ha favorecido una rama muy importante de la IA. B. F. Skinner publicó *Verbal Behavior* en 1957 presentando una visión extendida desde el punto de vista conductista al aprendizaje del lenguaje (Bem, 2017, pp. 17-22). La lingüística y la IA conforman la lingüística computacional o procesamiento del lenguaje natural. El entendimiento del lenguaje requiere la comprensión de la materia bajo estudio y de su contexto, y no solamente el entendimiento de la estructura de las sentencias (Hansen, McMahon, & Prat, 2018).

2.3 Historia de la inteligencia artificial

Los inicios de la inteligencia artificial se remontan a los años 40(1943-1955), Warren McCulloch y Walter Pitts (1943) han sido reconocidos como los autores del primer trabajo de IA, en el cual partieron de insumos: conocimientos sobre la filosofía básica y funcionamiento de las neuronas en el cerebro, el análisis formal de la lógica proposicional y la teoría computacional de Turing. Propusieron un modelo constituido por neuronas artificiales, en el que cada una de ellas se caracterizaba por estar activada o desactivada (Arbib, 2018, pp. 36-50). Otro de los primeros trabajos de IA es el de Alan Turin quien articuló primero una visión de la IA en su artículo *Computing Machinery and Intelligence* en 1950 (Deloitte, 2018, pp. 1-8).

En el nacimiento de la inteligencia artificial (1956), Princeton acogió a otras de las figuras señaladas de la IA como John McCarthy, Claude Shannon y Nathaniel Rochester para que le ayudaran a aumentar el interés de los investigadores americanos en la teoría de autómatas, las redes neuronales y el estudio de la inteligencia (Eklund, Galán, & Gähler, 2009, pp. 1-13). La IA desde el primer momento abarcó la idea de duplicar facultades humanas como la creatividad, la automejora y el uso del lenguaje. Así apoyándose de metodologías persigue la construcción de máquinas que funciones automáticamente en medios complejos y desafiantes (Agrawal, Gans, & Goldfarb, 2017).

En los sistemas basados en el conocimiento (1969-1979) el uso de conocimiento específico del dominio facilita el desarrollo de etapas de razonamiento más largas, pudiéndose así resolver casos concurrentes en dominios de conocimiento restringido. El programa DENDRAL constituye uno de los primeros ejemplos de este enfoque que fue diseñado en Standford. Los autores se enfocaron en inferir una estructura molecular a partir de la información proporcionada por un espectrómetro de masas (Lederberg et al., 1969). El crecimiento generalizado de aplicaciones para solucionar problemas del mundo real provocó el respectivo aumento en la demanda de esquemas de representación del conocimiento. Como el trabajo para la representación de situaciones estereotipo (Carbonell, Cullingford, & Gershman, 1981). la descripción de la organización de la memoria humana (Rieger, 1976, pp. 89-127) y la comprensión de planes y objetivos (Wilensky, 1983, pp. 579-591).

La inteligencia artificial seguidamente se convirtió en una industria a partir de los años 80 (desde 1980 hasta la presente). El primer sistema experto que tuvo éxito fue R1, que inició su actividad en Digital Equipment Corporation (Spechler, 2017, pp. 216-230). El programa se utilizaba en la elaboración de pedidos de nuevos sistemas informáticos. En 1986 representaba para la compañía un ahorro estimado de 40 millones de dólares al año. En el 1988 el grupo de inteligencia artificial de DEC había distribuido ya 40 sistemas expertos, y había más en camino. En su conjunto, la industria de la IA creció rápidamente, pasando de unos pocos millones de dólares en 1980 a billones de dólares en 1988 (Fischetti, 2013, pp. 76-82).

El regreso de las redes neuronales (desde 1986 hasta el presente) ha supuesto una revolución tecnológica en los últimos años. Aunque la información había sido abandonada de manera general el campo de las redes neuronales a finales de los años 70. el trabajo continuó con otros

trabajos como John Hopfield (1982) donde utilizaron técnicas de la mecánica estadística para analizar las propiedades de almacenamiento y optimización de las redes, tratando colección de nodos como colecciones de átomos (Núñez Cardenas, Tomás Mariano, Crespo Hernandez, & Andrade Hernández, 2017, pp. 120-135).

La IA se convierte en una ciencia (desde 1987 hasta el presente) que en nuestra última década se ha visto una revolución en el contenido como en la metodología de trabajo en el campo de la inteligencia artificial (LExcellent, 2019, pp. 23-27). La IA se fundó en parte en el marco de una rebelión en contra de las limitaciones de los campos de la teoría de la estadística y abarca campo como lo que indica David W Ertel (Ertel, 2017, pp. 89-95). En términos ilustrativos se indica que la IA forma parte del concepto en los métodos científicos y los resultados se deben analizar como estadística para identificar su relevancia (Price, Gerke, & Cohen, 2019, pp. 1574-1577). La popularidad del uso de internet y poder compartir los datos de prueba y código fuente han hecho posible que tengamos al alcance poder contrastar los experimentos.

Otro campo importante de la IA es el campo del reconocimiento del habla. Sobre la década de los años 70 se puso a prueba una variedad de arquitecturas, muchas de ellas fueron un tanto frágiles. En estos años los experimentos basados en los modelos de Markov ocultos o también llamados MMO (Franzese & Iuliano, 2018), han tenido mayor relevancia. Primero se basan en una rigurosa teoría matemática. Segundo, los modelos se han generado mediante procesos de aprendizaje sobre corpus de datos de lenguaje real. La tecnología del habla y el campo relacionado del reconocimiento de caracteres manuscritos están en transición hacia una generalidad utilizada en aplicaciones industriales y consumo (Barker, Marxer, Vincent, & Watanabe, 2016). El uso de metodologías y marcos teóricos, han permitido que este campo alcance un grado de conocimiento que han permitido que las redes neuronales se puedan comparar con otras técnicas similares de campos como la estadística, el reconocimiento de patrones y el aprendizaje automático. El resultado de estas aplicaciones es denominada minería de datos (Van der Aalst, 2016).

2.4 El lenguaje natural

El lenguaje natural es el medio que utilizamos los seres humanos para comunicarnos y expresarnos. Es la herramienta que ha ido evolucionando a través del tiempo como por ejemplo

español, inglés, alemán o cualquier otro idioma o dialecto. Los lenguajes se conceptualizan a través de la gramática; la cual, permite evolucionar y hacer enriquecimientos progresivos para la expresión de comunicación. El lenguaje natural evoluciona sin tener presente las reglas gramaticales, a diferencia del lenguaje formal, el cual se rige de reglas preestablecidas. En la **Figura 2** Concepto de lenguaje se muestra el lenguaje formal donde se determina los componentes léxicos, reglas gramaticales y un enfoque semántico; la principal ventaja es que se elimina la ambigüedad al utilizar expresiones precisas. Un ejemplo, es el lenguaje de programación en el cual se define los pasos de instrucciones ordenadas que son procesadas por el computador (De Asis, Fernando Hernandez-Garcia, Zacharias Holland, & Sutton, 2018).

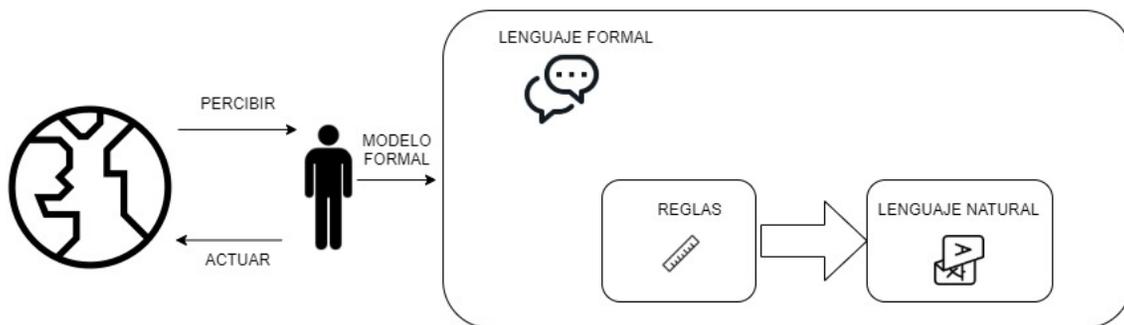


Figura 2 Concepto de lenguaje

2.5 Procesamiento del lenguaje natural

El Procesamiento del Lenguaje Natural (PLN, o NLP, por sus siglas en inglés) es una rama de la Inteligencia Artificial que sirve como nexo entre el lenguaje natural y la Informática de forma que una computadora sea capaz de entender, interpretar y generar lenguaje humano (Gatt & Krahmer, 2018).

Esta rama del conocimiento no es nueva. En concreto, los primeros estudios de PLN datan de los años 70. Sin embargo, gracias al interés que despierta este nexo entre humanos y computadores, unido a los avances en las técnicas de aprendizaje por computador y el acceso a grandes volúmenes de datos, ha llevado a que el PLN esté en una época dorada. Prueba de ello, son los asistentes virtuales como Alexa o Siri, que permiten reproducir recursos

multimedia en su hogar o acceder a datos desde Internet, mediante comandos de voz que una máquina tiene que transcribir, desambiguar y procesar (Hirschberg & Manning, 2015, pp. 55-60).

Una computadora, a bajo nivel, es sólo capaz de trabajar con un sistema binario. Sin embargo, las personas pueden utilizar un gran número de idiomas, variantes de este y, en muchas ocasiones, este uso del lenguaje tiene un importante componente cultural y de contexto que dificulta enormemente la traducción entre ambos. El PLN se encarga de poder entender y catalogar la estructura de un texto a partir de aplicar modelos tanto probabilistas como basados en conocimiento experto. Esto permite extraer información a nivel morfológico, sintáctico, semántico y pragmático con el que se un ordenador es capaz de entender el lenguaje natural (Rubinstein & Ami Podrebarac, 2019).

Como aplicaciones del PLN podemos nombrar, grosso modo, las siguientes:

- **Catalogar y clasificar contenido.** Permite clasificar un texto dentro de un conjunto de categorías. Por ejemplo, esto permite determinar si un email es spam.
- **Modelado de temas.** Permite capturar y clasificar los temas de los que trata un texto.
- **Extracción contextual.** Permite extraer automáticamente información estructurada de fuentes basadas en texto.
- **Minería de opiniones.** Permite extraer la polaridad subjetiva en un texto. Ciertos enfoques permiten también catalogar la emoción del hablante cuando la entonación también está recogida. Aunque esta aplicación del PLN también puede entenderse como un tipo de catalogar y clasificar contenido, es posible ir más allá aplicando lo que se conoce como análisis de sentimientos basado en aspectos, que además de catalogar un texto como positivo, neutro, o negativo, es capaz de extraer información acerca de qué partes del objeto estudiado evocan tales sentimientos, así como extraer también a qué tiempo se refiere y quién expresó dicha opinión.
- **Procesamiento del habla.** Permite convertir un texto a una onda de sonido y viceversa. Se encarga de los fenómenos físicos y lingüísticos que tienen que ver con la voz.
- **Resumen de documentos.** Consiste en la generación automática de un resumen de un texto. En este sentido hay sistemas que son capaces de detectar los párrafos relevantes

y otros sistemas que son capaces de sintetizar un nuevo resumen a partir del texto analizado.

- **Traducción automática.** Sistemas de traducción entre varios idiomas, incluso usando lenguajes pivote cuando no existe un sistema traducción directa entre dos idiomas; lo cual favorece enormemente a idiomas que cuentan con pocos recursos.

Dentro del PLN, otra rama de estudio es la comprensión del lenguaje natural (NLU, por sus siglas en inglés Natural Language Understanding), cuya idea es ir un paso más en cuanto a poder interpretar intención del hablante y resolver la ambigüedad de contexto detectando intenciones, tiempo, localizaciones y sentimientos. Podemos ver visualmente como se relacionan estos términos en la **Figura 3**. Aunque ambos conceptos pueden verse solapados, se sugiere que el entendimiento del lenguaje natural puede verse como una tarea al PLN.

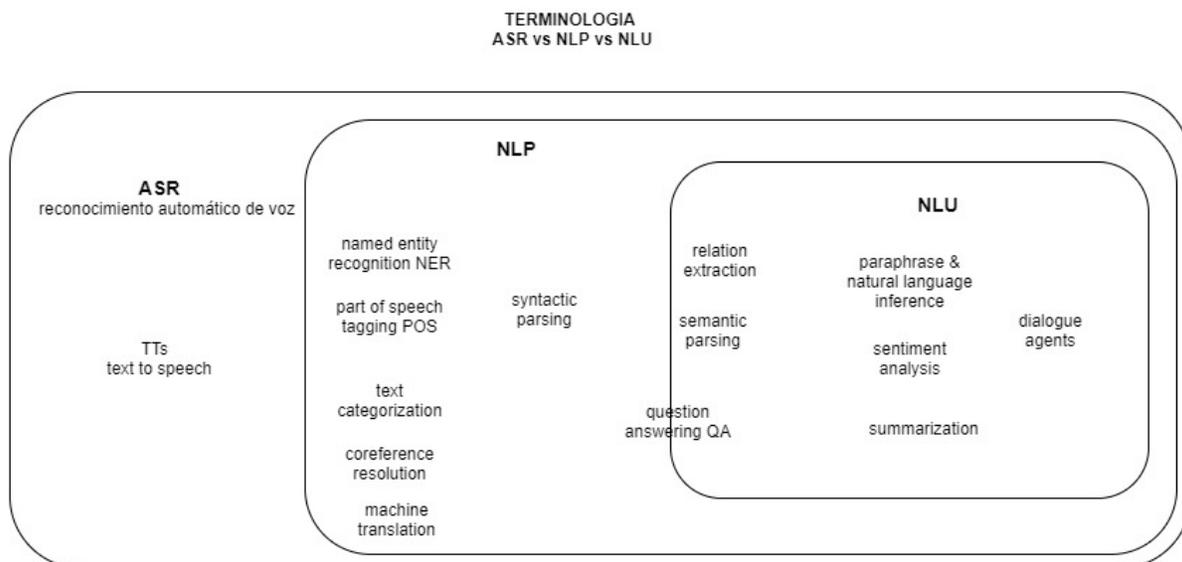


Figura 3 Diferencias entre NLU y NLP

Como hemos mencionado anteriormente, existen diversas técnicas para poder trabajar con el lenguaje natural, y hemos destacado las basadas en conocimiento experto y las basadas en modelos estadísticos.

En primer lugar, para poder trabajar con conocimiento experto, es indispensable el trabajo de los lingüistas computacionales, que escriben reglas que permiten el reconocimiento de patrones gramaticales y que se aplican para solucionar una tarea concreta, como el análisis léxico,

morfológico o la traducción automática. Sin embargo, crear y mantener estas reglas es un trabajo arduo y que requiere un mantenimiento continuo (Kumar et al., 2016, pp. 1-8).

En segundo lugar, y en contraposición a los modelos expertos, los modelos estadísticos se basan en analizar textos sobre un conjunto de ejemplos etiquetados, conocidos como corpus. A partir de medir la frecuencia de diferentes unidades lingüísticas, tales como palabras o expresiones se permiten crear modelos estadísticos y basados en aprendizaje computacional que permiten inferir conocimiento. Además, los recientes avances en minería de datos, que permiten compilar y analizar grandes volúmenes de datos, unido a las facilidades de estos métodos estadísticos, ha resultado que sean estos modelos estadísticos los más usados frecuentemente en diversas tareas relacionadas con el PLN (Bhandare, Bhide, Gokhale, & Chandavarkar, 2016, pp. 611-629).

2.6 Machine Learning y Deep Learning

2.6.1 Machine Learning.

El aprendizaje computacional o Machine Learning (en inglés) es una rama de la inteligencia artificial encargada de crear programas de software capaces de generalizar comportamientos a partir de los datos recibidos (Valencia-Garcia, 2005). Otros autores definen este concepto de la siguiente manera: “Enseñar a un computador a aprender conceptos usando datos, sin ser explícitamente programado para ello. Este campo de estudio que da a los ordenadores la habilidad de aprender sin la necesidad de ser explícitamente programados”. Se dice que un programa de ordenador aprende por medio de la experiencia E con respecto a alguna clase de tareas T y medida de rendimiento P, si su desempeño en tareas en T, medida por P, mejora con la experiencia E. Dentro del aprendizaje computacional existen tres métodos de aplicación diferenciados, los cuales se nombrarán a continuación de manera específica: Método de regresión, método de clasificación, y método de agrupación. Los algoritmos de aprendizaje computacional se pueden dividir en dos tipos: aprendizaje supervisado que para un conjunto de datos de entrada conocemos de antemano los datos correctos de salida y el aprendizaje no supervisado para un conjunto de datos de entrada, no conocemos de antemano los datos de salida (Larsson, Maire, & Shakhnarovich, 2016, pp. 577-593).

2.6.2 Clasificación de algoritmos de aprendizaje.

De manera general, podemos clasificar los algoritmos de aprendizaje automático en dos grandes grupos: supervisados y no supervisados. Por una parte, los algoritmos de aprendizaje automático supervisados se basan en aprender sobre un conjunto de ejemplos etiquetados previamente para predecir eventos futuros o para clasificar nuevos datos. En este sentido, el algoritmo produce y ajusta una función matemática capaz de hacer predicciones sobre un nuevo conjunto de valores de entrada. Además, estos algoritmos pueden comparar su salida con la salida prevista correcta y encontrar errores para modificar el modelo en consecuencia. Por el contrario, los algoritmos de aprendizaje automático no supervisados se utilizan cuando la información utilizada para entrenar no ha sido clasificada ni etiquetada previamente. En este sentido, el aprendizaje no supervisado estudia cómo un sistema puede modelar una función matemática que describa la estructura subyacente a partir de dichos datos.

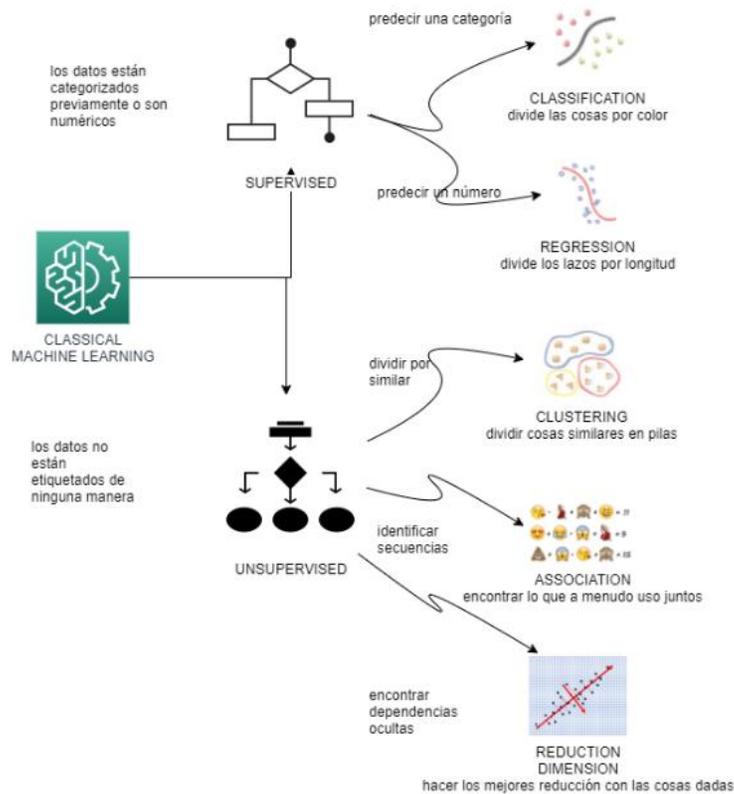


Figura 4 Clasificación típica de modelos de aprendizaje por computador

Además de estos dos tipos de aprendizaje (supervisados y no supervisados) podemos considerar también una serie de modelos mixtos, también conocidos como semi-supervisados,

en los cuales existe sólo una pequeña cantidad de los datos etiquetados; mientras que el resto no están etiquetados. Los sistemas que utilizan este método pueden mejorar considerablemente la precisión del aprendizaje. Este enfoque también es muy práctico cuando etiquetar los datos requiere una gran cantidad de recursos como lo indica la

Figura 4.

Los algoritmos supervisados de clasificación se encargan de encontrar una ecuación matemática que separe de la mejor manera posible dos o varias clases. Por ejemplo, podríamos hacer un algoritmo de clasificación que dijera si mañana va a ser un día muy caloroso o con mucho frío en base a un histórico. Por otro lado, los algoritmos de regresión tratan de generar una función matemática que describa de la mejor manera posible los datos. Siguiendo el mismo ejemplo, este tipo de algoritmos trataría de predecir cuál es la temperatura exacta de mañana, en lugar de predecirlo dentro de un conjunto de etiquetas preestablecidas como lo muestra la Figura 5.



Figura 5 Diferencias entre algoritmos de clasificación y regresión

De acuerdo con los algoritmos no supervisados, podemos encontrar también varias alternativas. Por otro lado, los algoritmos de agrupamiento o de clustering tratan de clasificar el conjunto

de datos de manera que se maximicen los elementos similares a la vez que se maximice también la distancia entre los grupos (ver la Figura 6). El agrupamiento es una técnica exploratoria de análisis de datos, que se usa para organizar información en grupos con significado sin tener conocimiento previo de su estructura. Cada grupo es un conjunto de objetos similares que se diferencia de los objetos de otros grupos. El objetivo es obtener un número de grupos de características similares. Un ejemplo de aplicación de este tipo de algoritmos puede ser para establecer tipos de consumidores en función de sus hábitos de compra, para poder realizar técnicas de marketing efectivas y “personalizadas”.

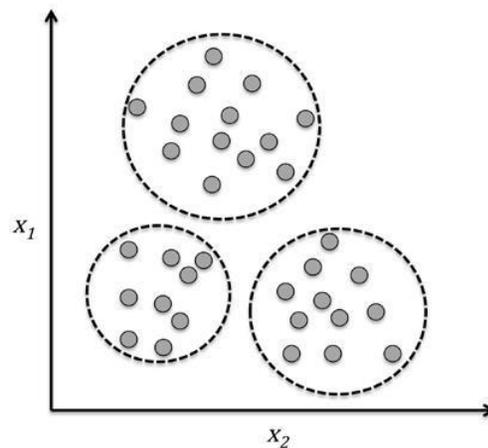


Figura 6 Ejemplo de algoritmo de agrupamiento

La reducción de la dimensionalidad son algoritmos de generalización que se encargan de encontrar cuáles son las mejores instancias de los datos y son útiles para eliminar el ruido en los datos. Es común que tengamos que trabajar con datos sobre los que podemos obtener un alto número de características, lo que supone un enorme gasto de recursos en tiempo y memoria para los algoritmos de aprendizaje. Este tipo de algoritmos nos permite encontrar una correlación entre estas las características con el objetivo de identificar y filtrar información redundante, así como el ruido.

Por último, otro tipo de algoritmo no supervisado son los algoritmos de asociación, que tratan de agrupar cuáles son los elementos que se utilizan de manera conjunta.

2.6.3 Deep Learning.

El aprendizaje profundo (Deep learning) es un tema que cada vez adquiere mayor relevancia en el campo de la IA. Siendo una subcategoría del aprendizaje automático, el aprendizaje profundo trata del uso de redes neuronales para mejorar cosas tales como el reconocimiento de voz, la visión por ordenador y el procesamiento del lenguaje natural. Rápidamente se está convirtiendo en uno de los campos más solicitados en informática. (Lecun, Bengio, & Hinton, 2015, pp. 436–444).

Curiosamente a lo que pueda parecer, el término de aprendizaje profundo se remonta hasta los años 40. Durante los primeros años se hicieron propuestas para imitar modelos basados en aprendizaje biológico que dieron lugar al modelo de Perceptrón que permitía el entrenamiento de una única neurona. Años después, este tema fue retomado a partir de la idea de retro-propagación, que es una idea que permite interconectar distintas redes con el objetivo de calcular los pesos de las neuronas situadas en las capas intermedias. Sin embargo, este tipo de cálculos era todo teórico, pues requerían grandes cantidades de computación que eran prohibitivos en aquella época. Actualmente, el aprendizaje profundo se centra en aprender a partir de múltiples niveles de composición con el objetivo de buscar sistemas capaces de superar la mente humana en tareas formales o abstractas.

La arquitectura típica de este tipo de redes puede verse en la Figura 7. Como podemos observar, las redes neuronales se basan en una arquitectura de capas jerarquizadas (como una pila) de redes neuronales artificiales conectadas entre sí y que permiten procesar datos con un enfoque no lineal. A partir de este ejemplo, podemos ver como la capa más a la izquierda de la red se llama capa de entrada (Input Layer) y la capa más a la derecha la capa de salida (Output Layer). Las capas intermedias se denomina capa oculta (Hidden Layers) porque sus valores no se observan durante el proceso de entrenamiento.

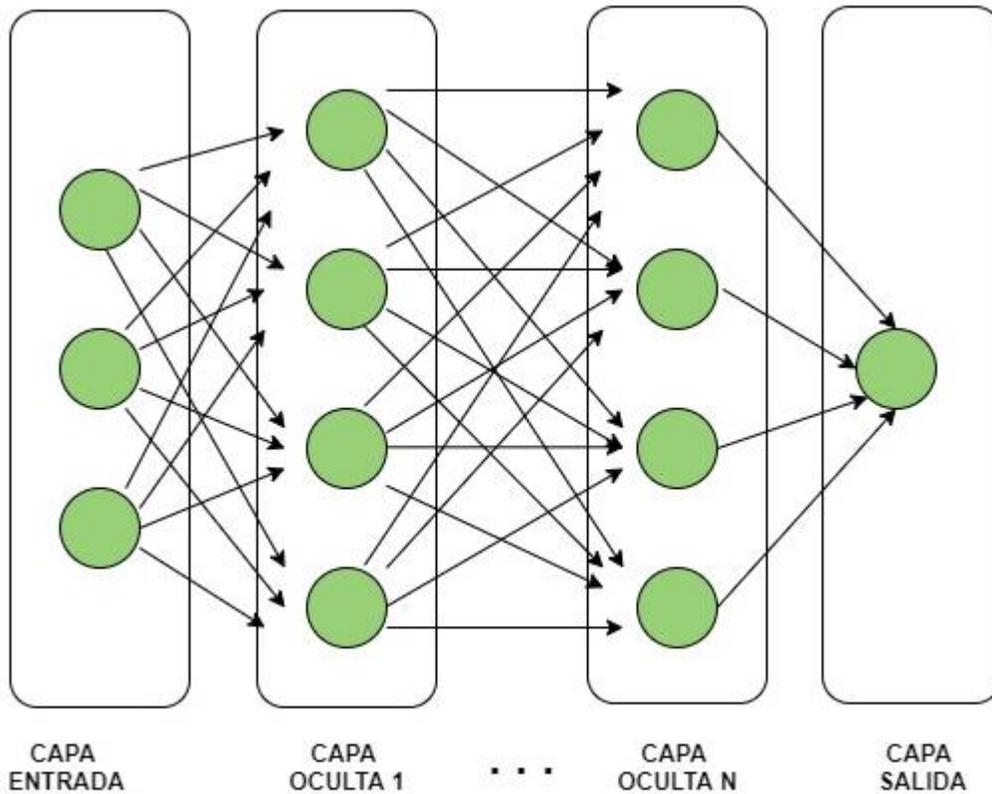


Figura 7 Arquitectura típica de una red neuronal

En concreto, la red del ejemplo es una red profunda, porque hay varias capas de entradas y salidas lo que permite usar múltiples capas de procesamiento, compuestas por múltiples transformaciones lineales y no lineales.

Conforme el modelo va aprendiendo, suele comparar la salida generada con la salida prevista. En este sentido, una vez que es capaz de ver la diferencia, lo que hace el modelo es ajustarse. Para ello, se hace uso de algoritmos de retro propagación los cuales calculan cómo se propaga el error a través de las capas de la red neuronal de forma que minimice cierta función de error.

Las redes neuronales tienen una serie de ventajas sobre los enfoques tradicionales:

- Escalabilidad. El rendimiento de las redes neuronales puede incrementarse a partir de la construcción de redes neuronales más complejas, capaces de inferir conocimiento sobre una gran cantidad de datos. Por el contrario, los enfoques más tradicionales suelen alcanzar una cota máxima en la cual no siguen aprendiendo.

- El aumento de disponibilidad de datos se debe tanto al aumento de la capacidad de los ordenadores como al abaratamiento de las tecnologías de la información. De hecho, se calcula que el 90% de los datos disponibles actualmente en el planeta se ha creado en los últimos dos años, produciéndose actualmente en torno a 2,5 quintillones de bytes por día.
- Precisión. El aprendizaje profundo logra una precisión de detección de patrones más avanzado que los sistemas tradicionales. Esto ayuda a tener mejores sistemas que se emplean en sistemas críticos, como la detección de patrones en radiografías o mejores sistemas de conducción en automóviles automáticos.
- Complejidad y flexibilidad de los modelos. La jerarquía de conceptos permite que un sistema aprenda conceptos complicados a partir de conceptos más simples. Esta idea se puede ver también con que existen distintos tipos de redes neuronales capaces de adaptarse a distintos tipos de características. Por ejemplo, las redes neuronales convolucionales son capaces de aprender sobre características que están cerca en el espacio de forma que se pueden ampliar para clasificar imágenes a partir de detectar bordes, o patrones comunes como pueden ser los ojos o la boca.
- Transferencia del aprendizaje. Los sistemas de aprendizaje profundo permiten extraer el conocimiento de capas intermedias y utilizarlo para otros dominios. Este es el éxito de algunos modelos como los word-embeddings en tareas de PLN.

Otra ventaja de estos sistemas es que algunas de las características o datos se pueden generar automáticamente durante el mismo procesamiento. Esto pasa en los modelos basados en procesamiento del lenguaje natural o en modelos de procesamiento de imágenes. Sin embargo, con los enfoques tradicionales de aprendizaje por computador, la extracción de estas características se realiza manualmente mientras que el modelo de aprendizaje profundo se resuelve por sí mismo.

También es importante añadir que los enfoques basados en aprendizaje por computador, sin que sean redes neuronales, pueden funcionar mejor en algunos casos, como por ejemplo cuando el número de características no es muy elevado.

Este tipo de sistemas está tan extendido que es posible el uso de distintos frameworks. A continuación, nombramos los más populares.

TensorFlow

TensorFlow es un framework desarrollado por Google y escrito para Python y R. Una de las ventajas de este sistema es que usa diagramas de flujo de datos para procesar datos lo que permite que, a medida que construye la red, se puede observar cómo fluyen los datos. Con TensorFlow, puedes usar TensorBoard para la visualización de datos, que es un paquete grande que generalmente pasa desapercibido. TensorBoard simplifica el proceso para mostrar datos visualmente. TensorFlow es de uso gratuito y más bastante confiable. Su principal ventaja es que hace que facilitar aplicar estos modelos.

Keras

Keras⁷ es un framework construido sobre TensorFlow, y permite facilitar todavía más el proceso de construir modelos de redes neuronales a partir de modelos secuenciales o del sistema funcional. La principal ventaja de Keras es que requiere de poco código y ofrece flujos de trabajo de backend de TensorFlow.

Al estar Keras especialmente diseñada para posibilitar la experimentación rápida con redes de aprendizaje profundo, la arquitectura sobre la que está construida se centra en ser amigable para el usuario, modular y extensible, de forma que los usuarios puedan construir e intercambiar bloques de manera sencilla. Por este motivo, Keras contiene además varias implementaciones de los bloques constructivos de las redes neuronales como por ejemplo layers, funciones objetivo, funciones de activación y optimizadores matemáticos.

Pytorch

Pytorch es un framework de Python diseñado para realizar cálculos numéricos haciendo uso de la programación de tensores. PyTorch es usado tanto para procesar los cálculos en GPU. PyTorch trabaja con grafos dinámicos, por lo que puede modificar en tiempo de ejecución el cálculo del gradiente.

A continuación, se muestra un esquema de estos tres sistemas en la Tabla 1 se describe la ficha técnica de los principales ítems que se toman en cuenta para el uso de framework dentro del aprendizaje profundo; así podemos indicar que el framework Keras es la mejor opción cuando

el dataset a procesar es relativamente pequeño, pero con unas características que su popularidad no ha escalado a los niveles de los frameworks como Pytorch o Tensorflow.

Tabla 1 Comparación de frameworks de aprendizaje profundo

Características	Keras	Pytorch	TensorFlow
Capacidades de la API	Alto	Bajo	Alto y bajo
Arquitectura	Sencilla	Compleja	Compleja
Datasets	Pequeños	Grandes	Grandes
Debug	No es necesario	Buenas capacidades	Difícil
Modelos entrenados	Sí	Sí	Sí
Popularidad	1	3	2
Velocidad	Lento	Rápido	Rápido
Desarrollado en	Python	Lua	C++, CUDA, Python

Al margen de estos frameworks, es necesario destacar el paquete de Scikit-learn de Python que es transversal a todos estos frameworks. Este paquete o módulo de Python integra una amplia gama de algoritmos de aprendizaje de máquinas de última generación para problemas de mediana escala supervisados y no supervisados. Este paquete se centra en llevar el aprendizaje automático a personas no especializadas utilizando un lenguaje de alto nivel de propósito general. Hace hincapié en la facilidad de uso, el rendimiento, la documentación y la coherencia de la API. Se considera un sistema transversal porque este paquete contiene una gran variedad de librerías relacionadas con los procesos de construcción de modelos tales como la evaluación de modelos y parámetros o la selección de conjuntos de entrenamiento basados en divisiones normales, estratificadas, o cruzadas.

2.7 Niveles del Lenguaje Natural

El lenguaje humano se puede sustentar en distintos niveles para su análisis, estos niveles se presentan entrelazados y son necesarios para la correcta interpretación de la estructura del lenguaje (Dascalu et al., 2017, pp. 495-499).

- Nivel Fonológico: se refiere de como las palabras se relacionan con los sonidos o fonemas que representan.
- Nivel Léxico. Se refiere al conjunto de información sobre cada palabra que el sistema utiliza para el procesamiento. Las palabras que forman parte de un diccionario que está representado por una entrada léxica, y en caso de que ésta tenga más de un significado o diferentes categorías gramaticales, tendrá asignada diferentes entradas.
- Nivel Morfológico: Se refiere de como las palabras se componen a partir de unas unidades de significado más pequeñas denominadas morfemas; este nivel tiene mucha relación con el nivel léxico.
- Nivel sintáctico: Se refiere al orden y la relación de las palabras comprendidas en una oración y la función que cumplen. Se analiza de como las palabras pueden unirse para formar oraciones, fijando el papel estructural que cada palabra juega en las oraciones en formas de árbol o red.
- Nivel semántico: Se refiere al significado de las palabras y de cómo los estos significados se unen para dar una nueva oración. También se refiere al significado independiente del contexto.
- Nivel contextual o discurso: se refiere a la interpretación del texto de manera que se comprenda no solo las palabras como conjuntos independientes de sentencias sino como un conjunto de oraciones interrelacionadas en función de las circunstancias.
- Nivel pragmático: Se refiere de como las oraciones se utilizan en distintas situaciones y de cómo el uso de afectan al significado de las oraciones.

En la Tabla 2 se representan las características de los niveles antes descritos detallando el nivel de procesamiento del lenguaje natural, su ámbito de acción y las características principales.

Tabla 2 Características de los Niveles del lenguaje natural

Niveles de procesamiento del lenguaje natural	Ámbito de acción	Características principales
Fonético	Sonidos	Corpus de aprendizaje Modelos acústicos
Léxico	Formas	Etiquetado POS diccionarios Lexicón computacional
Morfológico	Formas	Etiquetado POS morfemas Lexicón computacional
Sintáctico	Estructuras	Bases de datos sintácticas Árboles de decisión
Semántico	Significados	Bases de datos semánticas
Discurso	Comunicación	Bases de datos semánticas Ontologías
Pragmático	Comunicación	Bases de datos semánticas Ontologías

2.7.1 Recursos Lingüísticos en sistemas de procesamiento de lenguaje natural.

Dentro de los recursos lingüísticos en los sistemas de procesamiento de lenguaje natural se destaca el léxico que incluye la información morfológica, la categoría gramatical, irregularidades sintácticas y representación del significado. Normalmente el léxico sólo contiene la raíz de las palabras con formas regulares, siendo el analizador morfológico el que se encarga de determinar si el género, número o flexión que componen el resto de la palabra son adecuados. Existen distintos tipos de recursos para los sistemas de PLN. Entre estos recursos podemos destacar el los lexicones (Adams, Makarucha, Neubig, Bird, & Cohn, 2017, pp. 1-11). Estos lexicones se pueden clasificar en las siguientes categorías:

- Base de datos léxica: son listas de términos, junto a su significado, en un contexto específico. Por ejemplo WordNet es una base de datos léxica con especificación del idioma inglés y EuroWordNet que incluye múltiples lenguajes (Tian, Lo, & Lawall, 2014, pp. 1-8).
- Diccionario electrónico: Es una digitalización del diccionario, a la que se le incorpora una lematización automática que permite la búsqueda de palabras (He & Deng, 2015, pp. 40-45).

- Tesoros: son vocabularios en los que se representan relaciones semánticas con otras palabras y con sus significados (Duwairi, Ahmed, & Al-Rifai, 2015, pp. 1-8).
- Red semántica: es una representación que las personas tienen con respecto a un objeto en particular. Los mapas o redes semánticas, se utilizan para describir y comunicar los conceptos que los sujetos tienen incorporados en su memoria (Wulff, Hills, Lachman, & Mata, 2016, pp. 1-4).

2.8 Ingeniería de características.

La ingeniería de características es el proceso por el cual se extraen características de datos en crudo a partir de aplicar técnicas de minería de datos. Con respecto a extraer características de textos, podemos distinguir entre tres grandes bloques: (1) características estadísticas, (2) características lingüísticas, y (3) características contextuales.

2.8.1 Características estadísticas.

Las características estadísticas se basan en medir la frecuencia de ciertas palabras, caracteres o expresiones. Los modelos estadísticos más básicos se basan en representar un documento como un vector donde se cuenta la frecuencia de ciertas palabras de un vocabulario conocido. Los enfoques más básicos son el de one-hot encoding para la representación de palabras y el modelo de “bolsa de palabras” para la representación de documentos. El modelo de one-hot encoding es un sistema para representar palabras como vectores (ver *Figura 8*).



Figura 8 One-hot encoding²

Por ejemplo, si nuestro vocabulario consistiera en dos únicas palabras: contento y triste. La representación de estas mediante one-hot encoding sería [0, 1] para contento y [1, 0] para

triste. Este sistema de representación permite trabajar a un ordenador con redes neurales, ya que son sistemas que enfocados a trabajar con matrices y vectores. Sin embargo, esta manera de representación de palabras representa ciertos inconvenientes. En primer lugar, y más importante, es el problema de la carencia de semántica por el que todas las palabras son linealmente independientes de las otras. Esto es un problema porque es sensato pensar que ciertas palabras en un lenguaje tienen una semántica parecida. Por ejemplo, las palabras “gato”, “perro”, “pájaro” están conceptualmente más cercanas entre sí que con otras relacionadas con utensilios de cocina: “tenedor”, “cuchara” o “cuchillo”. En segundo lugar, está el problema de dimensionalidad, ya que conforme aprendemos y añadimos nuevas palabras a nuestro diccionario, el tamaño del vector crece linealmente. Por ejemplo, si nuestro vocabulario fuera de 10.000 palabras, la representación de una palabra aplicando one-hot encoding sería una matriz con 9.999 valores a 0; y sólo un valor a 1. Esto lleva a matrices muy dispersas y con problemas de memoria.

El modelo de bolsa de palabras es un modelo que permite representar documentos como un vector que mide la frecuencia de las palabras que lo componen y que aparecen en un determinado diccionario. Una representación gráfica de este modelo puede ver en la figura 11.

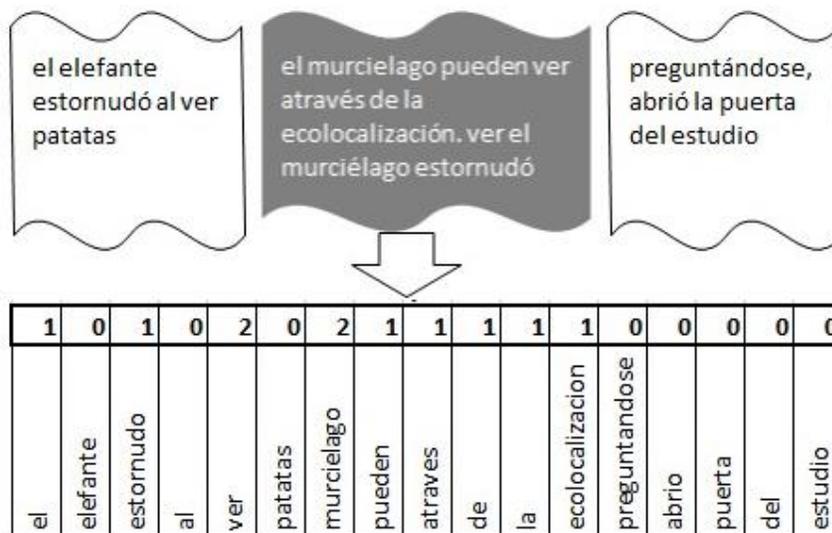


Figura 9 Modelo de bolsa de palabras

Básicamente, el modelo de bolsa de palabras consiste con contar términos que se repiten en un texto. Este modelo debe su nombre a unas estructuras de datos usadas en informática conocidas

como “bags”. Aunque es extremadamente sencillo, este modelo es ampliamente utilizado para llevar a cabo diferentes tareas de clasificación de textos. Sin embargo, los inconvenientes de este modelo son similares a los identificados para one-hot encoding. En primer lugar, el modelo no contempla el orden de las palabras por lo que éstas pierden contexto. Por ejemplo, la palabra “banco” es considerada de igual manera tanto si se refiere a una institución bancaria, a un asiento, o a un banco de peces; o bien, palabras que pierden su significado porque están dentro de frases hechas. Por ejemplo, no es lo mismo “fiebre” para referirse a un estado febril de un paciente, que “fiebre del sábado noche”, o bien la palabra “nueva” que cuando está acompañada de “york” para referirse al nombre de la ciudad. Además, si el corpus es muy grande, el número de características que tendríamos que representar para un mismo documento crecería conforme añadimos más palabras (y todavía más cuando hay palabras mal escritas en el documento) lo que causa el problema de la dimensionalidad.

Para solucionar los problemas del modelo de bolsa de palabras ha habido diferentes mejoras. En primer lugar, en lugar de contar las palabras, se pueden utilizar otra serie de métricas que tengan en cuenta potenciar palabras poco frecuentes y palabras raras a la vez que se descartan palabras que aportan poco valor a la clasificación. Para ello, algunos investigadores han propuesto usar el TF-IDF (Term Frequency-Inverse Document Frequency) que es una métrica que pondera lo frecuente que es un término con si éste aparece con frecuencia en el resto de los documentos. Este enfoque es muy bueno para restar valor a palabras con poco valor semántico, como podrían ser el uso de conjunciones; y también pone en valor a palabras que aparecen menos frecuentemente en el conjunto de documentos. Por otro lado, se ha propuesto que, en lugar de manejar las palabras de manera individual, se traten de manera conjunta en grupos de dos, tres, etc. Esto es útil porque permite tratar conjuntos de palabras con significado propio como “Nueva York”. Este enfoque se conoce como n-gramas. Incluso otros autores han propuesto el uso de character n-grams, que consiste en separar un texto en cadenas de caracteres de longitud fija, en lugar de por palabras. Este enfoque tiene la ventaja de que es más sensible a palabras que contienen errores y, por tanto, tienen letras cambiadas de orden, o bien a es tolerante con palabras inventadas o compuestas.

Como hemos podido observar, los enfoques de one-hot encoding y el de bolsa de palabras tienen el problema del contexto y la gran dimensionalidad. Un enfoque moderno para tratar

estos problemas se conoce como word embeddings, o representaciones densas de palabras, que consiste en representar las palabras como un vector de longitud fija, donde los valores de cada vector han sido aprendidos mediante algún algoritmo de predicción de la siguiente palabra.

Por ejemplo, la frase “tengo muchas ganas de que llegue MASK_WORD para ver a mi familia”, unas palabras tienen más sentido que otras, como podría ser la palabra “Navidad” sobre “Silla”. Para ello, se aplica un modelo conocido como word2vec que consiste en entrenar a una red neuronal a encontrar cuáles son las palabras más probables según un contexto. Mientras esta red neuronal se va entrenando, los valores que va tomando van acercando palabras que comparten un mismo contexto; es decir, es una representación donde palabras semánticamente relacionadas están conceptualmente cerca.

Una propiedad interesante es que si representamos las palabras dentro del espacio (por ejemplo, en un espacio tridimensional) podemos observar que ciertos grupos de palabras mantienen una distancia similar con otra. Por ejemplo, la distancia entre “hombre” y “mujer” es similar a la distancia entre “rey” y “reina”.

La Figura 10 contiene un ejemplo visual de cómo estos word embeddings pueden visualizarse en un espacio tridimensional y de cómo las relaciones se mantienen constantes según características semánticas y de taxonomía.

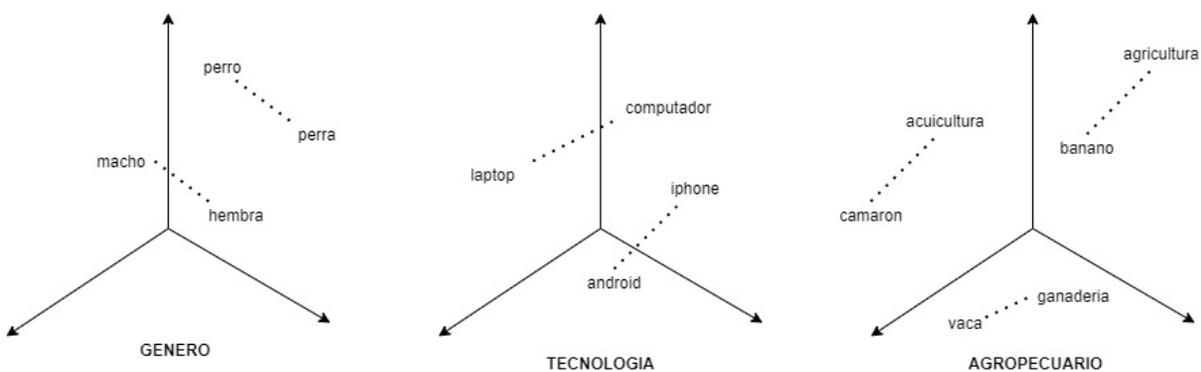


Figura 10 Representación de word embeddings en un espacio tri-dimensional

Además de word2vec hay otras propuestas para aprender word embeddings, tales como GloVe (Dev, Li, Phillips, & Srikumar, 2020) o fastText (Salur & Aydin, 2020). Por un lado, fastText lo que hace es aprender los valores de las palabras a partir de dividir una palabra en trozos más

pequeños, por lo que tiene en cuenta sufijos, prefijos, y otras partes de la palabra. GloVe, por otro lado, trata de formar una matriz de coocurrencia e intentar representar mejor las palabras análogas.

Una de las características más interesantes de los word embeddings y las redes neuronales es la capacidad de transferir el aprendizaje entre tareas. Esto quiere decir que es posible utilizar una red neural para aprender una representación de las palabras a través de utilizar grandes corpus no etiquetados con textos, tales como enciclopedias o redes sociales. Esto se conoce como word embeddings pre entrenados y, actualmente existen distintos recursos a partir de distintos modelos y distintas fuentes de datos tales como enciclopedias o webs de noticias.

2.8.2 Características lingüísticas.

Las características lingüísticas incluyen información experta relacionada con la lingüística para poder clasificar un texto. Por ejemplo, se pueden contabilizar verbos, pronombres, adverbios, frases hechas, o marcadores del discurso. También es posible incluir frases hechas, o características de estilometría para identificar el uso de lenguaje arcaico.

Existen herramientas como Linguistic Inquiry Word Count 2015 (LIWC2015) (Dudău & Sava, 2020, pp. 1-19) (0) que permiten obtener este tipo de variables. Específicamente, LIWC es la herramienta de-facto ya que ha sido validada sobre distintos dominios, tales como el análisis de notas de suicidio, ciber-acoso escolar, detección de la sátira y que cuenta las palabras en distintas categorías, incluyendo palabras y expresiones que miden qué y cómo se expresan los individuos.

Una ventaja de las características lingüísticas es que permiten generar modelos altamente interpretables, pues el reducido número de características permite deducir qué combinaciones de características tienen peso a la hora de obtener determinado sentimiento. El problema de este tipo de características es que son difíciles de obtener y mantener.

Otro ejemplo de herramientas para la extracción de características a partir de texto es la proporcionada por UMUTextStats que ha sido utilizada en distintas tareas de clasificación de texto (García-Díaz, Cánovas-García, Colomo-Palacios, & Valencia-García, 2021, pp. 506-518) (García-Díaz, Cánovas-García, & Valencia-García, 2020, pp. 43-57).

2.8.3 Características contextuales.

Por último, es posible extraer características relacionadas con el contexto de la conversación e información de los interlocutores. Por ejemplo, si estuviéramos extrayendo información desde Twitter, podríamos mirar la fecha en la que se escribió el comentario, ver si la cuenta está verificada o no, si es un mensaje nuevo o la respuesta a otro, etc. De acuerdo Sánchez- Rada (Sánchez-Rada & Iglesias, 2019, pp. 7-13), las características contextuales pueden a su vez subdividirse en: micro características, en el caso de incluir información sobre el autor de este (edad, género) y en conversaciones previas. (2) meso características, en el caso de incluir características que midan como los usuarios interactúan entre sí, y, por último (3) macro características, que incorporan información de fuera de la red social o entorno de la conversación.

Las ventajas de aplicar este tipo de características es que permiten trabajar de mejor manera con el sarcasmo o la ironía, o cualquier forma en el que para entender un texto sea necesario un contexto del exterior. El problema es que estas características son difíciles de extraer y no están siempre disponibles.

2.9 Modelos de aprendizaje profundo para el procesamiento del lenguaje natural

Existe una tendencia de modelos de aprendizaje profundo aplicados en el campo del procesamiento del lenguaje natural (PLN). Las redes neuronales profundas (DNN) están formadas por redes neuronales artificiales que tienen múltiples capas ocultas entre la capa de entrada y la capa de salida. Esta sección proporciona una discusión detallada sobre algunos de los modelos de aprendizaje profundo más populares.

2.9.1 Redes neuronales convolucionales (CNN).

Las CNN pertenecen a la clase de redes neuronales y han mostrado un éxito e innovación significativos en la visión por computadora y el procesamiento de imágenes. La arquitectura fundamental de CNN se muestra en la Figura 11. Como es evidente en la figura, CNN consta de varias capas, como la capa de entrada, la capa convolucional, la capa de agrupación y la capa completamente conectada. La tarea de la capa de entrada es tomar el valor de píxel de la

imagen como entrada. A continuación, la capa de convolución tiene la responsabilidad de producir resultados basados en sus valores de kernel o filtro. La salida obtenida mediante una operación de convolución y Pooling Layer se utiliza para reducir el tamaño de la representación (dimensionalidad) y acelerar el cálculo.

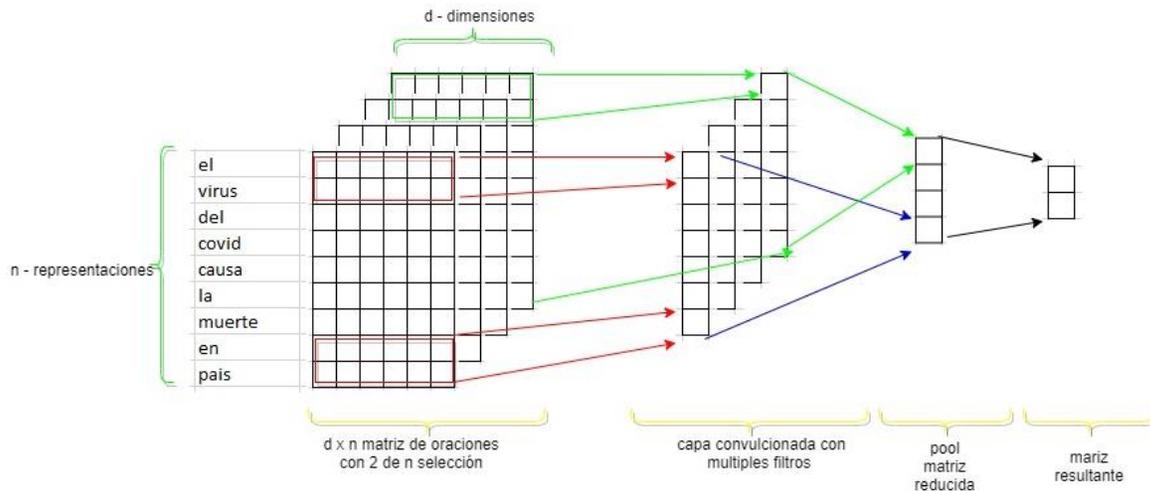


Figura 11 Redes neuronales CNN

El tipo más popular de agrupación es la agrupación máxima, en la que se toma el valor máximo de cada ventana. La capa totalmente conectada conecta todas las neuronas de esta capa con todas las activaciones de la capa anterior, como se ve en las redes neuronales ordinarias. Cada vez más investigadores utilizan activamente las CNN en el campo del análisis de sentimientos. El modelo más popular de CNN para la clasificación de sentimientos a nivel de oración es el trabajo realizado por Kim (Liu et al., 2019, pp. 1091-1107). El autor realizó un experimento con CNN construido sobre word2vec previamente entrenado. Los resultados experimentales muestran que los vectores entrenados previamente pueden servir como un excelente extractor de características para tareas relacionadas con la PNL mediante el aprendizaje profundo.

2.9.2 Word embedding.

Inspiradas por las ideas, la arquitectura y los resultados de las CNN en el campo de la visión por computadora, las CNN también están ganando popularidad en el dominio de la PLN. En las tareas relacionadas con la PLN, una capa de entrada consiste en la representación matricial de oraciones o documentos, en lugar de los píxeles de la imagen. Cada fila de la matriz es una

representación vectorial de una palabra o un carácter. Estos vectores se denominan incrustación de palabras o incrustación de caracteres. El enfoque anterior utilizado para representar el vocabulario de un documento era la codificación one-hot. El problema con este enfoque es que el tamaño del vector aumenta con el tamaño del cuerpo. Además, esta codificación no puede capturar la relación entre palabras. Por lo tanto, las incrustaciones de palabras se desarrollaron como una de las técnicas más populares para representar el vocabulario de un documento. Contienen un conjunto de métodos de selección de características o un conjunto de modelos de lenguaje que mapea la palabra textual en sus representaciones vectoriales densas y de baja dimensionalidad equivalentes. Pueden capturar el contexto de la palabra y pueden proporcionar información sobre la relación de una palabra con otras palabras. Por lo tanto, el significado de una palabra se puede predecir con precisión, ya que puede capturar información sintáctica y semántica sobre las palabras.

Word2Vec (Mikolov (Mikolov, Chen, Corrado, & Dean, 2013, pp. 1-9)) es una de las técnicas famosas para aprender la inserción de palabras. ya que utilizan una red neuronal superficial para procesar un texto antes de pasarlo a un algoritmo de aprendizaje profundo. Las incrustaciones se pueden obtener utilizando el modelo Skip Gram y el modelo Common Bag of words (CBOW). El modelo CBOW predice la palabra actual de los alrededores palabras de contexto, mientras que el modelo Skip Gram predice las palabras de contexto circundantes a partir de la palabra actual. Las palabras se mapean en una matriz de palabras y se convierten en vectores en un espacio vectorial n-dimensional al representar palabras similares cerca unas de otras.

Por otro lado, en lugar de aprender a incrustar la palabra completa, char2vec (Cao y Rei (Cao & Rei, 2016, pp. 1-9)) puede aprender a incrustar asociado con cada carácter de una palabra. Más recientemente, muchos investigadores están ideando nuevos enfoques sobre la representación de palabras para el análisis de sentimientos. Los métodos tradicionales de incorporación de palabras aprenden distribuciones de palabras que son independientes de cualquier tarea específica. Para el análisis de sentimientos, esto se puede superar utilizando el conocimiento previo que está disponible en forma de etiquetas de sentimientos o palabras obstinadas de léxicos de sentimientos.

Yu et al. (L. C. Yu, Wang, Robert Lai, & Zhang, 2018, pp. 5-9) propuso un modelo que aprende las incorporaciones de sentimientos mediante el uso de puntuaciones de intensidad de sentimientos de léxicos de sentimientos. Esto mejora los vectores de palabras, ya que están semántica y sentimentalmente más cercanos a palabras similares.

2.9.3 Redes neuronales recurrentes (RNN).

RNN es una variante de Rec NN y se utiliza para modelar los datos secuenciales. Los datos secuenciales se aplican en una variedad de aplicaciones. Por ejemplo, en la traducción de un idioma, una secuencia de oración se traduce de un idioma a otro, en el reconocimiento de voz, un clip de audio (secuencia que se reproduce durante un tiempo) se asigna a un guion de texto (secuencia de palabras), y en el reconocimiento de actividad de video, La secuencia de fotogramas de video se convierte en texto que describe la actividad mostrada en el video. La principal diferencia entre RNN y Rec NN es que, a diferencia de Rec NN, RNN considera el factor de tiempo para procesar los elementos en una secuencia. Por lo tanto, la salida en RNN depende no solo de la entrada actual, sino también de la salida calculada a partir del estado previamente oculto de una red. RNN almacena los estados internos de las entradas procesando cada palabra en una oración de forma recurrente.

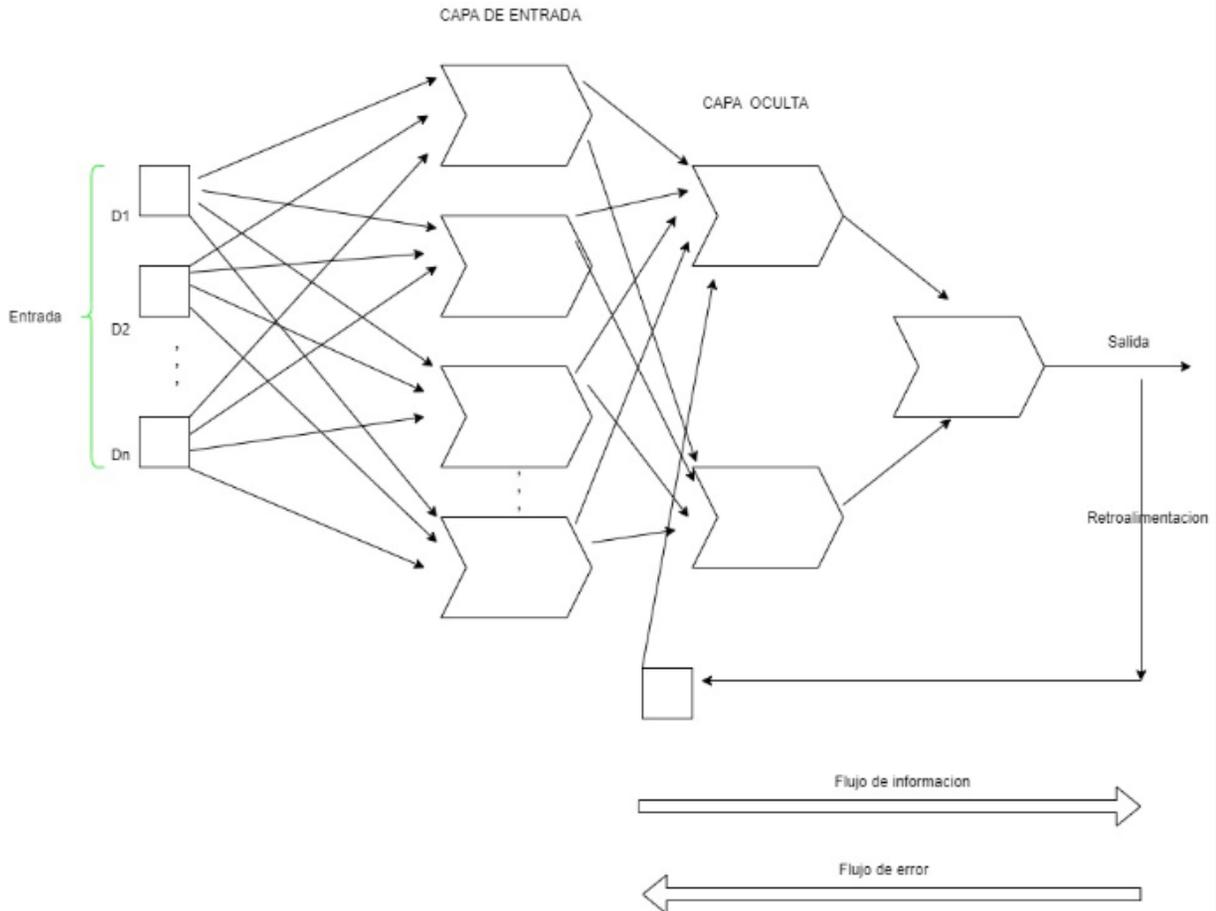


Figura 12 Redes neuronales RNN

Figura 12 muestra una RNN con un bucle que conserva toda la información, la versión desenrollada indica que hay múltiples copias de RNN conectadas y se comunican pasando información de un estado a otro. Por lo tanto, RNN con un bucle significa que está compuesto por múltiples copias de RNN. El estado actual se calcula utilizando el valor del estado anterior y la entrada actual. Los RNN son muy populares para la clasificación de sentimientos, donde la red clasifica un fragmento de texto en varias tareas de clasificación de sentimientos. Las variantes de RNN de uso común son LSTM, GRU, RNN bidireccionales y RNN profundos.

2.9.4 Memoria a corto plazo (LSTM).

LSTM es una de las variantes más populares de RNN, que posee la capacidad de manejar el problema del gradiente de desaparición en RNN estándar y puede detectar dependencias a largo plazo. Esto los hace más poderosos y flexibles. La arquitectura de LSTM se muestra en la **Figura 13**.

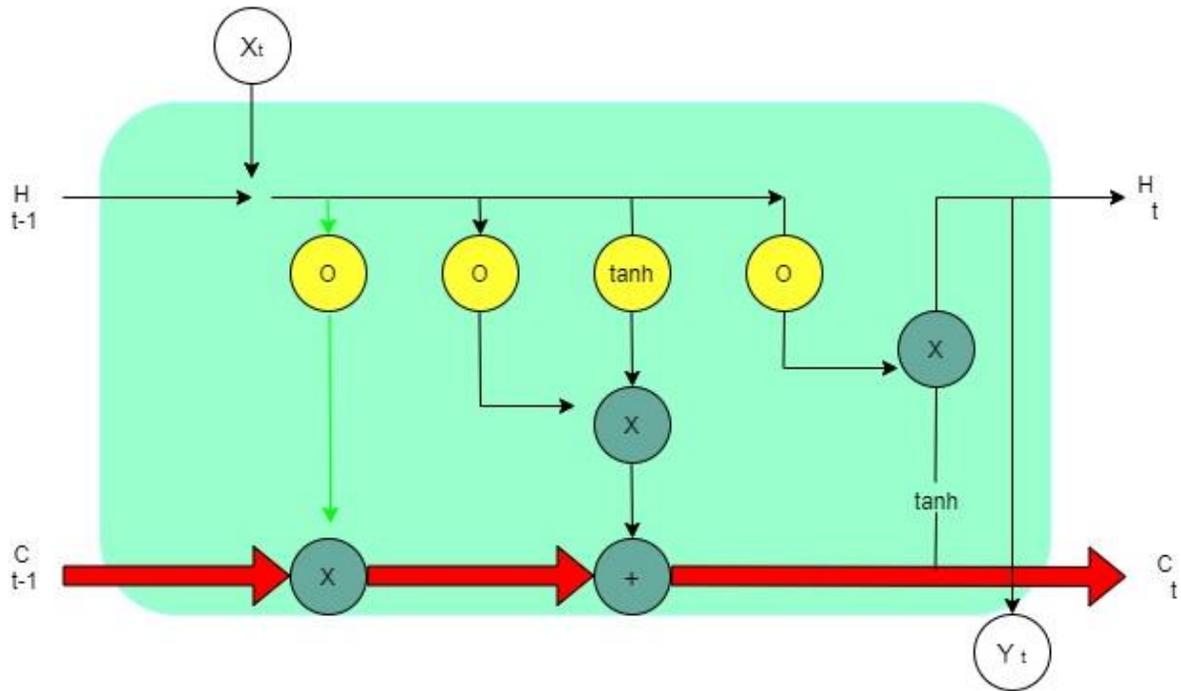


Figura 13 Arquitectura LSTM

Los investigadores también están combinando varias técnicas de aprendizaje profundo para la clasificación de sentimientos. Huang y col. (Huang, Chen, Zheng, & Dong, 2017, pp. 12-17) combinó dos redes populares de aprendizaje profundo llamadas LSTM y CNN (**Figura 14**) y propuso una arquitectura que se compone de una capa de CNN y dos capas de LSTM apiladas en CNN. Del mismo modo, Hassan y Mahmood (2017b) propuso una arquitectura llamada ConvLstm, que nuevamente combinó CNN y LSTM para clasificar textos cortos en la parte superior de word2vec.

Existe una variante de este algoritmo que se denomina BiLSTM (Bi-direccional Long Short-Term Memory), que es una combinación de LSTM hacia adelante y LSTM hacia atrás. BiLSTM alimenta el algoritmo de aprendizaje con los datos originales una vez de principio a fin y otra vez de fin hacia el principio. Esto hace que el algoritmo aprenda más rápido en comparación con el enfoque unidireccional.

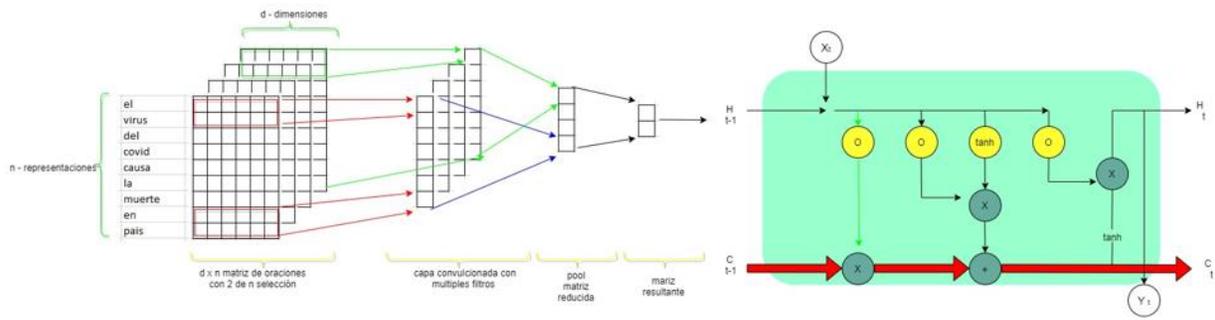


Figura 14 Arquitectura combinada de LSTM y CNN

2.9.5 Unidades recurrentes cerradas (GRU).

Los GRU son variantes de LSTM y se consideran LSTM sin puerta de salida. Se ocupan de dos tipos de puertas: actualizar la puerta y restablecer la puerta. La arquitectura de GRU se muestra en la Figura 15.

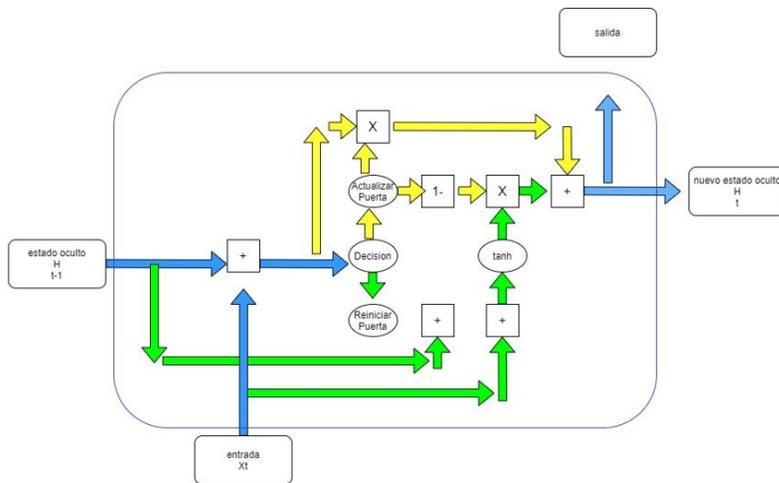


Figura 15 Arquitectura GRU.

Las redes GRU basadas en la atención también se aplican a varias tareas de análisis de sentimientos, como la clasificación de sentimientos basada en objetivos. Zhang y col. (Z. Zhang, Wang, Zou, & Gan, 2018, pp. 5-13) propuso un enfoque para modelar la clasificación del sentimiento objetivo en un sistema de preguntas y respuestas utilizando Dynamic Memory Networks (DMN). El DMN constaba de cuatro módulos: módulo de memoria que consta de un

módulo de entrada que codifica la oración de entrada y utiliza Bi-GRU de una sola capa para aprender la semántica de cada palabra, el módulo de preguntas enumera las preguntas de acuerdo con diferentes tipos de objetivos y el módulo de respuesta muestra el sentimiento expresado hacia el objetivo.

Una Bi-GRU (Bidireccional GRU) es una variante del algoritmo de procesamiento secuencia que consta de dos GRU, uno toma la entrada hacia adelante y el otro hacia atrás. Es una red neuronal bidireccional recurrente con solo las puertas de entrada y de olvido.

2.10 Técnicas de clasificación del procesamiento del lenguaje natural

En esta sección se describen algunos de los trabajos más importantes sobre la clasificación del procesamiento del lenguaje natural relacionados con esta tesis doctoral.

2.10.1 Clasificación del procesamiento del lenguaje natural a nivel de documentos.

Este nivel de clasificación toma todo el documento como una unidad primaria de información que se centra en un tema u objeto. El documento se clasifica además en polaridad positiva o polaridad negativa. Por lo tanto, se puede generar un sentimiento general de texto. Yang y col. (Z. Yang et al., 2016), propusieron un modelo de red de atención jerárquica que se centra en el contenido vital para construir la presentación del documento. Los resultados experimentales basadas en texto demuestra que el modelo superó los resultados del estado de la técnica por una margen superior, ya que puede capturar los conocimientos sobre la estructura del documento.

Un desafío importante es la clasificación de sentimiento a nivel de documento es el modelo de textos largos para generar relaciones semánticas entre oraciones. Este problema fue manejado por Huang et al. (Rao, Huang, Feng, & Cong, 2018, pp. 21-30), quien propuso un modelo llamado SR-LSTM en el que la primera capa usaba LSTM para aprender los vectores de oraciones, y la segunda capa codifica las relaciones entre las oraciones.

Ghosh et al (Ghosh, Ravi, & Ravi, 2016, pp. 1-6), proponen un enfoque híbrido de RBM y red neuronal probabilística(PNN), en el que se utiliza RBM para la reducción de dimensionalidad y PNN realiza una clasificación de sentimiento. El experimento se lleva a cabo en cuatro pasos:

El primero, se recopilan datos de varios dominios que contienen reseñas sobre películas, libros entre otros elementos; luego, los datos se procesan previamente mediante tokenización, derivación y eliminación de palabras vacías. El siguiente paso, incluye la reducción de la dimensionalidad en el conjunto de datos utilizando RBM y, para terminar, se utiliza PNN para la clasificación binaria de sentimientos.

2.10.2 Clasificación de procesamiento del lenguaje natural a nivel de oraciones.

La desventaja de la clasificación de sentimiento a nivel de documento es que es difícil extraer la polaridad o sentimiento acerca de distintas entidades por separado. Por lo tanto, en el nivel de la oración un sentimiento se clasifica en tipo subjetivo o tipo objetivo. Un tipo subjetivo es una declaración expresa de una opinión hacia una entidad. Por ejemplo, “Tengo una hermosa esposa”, significa de polaridad positiva sobre esposa. Por lo tanto, se considera como una declaración subjetiva que puede clasificarse en diferentes polaridades. Desde otro punto, las declaraciones fácticas se denominan declaraciones objetivas. Una declaración como “la botella es de color blanco”. No muestra ningún sentimiento, por lo que se clasifica como una declaración objetiva.

Zhao y Col. (Zhao et al., 2018, pp. 205-213), propuso un marco llamado Deep Embedded débilmente supervisado (WDE), que emplea calificaciones de revisión para capacitar a un clasificador de sentimientos. Utilizado CNN para construir WDE-CNN y LSTM para construir WDE-LSTM para extraer vectores de características de oraciones de revisión. El modelo se evaluó en un conjunto de datos de Amazon de tres dominios. La precisión obtenida en el modelo WDE-CNN fue de 87.7% y en el modelo WDE-LSTM fue de 87.9%, o que muestra que los modelos de aprendizaje profundo ofrecen la mayor precisión en comparación con los modelos de referencia.

Xiong y Col (Xiong, Lv, Zhao, & Ji, 2018, pp. 2459-2466), desarrollaron un modelo llamado word-embeddings enriquecidas con sentimiento multinivel (MSWE), que utiliza un perceptrón multicapa (MLP) para modelar información de sentimiento a nivel de palabra y CNN para modelar información de sentimiento a nivel de tweet. El modelo aprende word-embeddings

específicas de sentimiento y SVM se utiliza para la clasificación de sentimientos. Se procesó en el conjunto de datos SemEval2013 y el conjunto de datos de Twitter sensible al contexto (CST), que son los conjuntos de datos de referencia para la tarea de clasificación de sentimientos. El F1 score obtenido en el conjunto de datos SemEval2013 fue 85.75% y el conjunto de datos CST fue de 81.34%.

2.10.3 Clasificación del procesamiento del lenguaje natural a nivel de aspectos.

El análisis de sentimiento a nivel de aspecto se denomina comúnmente análisis de sentimientos basado en características o análisis de sentimiento basados en entidades. Esta tarea de análisis de sentimiento incluye la identificación de características o aspectos en una oración y categorizar las características como positivas o negativas. Peng et. al. (Peng, Ma, Li, & Cambria, 2018, pp. 167-176), estudiaron los objetivos de aspecto en chino en tres niveles de granularidad: radical, carácter y palabra al proponer un modelo llamado “Modelo de secuencia de objetivo de aspecto para a granularidad única (ATSM-S)”. El trabajo anterior estaba relacionado con el procesamiento de solo un aspecto a la vez, por lo que abordaron este tema y presentaron un enfoque para procesar dos aspectos a la vez centrándose en el aspecto objetivo en sí.

Últimamente, los mecanismos LSTM basados en redes recurrentes que se están utilizando para el análisis de sentimiento basado en aspectos. Wang y col (Wang, Huang, Zhao, & Zhu, 2016, pp. 1-10), propuso un modelo LSTM basado en la atención, que puede centrarse en diferentes partes de una oración cuando se trata de varios aspectos. Los pesos de atención se calculan concatenando en diferentes partes de una oración cuando se trata de varios aspectos. Los pesos de atención se calculan concatenando el vector de aspecto en la representación oculta de la oración (modelo AE-LSTM) o agregando el vector de aspecto incrustando en cada vector de entrada de palabra (modelo ATAE-LSTM). Los resultados demuestran que ambos modelos lograron un rendimiento superior sobre los modelos de referencia, lo que demuestra que los modelos LSTM basados en la atención mejoran el rendimiento. Yu et. al. (J. Yu, Jiang, & Xia, 2019, pp. 2111-2124), propusieron un marco de trabajo que utiliza Bi-LSTM y redes de

atención multiplicada para la extracción de términos de aspectos y opinión. Al-Smai y aspecto (OTEL), junto con LSTM basado en aspecto, donde las OTE de aspecto se tratan como expresión de atención para clasificación de polaridad de sentimiento de aspecto.

El enfoque más reciente es el propuesto por Yang et al (C. Yang, Zhang, Jiang, & Li, 2019, pp. 463-478), que reemplaza los modelos de atención convencionales con un mecanismo de atención de capa al introducir una red Coattention-LSTM que puede modelar la atención a nivel de contexto y a nivel de objetivo de forma alternativa al aprender las representaciones no lineales del objetivo y el contexto simultáneamente. Desde otro punto, el modelo propuesto puede extraer características de sentimiento más eficiente para el análisis de sentimiento basado en aspectos.

2.10.4 Clasificación del procesamiento del lenguaje natural multidominio.

La palabra dominio se refiere a un conjunto de documentos relacionados con un tema específico. La clasificación de sentimiento multidominio se centra en la transferencia de información de un dominio al siguiente. Los modelos se entrenan primero en el dominio de la fuente y luego, el conocimiento se transfiere y se explora en otro dominio. Dragoni y Petrucci (Dragoni & Petrucci, 2017, pp. 186-197), incorporaron word-embeddings con un modelo de aprendizaje profundo para implementar una herramienta NeuroSent para construir un modelo de sentimiento multidominio. Yuan y col. (Yuan, Wu, Wu, Liu, & Huang, 2018, pp. 1-10), propuso un modelo de atención de dominio (DAM) para modelar las tareas a nivel de característica utilizando un mecanismo de atención para la clasificación de sentimientos de múltiples dominios. DAM se compone de dos módulos: módulo de dominio y módulo de sentimiento. El módulo de dominio predice el dominio al que pertenece el texto usando bi-LSTM, y el módulo de sentimiento selecciona las características importantes relacionadas con el dominio usando otro bi-LSTM con mecanismo de atención. El vector así obtenido del módulo de sentimiento se alimenta a un clasificador para predecir la polaridad de los textos.

2.10.5 Clasificación del procesamiento del lenguaje natural multimodal.

Las personas expresan sus sentimientos u opiniones de diferentes maneras. Anteriormente, el texto se consideró como el medio principal para expresar una opinión. Esto se conoce como enfoque unimodal. Con el avance de la tecnología y la ciencia, las personas ahora están

cambiando hacia modalidades visuales (videos, imágenes o clips) y de audio (habla) para expresar sus sentimientos. Combinar o fusionar más de una modalidad para detectar la opinión se conoce como análisis de sentimiento multimodal. Por lo tanto, los investigadores ahora se están enfocando en esta dirección para mejorar el proceso de clasificación de sentimientos.

Chen et. al. (Chen, Ji, Su, Cao, & Gao, 2018, pp. 1634-1646), plantearon un aprendizaje profundo multimodal débilmente supervisado (WS-MDL) para predecir sentimientos multimodales para tweets. El modelo usa CNN y Dynamic CNN (DCNN) para calcular puntajes de predicción multimodal y puntajes de consistencia de sentimiento. Debido a la enorme cantidad de datos disponibles en las redes sociales en diferentes formas, como videos, audios, fotos para expresar sentimientos en las plataformas de redes sociales, el enfoque convencional para el análisis de sentimientos basado en texto se convirtió en modelos compuestos de análisis de sentimientos multimodal. Por lo cual, extraer las opiniones expresadas en diferentes modalidades se convirtió en un enfoque crucial.

Poria et. al. (Poria, Cambria, Howard, Huang, & Hussain, 2016, pp. 50-59), propusieron una metodología novedosa para fusionar la información afectiva extraída de las modalidades auditiva, visual y textual. Discutieron cómo las diferentes modalidades se combinaron para mejorar el proceso general de análisis de sentimientos. exploró tres arquitecturas de aprendizaje profundo para la clasificación de sentimientos unimodal, bimodal y multimodal (trimodal). Los resultados experimentales mostraron que los modelos bimodales y trimodales han mostrado una mejor precisión en comparación con los modelos unimodales, lo que muestra la importancia de utilizar características de todas las modalidades para mejorar el rendimiento de los modelos de análisis de sentimientos.

2.11 Análisis del lenguaje subjetivo

En la teoría del conocimiento la subjetividad es la propiedad de las percepciones, argumentos lenguajes basados en el punto de vista del ser humano (Slaby & Gallagher, 2015, pp. 179–193). El factor opuesto es la objetividad que está basada en punto de vista verificables.

En la sociología la subjetividad indica de la acción y representación del ser humano siempre condicionados a una circunstancia (Thompson, 2014). En la filosofía la subjetividad se refiere a la interpretación de la existencia de dos factores distintos sobre un mismo tema o cosa

(McLachlan & Garcia, 2015, pp. 195-210). En la tecnología social el ser humano como unidad de análisis (mente + cuerpo), está enfrentado con los diferentes temas de la sociedad; donde su uso puede ser visto o percibido para diferentes fines (Bury, 2017, pp. 627-642).

La subjetividad es una propiedad opuesta a la objetividad. Mientras que la primera se sustenta en opinión y los intereses propios del sujeto, la objetividad implica tratar a los conceptos como si fuesen cosas de manera distantes y con menor implicación personal posible. Así aquellos términos que explican o detallan una opinión son subjetivos; los que intentan limitarse a los datos concretos y facticos son objetivos (Gelman & Hennig, 2017, pp. 15-23).

En la lingüística existen los conceptos de objetividad y subjetividad para referirse a la intencionalidad de un determinado mensaje (Krawczak, 2016, pp. 705-727). Para identificarlo se realiza un análisis sobre las modalidades oracionales que han sido utilizadas en el mismo. La objetividad se vincula con un significado denotativo u objetivo, el cual es común a todos los hablantes y cuenta con una estructura tangible que permite un entendimiento exacto de lo que se dice. La subjetividad está relacionada con un significativo connotativo que responde a experiencias emocionales, ya sean positivas o negativas; En la comprensión de ese mensaje dependerá de quien lo exprese, en qué situación se encuentre y a quién va dirigido. Es un elemento que resalta en el lenguaje coloquial como el literario (Munro & Hardie, 2019, pp. 411-427).

Entre las características del mensaje subjetivo se encuentran las siguientes: (1) la modalidad oracional: que se utilizan en oraciones exclamativas, desideratas para expresar afectividad (Murdock, Horissian, & Crichlow-Ball, 2017, pp. 1-9); (2) el léxico valorativo: preponderancia o vocablos afectivos, verbos imperativos o que indiquen posibilidad, expresiones que manifiestan una duda y complementos que expresen una psicología ideológica (Rocklage, Rucker, & Nordgren, 2018, pp. 1327-1344); (3) recursos expresivos: que utilizan redundancia, interrogaciones de tipo retorico, palabras malsonantes, insultos, hipérbolas y eufemismos (Maiese, 2016, pp. 1-8); o (4) signos de puntuación: uso de puntos suspensivos con el objetivo de crear duda o expectativas, paréntesis en el lenguaje escrito para realizar aclaraciones y comillas para marcar una palabra que ha sido dicha con ironía (Smithers & Eaton, 2017, pp. 1-5).

2.12 Análisis de sentimientos.

La minería de opiniones, o análisis de sentimiento, se basa en aplicar técnicas de PLN, lingüística computacional y minería de textos con el objetivo de identificar la información subjetiva en un texto. El auge de la Web 2.0. y las redes sociales, que han dado a los usuarios un rol principal en cuanto a la creación de contenidos, ha descubierto a empresas e investigadores un nicho de datos sobre el que poder trabajar para obtener información y conocimiento que se puede aplicar a distintos dominios. Por ejemplo, mediante descubrir la opinión generalizada de productos, de su imagen de marca, o bien, de cómo se perciben servicios prestados por la empresa (Weiss, Indurkha, Zhang, & Damerau, 2005, pp. 16-24).

La Minería de Opiniones puede definirse formalmente como obtener de un texto una tupla formada por: (1) el sentimiento, (2) la característica concreta a la que alude dicho sentimiento, (3) la persona que expresa dicho sentimiento, y (4) el momento en el que el sentimiento ha sido expresado. Sin embargo, no todos los problemas requieren un fino nivel detalle y en algunos estudios esta tupla se simplifica eliminando algunos de estos elementos. Por ejemplo, hay estudios donde no nos interesa saber con detalle ni el tiempo ni la persona, sino simplemente el sentimiento en sí. En este sentido, podemos distinguir entre análisis de sentimientos orientado a documentos, cuando la polaridad se extrae como un todo, o bien análisis de sentimientos orientado a aspectos, cuando cada sentimiento se asocia a determinados aspectos o características del dominio evaluado. El análisis de sentimientos orientado a aspectos o características (ABSA, por sus siglas en inglés) es muy útil para evaluar análisis de usuarios en revistas especialidades o en catálogos de productos, donde los usuarios suelen evaluar cada característica por separado. Por ejemplo, cuando se evalúa un teléfono móvil, la opinión subjetiva del usuario sobre la pantalla y la duración de la batería pueden ser contradictorios (García-Díaz et al., 2018, pp. 229-239).

Para extraer el sentimiento de un determinado texto existen dos enfoques principales. Por un lado, algunos investigadores hacen uso de lexicones de sentimientos, que se basan en identificar dentro del texto palabras que expresan un sentimiento de manera general. Podemos pensar en aquellas palabras como “bueno”, “malo”, “sensacional” o “decepcionante”. Existen recursos que contienen listados de estas palabras, como Sentiwordnet (Esuli & Sebastiani, 2015, pp. 6-13), que tienen diccionarios de términos con una polaridad asociada a cada término. Una vez

identificadas dichas palabras, se pueden sumar todas para calcular la polaridad del documento, o bien buscar cuáles son las palabras cercanas para poder asociar esos sentimientos a aspectos en lugar del documento completo. Sin embargo, el uso de lexicones tiene inconvenientes. Por un lado, no funciona correctamente cuando se usa lenguaje figurado, como podría ser la ironía. Tampoco funciona bien en algunos dominios específicos donde algunas palabras pueden cambiar de polaridad, como pueden ser el uso de negaciones. Para solucionar estos problemas, existen técnicas que se basan en aplicar métodos estadísticos y de aprendizaje computacional. En este sentido, lo que se hace es extraer una serie de características de los textos y tratar de construir un clasificador automático a partir de un conjunto de ejemplos ya etiquetados. Debido al auge de este enfoque y el gran número de técnicas posibles, éstas se van a describir en las siguientes subsecciones en detalle (Luna-Aveiga et al., 2018, pp. 229-239).

2.13 Modelos populares para la clasificación de sentimiento

Los modelos populares aplicados a las tareas de SA se pueden resumir como:

- Para la clasificación de opiniones a nivel de documento, CNN seguido de LSTM ha mostrado más precisión en los distintos conjuntos de datos.
- Para la clasificación de sentimientos a nivel de oración y la clasificación de sentimientos a nivel de aspecto, los investigadores se han centrado principalmente en RNN (particularmente LSTM).
- Para la clasificación de sentimientos multidominio, LSTM ha dado buenos resultados, y para la clasificación de sentimientos multimodal, CNN y RNN son modelos populares de aprendizaje profundo. Por lo tanto, los modelos RNN son la opción más buscada y popular para el análisis de sentimientos entre investigadores.
- Además, podemos ver que LSTM se aplica popularmente para sentimientos basados en texto, y los modelos de CNN han mostrado buenos resultados para el sentimiento de la imagen. Para datos multimodales, CNN + LSTM seguido de fusión se convierte en el enfoque deseado.

Sin embargo, la elección de un modelo de aprendizaje profundo específico aún puede depender de varios factores, como la cantidad de datos disponibles, la cantidad de unidades ocultas (nodos) necesarios para el problema, etc.

2.14 Otras tareas de clasificación de textos: detección de la sátira

Además del análisis de sentimientos hay otras tareas de clasificación de texto como la detección de agresividad, ironía, sátira, sarcasmo o humor en el mensaje. En esta tesis doctoral nos interesa en particular la detección de la sátira ya que se han probado distintas tecnologías para la detección de esta.

“La sátira es un género literario que expresa indignación hacia alguien o algo, con propósito moralizador, lúdico o meramente burlesco. Se puede escribir en prosa, verso o alternando ambas formas (sátira menipea). Se inspira en la poesía yámbica griega y se desarrolló sobre todo en la literatura latina” (Morales Harley, 2017, pp. 219-230). De forma cerrada la sátira es un género literario, pero también es un recurso que encontramos en las artes gráficas y escénicas. En la sátira los vicios individuales o colectivos, las locuras, los abusos o las deficiencias se ponen de manifiesto por medio de la ridiculización, la farsa, la ironía y otros métodos; ideados todos ellos para lograr una mejora de la sociedad (Haro, 2018, pp. 37-47). Es común y casi característico que la sátira se encuentre fuertemente impregnada de ironía o sarcasmo; además la parodia, la burla, la exageración, las comparaciones, la analogía y las dobles son usados de manera frecuente en el discurso y la escritura satírica (Lysaker, 2018). La sátira también puede contraponer dos cuestiones que resultan opuestas o muy diferentes entre sí para desvalorizar una y otorgarle mayor importancia a la otra. Exagerar o minimizar algo real hasta transformarlo en una ridiculez es otra de las técnicas habituales de la sátira. Este mecanismo es habitual en las sátiras gráficas que apelan a las caricaturas. De la misma manera, es importante establecer que la historia de la literatura universal está llena de escritores que han practicado la sátira con gran maestría y de obras muy conocidas en las que juega una labor fundamental. Este sería el caso de “Don Quijote de la Mancha” (1605) de Miguel de Cervantes; “La vida de Lazarillo de Tormes” (1554), “La vida del Buscón” (1626) de Francisco de Quevedo, “Los viajes de Gulliver” (1726) de Jonathan Swift, “Rebelión en la granja” (1945) de George Orwell o “Un mundo feliz” (1932) de Aldous Huxley.

Como la sátira es un género alfabetizado, investigamos aquellas obras que consideran características lingüísticas y estilísticas para la identificación de la sátira. En el trabajo descrito en (Reganti, Maheshwari, Kumar, Das, & Bajpai, 2016), los autores identificaron componentes y características de valor clave para la detección automática de la sátira. Evaluaron esas características a partir de varios conjuntos de datos escritos en inglés, incluidos tweets, reseñas

de productos y artículos de noticias. Combinaron (1) características de referencia basadas en n-gramas de palabras; (2) características léxicas de diferentes léxicos; (3) amplificadores de sentimiento que incluyen citas, ciertas reglas de puntuación y ciertos emoticonos; (4) características del acto de habla con disculpas, agradecimientos, declaraciones y preguntas; (5) léxicos sensoriales relacionados con los cinco sentidos básicos; (6) características de interrupción de la continuidad de la opinión que miden las variaciones de la opinión en el mismo texto; y (7) características específicas de dispositivos literarios, como hipérbolos, onomatopeyas, aliteraciones o imágenes. Sus mejores resultados se lograron con un método de conjunto que combinaba clasificadores tradicionales de aprendizaje automático, como Random Forest, Logistic Regression, Support Vector Machines y Decision Trees. Otro trabajo sobre la identificación de la sátira en inglés se describe en (Ravi & Ravi, 2017, pp. 15-33), en el que los autores proponen un sistema capaz de detectar la sátira, el sarcasmo y la ironía a partir de noticias y reseñas de clientes. Este sistema se basó en la selección de características de texto ensamblado. La efectividad de su propuesta se demostró en tres conjuntos de datos, incluidos dos satíricos y uno irónico. Durante su investigación, los autores descubrieron algunas características comunes interesantes de la sátira y la ironía, como el proceso de afectación (emoción negativa), la preocupación personal (ocio), el proceso biológico (corporal y sexual), la percepción (ver), el lenguaje informal (jurar), el proceso (masculino), cognitivo (verdadero) y psicolingüístico (concreción e imaginación), que fueron de especial importancia. En (Tsonkov & Koychev, 2015, pp. 1-7), los autores presentaron un método para la detección automática de doble sentido de textos en inglés provenientes de redes sociales. A los efectos de este artículo, definieron el doble sentido como el de la ironía, el sarcasmo y la sátira. Calificaron seis características y evaluaron su precisión predictiva con tres clasificadores de aprendizaje automático diferentes: ingenuo Bayes, k-vecinos más cercanos y máquinas de vectores de soporte.

En (Barbieri, Ronzano, & Saggion, 2015^a, pp. 1-9), los autores evaluaron la detección automática de tweets que publicitan noticias satíricas en inglés, español e italiano. Con este fin, los autores combinaron características independientes del lenguaje que describen las propiedades léxicas, semánticas y de uso de palabras de cada Tweet. Evaluaron el rendimiento de su sistema mediante la realización de experimentos de clasificación monolingües y multilingües, calculando la eficacia de la detección de sátiras. En (Salas-Zárate, Paredes-

Valverde, Rodríguez-García, Valencia-García, & Alor-Hernández, 2017,, , pp. 20-33), los autores propusieron un método que emplea una amplia variedad de características psicolingüísticas que detecta tweets satíricos y no satíricos de un corpus compuesto de tweets de noticias satíricas y no satíricas compiladas de México y España. Usaron LIWC (Tausczik & Pennebaker, 2010, pp. 24-54) para extraer rasgos psicológicos y lingüísticos del corpus y utilizarlos para evaluar tres clasificadores de aprendizaje automático. Alcanzaron una medida F1 del 85,5% con el conjunto de datos español mexicano y una medida F1 del 84,0% con el conjunto de datos europeo España, ambos con Support Vector Machines.

En cuanto a otros idiomas además del inglés y el español, podemos encontrar algunas palabras sobre la identificación de la sátira. Por ejemplo, en (Sharma, Mridul, & Islam, 2019, pp. 1-5), los autores aplicaron una técnica híbrida para extraer características de documentos de texto que fusionan Word2Vec y TF-IDF mediante la aplicación de una red neuronal convolucional para la arquitectura de aprendizaje profundo, logrando una precisión de hasta el 96%. En (Toçoğlu & Onan, 2019, pp. 107-117), los autores presentaron un clasificador de aprendizaje automático para detectar sátiras en artículos de noticias turcos. Emplearon el recuento de términos basado en el TF-IDF de unigramas, bigramas y trigramas, y evaluaron con algunos métodos tradicionales. clasificadores de aprendizaje automático como Naïve Bayes, Support Vector Machines, Logistic Regression y C4.5.

La identificación de la sátira juega un papel crucial al discernir entre la diversión y los hechos. En este sentido, es posible aplicar técnicas de identificación de la sátira para distinguir entre noticias satíricas y noticias falsas. En (Rashkin, Choi, Jang, Volkova, & Choi, 2017, pp. 2931–2937), por ejemplo, los autores realizaron un estudio analítico sobre el lenguaje de los medios en el contexto de la verificación de hechos políticos y la detección de fake news. Compararon el lenguaje de las noticias reales con el de la sátira, los engaños y la propaganda para encontrar rasgos lingüísticos de un texto poco fiable. Para evaluar la viabilidad de la verificación automática de hechos políticos, también presentaron un estudio de caso basado en PolitiFact.com utilizando sus juicios de viabilidad en una escala de seis puntos.

Como la identificación de la sátira es una tarea desafiante que puede mejorar otras tareas de la PLN. En algunos talleres se ha propuesto la identificación automática de la sátira y la ironía. Por el alcance de nuestra propuesta, destacamos el foro de evaluación de lenguas ibéricas

(IberLEF 2019), en el que se realizó una tarea compartida (Ortega-Bueno et al., 2019, pp. 1-11). Se pidió a los participantes de esta tarea compartida que identificaran expresiones irónicas en diferentes variantes del español. Hubo un total de 12 equipos que participaron en la tarea y los tres mejores resultados logrados fueron un promedio macro de F1 de 0,7167, 0,6803 y 0,6596, respectivamente. Analizamos en detalle algunos de los equipos que participaron en la tarea. En (Cignarella & Bosco, 2019), los autores entrenaron una Máquina de Vector de Soporte Lineal con características que capturan información de morfología y sintaxis de dependencia. Su enfoque logró mejores resultados a partir de los conjuntos de datos español europeo y español mexicano, pero un poco peores a partir del conjunto de datos cubano. En (Miranda-Belmonte & López-Monroy, 2019, pp. 1-6), los autores extrajeron características de recuento de términos para alimentar una red neuronal. Esta propuesta logró el mejor desempeño en el conjunto de datos cubano y el segundo lugar general en la tarea compartida. Otras propuestas se basaron en transformadores. En (González, Hurtado, & Pla, 2019, pp. 1-7), con un modelo basado en Transformer Encoders y embeddings en español aprendidas de un gran conjunto de datos compilado de Twitter. En (García, Moctezuma, & Muñiz, 2019), aplicando ELMo y en (Iranzo-Sánchez & Ruiz-Dolz, 2019, pp. 1-7), utilizando BERT. Otros enfoques se centraron en las características lingüísticas. En (Frenda & Patti, 2019, pp. 1-9), los autores presentaron rasgos estilísticos, léxicos y afectivos. En (Deon & De Freitas, 2019, pp. 1-6), el enfoque se basó en siete patrones lingüísticos elaborados a partir de nueve mil textos escritos en tres variantes lingüísticas diferentes. Los patrones lingüísticos incluían listas de expresiones de risa, palabras en mayúsculas, comillas, oraciones exclamativas, así como léxicos compuestos por frases hechas. Los beneficios de las características lingüísticas y afectivas son que proporcionan resultados interpretables que pueden utilizarse para rastrear analogías y diferencias en la expresión de la ironía. Finalmente, en (Castro & Benavides, 2019, pp. 1-8), el autor empleó tres representaciones diferentes de información textual y medidas de similitud usando una combinación ponderada de estas representaciones.

2.15 Resumen del estado del arte para la clasificación de textos

En paralelo de las tareas de análisis de sentimientos, se ha identificado otras tareas que tienen poca investigación y se requiere tener más profundidad. Aquí uno de los pocos trabajos que se pudo encontrar como son el marcado temporal (Hafez, Ismail, & Karam, 2017, pp. 373-378),

resumen de sentimientos (Abdi, Shamsuddin, Hasan, & Piran, 2018, pp. 66-85), detección de sentimientos implícitos (Chandankhede, Devle, Waskar, Chopdekar, & Patil, 2017, pp. 357-361), detección de la postura (Krejzl, Hourová, & Steinberger, 2017, pp. 1-5).

Se pudo identificar que los idiomas populares utilizados para el análisis de sentimientos son: inglés, chino y árabe. Ninguno de ellos ha tocado el idioma español y específicamente en la sátira.

Los modelos de minería de opiniones creados se basan en métodos de Deep learning basados en métodos estadísticos basados en Word embeddings sobre distintas arquitecturas de aprendizaje profundo.

Los sistemas de aprendizaje profundo, o Deep-learning se basan en montar una serie de capas, incluyendo capas de entrada, salida y un número arbitrario de capas de aprendizaje en donde se ajustan los pesos de una matriz a fin de maximizar el resultado.

2.16 Objetivos de la tesis doctoral

2.16.1 Motivación.

Las redes sociales también conocidas como walled gardens son una analogía que se utiliza para hacer referencia a las plataformas que recogen información de los usuarios y su interacción en la sociedad es altamente evidente. El contenido en redes sociales abre campos para realizar investigación por medio de muchas ciencias combinadas como la informática, matemáticas, sociología, administración, psicología, etc. Y es primordial decir, que el desarrollo de las redes sociales ha proporcionado a las personas una ventana para compartir conocimientos e intereses en común. El estallido de estas ha permitido la comunicación en nuestra vida cotidiana y, al mismo tiempo, trae información sobre múltiples dominios desde el surgimiento de la WEB 2.0 (Angarita et al., 2019, pp. 1533-5399).

En este contexto, Twitter es una aplicación gratuita de microblogging que se usa como red social, mensajería instantánea y blog. Tiene más millones de usuarios que interactúan en tiempo real a través de textos llamados tuits y que tienen una extensión de 280 caracteres. Esta aplicación permite seguir a otros usuarios conocer sus publicaciones e interactuar con ellos. En

Twitter los usuarios pueden agrupar mensajes entre sí por palabras o temas utilizando lo que se llama hashtags que inicia con el símbolo de numeral #, pueden nombrar a otro usuario iniciando con el símbolo de arroba @ seguido del nombre del usuario, pueden indicar que les gusta un tuit y este se almacena en una lista o bien pueden compartir un tuit de otro usuario con sus propios seguidores dándole retuit. Cuando una palabra o frase se repite constantemente se dice que es tendencia, estos temas se clasifican por región y se actualizan constantemente. Estas características hacen de Twitter una herramienta muy útil para conocer de qué está hablando cierto sector de la población en diferentes momentos. Debido al dinamismo de la comunicación, Twitter ha tenido un papel muy relevante en el contexto de la infodemiología (Buzzelli, Holdan, Rota, & McCarthy, 2016, pp. 64-76).

En el ámbito de la salud la infodemiología nos describe las necesidades a profundizar sobre la información que generan los usuarios de internet en las redes sociales y tratar de entender lo escrito o expresado en un lenguaje natural que proviene del ser humano. Por un lado, los usuarios comunes de la red social como también autoridades sanitarias nacionales e internacionales han optado por emitir muchas de sus comunicaciones por Twitter lo mismo que los jefes de estado de diferentes países. Así mismo muchos profesionales médicos y científicos también han elegido este medio para difundir medidas de prevención, hallazgos científicos, iniciar colaboraciones y dar seguimiento prácticamente en tiempo real de los contagios y actualmente de los avances en la salud y es aquí donde radica su gran potencial. Ha sido tan grande su impacto que ha empezado a usarse en un área de estudio relativamente reciente que se llama infodemiología y que es básicamente la evaluación de la información que suben los usuarios de la internet o redes sociales para mejorar las políticas de salud pública o bien para realizar vigilancia epidemiológica (Organización Mundial de la Salud, 2020).

La ciencia de la ingeniería lingüística trata de ir creando herramientas que por medio de las computadoras que permitan investigar o entender lo expresado en el lenguaje natural sobre las redes sociales. El procesamiento del lenguaje natural es un área de inteligencia artificial enfocada especialmente del lenguaje, donde se trata de entender y modelar el cerebro humano, para que las computadoras puedan imitar los trabajos que consideramos inteligentes por parte del ser humano. En este mismo entorno, la minería de opiniones o análisis del sentimiento abarca una serie de aplicaciones técnicas del procesamiento del lenguaje natural, lingüística

computacional y minería de textos que tienen como objetivo la extracción de información subjetiva a partir de contenidos generados por los usuarios en las redes sociales (Nurmatova, 2019).

Del estudio del estado del arte se concluye que, aunque existen estudios relacionados sobre el procesamiento del lenguaje natural en la salud, no son suficientes para verificar las nuevas técnicas de clasificación de información para los corpus propuestos; que permitan maximizar modelos predictivos del análisis de sentimiento tanto por los investigadores como por los profesionales de la salud.

La motivación de este estudio es proporcionar nuevos recursos para el análisis de sentimientos en medicina, la creación de dos corpus; uno para las enfermedades infectocontagiosas como el ZIKA y otro para el COVID-19; aplicando así el estudio de distintas tecnologías para ver cómo se pueden clasificar los sentimientos en estos dominios y ampliar el estudio de esas mismas tecnologías para la detección de la SATIRA.

2.16.2 Objetivos.

El objetivo principal de esta tesis doctoral es la aplicación de técnicas para clasificación de sentimientos en modelos predictivos para el procesamiento del lenguaje de corpus en el dominio de enfermedades infectocontagiosas como ZIKA y COVID-19; extendiendo este mismo análisis sobre un corpus del género literario SATIRA para lograr una mejor precisión en la predicción de sentimiento sobre lo expresado en redes sociales y el entendimiento del lenguaje natural. Para conseguir este objetivo se plantearon los siguientes subobjetivos:

- Obtención de un corpus sobre el dominio de enfermedades infectocontagiosas Zika, dengue y chikungunya.
- Obtención de un corpus sobre el dominio de enfermedades infectocontagiosas Covid-19.
- Obtención de un corpus sobre el dominio en el género literario de la sátira.
- Obtención de modelos de clasificación para la predicción de sentimientos en cada uno de los corpus.
- Obtención del mejor clasificador de precisión en cada uno de los corpus por tipos de clasificadores.

2.16.3 Metodología.

Esta tesis doctoral se desarrolló mediante 3 fases principalmente: la primera fase fue el estudio del estado del arte que se ha mostrado en este capítulo; la segunda fue la elaboración de métodos y artefactos de clasificación y predicción que permitan el tratamiento de los corpus propuestos; y la tercera la validación de la propuesta.

- Estudio del estado del arte: Estudio de los conceptos y términos de la inteligencia artificial en el campo del lenguaje natural, técnicas del aprendizaje automático supervisado y no supervisado, modelos para la predicción de sentimiento y herramientas para la clasificación de texto. Además, se analizaron las distintas investigaciones en el ámbito de la minería de opiniones sobre modelos y su aplicabilidad en distintos dominios.
- Elaboración de laboratorios para la obtención de modelos que realicen la predicción de sentimientos sobre los corpus en el contexto de esta tesis doctoral; también se desarrollaron las matrices de confusión para evaluar del sentimiento real vs la predicción obtenida en cada uno de los corpus.

Validación de los laboratorios: Con el detalle de los hyper-parámetros utilizados y el resultado obtenido de las predicciones para cada una de las ejecuciones realizadas. Aquí encontraremos los resultados del mejor modelo usado para evaluar la partición de validación y test. Cada una tiene el reporte de clasificación con la accuracy, precision, recall y f1 de cada clase.

3 Capítulo 3. Materiales y métodos para la clasificación de los corpus

En esta sección contiene información detallada sobre los dominios que se centró este trabajo doctoral y que tenemos las siguientes secciones: (1) la arquitectura del sistema propuesto de monitoreo de redes sociales, (2) el pipeline de nuestra propuesta, (3) las técnicas de extracción de características, (4) pipeline de nuestra propuesta, (5) la optimización de los hiperparámetros y (6) los corpus detallados en esta tesis doctoral.

3.1 Arquitectura del sistema propuesto de Monitoreo de redes sociales.

En esta sección contiene

La arquitectura para procesar los algoritmos de minería de opiniones y clasificación de documentos propuestos en esta tesis doctoral se muestra en la **Figura 20**

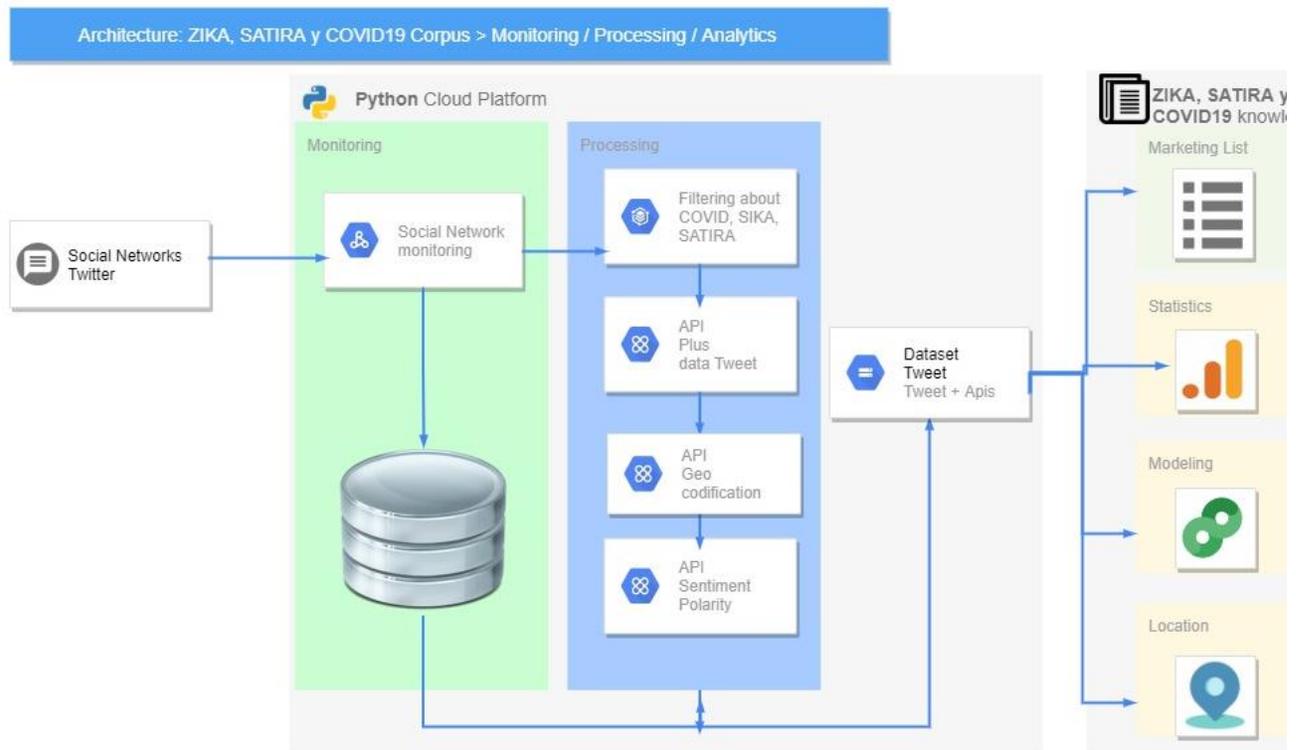


Figura 20 Arquitectura del módulo de monitorización de redes sociales

El funcionamiento de este sistema puede describirse como sigue: En primer lugar, se construye un modelo de aprendizaje supervisado en base a una colección de documentos extraídos de redes sociales y etiquetados manualmente por un grupo de usuarios. Estos documentos son procesados previamente eliminando palabras repetidas, corrigiendo errores ortográficos, o simplificando los textos para mejorar el procesamiento del modelo. En segundo lugar, estos ejemplos sirven para que la máquina sea capaz de distinguir y poder evaluar los datos como si fuera un humano, pero de forma automática. Los modelos se construyen a partir de aplicar y evaluar distintas técnicas de aprendizaje supervisado, aprendizaje profundo y sistemas de representación basados en el procesamiento del lenguaje natural descritos en el capítulo II. Los resultados del estudio de estos modelos se presentan en el capítulo IV de esta tesis doctoral. Finalmente, una vez los modelos ha sido creado y evaluado, se crearon gráficos comparativos para poder evaluar el sentimiento de futuros documentos.

Como hemos comentado, el sistema de minería de opiniones extrae, de manera global, si un determinado documento contiene un sentimiento positivo, neutral o negativo.

Para poder aplicar los conceptos de minería de opiniones los modelos se nutren de un conjunto grande de corpus anotados durante un periodo de tiempo. En la siguiente sección se describen los corpus que se han creado para esta tesis doctoral.

3.2 Pipeline de nuestra propuesta

Primero, extraemos y reprocesamos un corpus sobre la identificación del zika, covid-19 y sátira. Usamos el conjunto de datos de entrenamiento para obtener características de term-counting y embeddings de capas a partir de word embeddings previamente entrenadas. A continuación, usamos la selección de características para obtener los mejores n-gramas de caracteres y palabras y los usamos para entrenar a cuatro clasificadores de aprendizaje automático. A continuación, usamos la capa de embeddings para alimentar cuatro arquitecturas de aprendizaje profundo. Todos los clasificadores se ajustaron con precisión al realizar una búsqueda de cuadrícula aleatoria para encontrar los mejores hiperparámetros y evaluar diferentes arquitecturas de aprendizaje profundo. Finalmente, evaluamos cada modelo prediciendo sobre el conjunto de datos de prueba. Esta canalización se muestra en la **Figura 21** se describe en detalle en las siguientes secciones.

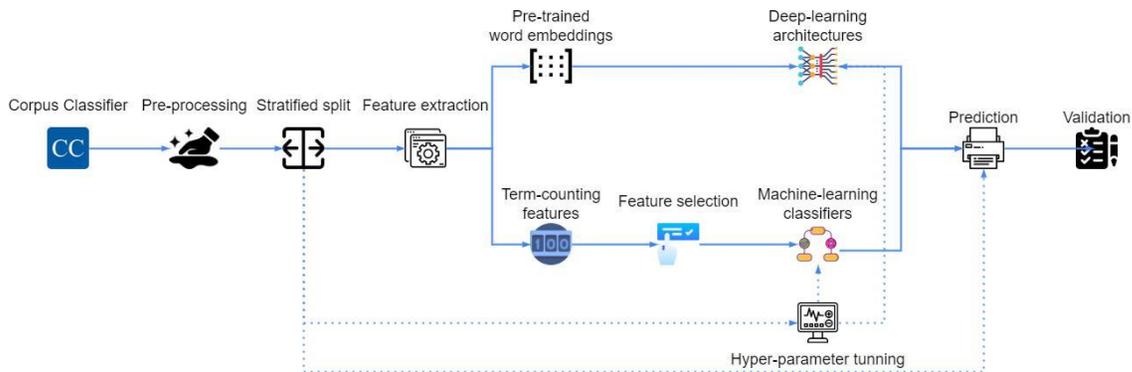


Figura 21 El pipeline de nuestra propuesta.

3.3 Extracción de características

El siguiente paso de nuestro pipeline consiste en la extracción de las características para realizar la clasificación supervisada. Estas características están organizadas en dos conjuntos principales: (1) características de recuento de términos, utilizadas con los clasificadores tradicionales de aprendizaje automático; y (2) word embeddings, empleadas para probar las arquitecturas de aprendizaje profundo.

3.3.1 Term-Counting Features.

Las funciones Term-Counting Features son la base del modelo de bolsa de palabras (BoW) y sus enfoques similares. El objetivo es la representación del lenguaje natural como vectores para facilitar que las computadoras y los algoritmos de aprendizaje automático manejen el lenguaje natural. En BoW, cada documento se representa como un vector compuesto por la frecuencia de sus palabras. Normalmente, el tamaño del vector depende de todas las palabras del corpus, pero también es posible utilizar léxicos de dominios específicos para categorizar el texto de acuerdo con algún dominio preestablecido. La ventaja del modelo BoW es que es fácil de implementar a la vez que ofrece resultados competitivos. El modelo BoW se ha utilizado como una sólida base en las tareas de clasificación de texto durante mucho tiempo. Sin embargo, el modelo BoW tiene algunos inconvenientes. Primero, este modelo ignora completamente el contexto de las palabras. Como cada palabra se trata por separado, el modelo BoW es débil contra la polisemia y la sinonimia. El segundo inconveniente es que en grandes conjuntos de datos y cuando el léxico está compuesto por todas las palabras de un corpus, los vectores resultantes son grandes y escasos, lo que provoca un fenómeno conocido como la maldición de la dimensionalidad, que dificulta la capacidad de algunos algoritmos

de aprendizaje automático y clasificadores para aprender al mismo tiempo que aumenta los requisitos de tiempo y memoria. El tercer inconveniente es que el modelo BoW tiende a sobrerrepresentar algunas palabras basadas en la ley de Zipf, que establece que las palabras tanto en natural como en artificial los idiomas siguen correlaciones de largo alcance (Mehri & Jamaati, 2017, pp. 2470-2477). Finalmente, el modelo BoW no es capaz de manejar palabras desconocidas, errores ortográficos y algunas formas populares de escribir en las redes sociales que incluyen el uso de abreviaturas, jerga y escritura informal.

Existen diferentes enfoques para resolver los inconvenientes del modelo BoW. El modelo de n-gram de palabras es una extensión del modelo BoW en el que se consideran palabras conjuntas (bígramas, trigramas). Las palabras conjuntas tienen dos beneficios principales: por un lado, pueden representar conceptos semánticos más altos que los unigramas y, por otro, pueden usarse para la desambiguación de palabras. Otro enfoque se conoce como n-grams de caracteres, que consiste en medir caracteres en lugar de palabras. Los n-gramas de caracteres son conscientes de errores ortográficos o palabras inventadas. Otra mejora del modelo BoW es la aplicación del Término Frecuencia-Frecuencia inversa del documento (ver Ecuación (1)) en lugar de frecuencias de recuento sin procesar. TF-IDF descarta las palabras comunes conocidas como palabras vacías y resalta otras palabras clave relevantes.

$$\text{TFIDF} = \text{TF} * \text{IDF} \quad (1)$$

$$\text{TF} = \text{number_of_occurrences}/\text{number_of_grams} \quad (2)$$

$$\text{IDF} = \log_2 \text{corpus_size}/\text{documents_with_terms} \quad (3)$$

3.3.2 Word Embeddings.

Como hemos observado en la sección 3.3.1, los métodos de recuento de términos (Term-Counting Features) representan documentos como frecuencias de tokens. En este sentido, ignoran la posición de cada término en la secuencia. En la comunicación humana, sin embargo, el orden de las palabras tiene un gran impacto en el significado de la frase. Las cláusulas negativas y las dobles negaciones, por ejemplo, pueden cambiar el significado de una oración en función de las palabras que modifican. Por otro lado, los modelos de recuento de términos (Term-Counting Features) consideran todas las palabras iguales. Es decir, no hay diferencia entre términos ni en su función sintáctica, semántica o pragmática. Estos problemas se resuelven mediante el uso de modelos de distribución, como las word embeddings, en las que las palabras se codifican como vectores densos con la idea subyacente de que palabras similares tienen una representación similar. Las word embeddings se pueden ver como una extensión de la codificación one-hot, que permite representar variables categóricas como vectores binarios. Con codificación one-hot, las palabras se codifican en vectores de longitud fija del

vocabulario completo, en los que cada vector tiene sólo una columna con un valor diferente a cero. En la codificación one-hot, por lo tanto, cada palabra es ortogonal con el resto de las palabras. Sin embargo, con las word embeddings, la representación numérica de cada vector se aprende típicamente entrenando una red neuronal con alguna tarea de propósito general, como la predicción de la siguiente palabra o la analogía de palabras. Uno de los principales beneficios de las word embeddings es que es posible aprender esas embeddings a partir de grandes conjuntos de datos no anotados, como redes sociales, sitios de noticias o enciclopedias, entre otros, para obtener vectores de palabras que transmiten un significado general que se puede aplicar para resolver otras tareas más rápido. por lo tanto, cada palabra es ortogonal con el resto de las palabras. Sin embargo, con las word embeddings, la representación numérica de cada vector se aprende típicamente entrenando una red neuronal con alguna tarea de propósito general, como la predicción de la siguiente palabra o la analogía de palabras. Uno de los principales beneficios de las word embeddings es que es posible aprender esas embeddings a partir de grandes conjuntos de datos no anotados, como redes sociales, sitios de noticias o enciclopedias, entre otros, para obtener vectores de palabras que transmiten un significado general que se puede aplicar para resolver otras tareas más rápido.

Por lo tanto, cada palabra es ortogonal con el resto de las palabras. Sin embargo, con las word embeddings, la representación numérica de cada vector se aprende típicamente. Los word embeddings han permitido el rendimiento superior de varias tareas de PLN y han significado un cambio de paradigma en el PLN. En este sentido, es posible utilizar los mismos principios para codificar las principales unidades lingüísticas, como las oraciones, para crear embeddings de oraciones (Krasnowska-Kieras & Wróblewska, 2020, pp. 1-11). Otros enfoques han tratado de enfocar el entrenamiento de word embeddings para resolver tareas específicas de PLN. Por ejemplo, en (Tang et al., 2014, pp. 1-11), los autores describen un método para aprender word embeddings centrado en el análisis de sentimientos. Los autores argumentan que algunas de las técnicas empleadas para aprender word embeddings pueden agrupar palabras con sentimientos opuestos, como bien o malo sólo porque tienen alguna función sintáctica similar. Desarrollan redes neuronales que consideran la polaridad del sentimiento en la función de pérdida y evaluaron su enfoque con algunos conjuntos de datos existentes centrados en el análisis del sentimiento.

En esta tesis, nos enfocamos en word embeddings en español previamente entrenadas con diferentes enfoques, incluidos word2vec, fastText y gloVe. Estos modelos y las incorporaciones de palabras pre-entrenadas resultantes se describen a continuación.

Word2Vec. Word2Vec fue uno de los primeros modelos para obtener word embeddings. Con word2vec, las word embeddings se aprenden entrenando una red neuronal, cuyo objetivo es la predicción de la siguiente palabra (Hellrich & Hahn, 2017, pp. 1-3). Específicamente, existen dos métodos para aprender las word embeddings con Word2Vec: (1) Modelo continuo de Bag of Words (CBOW), en el que el objetivo es predecir una palabra basada en las palabras de contexto; y (2) Skip-Grams, en el que el objetivo es todo lo contrario: predecir palabras de contexto a partir de una palabra objetivo. Independientemente del enfoque, ambas estrategias aprenden las representaciones de palabras subyacentes. La diferencia es que el modelo CBOW es más rápido que el mismo tiempo que proporciona una mayor precisión con palabras frecuentes y el modelo Skip-Gram es más preciso utilizando datos de entrenamiento más pequeños al mismo tiempo que proporciona una mejor representación de una palabra que aparece con poca frecuencia. Las word embeddings previamente entrenadas de Word2Vec utilizadas en estos experimentos fueron entrenados con el corpus español Billion Corpora.

GloVe. GloVe es una técnica para aprender con word embeddings que explotan estadísticas de información con respecto a las coincidencias de palabras que es más adecuada para realizar tareas de PLN como la analogía de palabras o el reconocimiento de entidades (Pennington, Socher, & Manning, 2014, pp. 1-12). Al igual que Word2Vec, donde las word embeddings se aprenden aplicando probabilidades de co-ocurrencia sin procesar, GloVe aprende la relación entre co-ocurrencias, lo que mejora para aprender detalles precisos en la relevancia de dos términos vinculados. Las word embeddings previamente entrenadas de GloVe utilizadas en este trabajo se entrenaron con el corpus español Billion Corpora [41].

FastText. FastText está inspirado en el modelo word2vec pero representa cada palabra como una secuencia de caracteres n-grams (Mikolov, Grave, Bojanowski, Puhersch, & Joulin, 2017, pp. 1-4). FastText es, por lo tanto, consciente de palabras desconocidas y errores ortográficos. Además, los n-grams de caracteres permiten capturar información semántica extra en diferentes tipos de idiomas. Por ejemplo, en idiomas afectados como el español, puede capturar información sobre prefijos y sufijos, incluida información sobre números y género gramatical. En los lenguajes aglutinantes, como el alemán, en el que las palabras pueden estar formadas por otras palabras, los n-grams de caracteres pueden incluir información de ambas palabras. Vale la pena señalar que estos n-grams de caracteres se comportan internamente en un modelo BoW, por lo que no tiene en cuenta el orden interno de los n-grams de caracteres. FastText tiene disponibles word embeddings previamente entrenadas de diferentes idiomas, incluido el español (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2019) entrenado con Wikipedia. Para estos experimentos, sin embargo, usamos las word embeddings

previamente entrenadas de fastText entrenadas con los corpus sin anotaciones en español. Esta decisión se tomó porque estas word embeddings previamente entrenadas han utilizado más fuentes, incluidos subtítulos, noticias y texto legislativo de la Unión Europea.

3.3.3 Clasificadores supervisados.

Como observamos durante la revisión de la literatura (ver Sección 2.14), hay multitud de obras que utilizan clasificadores de aprendizaje automático para realizar la identificación. Por ejemplo, en (Salas-Zárate et al., 2017, pp. 20-33), los autores emplearon máquinas de vectores de soporte, árboles de decisión y clasificadores bayesianos. En (Ravi & Ravi, 2017, pp. 15-33), los autores emplearon LibSVM, regresión logística, redes bayesianas y perceptrón multicapa. En (Reganti et al., 2016, pp. 1-8), los autores emplearon regresión logística, dos modelos de árboles de decisión y máquinas de vectores de soporte. Como el objetivo principal de este trabajo es evaluar los recursos de PLN novedosos en español aplicados a la clasificación de texto como el análisis de sentimientos y la identificación de la sátira, decidimos utilizar clasificadores de aprendizaje automático para las familias principales como línea de base.

Específicamente, en este trabajo evaluamos dos tipos diferentes de clasificadores de aprendizaje automático supervisados. Por un lado, la evaluación de las características de term-counting con cuatro clasificadores tradicionales de aprendizaje automático, incluidos árboles de decisión, máquinas de vectores de soporte, regresión logística y clasificadores bayesianos (consulte la Sección 3.3.4. Por otro lado, se utilizaron tres word embeddings previamente entrenadas para entrenar cuatro arquitecturas de aprendizaje profundo, incluidos perceptrones multicapa, redes neuronales convolucionales y diferentes variantes de embeddings de redes neuronales recurrentes.

3.3.4 Clasificadores de aprendizaje automático.

Evaluamos los siguientes clasificadores de aprendizaje automático:

Random Forest (RF). Pertenecen a la familia de los árboles de decisión. Los árboles de decisión son algoritmos que construyen una estructura de árbol compuesta por reglas de decisión en el formulario. if-then-else. Cada decisión dividida se basa en la idea de entropía, maximizando la homogeneidad de nuevos subconjuntos. Los árboles de decisión son populares porque proporcionan buenos resultados, se pueden utilizar tanto en problemas de clasificación como de regresión y, en conjuntos de datos más pequeños, proporcionan modelos interpretables. Sin embargo, presentan algunos inconvenientes. Primero, tienden a generar modelos sobre ajustados al crear árboles demasiado complejos que no generalizan el patrón subyacente. En segundo lugar, son muy sensibles a los datos de entrada y

pequeños cambios pueden dar como resultado árboles completamente diferentes. En tercer lugar, los árboles de decisión se ven afectados por el sesgo cuando el conjunto de datos está desequilibrado. En este trabajo, seleccionamos Random Forest (Liaw & Wiener, 2002, pp. 1-5). Random Forest es un método conjunto de aprendizaje automático que utiliza el ensacado para crear varios árboles de decisión y promediar sus resultados. Además, cada árbol forestal aleatorio considera solo un subconjunto de las características y un conjunto de ejemplos aleatorios, lo que reduce el ajuste excesivo del modelo.

Support Vector Machines (SVM). Son una familia de clasificadores basados en la distribución de las clases en un hiperespacio y determinan la separación que mejor distribuyó las clases. Support Vector Machines permite el uso de diferentes núcleos que resuelven problemas de clasificación lineales y no lineales. Algunos trabajos que han evaluado la identificación aplicando SVM se pueden encontrar en [(Reganti et al., 2016, pp. 1-8), (Salas-Zárate et al., 2017, pp. 1-10)].

Logistic regression (LR). Este clasificador se usa normalmente para problemas de clases binarias combinando linealmente los valores de entrada para crear una función sigmoidea que distingue entre la clase predeterminada. Se ha aplicado la regresión logística para la identificación en [(Reganti et al., 2016, pp. 1-8), (Ravi & Ravi, 2017, pp. 15-33) , (Salas-Zárate et al., 2017, pp. 1-10)] y detección de sátira (Ortega-Bueno et al., 2019, pp. 1-28).

Multinomial Naïve Bayes (MNB). Es un clasificador probabilístico que se basa en el teorema de Bayes. Específicamente, la variante ingenua de este clasificador asume una independencia entre todas las características y clases. Este clasificador ha sido evaluado para resolver tareas similares (Ortega-Bueno et al., 2019, pp. 1-28).

KNeighbors (K). Es un clasificador que se basa en estimar la función de densidad por clase; es decir, K es un hiperparámetro de los vecinos más cercanos y que es esencial para el reconocimiento de patrones. Esto implica que no hace suposiciones explícitas sobre la forma funcional de los datos (Ali, Neagu, & Trundle, 2019, pp. 1-15).

J48 (J48). Es un clasificador que genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta de mayor ganancia (Aljawarneh, Yassein, & Aljundi, 2019, pp. 10549–10565).

3.3.5 Arquitecturas de aprendizaje profundo.

Evaluamos las siguientes arquitecturas de aprendizaje profundo:

Multilayer Perceptron (MLP). Los modelos de aprendizaje profundo se componen de capas apiladas de perceptrones en las que cada nodo está completamente conectado con los demás y hay, al menos, una capa oculta. En este trabajo, hemos evaluado diferentes redes neuronales de vainilla que incluyen diferentes números de capas, neuronas por capa, tamaños de lotes y estructuras. Los detalles de este proceso de optimización de hiperparámetros se describen en la Sección 3.5

Convolutional Neural Networks (CNN). De acuerdo a (Kim, 2014), las redes neuronales profundas convolucionales emplean capas específicas que convierten los filtros que se aplican a las características locales. CNN se hizo popular por la visión por computadora, pero también ha logrado resultados competitivos para tareas de PLN como la clasificación de texto (Yin, Kann, Yu, & Schütze, 2017, pp. 1-7). La idea principal detrás de las CNN es que pueden gestionar de forma eficaz las características espaciales. En PLN, eso significa que CNN es capaz de entender palabras conjuntas. En este trabajo, apilamos una capa Spatial Dropout, una Convolutional Layer y una Global Max Pooling layer. Durante la evaluación de hiperparámetros, probamos para concatenar la red neuronal convolucional a varias redes neuronales de avance.

Recurrent Neural Networks (RNN). Las RNN son arquitecturas de aprendizaje profundo en las que la entrada se trata como una secuencia y la conexión entre las unidades es un ciclo dirigido. En las RNN, tanto la capa de entrada como la de salida están relacionadas de alguna manera. Además, los RNN bidireccionales pueden considerar estados y pesos pasados, pero también futuros. Los RNN se utilizan ampliamente en PLN porque manejan la entrada como una secuencia, que es adecuada para el lenguaje natural. Por ejemplo, BiLSTM se ha aplicado para realizar la clasificación de textos (Zhou et al., 2016, pp. 1-11) o detección de la sátira (S. Zhang, Zhang, Chan, & Rosso, 2019, pp. 1633-1644). Para este experimento, evaluamos dos RNN bidireccionales: unidades bidireccionales cerradas recurrentes (BiGRU) y unidades bidireccionales de memoria a largo y corto plazo (BiLSTM). BiGRU es una versión mejorada de los RNN que resuelve el problema del gradiente de desaparición mediante el uso de dos puertas (actualización y reinicio), que filtran la información dirigida a la salida. BiGRU puede mantener información de memoria larga. Como hicimos con la CNN, evaluamos conectar capas de RNN a diferentes capas de redes neuronales.

3.4 Implementación de los modelos

En este trabajo usamos Python junto con la plataforma Scikit-learn y Keras [51]. Como se comentó anteriormente durante el proceso de adquisición de datos, nuestros equipos de entrenamiento, evaluación y prueba son divisiones estratificadas (60% -20% -20%) resultando en un conjunto de datos casi equilibrado. Por lo tanto, todos los modelos se evalúan en nuestra propuesta utilizando el

Accuracy métrica (ver Ecuación (4)), que mide la relación entre las instancias clasificadas correctamente con todas las instancias.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

3.5 Optimización de hiperparámetros

En este paso de la canalización, evaluamos los hiperparámetros para los clasificadores de aprendizaje automático y las arquitecturas de aprendizaje profundo. Ambos se realizaron con una cuadrícula de búsqueda aleatoria con Sci-kit y Talos.

Primero llevamos a cabo un análisis para evaluar diferentes parámetros de modelos basados en n-grams de caracteres y n-grams de palabras. Evaluamos diferentes filtros de corte y diferentes enfoques de TF-IDF, incluida la escala de TF vainilla o TF sub-lineal (ver Ecuación (5)). Estas opciones se muestran en la **Tabla 3**. Debido a la gran cantidad de parámetros, utilizamos una búsqueda de cuadrícula aleatoria para evaluar 5000 combinaciones de hiperparámetros. Para interpretar correctamente esta tabla, se utilizó las mismas convenciones de nomenclatura que se usan en Python, en las que los corchetes representan una lista de elementos que la búsqueda aleatoria selecciona al azar.

$$\text{TF} = 1 + \text{LOG}(\text{TF}) \quad (5)$$

Tabla 3 Opciones de hiperparámetros para los clasificadores tradicionales de aprendizaje automático.

Hyper-Parameter	Options
word_n_grams	[(1, 1), (1, 2), (1, 3)]
character_n_grams	[(4, 4), (4, 5), (4, 6), (4, 7), (4, 8), (4, 9), (4, 10)]
min_df	[0.01, 0.1, 1]
sublinear_tf	[True, False]
use_IDF	[True, False]
strip_accents	[None, 'unicode']
rf_n_estimators	[200, 400, 800, 1600]
rf_max_depth	[10, 100, 200]
svm_kernel	['rbf', 'poly', 'linear']
lr_solver	['liblinear', 'lbfgs']
lr_fit_intercept	[True, False]

A continuación, realizamos la evaluación de hiperparámetros para las arquitecturas de aprendizaje profundo. Esta evaluación de parámetros incluye diferentes arquitecturas internas de las capas ocultas

de aprendizaje profundo, según el número de capas ocultas, su tamaño y su forma (ver **Tabla 4**, para obtener una lista de todos los parámetros). Dejamos fuera de este proceso el número de épocas, porque decidimos dejarlas fijas en 1000, pero incluyendo un método de parada anticipada con una paciencia de 15. Para el optimizador elegimos Adam (Ramalingam & Ragavendran, 2020, pp. 1-6) porque es el optimizador que nos proporciona el mejor resultado en las pruebas preliminares. Específicamente, automatizamos la arquitectura del aprendizaje profundo para evaluar los siguientes parámetros: (1) funciones de activación, (2) el número de capas ocultas, (3) el número de neuronas por capa, (4) el tamaño del lote, (5) la tasa de aprendizaje, y (6) la forma de las redes neuronales: si todas las capas ocultas tenían el mismo tamaño, nos referimos a ellas como forma brick-shape o, si las capas ocultas tienen la mitad de las neuronas a excepción de la primera capa, nos referimos a ellos como forma funnel-shape. Este proceso es similar a las funciones de Term-Counting. Por un lado, sintonizamos los hiperparámetros con el conjunto de evaluación y, por otro, realizamos una selección aleatoria para reducir el número de combinaciones. Los resultados se muestran en la **Tabla 4**.

Tabla 4 Opciones de hiperparámetros para las arquitecturas de aprendizaje profundo.

Hyper-Parameter	Options
Activation	[elu, relu, selu, sigmoid, tanh]
Batch size	[16, 32, 64]
Dropout	[False, 0.2, 0.5, 0.8]
Neurons per layer	[8, 16, 48, 64, 128, 256]
Learning rate	(0.5, 2, 10)
Numbers of layers	[1, 2, 3, 4]
Shape	['brick', 'funnel']
Adjust embeddings	[True, False]

El ajuste de hiperparámetros que se muestra en la **Tabla 4**, para las arquitecturas de aprendizaje profundo reveló que ajustar la capa de embeddings independientemente de la word embeddings previamente entrenadas o la arquitectura de aprendizaje profundo.

3.6 Corpus desarrollados en la tesis doctoral

La necesidad de compilar distintos corpus se debe a que se necesitan conjuntos de datos anotados manualmente para poder entrenar los modelos de análisis de sentimientos, que se basan en técnicas de aprendizaje supervisado y Deep learning.

En esta tesis doctoral compilamos dos corpus relacionados con el Zika y el COVID-19. Además, se utilizó otro corpus existente para la detección de la sátira que también se detalla en esta sección.

Para extraer los documentos de las redes sociales empleamos Twitter, que es una red social de microblogging en donde los usuarios pueden escribir pequeños textos de unos 280 caracteres de longitud máximo y en donde pueden incluir fotos y metadatos. Twitter es una red ampliamente utilizada para llevar a cabo este tipo de estudios por diferentes razones. En primer lugar, es una red pública que permite sacar datos de manera gratuita de los últimos 7 días; de esta forma, preparamos un script que obtiene nuevos tweets sobre un tema cada hora. En segundo lugar, Twitter tiene un mecanismo de suscripción basado en hashtags donde los usuarios pueden, manualmente, etiquetar sus contenidos, lo que proporciona un mecanismo sencillo para buscar contenido relacionado. En tercer lugar, el límite de caracteres hace que los usuarios tengan que ser escuetos y directos en sus escritos.

Por otro lado, se utilizaron otros servicios web disponibles por ID para enriquecer la información del conjunto de datos contenidos en los corpus.

El nivel de seguridad que los usuarios de la plataforma Twitter son los parámetros de alcance que permite leer la información pública he incorporarla a nuestro conjunto de análisis. Así hemos podido enriquecer con datos puntuales de cantidad de re-tweet, cantidad de me gusta del tweet, cantidad de seguidores de la cuenta, cantidad de amigos de la cuenta, cantidad de tweet generados, lenguaje del tweet y datos de geoposicionamiento como latitud, longitud, ciudad y país de origen de la cuenta que genero el tweet. Estos parámetros en su mayoría no están disponibles por los niveles de seguridad que los usuarios activan y minimizan la información pública de sus tweets.

Los dos corpus nos permiten conformar un corpus específico orientado para enfermedades infecciosas. Algunas de las estadísticas de los corpus compilados pueden verse en la **Tabla 5**.

Tabla 5 Estadísticas de los corpus compilados y etiquetados para el ZIKA, COVID y SATIRA.

	Corpus del zika	Corpus del COVID-19	Corpus de la Sátira
Positivos	13.272	67.449	4.790
Neutrales	13.272	59.497	-
Negativos	13.272	61.672	4.898
Total	39.816	188.618	9.688

Previo a este proceso de etiquetado, los tweets fueron analizados automáticamente y los tweets duplicados o con un alto grado de similitud, así como tweets que sólo tenían enlaces o hashtags, pero no información textual relevante, fueron descartados.

Los corpus compilados fueron etiquetados por un grupo de voluntarios y estudiantes de la Universidad de Guayaquil (Ecuador) supervisados por el doctorando de esta tesis doctoral. Este proceso se llevó a cabo en 4 iteraciones. Las dos primeras tuvieron que ver con el corpus del Zika, y las dos últimas fueron sobre el Covid-19. En cuanto se inició la clasificación de los tweets, se hicieron seguimientos para ver cómo están trabajando, los usuarios realizan un lote limitado por día y por hora de clasificación en los tweets para mantener el interés de las personas sobre el trabajo diario y no perder la motivación sobre el objetivo propuesto, en el caso que fue permitente se dio un feedback.

Los anotadores clasificaban los tweets en las siguientes categorías: muy positivo, muy negativo, neutral, muy positivo, muy negativo, no sabe no contesta y fuera del dominio. Para ello, se puso a disposición de los anotadores la herramienta de clasificación UMUCorpusClassifier (Garcia-Diaz, Almela, Alcaraz-Marmol, & Valencia-Garcia, 2020, pp. 1-8) y su trabajo se supervisaba semanalmente. En la Figura 22 se puede observar una captura de la herramienta de clasificación de tweets. UMUCorpusClassifier es una aplicación web que permite a diferentes usuarios valorar una serie de textos y crear un corpus representativo de un modelo. Para esto, los usuarios de la aplicación verán un listado previamente de frases relacionadas con el modelo y tendrán que clasificarlos dentro de una escala.



Figura 22 Captura de pantalla de la herramienta de clasificación de tweets.

Un aspecto muy importante es que un mismo tweet podía ser etiquetado por varios usuarios, lo cual permitía ver si había consenso entre las valoraciones o no. Una vez entramos en la aplicación donde podemos operar clasificando el corpus. La pantalla se divide en dos áreas:

En la parte superior tenemos un selector de corpus y un filtrado por los estados en lo que hemos clasificado previamente los tweets.

En la parte inferior veremos un listado de tweets con los siguientes atributos:

- Tweet (es un enlace al tweet original)
- Autor. Cuenta del usuario que escribió el tweet
- Fecha. Fecha de publicación del tweet
- Clasificación. La clasificación actual asignada

El modo normal de trabajo es ir aplicando una clasificación a cada uno de los tweets acorde a si el mensaje es positivo o negativo o si, en cambio, está fuera del dominio. Todos los datos se van guardando conforme se trabaja, por lo que dividimos nuestra tarea en distintos días, el uso de los filtros permitió ver sólo los tweets pendientes de clasificar.

En concreto, cada tweet fue etiquetado de media unas 2 veces y se alcanzó un porcentaje de acuerdo entre anotadores de 0,68369 para el corpus del Zika. Sin embargo, el corpus del COVID-19 tiene un porcentaje ligeramente inferior: 0,580735. Estos datos están calculados según el coeficiente Krippendorff's alpha

Una vez los tweets fueron compilados y clasificados, fueron preprocesados para facilitar el trabajo de los algoritmos de aprendizaje. En primer lugar, todas las palabras fueron pasadas a minúscula, lo que permite a un ordenador tratar palabras escritas de distinta manera como si fueran las mismas. También se utilizaron librerías específicas para corregir erratas sobre ciertas palabras y se eliminaron enlaces, saltos de línea o cualquier otra información no relevante.

A continuación, se muestran distintas estadísticas de los corpus desarrollados en esta tesis doctoral ZIKA, COVID-19 y SATIRA.

3.6.1 El corpus ZIKA.

En la tabla Tabla 6 Estadísticas generales del corpus ZIKA. se muestran algunas de las características generales del corpus ZIKA Entre las estadísticas generales podemos visualizar el número de tweets recolectados para el corpus con una cantidad de 102489; la media de tweets compilados a la semana fue 748,1168 (618,11922 σ), de entre todos los tweets la cantidad de cuentas(usuarios de la plataforma de Twitter) diferentes encontradas fue de 49,441; de estos se desprende que el número de tweets por cuenta generado fue de 2,07296 (7,1476 σ). Los usuarios que trabajaron sobre estos tweets realizaron una cantidad de 240014 anotaciones por lo cual fueron valorados 39816(38,84905 %), la cantidad de anotaciones por tweet fue de 1,6672 (1,80837 %) y la cantidad de anotadores fue de 74 que se dividieron el trabajo por tiempo de procesamiento en días y horas; la media de anotaciones por anotador fue de 3243 (296,27329 σ). Concluyendo con el porcentaje de acuerdo Krippendorff's Alpha 0,68369.

Tabla 6 Estadísticas generales del corpus ZIKA.

Métrica	Valor
Nº de tweets	102 489 tweets
Media de tweets compilados a la semana	748,1168 (618,11922 σ)
Nº de cuentas de Twitter diferentes	49 441 cuentas
Nº de tweets por cuenta	2,07296 (7,1473 σ)
Nº de anotaciones	240 014
Nº de tweets valorados	39 816 (38,84905 %)
Anotaciones por tweet	1,6672 (1,80837 %)
Nº de anotadores	74
Media de anotaciones por anotador	3243 (296,27329 σ)
Porcentaje de acuerdo Krippendorff's alpha	0,68369

En una primera visión de la información podemos resumir que hemos podido construir un corpus que contiene información representativa sobre el Zika y enfermedades relacionadas como Dengue. Así tenemos, por ejemplo, que los tweets tienen una vigencia significativa en los tiempos que van en un rango desde el año 2017 hasta el año 2018 y cuya clasificación aplican como positivos son el 13% de todo el conjunto de tuits su clasificación directamente opuesta son los que aplican como negativo y alcanzan el 19%. Entre los datos más representativos con los datos que se califican como fuera del dominio con el 24% y los que caen en la categoría neutral el 31%. La suma de las categorías negativo y muy negativo alcanza un 27% en comparación con la suma de las categorías positivo y muy positivo 16%.

Se comenzó a compilar tuits relacionados con el zika dentro de Latinoamérica. En primer lugar, el corpus Zika trata sobre las enfermedades del Zika, Dengue y el Chikungunya en América Central. Este corpus contiene 39.816 tweets etiquetados y contiene palabras y hashtags relacionados con estas enfermedades compilados en Ecuador y países limítrofes.

En la **Figura 23** indica cuáles son las palabras más comunes del corpus del Zika. Para sacar estos valores hemos quitado las palabras comunes, conocidas como stop-words. Podemos observar que las palabras más comunes son aquellas relacionadas con la enfermedad, como síntomas. Por ejemplo, con respecto al Zika, las palabras que más aparecen tienen que ver con las enfermedades dengue, zika chikungunya, zonas (Honduras), con términos relacionados: casos, salud, mosquito y con campañas de prevención (alertas, campañas, etc).

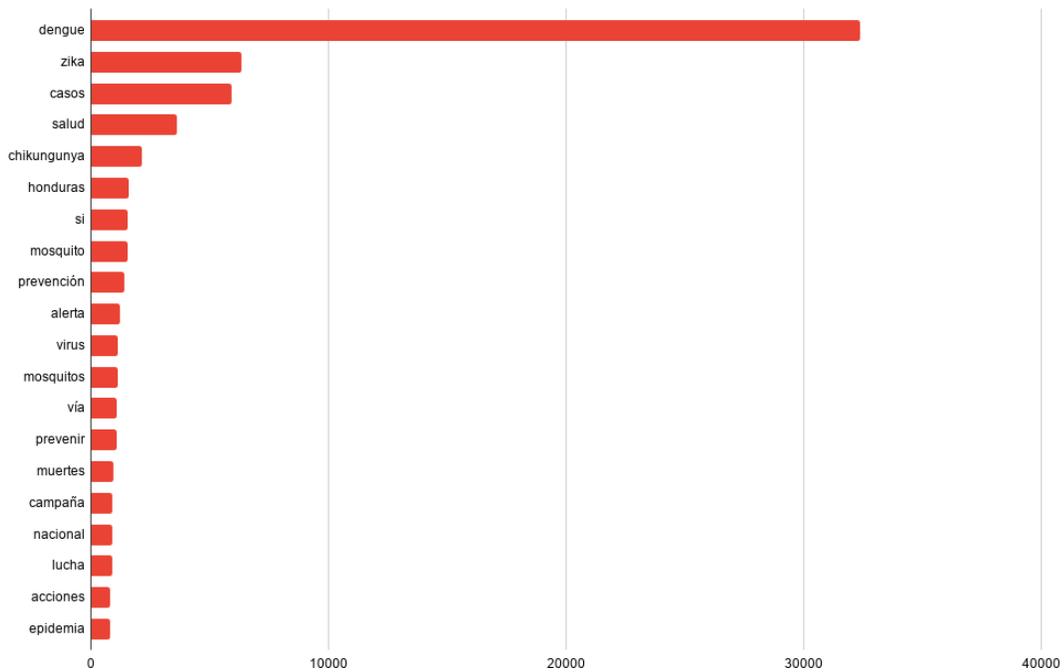


Figura 23 Palabras más frecuentes del corpus del Zika (rojo).

Para el análisis del corpus se ha incorporado estadísticas descriptivas que permiten ver de forma visual la estructuración y la composición contextual del conjunto de tweets. A continuación, los datos más representativos que forman parte de este corpus.

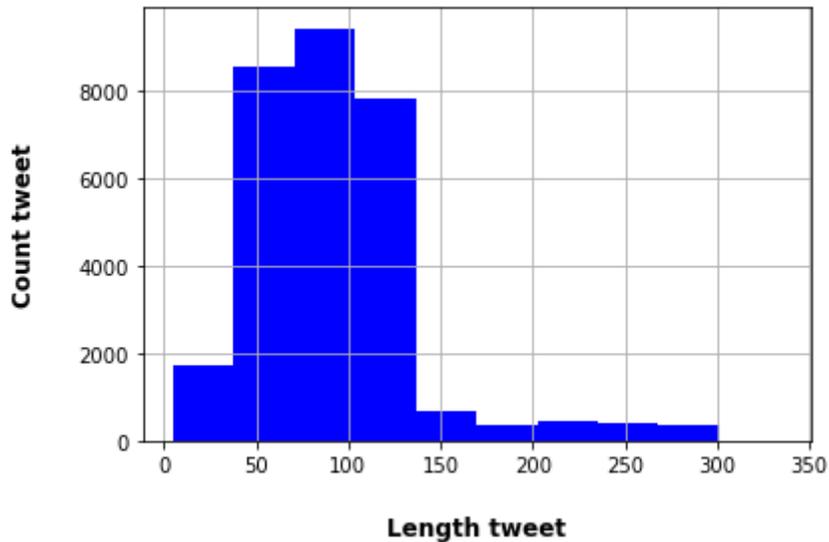


Figura 24 Histograma de longitud de textos del corpus ZIKA.

En la **Figura 24**, se muestra que los textos de los tweets oscilan entre 2 y 280 caracteres, en lo general, cada texto está en un rango de 45 a 145 caracteres. Y el pico máximo de la composición de caracteres

en un texto está en 75 a 100. En un inicio, Twitter se permitió una cantidad de caracteres con un límite de 140 caracteres y luego se amplió este límite en los tuits a 280.

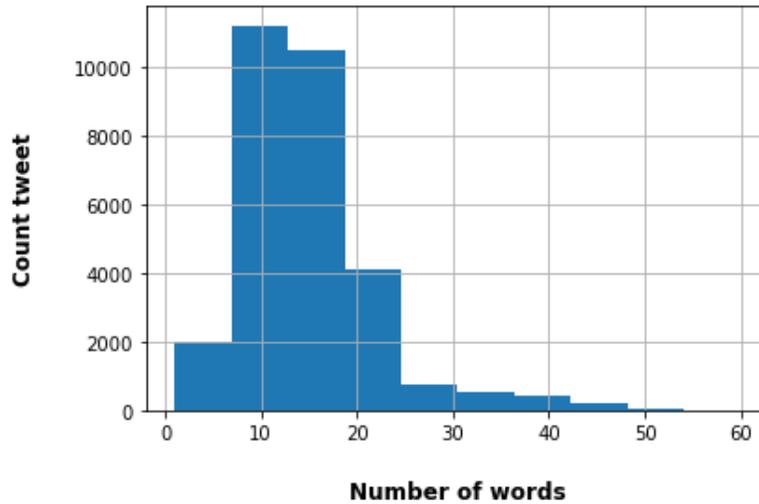


Figura 25 Histograma de números de palabras en los tweets ZIKA.

En la **Figura 25** Presenta la composición de la cantidad de palabras utilizadas en los tweets de este corpus. El número de palabras en los textos de los tweets varía de 2 a 55 y en su mayoría cae en un rango de 8 y 16 palabras. El límite de caracteres mantiene la importancia de saber usar cada letra para construir un texto interesante, llamativo, bien redactado y que despierte el interés de las personas, independiente de que el objetivo que sea.

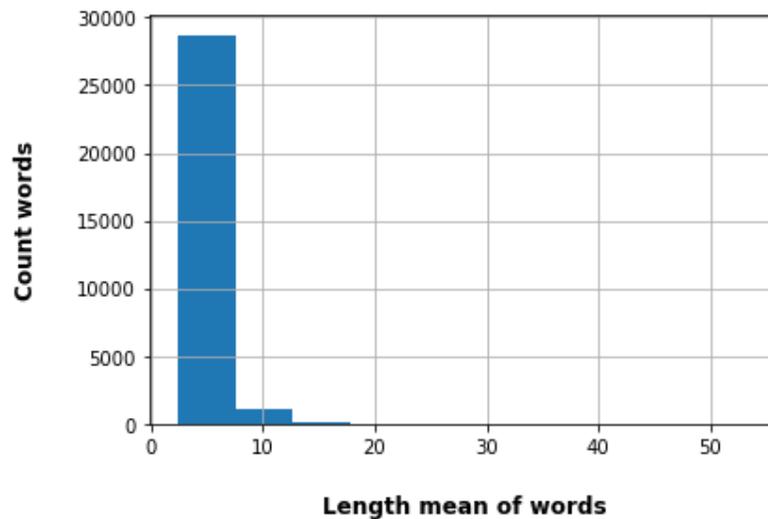


Figura 26 Histograma de longitud promedio de palabras en los tweets ZIKA.

En la En la **Figura 26** la longitud promedio de las palabras oscila entre 3 y 11 caracteres, siendo 7 la longitud más común. La materialización del grafico indica que las palabras son tan cortas debido a la gran cantidad de stop-words de los tuits. Los stop-words son las palabras que se usan con mayor frecuencia en cualquier idioma como "de", "en", "el", etc. Como estas palabras son probablemente de longitud pequeña, estas palabras pueden provocar que el gráfico anterior quede sesgado.

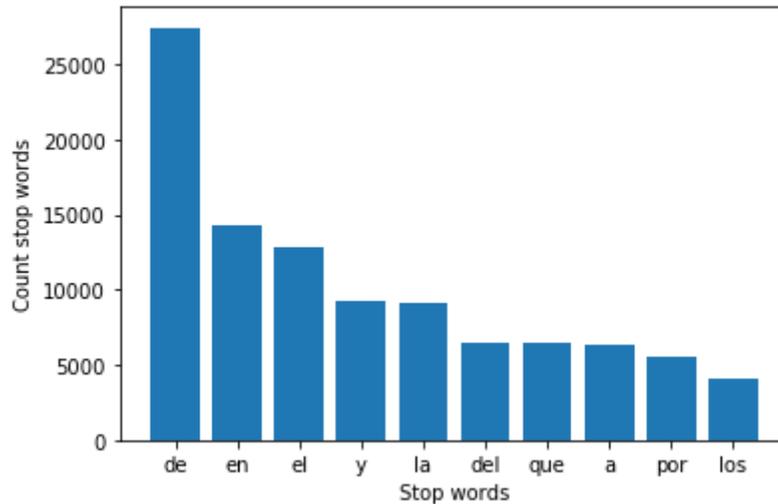


Figura 27 Histograma de s stop-word más utilizados en el corpus ZIKA.

En la **Figura 27** Histograma de s stop-word más utilizados en el corpus ZIKA., observamos que los stop-words como las preposiciones "de", "en" o "a" y "los" son determinantes como "el", "la" "los" que dominan en el texto dentro de los tweets. Por lo tanto, ahora podemos describir las palabras clave que aparecen con frecuencia en nuestro texto.

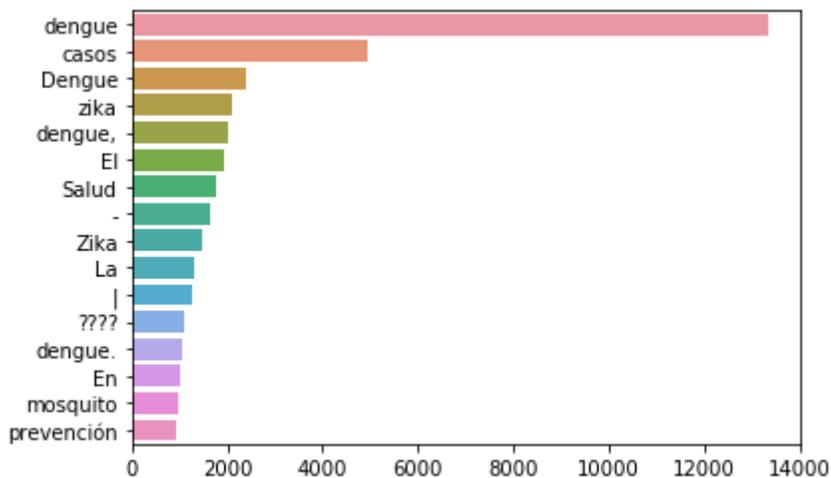


Figura 28 Histograma de Word más utilizados en el corpus ZIKA.

En la **Figura 28** Histograma de Word más utilizados en el corpus ZIKA. muestra las palabras "dengue", "casos" y la "zika" que tienen más representación en los textos en el corpus. Dando mayor

peso a la palabra “dengue” como pivote al identificativo de la aparición del virus y su correlación al corpus del ZIKA que contiene información relacionada con el dengue, zika y chikungunya.

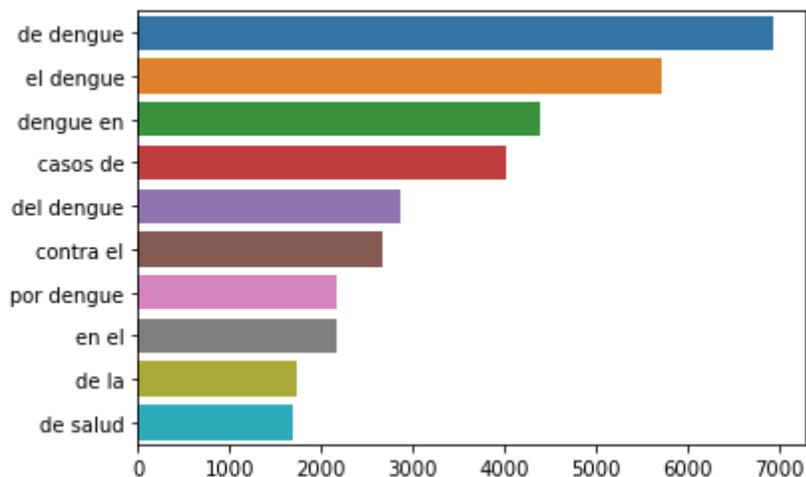


Figura 29 Histograma de bigramas más utilizados en el corpus ZIKA.

En la **Figura 29** Histograma de bigramas más utilizados en el corpus ZIKA. muestra que los bigramas como 'de dengue', 'el dengue' y 'dengue en' lleva la correlación como la combinación de palabras que están relacionados con el corpus del zika en los textos de los tweets.

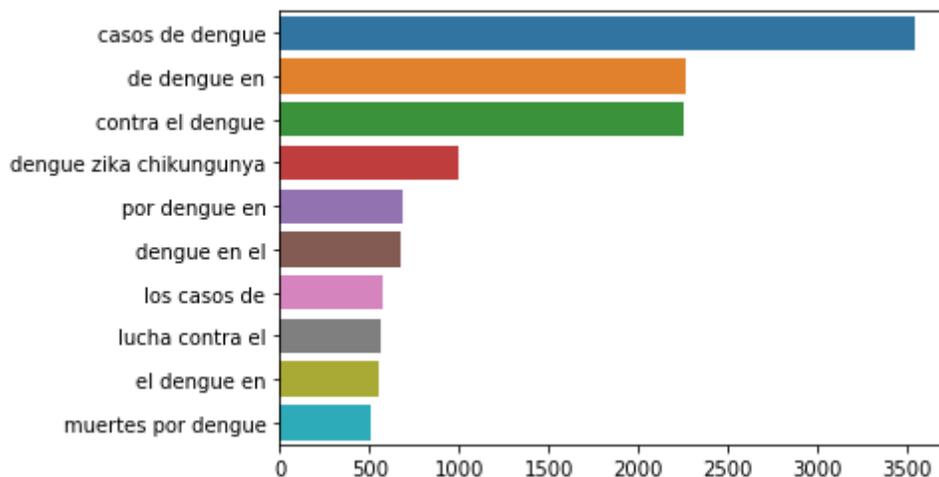


Figura 30 Histograma de trigramas más utilizados en el corpus ZIKA.

En la **Figura 30** muestra que muchos de estos trigramas son algunas combinaciones de "casos de dengue", "de dengue en " y "contra el dengue". Son trigramas que correlacionan a las expresiones que va describiendo el objetivo y expresiones de los textos en cada tweet.

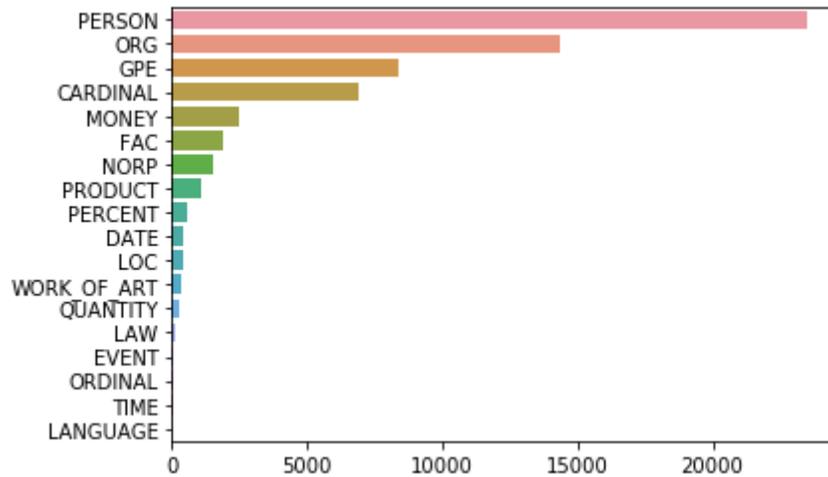


Figura 31 Histograma de entidades más utilizados en el corpus ZIKA.

En la **Figura 31** Muestra el reconocimiento de entidades es un método de extracción de información en el que las entidades que están presentes en el texto se clasifican en tipos de entidad predefinidos como "Persona", "Lugar", "Organización", etc. Al usar esta herramienta se puede obtener información sobre los tipos de entidades presentes en el conjunto de datos de texto. Así podemos indicar que el PERSON (personas, incluyendo ficticios) y el ORG (Empresas, agencias, instituciones, etc.) dominan los textos del tweet seguidos por la entidad GPE (Países, ciudades, estados).

La herramienta de exploración sobre las partes del etiquetado en el discurso nos permite entender qué tan legible (dificultad de lectura) es el texto y qué tipo de lector puede comprenderlo por completo. Así se ha detallado el índice de legibilidad en un documento o texto. El índice de legibilidad es un valor numérico que indica cuán difícil (o fácil) es leer y comprender un texto ver formula (6).

$$ARI = 4.71 + (\text{caracteres} / \text{palabras}) + 0.5 * (\text{palabras} / \text{oraciones}) - 21.43 \quad (6)$$

Índice de legibilidad automatizada (ARI) el resultado es una representación aproximada del nivel de grado de EE. UU. Necesario para comprender un texto. $ARI = 4.71 * (\text{caracteres} / \text{palabras}) + 0.5 * (\text{palabras} / \text{oración}) - 21.43$

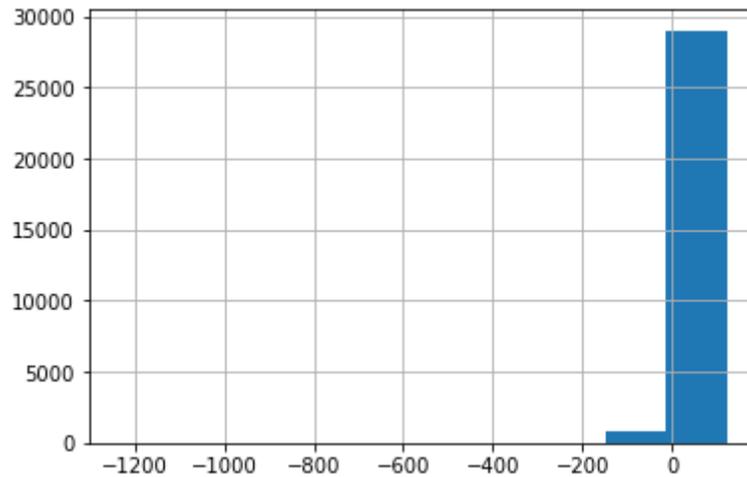


Figura 32 Histograma de índice de legibilidad ARI en el corpus ZIKA.

En **Figura 32** Muestra que el índice de legibilidad, en casi todos los puntajes de legibilidad caen por encima de 60. Esto indica que un estudiante promedio de 11 años puede leer y comprender los textos de los tweets.

3.6.1.1 Visualización del corpus ZIKA.

En esta sección se analiza la localización de los tuits del corpus utilizando mapas cartográficos para determinar su geoposicionamiento y tener una mejor perspectiva de la distribución de la información en gráficos de calor.

El servicio web que provee la plataforma de Twitter también tiene limitantes para su uso; por lo hemos utilizado parámetros que permiten leer una cantidad determinada de tweet y luego esperar un tiempo en segundos para poder hacer la siguiente lectura. Esto incrementó el tiempo de procesamiento para el enriquecimiento de los conjuntos de datos.

Los tuits contenidos en los corpus tienen distintas localizaciones y se trató de obtener los datos de geolocalización y la descripción ampliada de lo que se conoce sobre dirección de la cuenta que genero el tweet. Para obtener esta información se utilizó un servicio web de geo codificación.

El servicio web de geo codificación requiere de un dato descriptivo de entrada como son datos de dirección o nombres de ciudad o nombres de país. Del cual se identificaron como dato enriquecedor en el tweet; y su resultado son los datos de geoposicionamiento de latitud y longitud.

El conjunto de datos que se procesó y del cual se obtuvo la latitud y longitud es un subconjunto del corpus original; ya el resultado de permitió obtener de la información pública disponible y quedaron por fuera los tweets cuyos niveles de permiso no permitieron obtener algún dato descriptivo para la

identificación del posicionamiento, En esta **Figura 33** podemos visualizar la ubicación de tweet geolocalizados por el número de palabras en cada tweet tenemos el siguiente mapa.



Figura 33 Numero de mensaje por zona de salud ZIKA

Las zonas del Ecuador son para acercar el estado a toda la ciudadanía, a través de la prestación de servicios eficientes, se requirió de una planificación. Con este propósito, la Secretaría Nacional de Planificación y Desarrollo, SENPLADES, conformó niveles administrativos de planificación: zonas, distritos y circuitos a nivel nacional, que permitirán una mejor identificación de necesidades y soluciones efectivas para la prestación de servicios públicos en el territorio. Esta conformación no implica eliminar las provincias, cantones o parroquias. Las zonas están conformadas por provincias, de acuerdo con una proximidad geográfica, cultural y económica. Tenemos 9 zonas administrativas de planificación. Cada zona está constituida por distritos y estos a su vez por circuitos. Desde este nivel se coordina estratégicamente las entidades del sector público, a través de la gestión de la planificación para el diseño de políticas en el área de su jurisdicción:

- Zona 1: Esmeraldas, Imbabura, Carchi, Sucumbíos.
- Zona 2: Pichincha (excepto el cantón Quito), Napo, Orellana.
- Zona 3: Cotopaxi, Tungurahua, Chimborazo, Pastaza.
- Zona 4: Manabí, Santo Domingo de los Tsáchilas.
- Zona 5: Santa Elena, Guayas (excepto los cantones de Guayaquil Samborondón y Durán), Bolívar, Los Ríos y Galápagos.

- Zona 6: Cañar, Azuay, Morona Santiago.
- Zona 7: El Oro, Loja, Zamora Chinchipe.
- Zona 8: Cantones de Guayaquil, Samborondón y Durán.
- Zona 9: Distrito Metropolitano de Quito

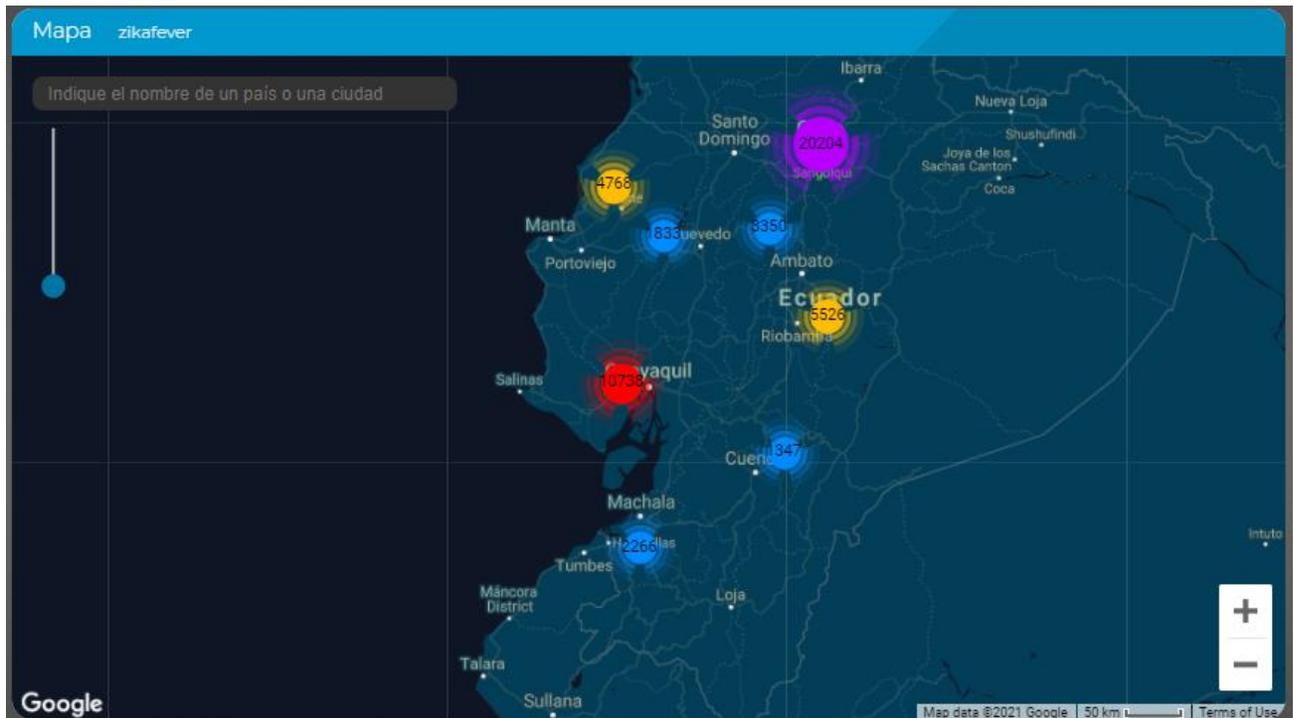


Figura 34 Mapa de calor de la información focalizada en Ecuador de ZIKA

El mapa de calor de la información en el Ecuador sobre el ZIKA describe la ubicación en el mapa que hace referencia del lugar del dueño de la cuenta que genero el tweet sobre el ZIKA, como la gráfica está basada en un mapa de calor, la dimensión del círculo y el color hace referencia a la cantidad de tweets generados en su geoposicionamiento. El Ecuador que es un país tropical tiene las características climáticas para que este tipo de enfermedades infectocontagiosas se presente y que ponen como un factor de riesgo a controlar por las autoridades sanitarias.

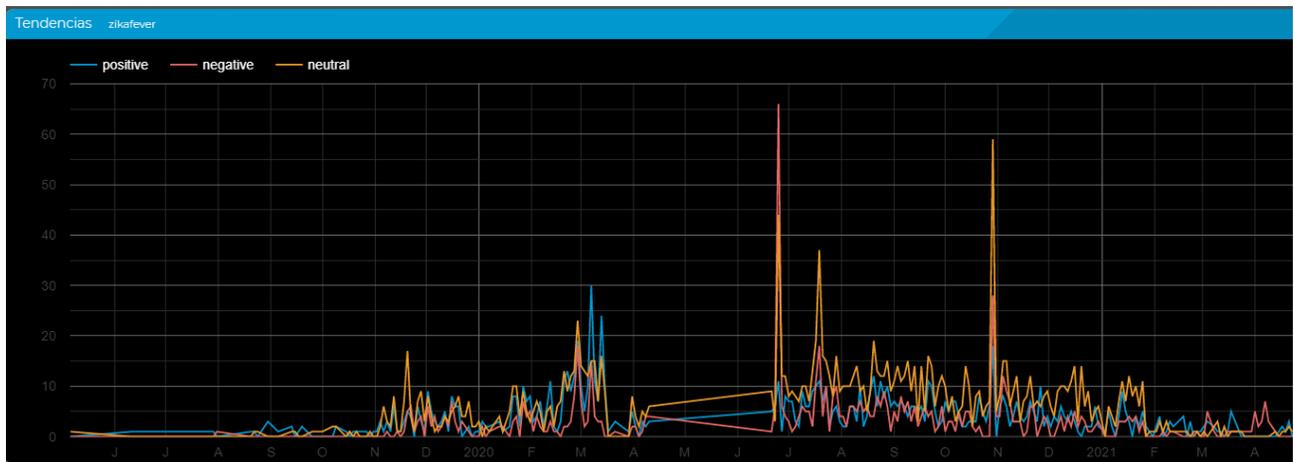


Figura 35 Tendencias de sentimientos de ZIKA

La tendencia de sentimientos sobre el ZIKA. indica que el virus afecta a muchos países de América. El Zika puede provocar parálisis progresiva de los músculos de tu cuerpo, en casos más graves hasta del sistema respiratorio, pudiendo causar la muerte: por lo cual, visualizamos en el grafico los sentimientos negativos. El riesgo es mayor para las mujeres embarazadas, ya que su contagio puede provocar que el bebé nazca con microcefalia: el tamaño de su cabeza puede ser más pequeño de lo normal, teniendo como efectos: retraso mental y afectación en la capacidad para moverse y hablar; por cual se van generando sentimientos neutros. El Zika es un virus transmitido por el mismo mosquito que causa “dengue y chikungunya”, vive en zonas menores a 1200m y se reproduce en sitios donde se almacena agua limpia, como: tanques, baldes y otros recipientes; es decir, mediante actividades específicas se puede contrarrestar generando sentimientos positivos.



Figura 36 Ejemplo de últimos tweets reportados ZIKA

El ejemplo de los últimos tweets reportado sobre el ZIKA son las opiniones expresadas que reflejan los puntos de vista de las personas que han contribuido a esta información y que referencian al sentimiento positivo según la categoría que se anotó para estos textos dentro del procesamiento del lenguaje natural. En las localidades donde la transmisión se mantiene de manera endémica durante todo el año, los ciclos epidémicos generalmente coinciden con la temporada de lluvias. La persistencia de la transmisión depende de distintos factores socioeconómicos, climáticos y ecológicos, principalmente de aquellos que permiten un aumento de la densidad de los mosquitos vectores en zonas donde está circulando el virus.

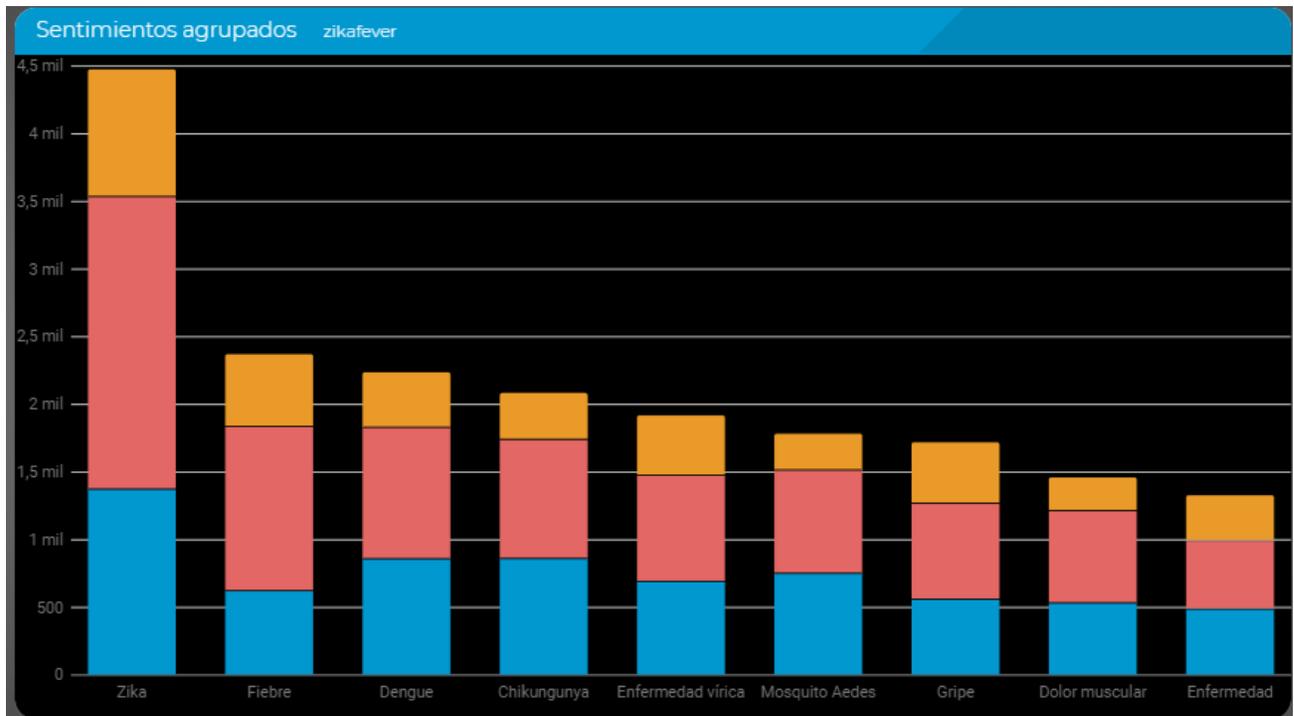


Figura 37 Sentimientos agrupados por conceptos en el ZIKA

Los sentimientos agrupados por conceptos del color azul son positivos, es color rosado es neutro y el color amarillo es negativo; así entre las enfermedades transmitidas por vectores representa un problema prioritario en salud pública debido al gran número de casos que se presentan cada año y se han registrado varios ciclos epidémicos, siendo las zonas tropicales y subtropicales de los países los que están en riesgo de transmisión de estos virus. El impacto de esta enfermedad depende de la distribución y densidad poblacional de los mosquitos vectores.

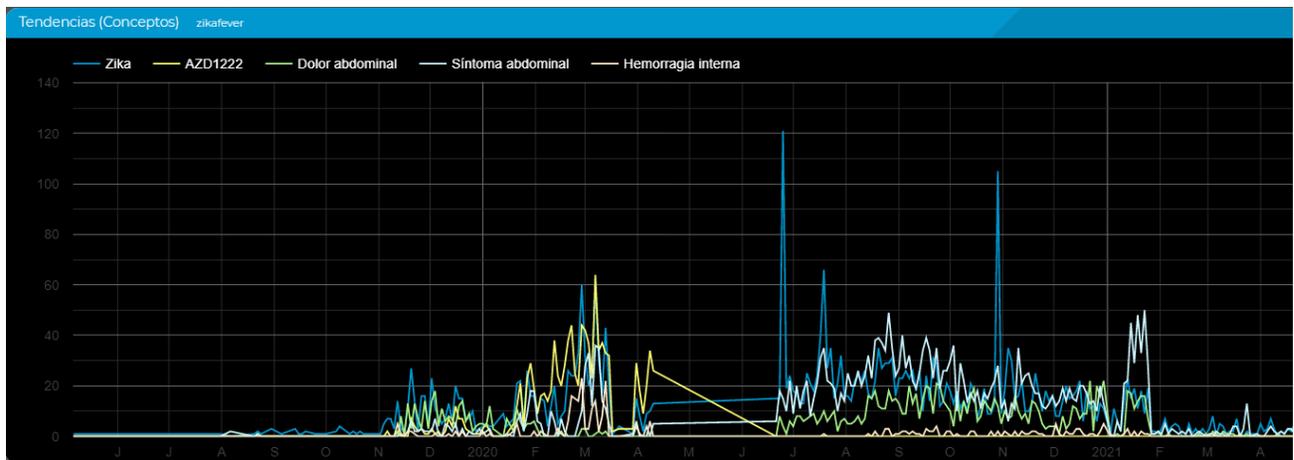


Figura 38 Sentimientos agrupados por conceptos en el ZIKA

Las tendencias de conceptos en una línea de tiempo nos demuestra que una persona afectado por estos virus podría padecer en la fase Crítica (24 a 48 horas posteriores a la disminución de la fiebre)

dolor abdominal intenso y continuo, vómito, hemorragia de las mucosas nasales y de las encías, sangrado vaginal en mujeres de edad fértil, letargia e irritabilidad, aumento de la permeabilidad capilar que podría generar derrame pleural (líquido en los pulmones) y la ascitis (líquido en cavidad abdominal), concentración de glóbulos rojos, estrechamiento de la presión arterial y sangrado digestivo.

3.6.1.2 Operacionalidad de los modelos para el corpus ZIKA.

Este trabajo (Apolinario-Arzube, García-Díaz, Luna-Aveiga, Medina-Moreira, & Valencia-García, 2020) correlaciono los principios de la infodemiología que consistió en la extracción y análisis de datos recopilados en Twitter sobre salud pública. Entre otras aplicaciones, la Infodemiología se puede utilizar para analizar tendencias en las redes sociales con el fin de determinar la prevalencia de brotes de enfermedades infecciosas en determinadas regiones. Este valioso dato aporta una mejor comprensión de la propagación de las enfermedades infecciosas, así como una visión sobre la percepción social de los ciudadanos hacia las estrategias que llevan a cabo las instituciones públicas de salud. En este trabajo, aplicamos técnicas de procesamiento del lenguaje natural para determinar el impacto de los brotes de enfermedades infecciosas como el Zika, el dengue o el chikungunya a partir de un conjunto de datos compilados con tweets escritos en español.

Para el corpus se utilizó la API de Twitter para consultar sobre tweets escritos en Idioma español que contiene palabras clave como Dengue y Zika en el ámbito geográfico zona de Ecuador. En esta primera etapa, filtramos automáticamente los tweets sin significado información, compuesta únicamente por menciones, enlaces o hashtags. En segundo lugar, se normalizaron los tweets realizando las siguientes acciones: (1) Eliminando el signo arroba de los nombres de usuario de Twitter, (2) Los enlaces se convirtieron en una etiqueta especial para facilitar la detección de tweets duplicados; debido a la mitificación de URL mecanismo de Twitter, la misma URL se codifica de manera diferente en copiar y pegar tweets y (3) retweets y tweets duplicados se eliminaron del corpus. En la **Figura 39.** se muestra un ejemplo de un tweet.



Figura 39 Ejemplos de tweets publicado en base al corpus del ZIKA.

En tercer lugar, se llevó a cabo la clasificación manual del corpus. Como aprendizaje automático aplicado al Análisis de Sentimiento se basa en técnicas supervisadas, todos los textos debían ser etiquetado como positivo o negativo. La clasificación manual es una tarea que requiere mucho tiempo. Sin embargo, debido al menor número de emoticonos en este corpus, se decidió el proceso de etiquetado manual. Contamos con la colaboración de un grupo de veinte voluntarios que todos los días durante cuatro semanas, etiquetaron un conjunto de tuits como positivos, negativo y fuera de dominio. Estos voluntarios realizaron un total de 39,816 manuales clasificaciones. Es importante resaltar que cada tweet fue revisado y clasificado por cada voluntario, por lo que se seleccionaron los tweets con mayor consenso de opinión, descartando los tweets más controvertidos.

La **Figura 40** detalla el flujo de trabajo de la clasificación manual. El color gris representa cuántos tweets se marcaron como fuera del dominio; los colores rojo y naranja representan tweets etiquetados como negativos o fuertemente negativos, y el amarillo y el verde representan tweets etiquetado como positivo o muy positivo.

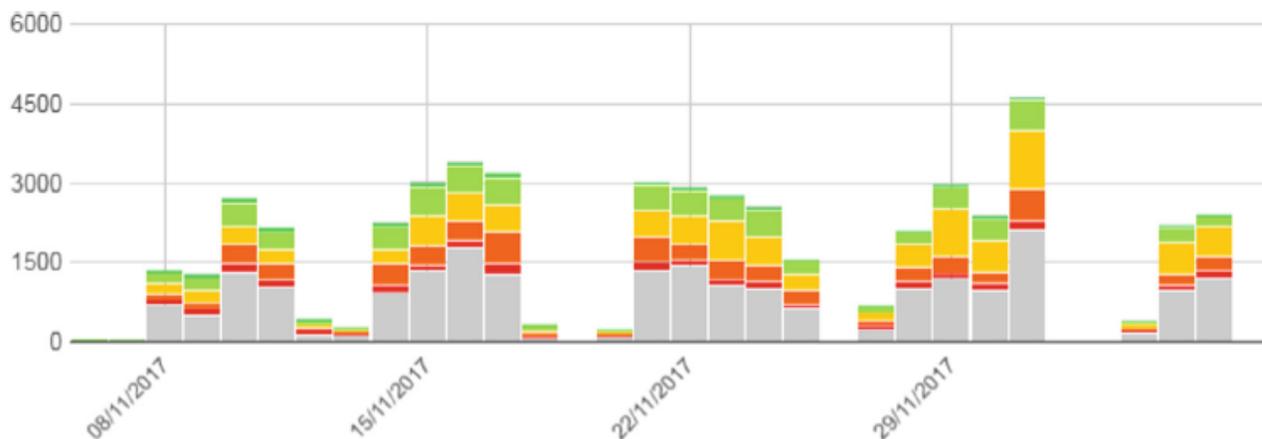


Figura 40 Ejemplos de tweets publicado en base al corpus del ZIKA.

Finalmente, una vez desarrollado el corpus, extrajimos características para realizar la clasificación de aprendizaje automático. Los modelos de los cuales se crearon para los experimentos son dos: uno para sentence word embeddings (obtención de vectores de palabras, clasificación de textos, cuantificación y autoajuste) y el segundo modelo para word embeddings (redes convolucionales, recurrentes, etc). En cada ejecución se obtuvo: (1) un fichero hyperparameters.csv que contiene el resumen de los modelos evaluados, (2) model-summary.txt, que contiene la arquitectura del mejor modelo, y (3) training_resume.json que contiene un resumen de la ejecución como cuál es la métrica usada para optimizar.

3.6.2 El corpus COVID-19.

Además, debido a la emergencia sanitaria causada por el COVID-19 que ocurrió a principios del año 2020, compilamos y etiquetamos tweets extraídos desde España y de Ecuador que denominamos corpus COVID.

En la **Tabla 7** se muestran algunas de las características generales del corpus COVID-19. Entre las estadísticas generales podemos visualizar el número de tweets recolectados para el corpus con una cantidad de 186499; la media de tweets compilados a la semana fue 3656,8431 ($665,16165 \sigma$), de entre todos los tweets la cantidad de cuentas(usuarios de la plataforma de Twitter) diferentes encontradas fue de 51,324; de estos se desprende que el número de tweets por cuenta generado fue de 3,63376 ($31,33563 \sigma$). Los usuarios que trabajaron sobre estos tweets realizaron una cantidad de 204885 anotaciones por lo cual fueron valorados 72638(38,9482 %), la cantidad de anotaciones por tweet fue de 0 (0 %) y la cantidad de anotadores fue de 28 que se dividieron el trabajo por tiempo de procesamiento en días y horas; la media de anotaciones por anotador fue de 7317 (0σ). Concluyendo con el porcentaje de acuerdo Krippendorff's Alpha 0,58073.

Tabla 7 Estadísticas generales del corpus COVID.

Métrica	Valor
Nº de tweets	186 499 tweets
Media de tweets compilados a la semana	3656,8431 (665,16165 σ)
Nº de cuentas de Twitter diferentes	51 324 cuentas
Nº de tweets por cuenta	3,63376 (31,33563 σ)
Nº de anotaciones	204 885
Nº de tweets valorados	72 638 (38,9482 %)
Anotaciones por tweet	0 (0 %)
Nº de anotadores	28
Media de anotaciones por anotador	7317 (0 σ)
Porcentaje de acuerdo Krippendorff's alpha	0,58073

En una primera visión de la información podemos resumir que hemos podido construir un corpus que contiene información representativa sobre el Covid-19 y enfermedades relacionadas como coronavirus. Así tenemos, por ejemplo, que los tweets tienen una vigencia significativa en los tiempos que van en un rango de meses en el año 2020 y cuya clasificación aplican como positivos son el 29.8% de todo el conjunto de tuits su clasificación directamente opuesta son los que aplican como negativo y alcanzan el 41.7%. Entre los datos más representativos con los datos que se califican como neutral el 28.5%.

Se comenzó a compilar tuits relacionados con el coronavirus dentro de Latinoamérica. En primer lugar, el corpus COVID-19 trata sobre las enfermedades del coronavirus, causas, síntomas y vacunas en América Central. Este corpus contiene 188618 tweets etiquetados y contiene palabras, hashtags relacionados con estas enfermedades compilados en Ecuador y países limítrofes.

En la **Figura 41**. podemos ver cuáles son las palabras más comunes del corpus del COVID. Para sacar estos valores hemos quitado las palabras conocidas como stop-words. Podemos observar que las palabras más representativas son aquellas relacionadas con la enfermedad, como síntomas. Por ejemplo, con respecto al COVID podemos ver palabras relacionadas con el estado de alarma, tales como cuarentena, palabras relacionadas con países como China, de donde se suponía que eran el origen del virus, y palabras que tienen que ver con la gestión de la pandemia como gobierno, crisis, país.

La OMS (organización mundial de la salud) realiza un seguimiento acerca del brote de enfermedad por coronavirus (COVID-19) que fue notificado por primera vez en Wuhan (China) el 31 de diciembre

de 2019. La propagación se da principalmente de persona a persona, siendo los adultos mayores de 60 años, personas con enfermedades preexistentes como diabetes o cardiopatías son quienes están en mayor riesgo de enfermarse gravemente. La transmisión es por el contacto personal cercano con una persona infectada o a través de personas infectadas al toser o estornudar; también al tocar objetos o superficies contaminadas y luego tocarse la boca, la nariz o los ojos.

La forma preventiva de prepararse a esta nueva enfermedad es cubrirse la boca con el antebrazo o cuando tosas o estornudes, o con un pañuelo desechable y limpiarse las manos frecuentemente con agua jabón y alcohol. Entre los síntomas más comunes son fiebre, cansancio, tos seca, neumonía y dificultad respiratoria; los síntomas pueden aparecer de 2 a 14 días después de la exposición al virus.

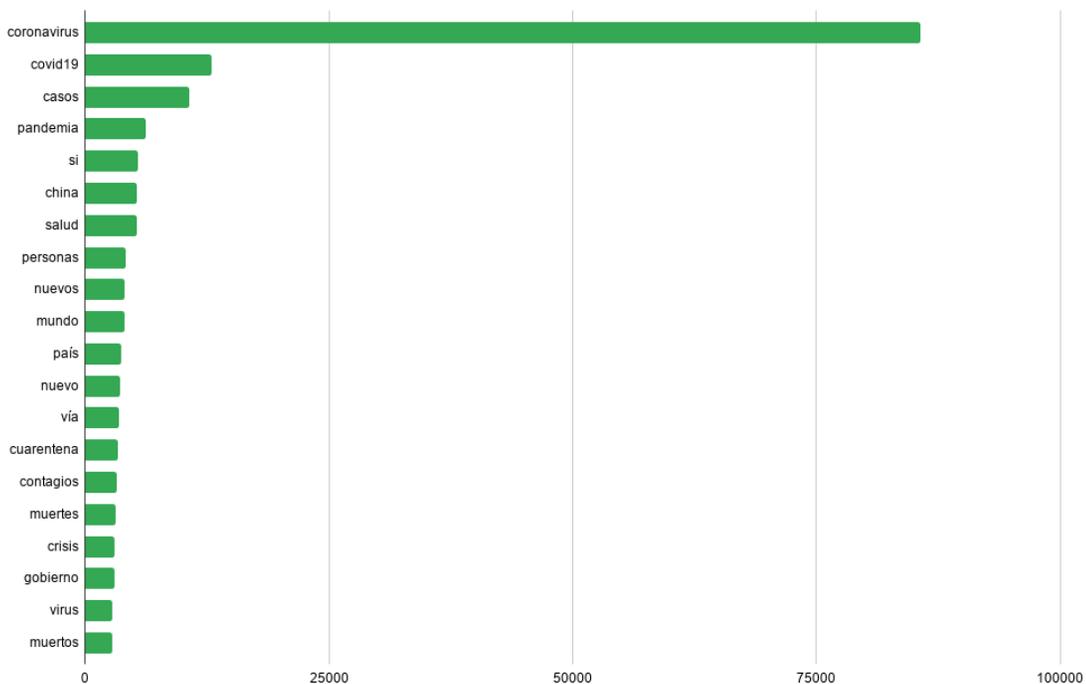


Figura 41 Palabras más frecuentes del corpus de COVID (verde).

Para el análisis del corpus se ha incorporado estadísticas descriptivas que permiten ver de forma visual la estructuración y la composición contextual del conjunto de tweets. A continuación, los datos más representativos que forman parte de este corpus.

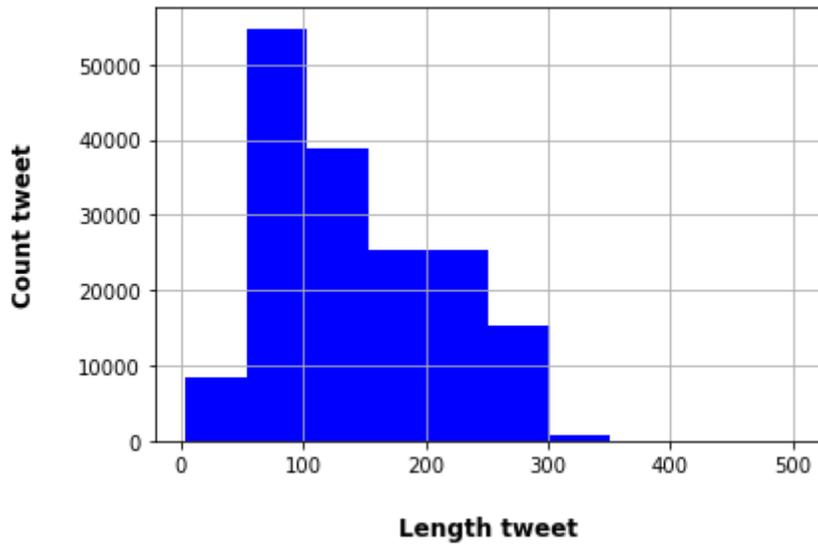


Figura 42 Histograma de longitud de textos del corpus COVID.

En la **Figura 42**, se muestra que los textos de los tweets oscilan entre 2 y 280 caracteres, en lo general, cada texto está en un rango de 45 a 280 caracteres. Y el pico máximo de la composición de caracteres en un texto está en 75 a 100. En un inicio, Twitter permitió una cantidad de caracteres con un límite de 140 caracteres y luego se amplió este límite en los tuits a 280.

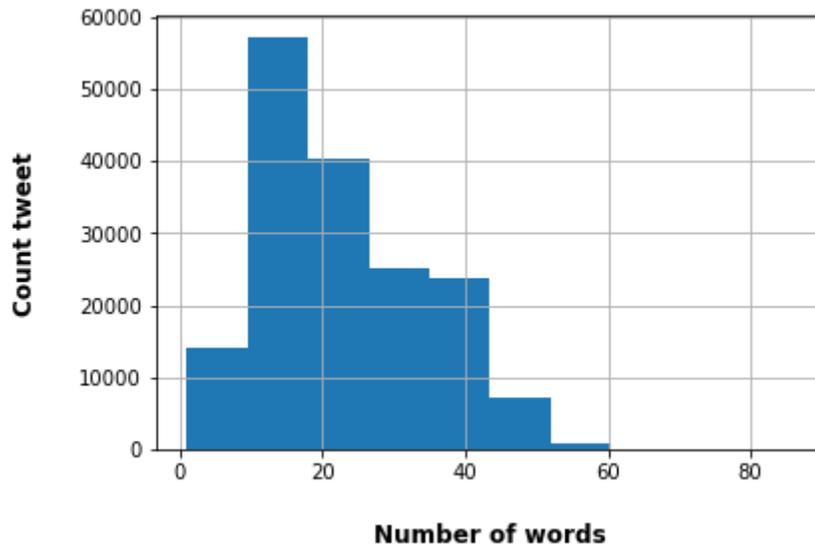


Figura 43 Histograma de números de palabras en los tweets COVID.

En la **Figura 43** presenta la composición de la cantidad de palabras utilizadas en los tweets de este corpus. El número de palabras en los textos de los tweets varía de 2 a 60 y en su mayoría cae en un rango de 8 y 18 palabras. El límite de caracteres mantiene la importancia de saber usar cada letra para

construir un texto interesante, llamativo, bien redactado y que despierte el interés de las personas, independiente de que el objetivo que sea.

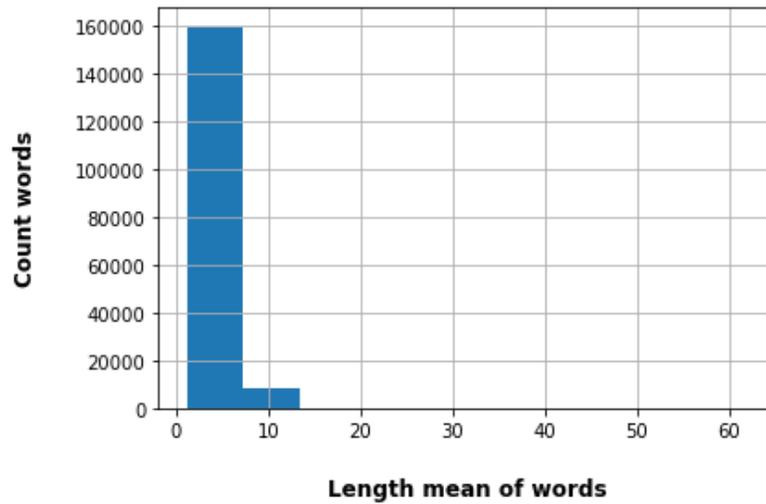


Figura 44 Histograma de longitud promedio de palabras en los tweets COVID.

En la **Figura 44** la longitud promedio de las palabras oscila entre 3 y 11 caracteres, siendo 7 la longitud más común. La materialización del gráfico indica que las palabras son tan cortas debido a la gran cantidad de stop-words de los tuits. Los stop-words son las palabras que se usan con mayor frecuencia en cualquier idioma como "de", "el", "la", etc. Como estas palabras son probablemente de longitud pequeña, estas palabras pueden provocar que el gráfico anterior quede sesgado.

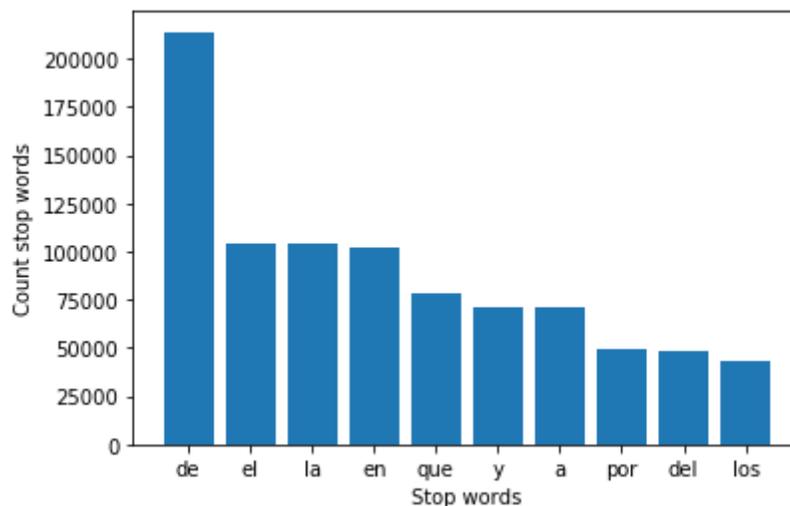


Figura 45 Histograma de stop-word más utilizados en el corpus COVID.

En la **Figura 45**, observamos que los stop-words como las preposiciones "de", "el" o "la" y "en" son determinantes como "el", "la" "los" que dominan en el texto dentro de los tweets. Por lo tanto, ahora podemos describir las palabras clave que aparecen con frecuencia en nuestro texto.

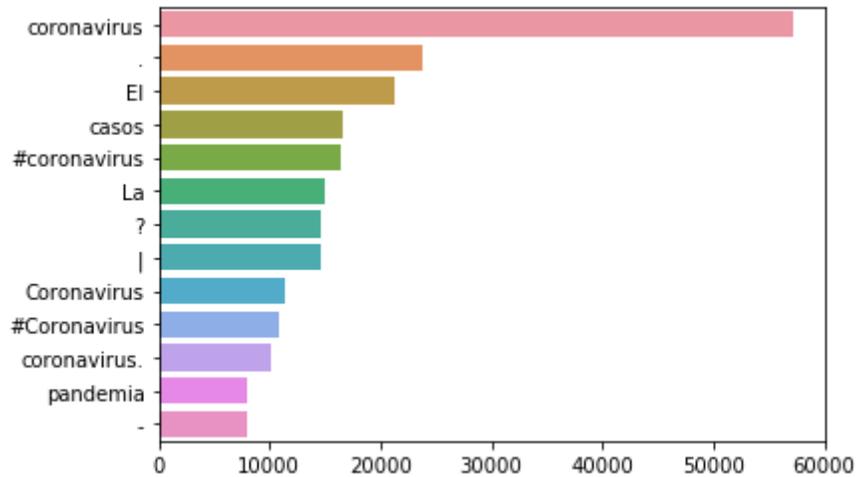


Figura 46 Histograma de Word más utilizados en el corpus COVID.

En la **Figura 46** muestra las palabras "coronavirus", "casos" y la "pandemia" que tienen más representación en los textos en el corpus. Dando mayor peso a la palabra "coronavirus" como pivote al identificativo de la aparición del virus y su correlación al corpus del COVID que contiene información relacionada con la pandemia del SARS-COV2.

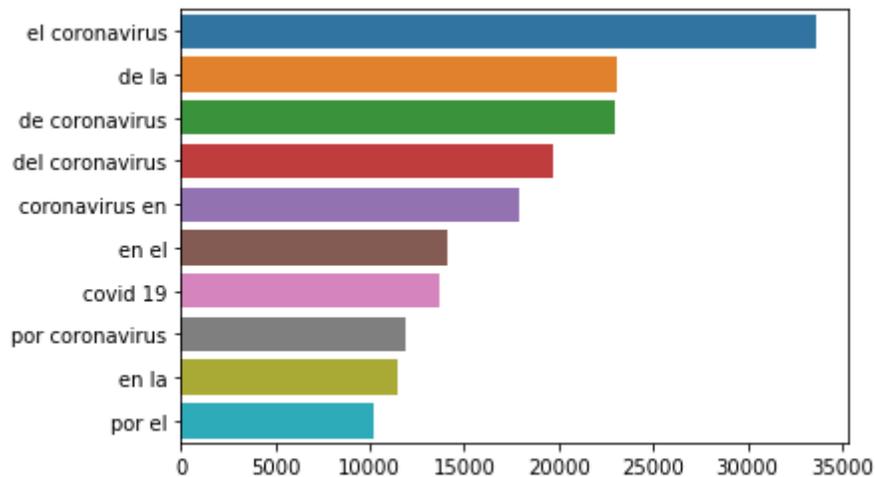


Figura 47 Histograma de bigramas más utilizados en el corpus COVID.

En la **Figura 47** muestra que los bigramas como 'el coronavirus', 'el coronavirus' y 'coronavirus en' lleva la correlación como la combinación de palabras que están relacionados con el corpus del covid en los textos de los tweets.

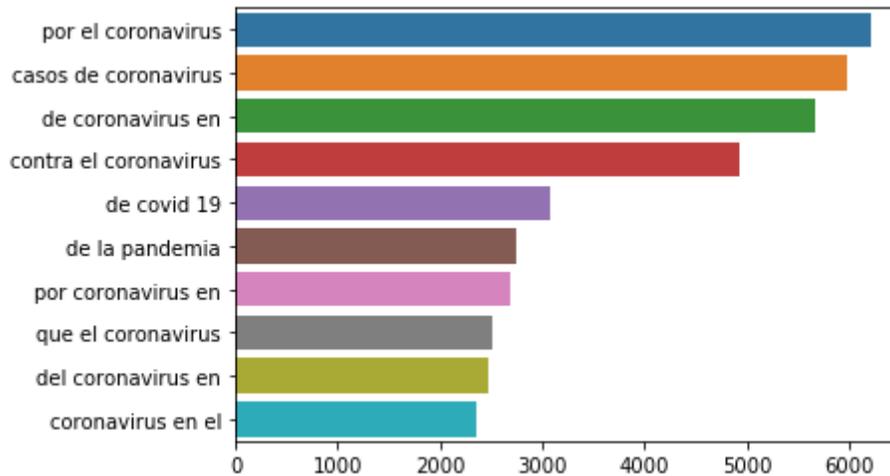


Figura 48 Histograma de trigramas más utilizados en el corpus COVID.

En la **Figura 48** muestra que muchos de estos trigramas son algunas combinaciones de "casos de coronavirus", "por el coronavirus " y "de coronavirus en". Son trigramas que correlacionan a las expresiones que va describiendo el objetivo y expresiones de los textos en cada tweet.

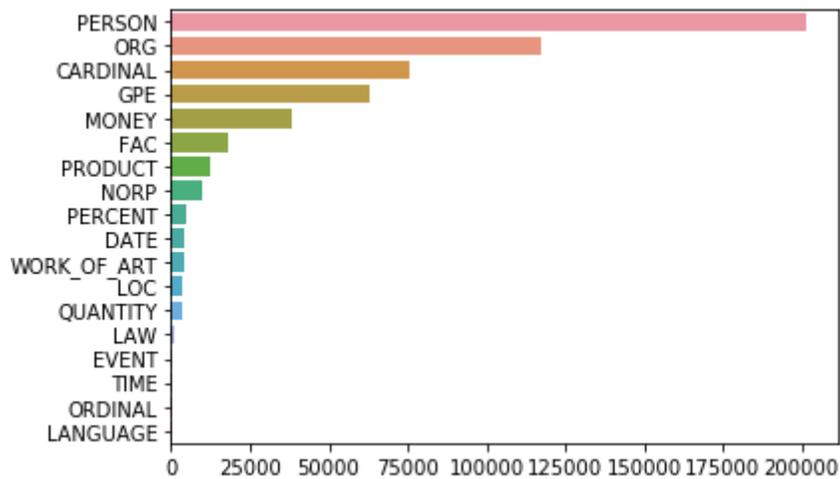


Figura 49 Histograma de entidades más utilizados en el corpus COVID.

En la **Figura 49** Muestra el reconocimiento de entidades es un método de extracción de información en el que las entidades que están presentes en el texto se clasifican en tipos de entidad predefinidos como "Persona", "Lugar", "Organización", etc. Al usar esta herramienta se puede obtener información sobre los tipos de entidades presentes en el conjunto de datos de texto. Así podemos indicar que el PERSON (personas, incluyendo ficticios) y el ORG (Empresas, agencias, instituciones, etc.) dominan los textos del tweet seguidos por la entidad CARDINAL (numerología de covid-19).

La herramienta de exploración sobre las partes del etiquetado en el discurso nos permite entender qué tan legible (dificultad de lectura) es el texto y qué tipo de lector puede comprenderlo por completo.

Así se ha detallado el índice de legibilidad en un documento o texto. El índice de legibilidad es un valor numérico que indica cuán difícil (o fácil) es leer y comprender un texto ver formula (6).

$$ARI = 4.71 + (\text{caracteres} / \text{palabras}) + 0.5 * (\text{palabras} / \text{oraciones}) - 21.43 \quad (6)$$

Índice de legibilidad automatizada (ARI) el resultado es una representación aproximada del nivel de grado de EE. UU. Necesario para comprender un texto. $ARI = 4.71 * (\text{caracteres} / \text{palabras}) + 0.5 * (\text{palabras} / \text{oración}) - 21.43$

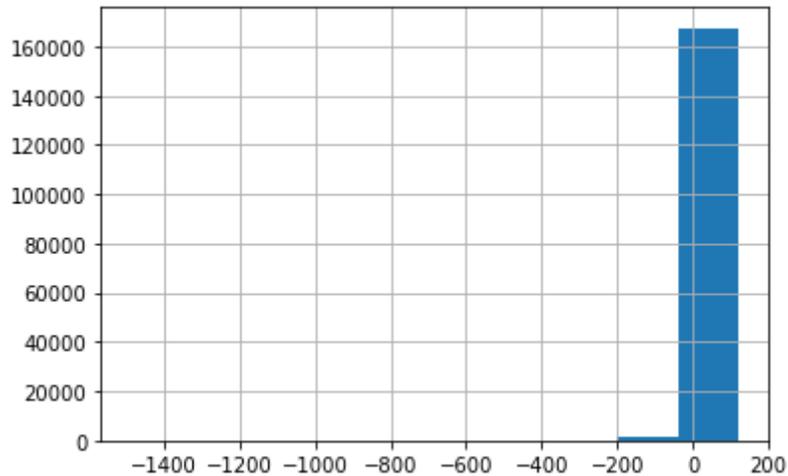


Figura 50 Histograma de índice de legibilidad ARI en el corpus COVID.

En Figura 50 Muestra que el índice de legibilidad, en casi todos los puntajes de legibilidad caen por encima de 60. Esto indica que un estudiante promedio de 11 años puede leer y comprender los textos de los tweets.

3.6.2.1 Visualización del corpus COVID.

En esta sección se analiza la localización de los tuits del corpus utilizando mapas cartográficos para determinar su geoposicionamiento y tener una mejor perspectiva de la distribución de la información en gráficos de calor.

El servicio web que provee la plataforma de Twitter también tiene limitantes para su uso; por lo hemos utilizado parámetros que permiten leer una cantidad determinada de tweet y luego esperar un tiempo en segundos para poder hacer la siguiente lectura. Esto incrementó el tiempo de procesamiento para el enriquecimiento de los conjuntos de datos.

Los tuits contenidos en los corpus tienen distintas localizaciones y se trató de obtener los datos de geolocalización y la descripción ampliada de lo que se conoce sobre dirección de la cuenta que genero el tweet. Para obtener esta información se utilizó un servicio web de geo codificación.

El servicio web de geo codificación requiere de un dato descriptivo de entrada como son datos de dirección o nombres de ciudad o nombres de país. Del cual se identificaron como dato enriquecedor en el tweet; y su resultado son los datos de geoposicionamiento de latitud y longitud.

El conjunto de datos que se procesó y del cual se obtuvo la latitud y longitud es un subconjunto del corpus original; ya el resultado de permitió obtener de la información pública disponible y quedaron por fuera los tweets cuyos niveles de permiso no permitieron obtener algún dato descriptivo para la identificación del posicionamiento, En esta **Figura 51** podemos visualizar la ubicación de tweet geolocalizados por el número de palabras en cada tweet tenemos el siguiente mapa.



Figura 51 Numero de mensaje por zona de salud COVID-19

Las zonas del Ecuador son para acercar el estado a toda la ciudadanía, a través de la prestación de servicios eficientes, se requirió de una planificación. Con este propósito, la Secretaría Nacional de Planificación y Desarrollo, SENPLADES, conformó niveles administrativos de planificación: zonas, distritos y circuitos a nivel nacional, que permitirán una mejor identificación de necesidades y soluciones efectivas para la prestación de servicios públicos en el territorio. Esta conformación no implica eliminar las provincias, cantones o parroquias. Las zonas están conformadas por provincias, de acuerdo con una proximidad geográfica, cultural y económica. Tenemos 9 zonas administrativas de planificación. Cada zona está constituida por distritos y estos a su vez por circuitos. Desde este nivel se coordina estratégicamente las entidades del sector público, a través de la gestión de la planificación para el diseño de políticas en el área de su jurisdicción:

- Zona 1: Esmeraldas, Imbabura, Carchi, Sucumbíos.
- Zona 2: Pichincha (excepto el cantón Quito), Napo, Orellana.
- Zona 3: Cotopaxi, Tungurahua, Chimborazo, Pastaza.
- Zona 4: Manabí, Santo Domingo de los Tsáchilas.
- Zona 5: Santa Elena, Guayas (excepto los cantones de Guayaquil Samborondón y Durán), Bolívar, Los Ríos y Galápagos.
- Zona 6: Cañar, Azuay, Morona Santiago.
- Zona 7: El Oro, Loja, Zamora Chinchipe.
- Zona 8: Cantones de Guayaquil, Samborondón y Durán.
- Zona 9: Distrito Metropolitano de Quito

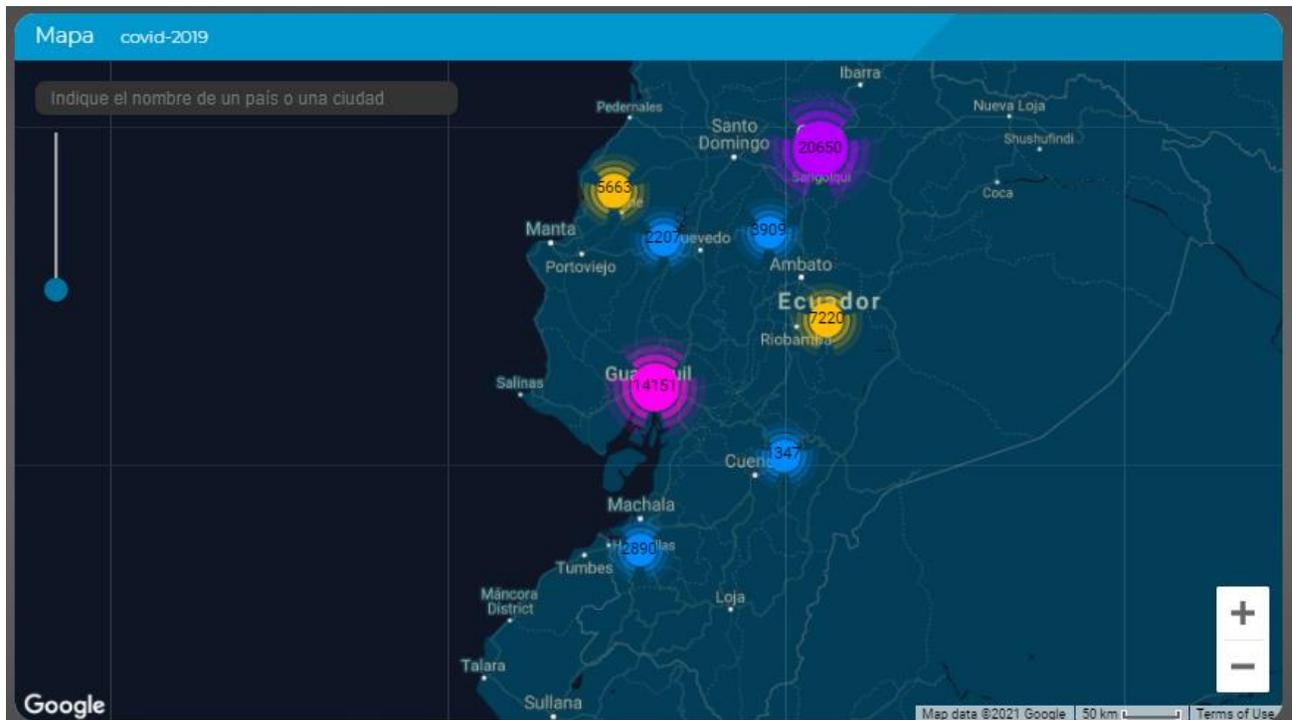


Figura 52 Mapa de calor de la información focalizada en Ecuador de COVID-19

El mapa de calor de la información en el Ecuador sobre el COVID-19 describe la ubicación en el mapa que hace referencia del lugar del dueño de la cuenta que genero el tweet sobre el COVID-19, como la gráfica está basada en un mapa de calor, la dimensión del círculo y el color hace referencia a la cantidad de tweets generados en su geoposicionamiento. El Ecuador que es un país tropical tiene las características climáticas para que este tipo de enfermedades infectocontagiosas se presente y que ponen como un factor de riesgo a controlar por las autoridades sanitarias.

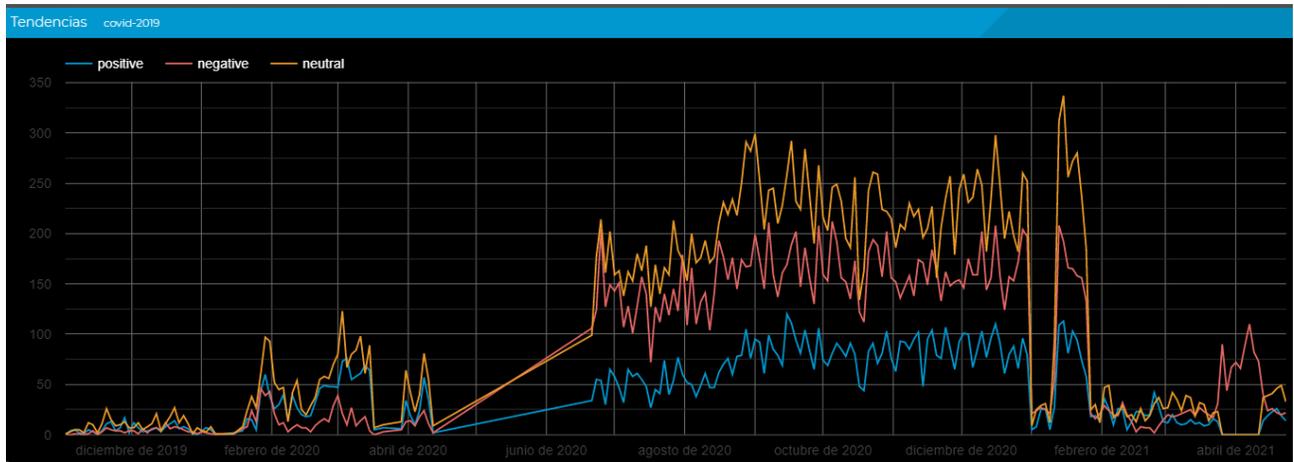


Figura 53 Tendencias de sentimientos de COVID-19

La tendencia de sentimientos sobre el COVID-19, indica que el virus afecta a muchos países de América. El Covid-19 puede provocar inflamación progresiva de los sistemas de tu cuerpo, en casos más graves hasta del sistema respiratorio, pudiendo causar la muerte: por lo cual, visualizamos en el grafico los sentimientos negativos. El riesgo es mayor para las personas de la tercera edad, ya que su contagio puede provocar hospitalización del paciente: el tratamiento comienza en una sala de triage, pudiendo pasar a salas UCI (unidad de cuidados intensivos) por falta de la saturación de oxígeno en la sangre; y los casos más graves pasan a un proceso de entubación de paciente; punto crítico desde el cual muchos pacientes no resisten a los tratamientos provocando el fallecimiento. El COVID-19 es un virus transmitido de persona a persona que causa una “anemia atípica”. El primer programa de vacunación colectiva se puso en marcha a principios de diciembre de 2020 y el número de dosis vacunales permite a la población que lo padecen tener esperanza de vida; es decir, mediante actividades de vacunación masiva se puede contrarrestar la enfermedad generando sentimientos positivos.

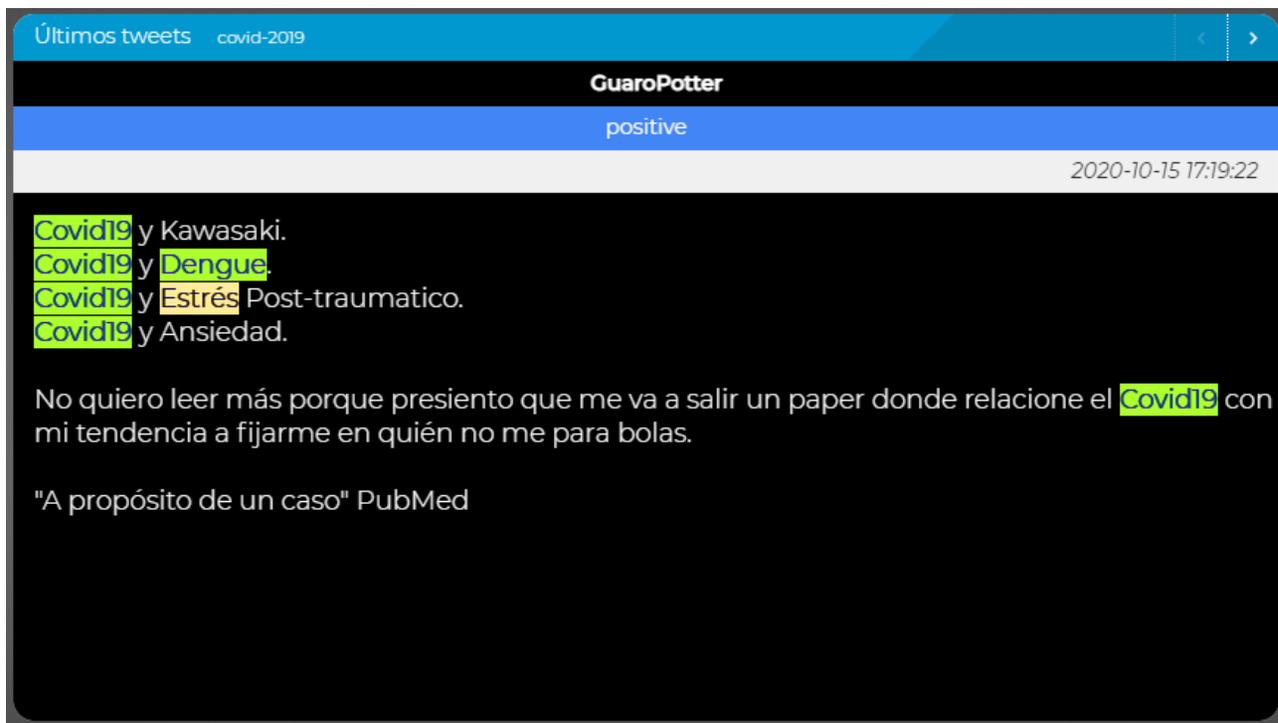


Figura 54 Ejemplo de últimos tweets reportados COVID-19

El ejemplo del último tweet reportado sobre el COVID-19 son las opiniones expresadas que reflejan los puntos de vista de las personas que han contribuido a esta información y que referencian al sentimiento positivo según la categoría que se anotó para estos textos dentro del procesamiento del lenguaje natural. Los síntomas de la enfermedad por coronavirus 2019 (COVID-19) pueden variar ampliamente. Algunas personas no presentan ningún síntoma, mientras que otras se enferman tanto que, a la larga, necesitan asistencia mecánica para respirar. El riesgo de tener síntomas peligrosos de la COVID-19 puede aumentar en las personas mayores y también en las personas de cualquier edad que tienen otros problemas de salud graves, como afecciones cardíacas o pulmonares, un sistema inmunitario debilitado, obesidad o diabetes. Esto es similar a lo que se ve con otras enfermedades respiratorias, como la influenza (gripe).

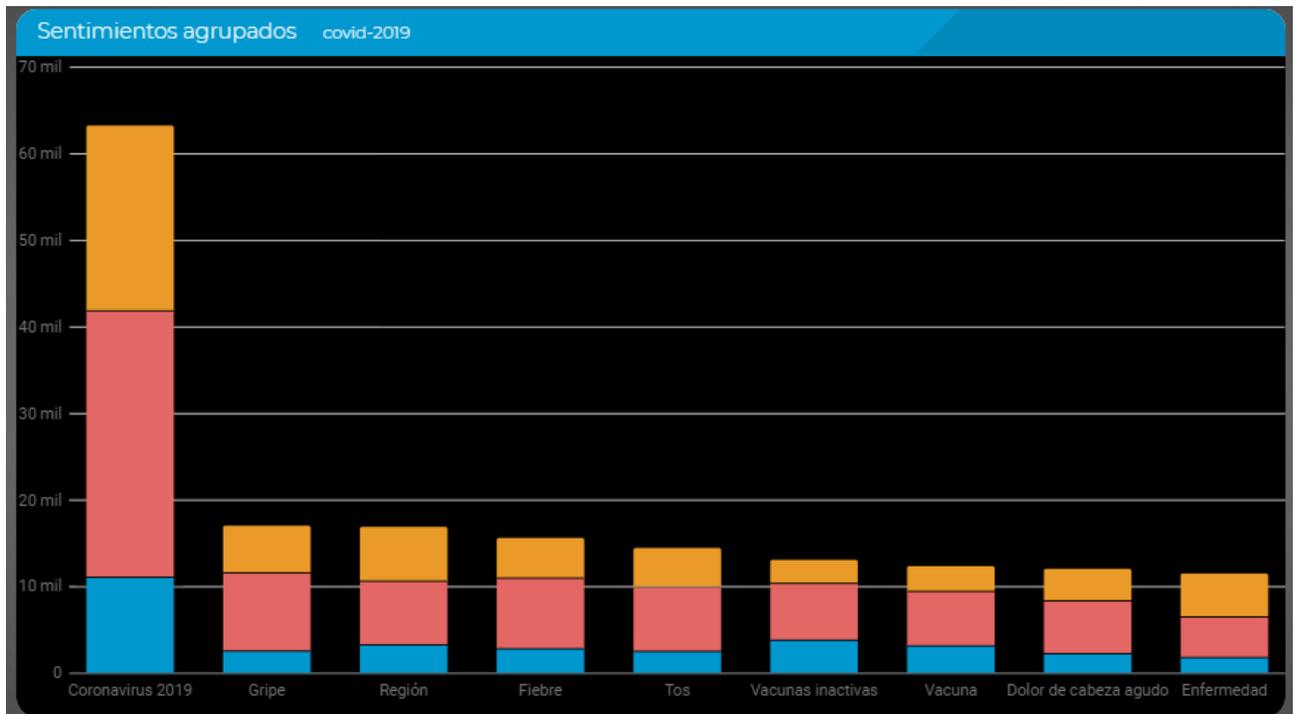


Figura 55 Sentimientos agrupados por conceptos en el COVID-19

Los sentimientos agrupados por conceptos del color azul son positivos, el color rosado es neutro y el color amarillo es negativo; así entre las personas de cualquier edad, incluso los niños, pueden contagiarse de COVID-19. Pero más comúnmente afecta a los adultos de edad mediana y mayores. El riesgo de presentar síntomas peligrosos aumenta con la edad, y los mayores de 85 años corren un riesgo más alto de tener síntomas de gravedad. La COVID-19 ataca los pulmones, así que es más probable que presentes síntomas graves si ya se tiene problemas pulmonares, tales como:

- Enfermedad pulmonar obstructiva crónica.
- Cáncer de pulmón
- Fibrosis quística
- Fibrosis pulmonar
- Asma moderada a grave.

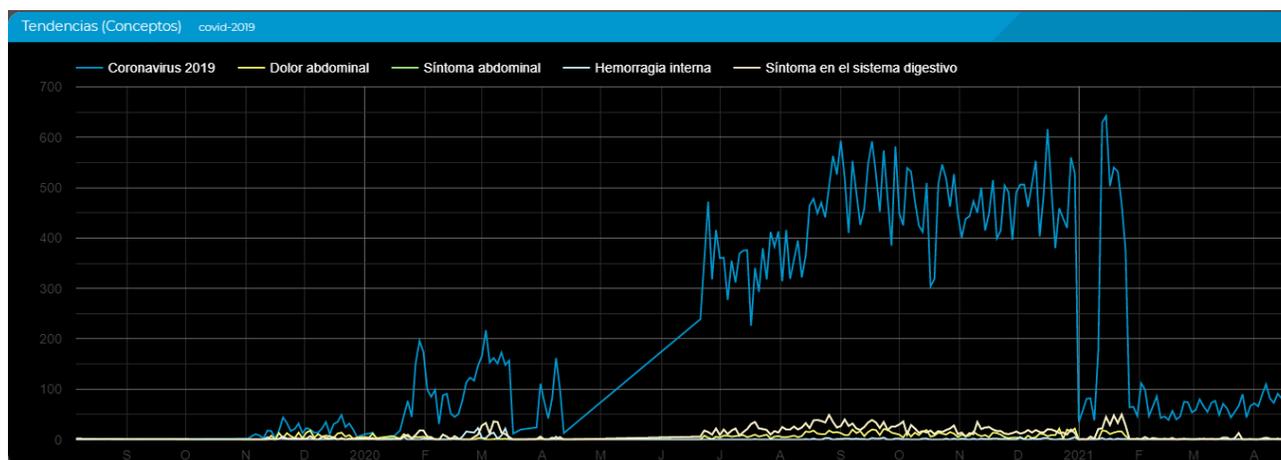


Figura 56 Sentimientos agrupados por conceptos en el COVID-19

Las tendencias de conceptos en una línea de tiempo nos demuestran que aunque la COVID-19 se considere una enfermedad que afecta principalmente a los pulmones, también puede dañar muchos otros órganos. Este daño a los órganos puede aumentar el riesgo de problemas de salud a largo plazo. Los órganos que pueden verse afectados por la COVID-19 incluyen: Corazón. Las pruebas por imágenes hechas meses después de la recuperación de la COVID-19 han mostrado daño duradero al músculo cardíaco, aun en el caso de las personas que presentaron solamente síntomas leves de COVID-19. Esto puede aumentar el riesgo de insuficiencia u otras complicaciones cardíacas en el futuro. Pulmones. El tipo de neumonía que suele asociarse a la COVID-19 puede causar daño duradero a los diminutos sacos de aire (alvéolos) de los pulmones. El tejido cicatricial resultante puede llevar a problemas respiratorios a largo plazo. Cerebro. Incluso en personas jóvenes, la COVID-19 puede causar accidentes cerebrovasculares, convulsiones y el síndrome de Guillain-Barré, una afección que causa parálisis temporaria. La COVID-19 también puede aumentar el riesgo de desarrollar la enfermedad de Parkinson y la del alzhéimer. Algunos adultos y niños presentan un síndrome inflamatorio multisistémico después de haber tenido COVID-19. En esta afección, algunos órganos y tejidos se inflaman gravemente.

3.6.2.2 Operacionalidad de los modelos para el corpus COVID-19.

Este trabajo (Apolinario-Arzube, García-Díaz, Luna-Aveiga, et al., 2020) correlaciono los principios de la infodemiología que consistió en la extracción y análisis de datos recopilados en Twitter sobre salud pública. Entre otras aplicaciones, la Infodemiología se puede utilizar para analizar tendencias en las redes sociales con el fin de determinar la prevalencia de brotes de enfermedades infecciosas en determinadas regiones. Este valioso dato aporta una mejor comprensión de la propagación de las enfermedades infecciosas, así como una visión sobre la percepción social de los ciudadanos hacia las estrategias que llevan a cabo las instituciones públicas de salud. En este trabajo, aplicamos técnicas

de procesamiento del lenguaje natural para determinar el impacto de los brotes de enfermedades infecciosas como el covid-19 a partir de un conjunto de datos compilados con tweets escritos en español.

Para el corpus se utilizó la API de Twitter para consultar sobre tweets escritos en Idioma español que contiene palabras clave como covid-19 y coronavirus en el ámbito geográfico zona de Ecuador. En esta primera etapa, filtramos automáticamente los tweets sin significado información, compuesta únicamente por menciones, enlaces o hashtags. En segundo lugar, se normalizaron los tweets realizando las siguientes acciones: (1) Eliminando el signo arroba de los nombres de usuario de Twitter, (2) Los enlaces se convirtieron en una etiqueta especial para facilitar la detección de tweets duplicados; debido a la mitificación de URL mecanismo de Twitter, la misma URL se codifica de manera diferente en copiar y pegar tweets y (3) retweets y tweets duplicados se eliminaron del corpus. En la **Figura 57**. se muestra un ejemplo de un tweet.



Figura 57 Ejemplos de tweets publicado en base al corpus del COVID-19.

En tercer lugar, se llevó a cabo la clasificación manual del corpus. Como aprendizaje automático aplicado al Análisis de Sentimiento se basa en técnicas supervisadas, todos los textos debían ser etiquetado como positivo o negativo. La clasificación manual es una tarea que requiere mucho tiempo. Sin embargo, debido al menor número de emoticonos en este corpus, se decidió el proceso de etiquetado manual. Contamos con la colaboración de un grupo de veinte voluntarios que todos los días durante cuatro semanas, etiquetaron un conjunto de tuits como positivos, negativo y fuera de dominio. Estos voluntarios realizaron un total de 188,618 manuales clasificaciones. Es importante resaltar que cada tweet fue revisado y clasificado por cada voluntario, por lo que se seleccionaron los tweets con mayor consenso de opinión, descartando los tweets más controvertidos.

La **Figura 58** detalla el flujo de trabajo de la clasificación manual. El color gris representa cuántos tweets se marcaron como fuera del dominio; los colores rojo y naranja representan tweets etiquetados

como negativos o fuertemente negativos, y el amarillo y el verde representan tweets etiquetado como positivo o muy positivo.

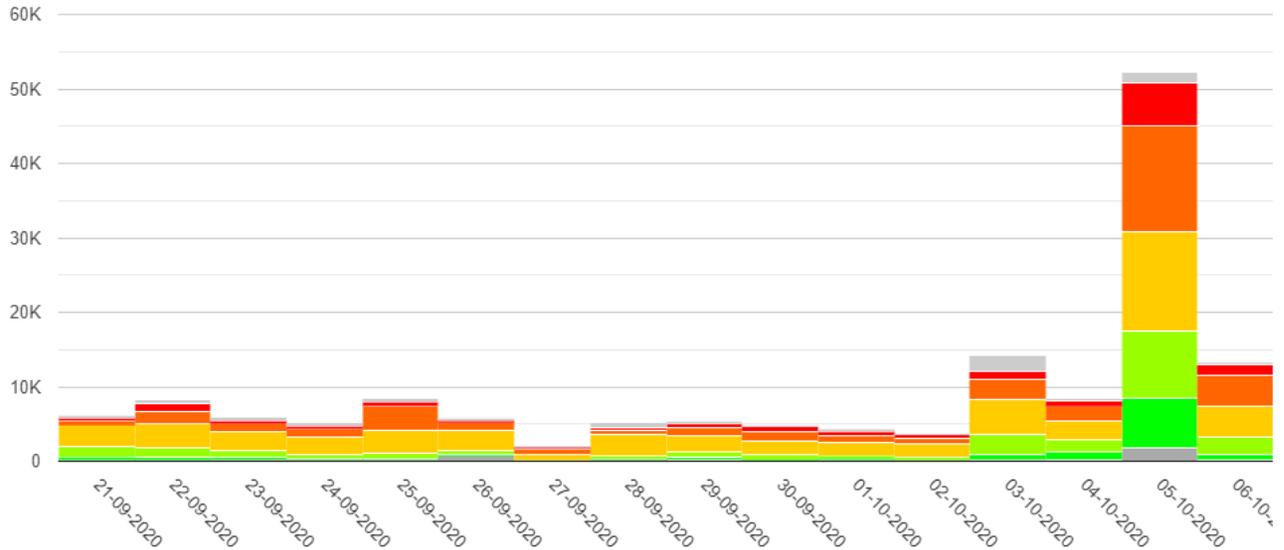


Figura 58 Ejemplos de tweets publicado en base al corpus del COVID-19.

Finalmente, una vez desarrollado el corpus, extrajimos características para realizar la clasificación de aprendizaje automático. Los modelos de los cuales se crearon para los experimentos son dos: uno para sentence word embeddings (obtención de vectores de palabras, clasificación de textos, cuantificación y autoajuste) y el segundo modelo para word embeddings (redes convolucionales, recurrentes, etc). En cada ejecución se obtuvo: (1) un fichero hyperparameters.csv que contiene el resumen de los modelos evaluados, (2) model-summary.txt, que contiene la arquitectura del mejor modelo, y (3) training_resume.json que contiene un resumen de la ejecución como cuál es la métrica usada para optimizar.

3.6.3 El corpus de la SATIRA

Por último, aunque esta tesis doctoral se centra en el análisis de sentimientos en tuits relacionados con enfermedades infecciosas, quisimos probar las tecnologías evaluadas para los anteriores corpus en el dominio de la sátira. Para eso utilizamos el corpus compilado en (Salas-Zárte et al., 2017, pp. 20-33) y que permitió condensar este trabajo en la siguiente publicación (Apolinario-Arzupe, García-Díaz, Medina-Moreira, Luna-Aveiga, & Valencia-García, 2020, pp. 1-23).

En la Tabla 8 se muestran algunas de las características generales del corpus SATIRA.

Tabla 8 Estadísticas generales del corpus SATIRA.

Feature-Set	Tweets	Satirical	Non-Satirical
European Spanish	4821	2488	2333
train (60%)	2892	1493	1400
evaluation (20%)	964	497	466
test (20%)	965	498	467
Mexican Spanish	4956	2488	2468
train (60%)	2974	1493	1481
evaluation (20%)	991	497	493
test (20%)	991	498	494

La identificación automática de la sátira puede ayudar a identificar textos en los que el significado pretendido difiere del significado literal, mejorando tareas como el análisis de sentimientos, la detección de noticias falsas o interfaces de usuario en lenguaje natural. Típicamente, la identificación de la sátira se realiza entrenando a un clasificador supervisado para encontrar pistas lingüísticas que puedan determinar si un texto es satírico o no. En consecuencia, en este trabajo se evaluó varias arquitecturas de aprendizaje profundo con word-embeddings en español que previamente fueron entrenadas y comparadas con los resultados de línea de base en term-counting features. Esta evaluación se realiza con dos conjuntos de datos que contienen tweets satíricos y no satíricos escritos en dos variantes de español: español europeo y español mexicano.

Para el análisis del corpus se ha incorporado estadísticas descriptivas que permiten ver de forma visual la estructuración y la composición contextual del conjunto de tweets. A continuación, los datos más representativos que forman parte de este corpus.

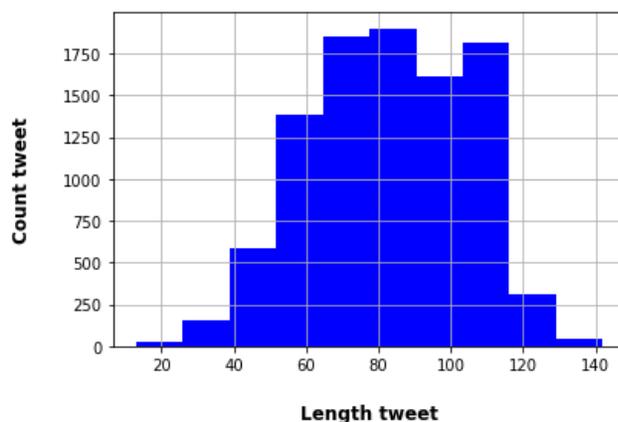


Figura 59 Histograma de cantidad de textos del corpus SATIRA.

En la **Figura 59**, se muestra que los textos de los tweets oscilan entre 2 y 145 caracteres y, en lo general, tiene entre 45 y 110 caracteres. Y el pico máximo de la composición de caracteres en un texto

está en 75 a 90. En un inicio, Twitter permitió una cantidad de caracteres con un límite de 140 caracteres y luego se amplió este límite en los tuits a 280

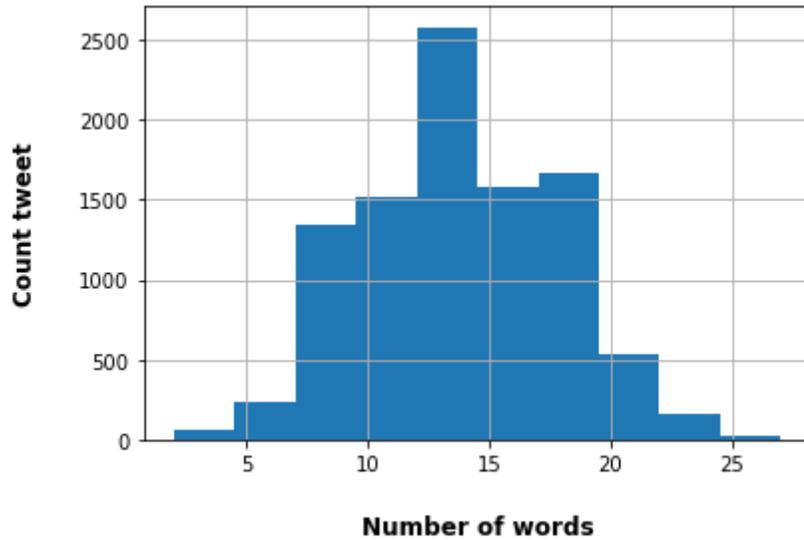


Figura 60 Histograma de números de palabras en los tweets SATIRA.

En la **Figura 60** se muestra que el número de palabras en los textos de los tweets varía de 2 a 25 y la mayoría de tuits tiene entre 11 y 14 palabras. El límite de caracteres mantiene la importancia de saber usar cada letra para construir un texto interesante, llamativo, bien redactado y que despierte el interés de las personas, independiente de que el objetivo que sea.

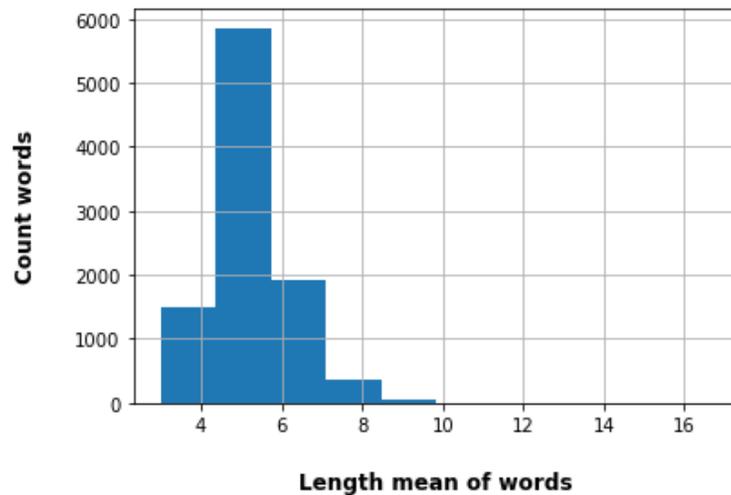


Figura 61 Histograma de longitud promedio de palabras en los tweets SATIRA.

En la **Figura 60** Histograma de números de palabras en los tweets SATIRA. la longitud promedio de las palabras oscila entre 3 y 9, siendo 5 la longitud más común. La materialización del grafico

indica que las palabras son tan cortas debido a la gran cantidad de stop-words de los tuits. Los stop-words son las palabras que se usan con mayor frecuencia en cualquier idioma como "de", "en", "el", etc. Como estas palabras son probablemente de longitud pequeña, estas palabras pueden haber provocado que el gráfico anterior quede sesgado. Analizar la cantidad y los tipos de palabras clave puede darnos algunas buenas ideas sobre los datos.

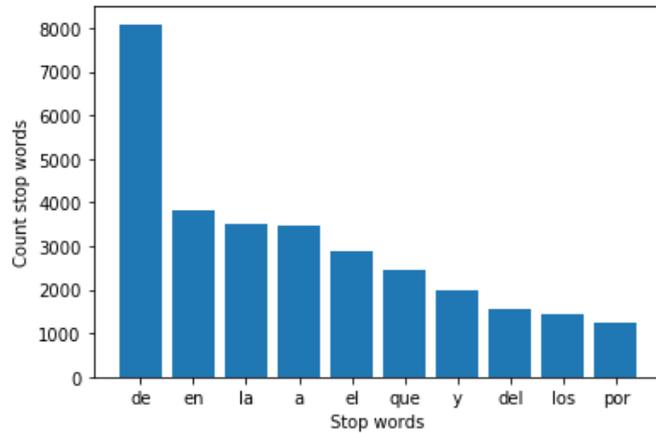


Figura 62 Histograma de stop Word más utilizados en el corpus SATIRA.

En la **Figura 62** observamos que los stop-words como las preposiciones "de", "en" o "a" y los determinantes como "el", "la" "los" dominan en los textos de los tuits; Por lo tanto, ahora podemos describir las palabras clave que aparecen con frecuencia en nuestro texto.

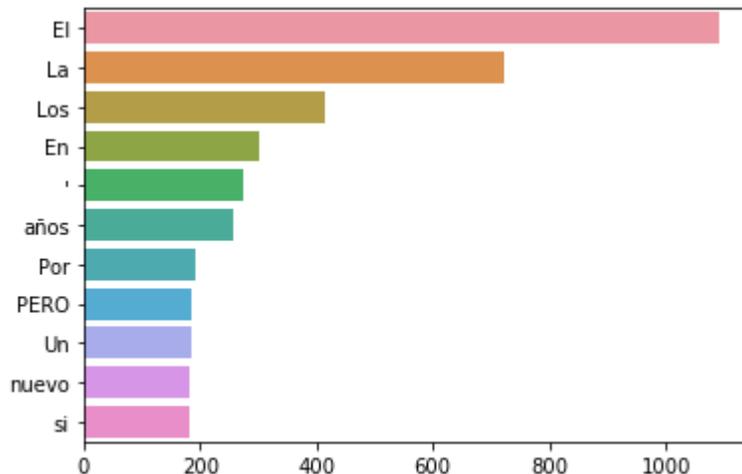


Figura 63 Histograma de Word más utilizados en el corpus SATIRA.

En la **Figura 63** las palabras "El", "La" y la "Los" que tienen más representación en los textos en el corpus. Dando mayor peso a la palabra "años" como pivote al identificativo de lo que denominamos la composición literaria en prosa o verso en que se critican agudamente las costumbres o vicios y su correlación al corpus de la SATIRA

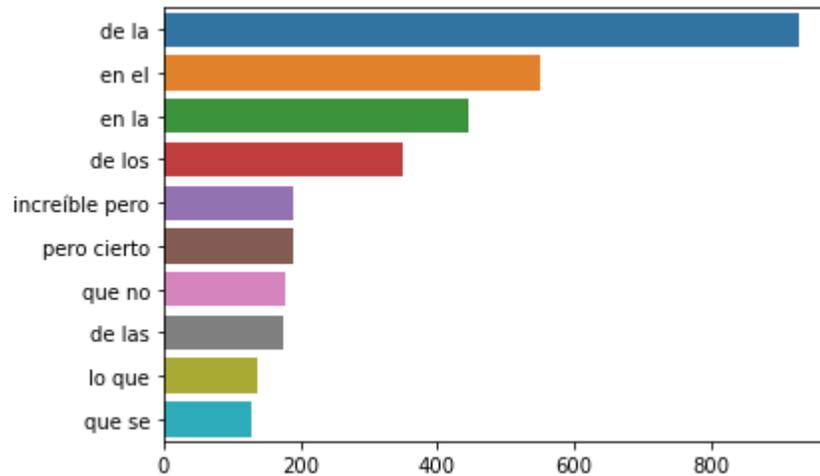


Figura 64 Histograma de bigramas más utilizados en el corpus SATIRA.

En la Figura 64 muestra que los bigramas como 'increíble pero', y 'pero cierto' lleva la correlación como la combinación de palabras que están relacionados con el corpus del sátira en los textos de los tweets.

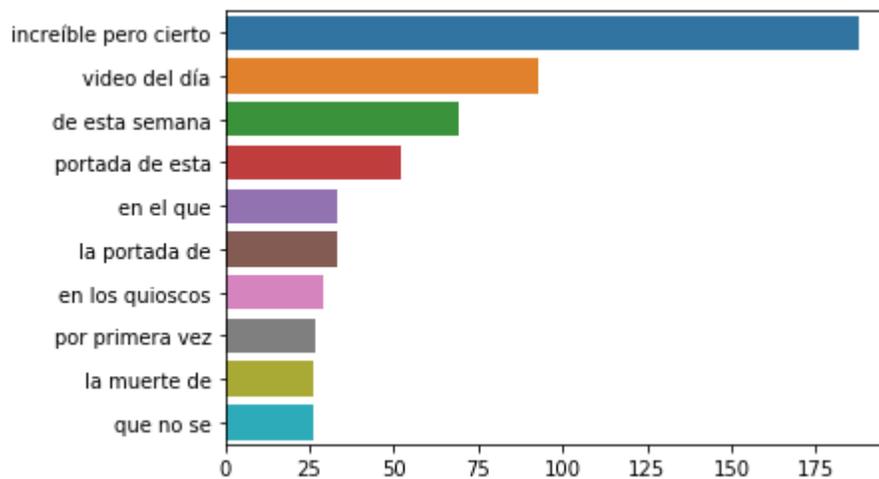


Figura 65 Histograma de trigramas más utilizados en el corpus SATIRA

En la **Figura 65** podemos ver que muchos de estos trigramas son algunas combinaciones de "increíble pero cierto" y "video del día". Son trigramas que correlacionan a las expresiones que va describiendo el objetivo y expresiones de los textos en cada tweet.

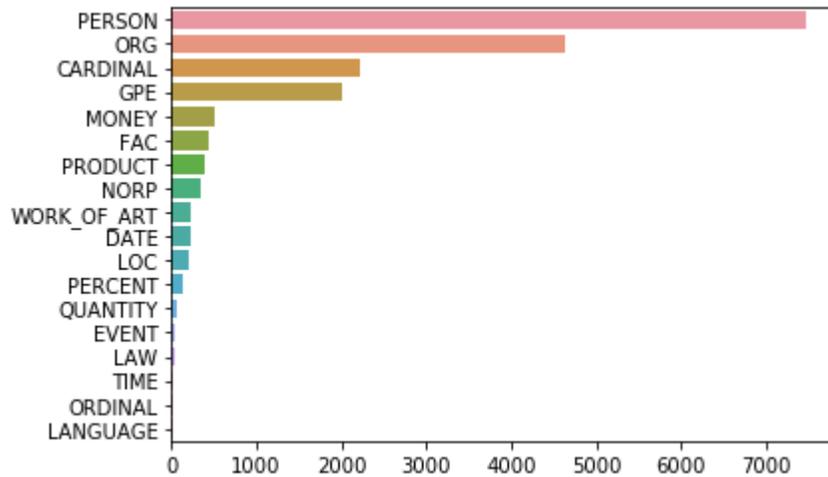


Figura 66 Histograma de entidades más utilizados en el corpus SATIRA.

En la **Figura 66** El reconocimiento de entidad con nombre es un método de extracción de información en el que las entidades que están presentes en el texto se clasifican en tipos de entidad predefinidos como "Persona", "Lugar", "Organización", etc. Al usar este análisis podemos obtener grandes conocimientos sobre los tipos de entidades presentes en el conjunto de datos de texto dado. Ahora podemos ver que el PERSON (personas, incluyendo ficticios) y el ORG (Empresas, agencias, instituciones, etc.) dominan los textos del tweet seguidos por la entidad GPE (Países, ciudades, estados).

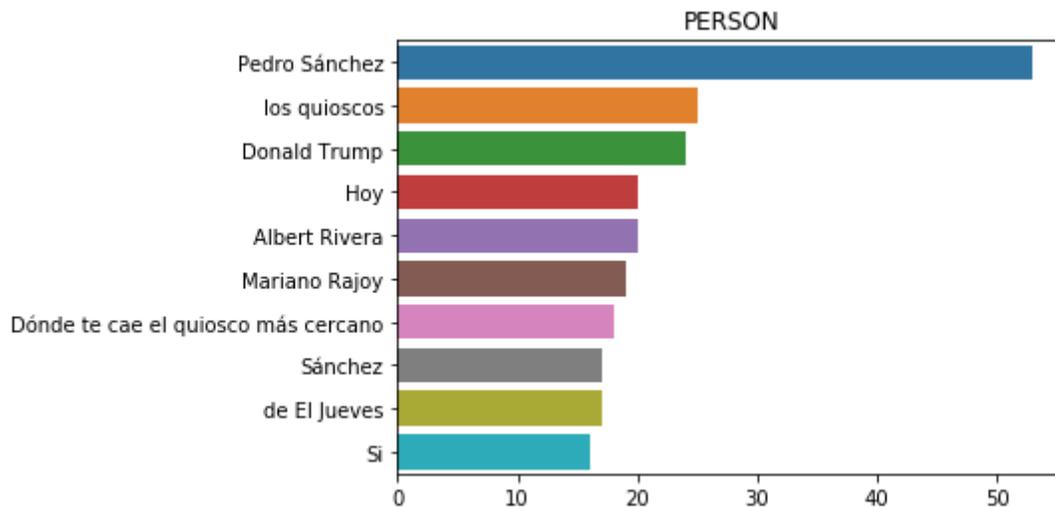


Figura 67 Histograma de entidades como PERSONO más utilizados en el corpus SATIRA

En la **Figura 67** Pedro Sanchez, los quioscos, Donald Trump y Hoy, fueron los personajes más nombrados. En el ámbito político el presidente del gobierno español, Pedro Sánchez ha tenido numerosos tuits que critican agudamente el ámbito social y decisiones que ha tomado, En el ámbito

de las publicaciones literarias como revistas que rayan en sátira; se exponen en quioscos físicos o digitales y Donald Trump hace poco presidente de EEUU donde su gestión o acciones es criticado con el uso de la sátira específicamente de un sector de la población o clases culturales que componen en el país

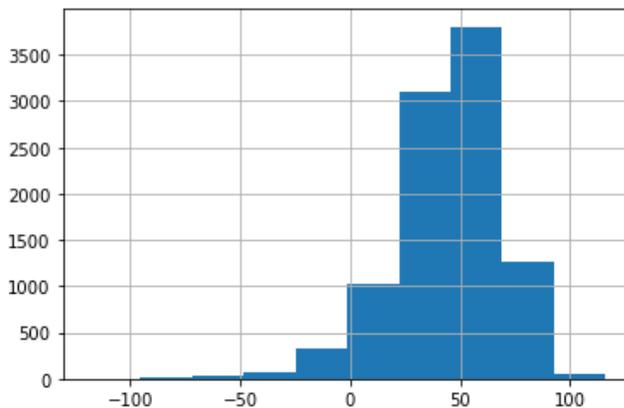


Figura 68 Histograma de índice de legibilidad ARI en el corpus SATIRA.

En Figura 68 Casi todos los puntajes de legibilidad caen por encima de 70. Esto significa que un estudiante promedio de 11 años puede leer y comprender los textos de los tweets del corpus sobre la sátira (Salaz Zárate, 2018, pp. 79-98).

Identificamos los siguientes conjuntos de datos para realizar identificaciones de sátira en español: [(Salas-Zárate et al., 2017, pp. 20-33), (Barbieri, Ronzano, & Saggion, 2015b, pp. 1-8)]. El corpus de (Barbieri et al., 2015b) fue compilado de Twitter y está compuesto por tweets de sitios de noticias en español. Sin embargo, el corpus no estaba disponible en el momento en que realizamos este trabajo. En (Salas-Zárate et al., 2017, pp. 20-33), los autores llevaron a cabo un enfoque similar desde (Barbieri et al., 2015b), pero incluyendo tweets satíricos y no satíricos compilados de Twitter de cuentas de Twitter de sitios de noticias mexicanos en español. El conjunto de datos final contiene tweets de medios satíricos, como El Mundo Today (<https://www.elmundotoday.com/>) o El Dizque (<https://www.eldizque.com/>) y nuevos sitios tradicionales como El País (<https://elpais.com/>) o el Universal <https://www.eluniversal.com.mx/>). El conjunto de datos resultante estaba compuesto por 5000 tweets satíricos en los que la mitad de ellos fueron compilados de México y la otra mitad de España; y 5000 tuits no satíricos con las mismas proporciones. Podríamos recuperar este corpus. Sin embargo, de acuerdo con las pautas de Twitter, (<https://developer.twitter.com/en/de>) este corpus contiene solo los identificadores de los tweets para respetar los derechos de los usuarios del contenido.

Usamos la herramienta UMUCorpusClassifier (Garcia-Diaz et al., 2020, pp. 1-5) para recuperar el conjunto de datos original. Aunque algunos de los tweets no estaban disponibles, pudimos recuperar

4821 tweets para el conjunto de datos español europeo y 4956 tweets para el conjunto de datos español mexicano, que representan el 96,42% y el 99,12% del conjunto de datos original, respectivamente.

Dividimos cada conjunto de datos en conjuntos de entrenamiento, evaluación y prueba en una proporción de 60% -20% -20%. Usamos el conjunto de entrenamiento y los conjuntos de evaluación para entrenar los modelos y evaluar los hiperparámetros, y el conjunto de prueba para evaluar los modelos finales. Los detalles se explican en la Sección 3.5. Las estadísticas del corpus recuperado se describen en la Tabla 9 Estadísticas de corpus de la Sátira.. En este sentido, ambos conjuntos de datos están casi equilibrados con un número ligeramente superior de tweets satíricos que es más significativo en el conjunto de datos español europeo (51,607% vs 48,392%) que en el conjunto de datos español mexicano (50,202% vs 49,798%). Vale la pena señalar que los autores de este estudio realizaron un enfoque de supervisión a distancia basado en la hipótesis de que todos los tweets compilados a partir de cuentas satíricas de Twitter pueden considerarse satíricos y aquellos tweets compilados a partir de cuentas no satíricas pueden considerarse seguros. Sin embargo, después de una revisión manual del corpus, descubrimos que algunos tweets etiquetados como satíricos contienen anuncios propios o incluso noticias no satíricas. Decidimos dejar los tweets tal y como fueron clasificados por los autores para poder realizar una comparación justa.

Cada conjunto de datos contiene tweets de cuatro cuentas: dos satíricas y dos no satíricas. Sin embargo, como podemos observar en la Figura 69, estas cuentas no están equilibradas logrando un promedio de 1205.25 tweets por cuenta con una desviación estándar de 216.483 para el conjunto de datos español europeo (ver Figura 69, izquierda) y un promedio de 1239 pero con una desviación estándar de 763.743 para el conjunto de datos de México en español (Figura 69, derecho). La alta desviación estándar del conjunto de datos en español mexicano puede causar patrones estilísticos sobre ajustados y sobre ponderados de las cuentas predominantes (@ eldeforma y @ eluniversalmx) sobre el conjunto de datos completo.

Tabla 9 Estadísticas de corpus de la Sátira.

Feature-Set	Tweets	Satirical	Non-Satirical
European Spanish	4821	2488	2333
train (60%)	2892	1493	1400
evaluation (20%)	964	497	466
test (20%)	965	498	467
Mexican Spanish	4956	2488	2468
train (60%)	2974	1493	1481
evaluation (20%)	991	497	493
test (20%)	991	498	494

Cifras 3 y 4 contienen cuatro ejemplos de un tweet satírico y no satírico para ambos conjuntos de datos. En inglés, la traducción es: 41 hospitales tienen unidades de custodia hospitalaria para presos (ver Figura 70, izquierda), Josep Guardiola insiste en regar él mismo el césped de todos los estudios europeos (ver Figura 70, derecho), # Encuesta veracruz Si hoy fueran las elecciones, habría un empate técnico entre los #Yunes (ver Figura 71, izquierda), y ¡Increíble pero cierto! Colaboradora de ElDeforma capacitada # LadyMatemáticas (ver Figura 71, derecho).

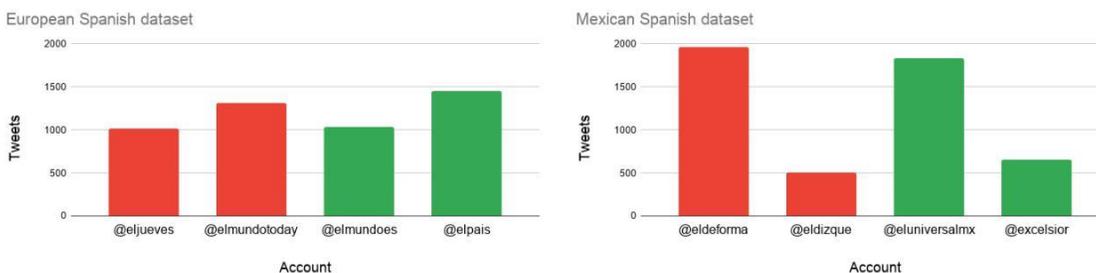


Figura 69 Distribución de corpus por cuentas para el español europeo (izquierda) y el español mexicano (derecha) conjuntos de datos compilados en (Salas-Zárate et al., 2017). En rojo, relatos satíricos. En verde, relatos no satíricos.
 Figura 3. Ejemplos de español europeo no satírico (izquierda) y satírico español europeo (derecho) tweets.



Figura 70 Ejemplos de español europeo no satírico (izquierda) y satírico español europeo (derecho) tweets.



Figura 71 Ejemplos de español mexicano no satírico (izquierda) y satírico español mexicano derecho) tweets.

Una vez compilado el corpus, realizamos una fase de normalización, que consistió en: (1) eliminar líneas en blanco y colapsar espacios y pestañas. También eliminamos el alargamiento intencional de algunas palabras que se utilizó como recurso retórico enfatizado; (2) resolver algunas formas de contracciones informales que son populares en los medios de mensajes de texto, como mñn dentro mañana En inglés: tomorrow); (3) eliminar hipervínculos y menciones; (4) codificar emoticonos como formas textuales (por ejemplo,: smile :) para facilitar su identificación; (5) corregir errores ortográficos en los textos utilizando la biblioteca ASPELL (<http://aspell.net/>); y, finalmente, (6) transformar los textos a su forma minúscula. Tenga en cuenta que algunas de estas técnicas también se realizaron en el experimento original (Salas-Zárate et al., 2017, pp. 20-33).

Para corregir automáticamente los errores ortográficos, dividimos cada tweet en palabras y verificamos si cada palabra no comenzaba con una letra mayúscula para evitar corregir los nombres propios. Luego, usamos la biblioteca ASPELL para verificar si la palabra estaba bien escrita. Si la palabra tenía errores ortográficos, buscamos la lista de sugerencias y reemplazamos la palabra mal

escrita con su primera sugerencia si la distancia entre la palabra sugerida y la palabra original era menor o igual al 90% (este umbral se estableció mediante un método de prueba y error). Seguimos este enfoque para tener la confianza de que sólo se corrigieron las palabras mal escritas obvias.

4 Capítulo 4. Evaluación

4.1 Introducción

En este capítulo se realiza el análisis de los resultados que se obtienen producto de la evaluación de los tres corpus ZIKA, COVID y SÁTIRA. En cada corpus se han realizado evaluaciones con modelos tradicionales de machine learning y modelos de deep learning. Las evaluaciones se han realizado sobre un conjunto de palabras o combinación sobre proporción de caracteres que permiten determinar la efectividad de los modelos aplicados.

Las métricas de evaluación que se aplican son la medición de precisión, la sensibilidad o exhaustividad o (“Recall”), la medida F y la exactitud (accuracy). Además, se también se muestra la matriz de confusión, que se utiliza como caja de herramientas para saber el mejor resultado en el rendimiento de cada modelo sobre cada uno de los corpus.

El análisis de sentimiento, lo aplicaremos sobre el primer corpus del ZIKA que contiene información de tuits recolectados en la red social twitter en un periodo limitado. Este corpus contiene textos que permitirán analizar las palabras u oraciones que relacionan estas enfermedades infectocontagiosas como el zika. El segundo análisis de sentimiento se realiza sobre el corpus del COVID que permitirán profundizar sobre esta última enfermedad declarada como pandemia mundial y por último aplicaremos la caja de herramienta de las métricas de evaluación sobre el corpus de la SÁTIRA que recoge distintos tuits sobre este género literario.

En el apartado 4.2 se realiza una explicación de lo que son las métricas de evaluación; para luego, pasar a los apartados 4.3 donde realizaremos la evaluación del corpus ZIKA, en el apartado 4.4 se ejecuta la evaluación del corpus COVID y finalizaremos con el apartado 0 con la evaluación del corpus SÁTIRA . Cada apartado está compuesto por tablas resultantes de los conjuntos de palabras o proporción de caracteres, modelos y métricas de evaluación, comparación de los dos mejores modelos y una discusión de los resultados.

4.2 Métricas de evaluación

Las métricas utilizadas en esta tesis son la precisión, la sensibilidad o exhaustividad o recall, la medida F(F1) y la exactitud (accuracy). Se procede a la explicación de cada uno de ellos y ver su utilidad práctica que utilizaremos en cada corpus.

En la precisión podemos medir la calidad del modelo en las operaciones de clasificación. Se refiere a lo cerca que está el resultado de una medición del valor verdadero. En otras palabras, la precisión es la proporción de casos positivos entre lo que redice el modelo y lo real (ecuación 7).

$$\text{Precision} = \text{Verdaderos positivos} / (\text{Verdaderos positivos} + \text{Falsos positivos}) \quad (7)$$

En la métrica de sensibilidad o exhaustividad o recall, nos dará informan sobre la cantidad de tuits que el modelo es capaz de identificar. También se conoce como tasa de verdaderos positivos (TP true positive o TN true negative rate) que se refiere a los casos positivos que fueron correctamente identificados por el modelo. Y así también los falsos negativos que no fueron correctamente identificados (FP falso positivo y FN falso negativo) (ecuación 8).

$$\text{Exhaustividad} = \text{Verdaderos positivos} / (\text{Verdaderos positivos} + \text{Falsos negativos}) \quad (8)$$

La métrica F1, se utiliza para combinar las medidas de precisión y recall en un solo valor. Esto es utiliza para hacer más sencillo el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre las soluciones (ecuación 9).

$$\text{Medida F(F1)} = 2 * [(\text{precision} * \text{exhaustividad}) / (\text{precisión} + \text{exhaustividad})] \quad (9)$$

En la exactitud (accuracy) mide el porcentaje de casos que el modelo ha acertado. Sobre esta métrica es que nos basaremos para determinar si un modelo es identificado como mejor. Para que esta métrica funcione correctamente hemos balanceado los corpus (ecuación 10).

$$\text{Accuracy (Acc)} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (10)$$

La matriz de confusión representa las n clases que contenga como objetivo el corpus. En las filas se representan las clases reales y en las columnas, el resultado de los resultados obtenidos por la evaluación del modelo.

Por último, para ayudar a la visualización en las tablas se ha creado una nomenclatura para representar los nombres de modelo en machine y deep learning. Así optimizamos el uso de espacio y le damos más relevancia en los resultados de cada métrica cuyo valor numérico es expresado en porcentaje con varios números decimales. Para lo cual, los modelos utilizados en machine learning son: SVM (Support vector machines), RF (Random Forest), MNB (Multinomial Naive Bayes), LR (Logistic Regression), K (k-nearest) y J48 (Decision Tree). En los modelos de deep learning son: CNN (convolutional neural network), DENSE (fully connected layer), GRU (Gated Recurrent Unit), LSTM (Long short-term memory), BiGRU (Bidirectional GRU) y BiLSTM (Bidirectional LSTM)

4.3 Evaluación del corpus ZIKA

En la sección 0 se realiza la evaluación para los modelos tradicionales de machine learning, en la sección 4.3.2 la evaluación para modelos deep learning y en la sección 4.3.3 la discusión de los mejores dos modelos de cada sección sobre el corpus del ZIKA.

4.3.1 Evaluación para modelos tradicionales de machine learning del corpus ZIKA.

En la Tabla 10 tenemos los resultados para la evaluación sobre cada modelo con sus métricas en el corpus ZIKA. Para realizar la evaluación se crearon conjuntos de combinaciones entre palabras que le denominamos n-gramas y proporción de caracteres que los identificamos como char-gramas. Así la operación empezó tomando de cada tuit una sola palabra y combinaciones de palabras desde un rango entre el inicio y fin en la cantidad de palabras: Por lo cual, nuestro análisis incluyó unigramas (solo una palabra), bigramas (dos palabras) y trigramas (tres palabras). La nomenclatura de las combinaciones de palabras es (1-1) que se refiere solo a unigramas, palabras (1-2) que indica que tomamos unigramas + bigramas, palabras (1-3) que incluye unigramas + bigramas + trigramas, palabras (2-2) solo bigramas, palabras (2-3) que toma bigramas + trigramas y palabras (3-3) solo trigramas.

También realizamos combinaciones con proporción de caracteres (char-gramas) empezando con selecciones en la misma modalidad de rango de inicio y fin; pero esta vez, con la proporción de caracteres sobre el tuit y se empezó con la combinación caracteres (4-4) que indica los cuatro primeros caracteres que aparecen en el texto dentro del tuit, seguimos con

los siguientes subconjuntos de caracteres (4-5, 4-6, 4-7, 4-8, 4-9 y 4-10). Estas combinaciones tienen en cuenta los conjuntos de 4 hasta 10 caracteres según sus combinaciones.

La columna modelo representa los algoritmos de machine learning y las columnas Precision, Recall, F1 y Acc (accuracy) las métricas; de las cuales evaluamos la eficiencia en la clasificación que realiza cada algoritmo (modelo).

Para el caso de los valores resultantes de cada métrica, éstos son representados en porcentaje y ampliado con una parte de decimal para proporcionar una mejor visibilidad en las predicciones que realizan los modelos. Las métricas de precisión, recall y F1 manejan diez dígitos decimales y el Acc una cantidad de ocho decimales.

Para identificar el mejor Acc (accuracy) se tomará como criterio de selección el valor resultante de esta métrica como el mayor valor que se acerque a la unidad (100%) y lo remarcamos en este documento con el resaltador “negritas”. En la tabla 28 se muestran los resultados de las evaluaciones de todas las combinaciones anteriormente descritas en los algoritmos de machine learning como el modelo SVM cuya función principal es crear una línea o hiperplano que separa los datos de las clases negative, neutral y positive, el modelo RF crea árboles de decisión sobre muestras de datos y luego obtiene la predicción de cada uno de ellos y finalmente selecciona la mejor solución mediante votación, el modelo MNB utiliza el enfoque de la bag of words (bolsa de palabras); donde las palabras individuales del documento constituyen sus características y se ignora el orden de las palabras, el modelo LR la regresión lineal realiza la tarea de predecir un valor de variable dependiente basado en una variable independiente, el modelo K (K-Nearest) es el algoritmo que detecta el número de vecinos más cercanos a una nueva variable y el modelo J48 es un algoritmo utilizado para generar un árbol de decisiones. Como se puede observar para el corpus ZIKA los dos mejores modelos para todas las evaluaciones han sido los modelos SVM (Support vector machines) y MNB (Multinomial Naive Bayes).

El primer mejor modelo con la combinación del conjunto de palabras (1-3) con SVM con un Acc (accuracy) de **61,20584653** y la segundo mejor modelo con la combinación la encontramos en el conjunto de palabras (1-2) con MNB con un Acc (accuracy) de **60,71863581**.

Esto indica que en el corpus ZIKA en el primer mejor modelo palabras (1-3) el uso de unigramas + bigramas + trigramas hace que la dispersión del conjunto de valores obtenidos sea menor. Por consiguiente, la proporción de predicciones correctas (tanto positivas como negativas) sobre el total de predicciones, dan como resultado un mejor accuracy. Por ejemplo, al tomar un tuit y seleccionar un trigramas permite al algoritmo SVM tener como resultado el mejor accuracy. Para el segundo mejor modelo de evaluación, tomar unigramas y bigramas (combinación de palabras 1-2) dan los siguientes mejores valores más altos en la exactitud.

Por otro lado, el menor accuracy que se reporta de entre todas las evaluaciones realizadas, es el modelo K (k-nearest) con el conjunto de combinación de palabras (3-3) que dan un valor de **44,33617540**. En esta misma evaluación el recall (exhaustividad) informa que la predicción de nuestro modelo en lo que respecta a falsos negativos (FN) y verdaderos positivos (TP); ha sido capaz de identificar un **44,3361753959** porcentaje sobre los tuits que han sido identificados como positivos en el corpus.

Tabla 10 Resultados de las evaluaciones en modelos de machine learning para el corpus ZIKA

Conjunto	Modelo	Precision	Recall	F1	Acc
Palabra 1-1	SVM	62,0200989349	60,9216402761	61,2791721437	60,92164028
	RF	61,3649594743	60,6171335769	60,8745304631	60,61713358
	MNB	59,6441907517	59,9878197320	59,7857900037	59,98781973
	LR	59,9514142356	59,7645148193	59,8507920351	59,76451482
	K	53,2643341015	50,8729192042	49,7911868159	50,87291920
	J48	50,2680721048	49,8984977670	50,0519447759	49,89849777
	Palabras 1-2	SVM	61,7661960521	60,9419407227	61,2365251599
RF		60,7629724169	59,6630125863	59,9924577869	59,66301259
MNB		60,3191610454	60,7186358100	60,4665195748	60,71863581
LR		60,6027328180	60,1502233049	60,3362034512	60,15022330
K		53,1180782015	51,0353227771	50,3436016315	51,03532278
J48		49,1067831809	49,0864799026	49,0920312254	49,08647990
Palabras		SVM	61,9365900287	61,2058465286	61,4783112417

Conjunto	Modelo	Precision	Recall	F1	Acc
1-3	RF	60,3340224799	59,7036134795	59,9319026918	59,70361348
	MNB	60,1393739765	60,5968331303	60,2879999376	60,59683313
	LR	60,6282506823	59,9675192854	60,2193278634	59,96751929
	K	52,5303310486	50,5481120585	49,6742747079	50,54811206
	J48	49,1935235442	49,1676816890	49,1760082008	49,16768169
Palabras	SVM	59,4309549819	58,8103938287	59,0491089015	58,81039383
2-2	RF	59,3500797409	57,3284612261	57,7527701095	57,32846123
	MNB	58,7415952476	59,2773041007	58,9042398635	59,27730410
	LR	58,9357826234	58,3840844499	58,6006875355	58,38408445
	K	48,2626082791	45,4730004060	42,8193422152	45,47300041
	J48	51,4691870191	51,2992285830	51,3782650509	51,29922858
Palabras	SVM	59,8467468641	59,2570036541	59,4815168687	59,25700365
2-3	RF	59,7124300092	57,2269589931	57,7148587475	57,22695899
	MNB	58,5660074021	59,1555014210	58,6976254148	59,15550142
	LR	59,1435034496	58,4855866829	58,7251942532	58,48558668
	K	47,8807365564	44,8842874543	41,8286730082	44,88428745
	J48	51,4629828903	50,9541209907	51,1582045507	50,95412099
Palabras	SVM	57,5106609191	56,4352415753	56,8020026053	56,43524158
3-3	RF	59,6697341983	55,0345107592	55,4388633135	55,03451076
	MNB	56,7404055141	57,1457572067	56,8874553420	57,14575721
	LR	58,1562746539	56,6179455948	57,0506723540	56,61794559
	K	47,3257651817	44,3361753959	40,5746142907	44,33617540
	J48	52,7020232962	50,6090133983	51,0175685937	50,60901340
Caracteres	SVM	61,8528547638	60,9622411693	61,2850929201	60,96224117
4-4	RF	60,4918006841	59,6021112464	59,8903485653	59,60211125
	MNB	59,8447165157	60,0893219651	59,9169185787	60,08932197
	LR	59,7688425481	59,5615103532	59,6560655596	59,56151035
	K	52,9497976660	50,6090133983	49,4751542328	50,60901340
	J48	49,6853927503	49,5330897280	49,6003783958	49,53308973
	SVM	62,1179696236	61,0434429557	61,4146294887	61,04344296

Conjunto	Modelo	Precision	Recall	F1	Acc
Caracteres	RF	61,2004328631	60,5156313439	60,7541145048	60,51563134
4-5	MNB	59,7452889421	60,0893219651	59,8233271720	60,08932197
	LR	60,2593591127	59,9878197320	60,1080383377	59,98781973
	K	52,2276139343	50,0203004466	48,7813408015	50,02030045
	J48	50,9156196423	50,7714169712	50,8381617397	50,77141697
Caracteres	SVM	62,0258673576	61,0637434024	61,4040133474	61,06374340
4-6	RF	60,9679283329	60,2111246447	60,4740524287	60,21112464
	MNB	59,5854015265	60,0487210719	59,6786479953	60,04872107
	LR	60,1890213083	59,8863174990	60,0188687029	59,88631750
	K	52,9397014589	51,0353227771	49,8767039254	51,03532278
	J48	50,3181994982	50,1015022330	50,2011684245	50,10150223
Caracteres	SVM	61,9514288274	61,0028420625	61,3405190077	61,00284206
4-7	RF	61,2501925199	60,4953308973	60,7535725141	60,49533090
	MNB	59,4417209250	60,0284206253	59,5297290713	60,02842063
	LR	53,0206215005	50,9338205441	49,7559139156	50,93382054
	K	53,0206215005	50,9338205441	49,7559139156	50,93382054
	J48	50,1130438573	50,0812017864	50,0955265106	50,08120179
Caracteres	SVM	61,9216649911	60,9622411693	61,3014694952	60,96224117
4-8	RF	61,0203358637	59,9878197320	60,3129010656	59,98781973
	MNB	59,3730964328	59,9878197320	59,4403271206	59,98781973
	LR	59,9930233148	59,7036134795	59,8318831958	59,70361348
	K	51,9533325683	49,5939910678	48,3639273335	49,59399107
	J48	50,6419279757	50,2842062525	50,4234112478	50,28420625
Caracteres	SVM	61,9659939849	60,9825416159	61,3274143796	60,98254162
4-9	RF	61,2501925199	60,4953308973	60,7535725141	60,49533090
	MNB	59,1457513844	59,7848152659	59,2193685754	59,78481527
	LR	59,8267286202	59,5006090134	59,6437335800	59,50060901
	K	52,5005248401	50,2639058059	49,0335599777	50,26390581
	J48	49,4698501753	49,2691839220	49,3614391764	49,26918392
	SVM	61,8116723554	60,8404384896	61,1813092492	60,84043849

Conjunto	Modelo	Precision	Recall	F1	Acc
Caracteres 4-10	RF	59,7264688417	58,7494924888	59,0705526196	58,74949249
	MNB	59,1014190971	59,7442143727	59,1862793592	59,74421437
	LR	59,8447585593	59,5006090134	59,6505529294	59,50060901
	K	52,0409630593	49,7969955339	48,5751396578	49,79699553
	J48	50,3233417926	50,2030044661	50,2586333348	50,20300447

En la Tabla 11 presentamos la comparación de los dos mejores resultados que han arrojado las evaluaciones; en primer lugar, el modelo SVM con el conjunto de palabras (1-3) que tiene un Acc **61,20584653** y en segundo lugar el modelo MNB con el conjunto de palabras (1-2) que tiene un Acc **60,71863581**. Aquí, se detallan los resultados para cada clase (negative, neutral y positive) y el resultado en cada métrica. También se muestra macro-avg (macro promedio) como weighted avg (promedio ponderado). El macro promedio calcula la métrica independientemente para cada clase (negative, neutral y positive) y luego tomará el promedio. El promedio ponderado como medida de tendencia central es apropiada para los conjuntos de datos donde la clase tiene una importancia relativa. Así las clases negative y positive sobre los tuis del corpus ZIKA empiezan a tener más importancia en los algoritmos que utilizan los modelos. Por ejemplo, en el primer mejor modelo evaluado sobre la clase negative y la métrica recall nos indica que la proporción entre los falsos negativos y verdaderos positivos alcanzo un (65,21219366); es decir, el modelo es capaz de identificar un 65,21% (65,21219366) sobre los tuis que mencionan al zika y que deben tener una predicción de clase negative; Por consiguiente, si tomamos un tuit y del cual esperamos que nuestro algoritmo lo prediga como negative, tendremos un 65,21% de lo relacionado como verdaderos positivos (TP) vs los falsos negativos(FN).

Ahora en el segundo mejor modelo evaluado sobre la clase negative y la métrica recall nos indica que la proporción entre los falsos negativos y verdaderos positivos alcanzo un (67,72265392); es decir, el modelo es capaz de identificar un 67,72% (67,72265392) de la proporción entre los falsos negativos y verdaderos positivos; pero con una precisión de 65,91041303 vs a la precisión del primer modelo 69,13814956. Para terminar la comparación de los mejores modelos veremos que la medida f1-score asume que nos importa de igual

forma la precisión como la exhaustividad (recall); así para el primer mejor modelo tenemos un f1-score (67,1178099) en comparación al f1-score (66,80424528) del segundo mejor modelo.

Tabla 11 Comparación de los dos mejores modelos de machine learning en el corpus ZIKA

	Primer mejor modelo Acc 61,20584653				Segundo Mejor modelo Acc 60,71863581			
Modelo	SVM				MNB			
Conjunto	Palabras 1-3				Palabras 1-2			
	precision	recall	f1-score	Support	precision	recall	f1-score	support
negative	69,1381496	65,2121937	67,1178099	69,1381496	65,910413	67,7226539	66,8042453	1673
neutral	49,1563555	55,0724638	51,9465082	49,1563555	50,4485852	46,0617517	48,1554677	1587
positive	66,8789809	63,0252101	64,894932	66,8789809	64,1069397	67,6470588	65,8294393	1666
macro avg	61,7244953	61,1032892	61,31975	61,7244953	60,1553127	60,4771548	60,2630508	4926
weighted avg	61,93659	61,2058465	61,4783112	61,93659	60,319161	60,7186358	60,4665196	4926

En la **Figura 72** presentamos la matriz de confusión comparando los dos mejores resultados que han arrojado las evaluaciones; en primer lugar, el modelo SVM con el conjunto de palabras (1-3) y el **Acc 61,20584653** y en segundo lugar el modelo MNB con el conjunto de palabras (1-2) y **Acc 60,71863581**. La matriz de confusión nos permite ver en las filas los tuis del corpus zika identificados como negative, neutral y positive. En las columnas obtendremos la predicción resultante de evaluar el modelo. Así en el primer mejor modelo vemos que SVM alcanza a predecir el 65% en verdaderos negativos (TN), 63% sobre verdaderos positivos (TP), 8% sobre falsos negativos (FN) y un 9% en falsos positivos (FP). En el segundo mejor modelo vemos que MNB alcanza a predecir el 68% en verdaderos negativos (TN), 68% sobre verdaderos positivos (TP), 10% sobre falsos negativos (FN) y un 11% en falsos positivos (FP).

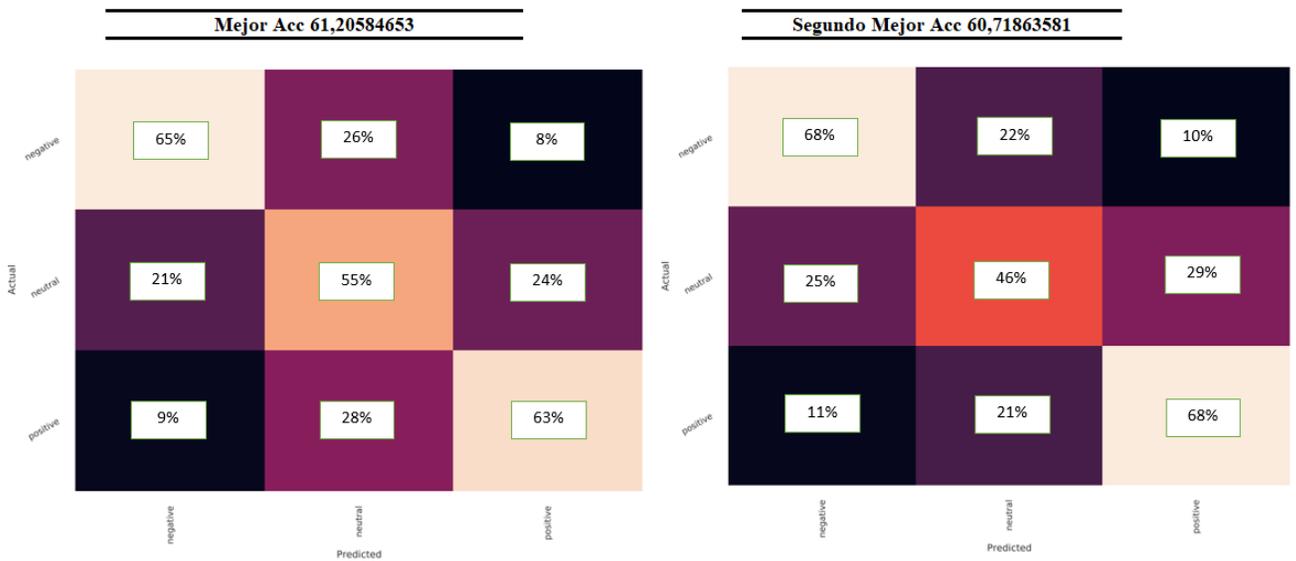


Figura 72 Matriz de confusión de los dos mejores modelos de machine learning en el corpus ZIKA

En la Tabla 12, realizamos la comparación entre los dos mejores modelos para los n-gramas y los dos mejores modelos para los char-gramas. Y resulta que de esta lista de cuatro elementos el Acc (accuracy) más alto es de 61,20584653 que corresponde a los n-gramas para el conjunto Palabras (1-3) que implica tomar unigramas + bigramas + trigramas sobre el modelo SVM. Este mismo modelo SVM tiene el siguiente mejor resultado en los char-gramas para el conjunto de caracteres (4-6) con un Acc (accuracy) de 61,06374340.

Dentro de los n-gramas el segundo mejor modelo es el MNB que tiene un Acc (accuracy) de 60,71863581 para el conjunto de Palabras (1-2) o unigramas + bigramas. Para los char-gramas el segundo mejor modelo tenemos a RF (Random Forest) con un Acc(accuracy) de 60,51563134.

Es así como el modelo SVM tiene como efecto una mejor predicción sobre los dos conjuntos de n-gramas y char-gramas; luego en orden de preferencia o efectividad en la clasificación está el modelo MNB que funciona bien para los n-gramas y char-gramas; para terminar de indicar que el tercer modelo más efectivo-sería RF que daría un nivel de exactitud aceptable pero enfocado en los char-gramas.

Tabla 12 Comparación por modelo de los dos mejores n-gramas y los dos mejores char-gramas en el corpus ZIKA

Modelo	Conjunto	Precision	Recall	F1	Acc
SVM	Palabras 1-3	61,93659	61,205847	61,478311	61,20584653
MNB	Palabras 1-2	60,319161	60,718636	60,46652	60,71863581
SVM	Caracteres 4-6	62,025867	61,063743	61,404013	61,06374340
RF	Caracteres 4-5	61,200433	60,515631	60,754115	60,51563134

En la **Figura 73**, permite ver la comparación de las matrices de confusión en los resultados que arrojan el mejor modelo para los n-gramas y el mejor modelo para los char-gramas. En los n-gramas tenemos un 65% en la predicción de los verdaderos negativos (TF) como los mayores aciertos y los char-gramas tenemos un 64% en los mejores aciertos que también corresponde a los verdaderos negativos (TF). Para el contexto del corpus ZIKA los usuarios escriben cosas que normalmente causan un análisis de sentimiento en clase “negative”.

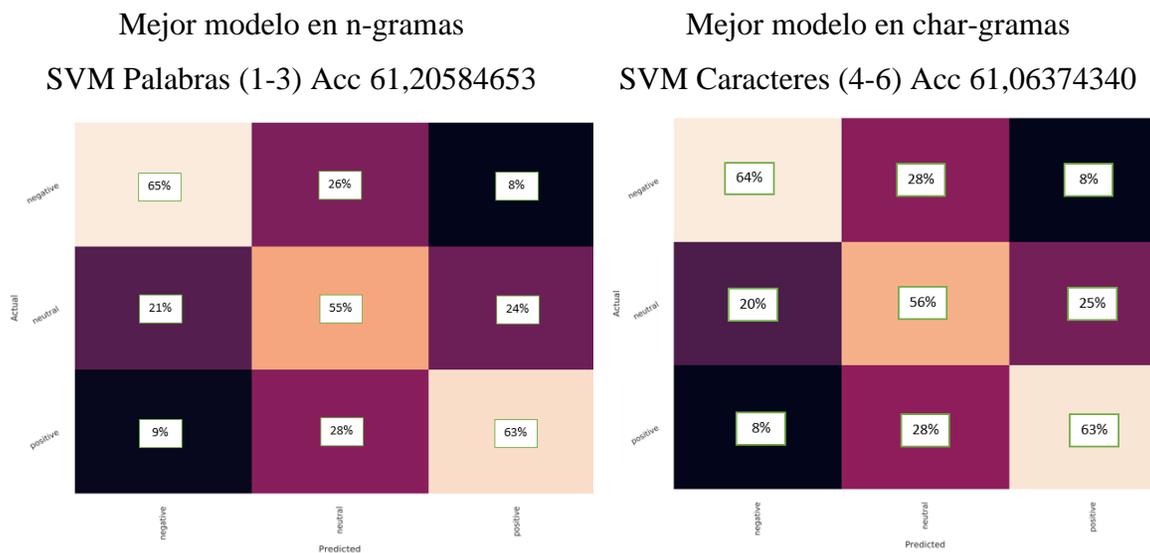


Figura 73 Matriz de confusión del mejor modelo para n-gramas y el mejor modelo para char-gramas para el corpus ZIKA

4.3.2 Evaluación para modelos de deep learning del corpus ZIKA.

Como se indicó anteriormente ampliamos la evaluación del corpus con las técnicas de aprendizaje automático o deep learning. Estos modelos también nos ayudaran en predicción de tuits sobre estas enfermedades infectocontagiosas que están correlacionadas con el zika. Los modelos que utilizamos son CNN (redes neuronales convulsionadas), LSTM (redes neuronales recurrentes, Long Short Term Memory), GRU (redes neuronales recurrentes, Gated Recurrent Units), Bi-GRU (Bidireccional GRU), BILSTM (Bidireccional LSTM) y DENSE (fully-connected layer). Para cada modelo ejecutaremos las mismas métricas de Precisión, Recall, F1 y Acc(Accuracy) que para los modelos anteriores. Aquí también escogeremos los dos mejores Acc(acurracy) para determinar la predicción que se realiza sobre los tuits.

En la Tabla 13 se muestran los resultados de las ejecuciones en los modelos de deep learning y las métricas resultantes. Por consiguiente, los dos mejores modelos son: en primer lugar y con una exactitud más alta el modelo DENSE (**60,65773447**); luego está en segundo lugar por tener la siguiente exactitud más alta es el modelo BILSTM (**59,70361348**).

El modelo DENSE es una red neuronal multicapa que interactúa como un modelo con capas en secuencia y se añaden capas hasta que la red tenga resultados óptimos en la topología de red. Se alcanzan métricas de precisión con 61,258080 para identificar las predicciones en la proporción de verdaderos positivos (TP) con falsos positivos (FP), la exhaustividad con 60,657734 que permite tener proporción de los verdaderos positivos (TP) con los falsos negativos (FN) y la medida F1 con 60,888571 que realiza una media entre la precisión y la exhaustividad.

El modelo BILSTM facilita el aprendizaje de elementos como la entonación de una pregunta que se realiza en los tuits obteniendo una precisión de 62,071704 para identificar las predicciones en la proporción de verdaderos positivos (TP) con falsos positivos (FP), la exhaustividad con 59,703613 que permite tener proporción de los verdaderos positivos (TP) con los falsos negativos (FN) y la medida F1 con 60,283021 que realiza una media entre la precisión y la exhaustividad.

Tabla 13 Resultados de las evaluaciones en modelos de deep learning para el corpus ZIKA

Modelo	Precision	Recall	F1	Acc
CNN	59,456498	58,729192	58,932637	58,72919204
LSTM	58,630257	58,749492	58,657563	58,74949249
GRU	58,161292	57,856273	57,930112	57,85627284
BI-GRU	55,247341	55,907430	54,681668	55,90742996
BILSTM	62,071704	59,703613	60,283021	59,70361348
DENSE	61,258080	60,657734	60,888571	60,65773447

En la Tabla 14, realizamos la comparación de los dos mejores modelos de deep learning ampliando en cada clase objetivo (negative, neutral y positive) y las métricas de evaluación como las medidas de macro avg (macro promedio) y weighted avg (promedio ponderado). Para el mejor modelo, DENSE con una exactitud de 60,65773447 tenemos en la clase objetiva negative tiene una precisión de 68,207024 que nos permite saber la proporción de los verdaderos positivos (TP) con los falsos positivos (FP) y la exhaustividad 66,168559 que visualiza los verdaderos positivos (TP) y falsos negativos (FN). En la misma clase objetivo negative comparamos el segundo mejor modelo BILSTM con una exactitud de 59,70361348 obteniendo una precisión de 72,451994 para identificar verdaderos positivos (TP) con los falsos positivos (FP) y la exhaustividad 58,637179 con los verdaderos positivos (TP) y falsos negativos (FN). Para una comparación completa vemos ahora los valores resultantes de F1 (media entre la precisión y recall) para misma clase negative y el primer mejor modelo DENSE da en F1 67,172330 y para el segundo mejor modelo el F1 da BILSTM 64,816650.

Tabla 14 Comparación de los dos mejores modelos de deep learning en el corpus ZIKA

	DENSE Acc 60,65773447				BILSTM Acc 59,70361348			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
negative	68,207024	66,168559	67,172330	1673	72,451994	58,637179	64,816650	1673
neutral	48,569794	53,497164	50,914543	1587	46,328502	60,428481	52,447361	1587
positive	66,366559	61,944778	64,079478	1666	66,644474	60,084034	63,194444	1666
macro avg	61,047793	60,536834	60,722117	4926	61,808324	59,716565	60,152819	4926
weighted avg	61,258080	60,657734	60,888571	4926	62,071704	59,703613	60,283021	4926

En la **Figura 74** presentamos la matriz de confusión comparando los dos mejores resultados que han arrojado las evaluaciones; en primer lugar, el modelo DENSE (**Acc 60,65773447**) y en segundo lugar el modelo BILSTM (**Acc 59,98781973**). La matriz de confusión nos permite ver en las filas los tuis del corpus zika identificados como negative, neutral y positive; en las columnas obtendremos la predicción resultante de evaluar el modelo. Así en el primer mejor modelo vemos que nuestro modelo DENSE alcanza a predecir el 66% en verdaderos negativos (TN), 62% sobre verdaderos positivos (TP), 8% sobre falsos negativos (FN) y un 10% en falsos positivos (FP). En el segundo mejor modelo vemos que nuestro modelo MNB alcanza a predecir el 59% en verdaderos negativos (TN), 60% sobre verdaderos positivos (TP), 8% sobre falsos negativos (FN) y un 7% en falsos positivos (FP).

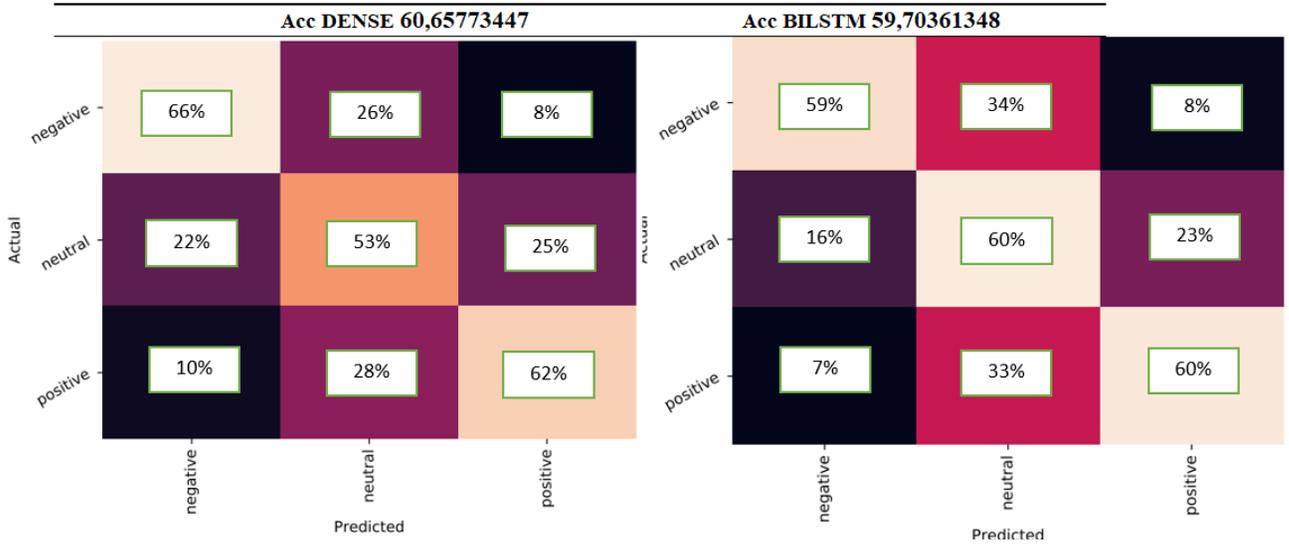


Figura 74 Matriz de confusión de los dos mejores modelos de deep learning en el corpus ZIKA

4.3.3 Discusión de los resultados del corpus ZIKA.

Los resultados de clasificación para el corpus ZIKA con técnicas tradicionales de machine learning y deep learning; indican que, la mejor exactitud (Acc accuracy) recae en las técnicas tradicionales de machine learning específicamente en el modelo SVM con un conjunto de palabras (1-3) o unigramas + bigramas + trigramas con una exactitud de 61,20584653. Los comentarios sobre este tipo de enfermedades infectocontagiosas como el zika, dengue o chikungunya tienden a generar texto en los tuits que contienen palabras o combinación de palabras que describen lo que están escribiendo los usuarios en la red social. Son palabras, que por su peso en el texto suelen ser encontradas en la mayoría de los tuits y van desde los síntomas de la enfermedad, casusas, efectos y hasta los tratamientos que las personas comentan o escriben en su red social. Este mismo modelo SVM tiene una buena predicción para la combinación aplicable en los char-gramas con la combinación (4-6) con un accuracy de 61,06374340.

La siguiente mejor exactitud (accuracy) lo tiene también otro modelo de las técnicas tradicionales de machine learning que es el algoritmo MNB (Multinomial Naive Bayes) que utiliza la relación de dependencia que tienen las variables entre ellas; Es decir, que de nuestros resultados la combinación de palabras (1-2) permite que el algoritmo aplicado de los siguientes mejores resultados sobre los tuits con una exactitud de 60,71863581 en el

corpus ZIKA. La combinación de palabras (1-2) indica que unigramas + bigramas seleccionados; dan una mayor probabilidad de predicción. En un contexto específico como las enfermedades infectocontagiosas los usuarios de la red social suelen utilizar un o dos palabras que casi siempre correlacionan a este tipo de enfermedades.

Entre los mejores resultados está el modelo DENSE que pertenece a las técnicas de Deep learning. La exactitud (accuracy) alcanzada es de 60,65773447; esto implica que el modelo de multicapas completamente conectadas dará una predicción aceptable sobre tuits del corpus ZIKA. La técnica de crear capas conectadas a partir de un texto; tiende a dar resultados positivos para poder predecir el análisis de sentimiento de lo contenga el texto de este tipo de enfermedades.

Cabe destacar que el modelo RF con las combinaciones de proporción de caracteres que van del conjunto (4-5) obtiene también buenos resultados con un accuracy de 60,51563134.

Entre los mejores modelos resultados el modelo BILSTM alcanza una exactitud (accuracy) de 59,70361348 que dista un poco de los resultados obtenidos por otros modelos. Sin embargo, es importante notar que aplicar técnicas de redes neuronales recurrentes bidireccionales que conectan dos capas ocultas de direcciones opuestas a la misma salida da mejores resultados que usar redes LSTM. Analizar el texto en ambas direcciones; de izquierda a derecha como de derecha a izquierda en cada tuit, dan un buen resultado en este contexto de enfermedades como el zika, dengue y chikungunya.

En la Tabla 15 presentamos un resumen de los mejores modelos con técnicas de machine learning y con modelos de deep learning que dieron como resultado la evaluación de tuits para el corpus ZIKA.

Tabla 15 Mejores accuracy de los modelos en machine learning y Deep learning en el corpus ZIKA

Clasificador	Conjunto	Accuracy
SVM	Palabras 1-3	61,20584653
SVM	Caracteres 4-6	61,06374340
MNB	Palabras 1-2	60,71863581
DENSE	FastText	60,65773447
RF	Caracteres 4-5	60,51563134
BILSTM	FastText	59,70361348

4.4 Evaluación del corpus COVID

En la sección 0 se realiza la evaluación para los modelos tradicionales de machine learning, en la sección 4.4.2 la evaluación para modelos deep learning y en la sección 4.4.3 la discusión de los mejores dos modelos de cada sección sobre el corpus del COVID

4.4.1 Evaluación para modelos tradicionales de machine learning del corpus COVID.

En esta sección se presentan los resultados para cada evaluación; producto de aplicar técnicas tradicionales de machine learning en el corpus COVID. La estructuración de estos resultados, modelos y métricas; ya fue descrita en las secciones 4.1 y 4.2 de este capítulo. Además de los resultados detallados en la sección 0 que permitió descubrir que modelos y combinaciones dan mejores predicciones para el corpus que se analizó.

El corpus del COVID se incorpora a esta tesis, producto de que en la ciencia de la infodemiología comienza a presentar un aumento de la información relacionada con la salud en el internet. La OMS declara a este coronavirus como pandemia que afecta con riesgo de muerte a la humanidad. Por lo cual, incluimos tuits relacionados al coronavirus, covid o covid-19 como el siguiente objetivo para aplicar las técnicas de machine learning y analizar sus resultados.

En la Tabla 16 presentamos los resultados obtenidos por los modelos de machine learning en combinación con el conjunto de palabras o porción de caracteres más la visualización de los resultados en las métricas, que nos permite evaluar el rendimiento de la predicción en el procesamiento de los textos que componen el corpus COVID. Así en la columna Acc (accuracy) presentamos el valor resaltado en negritas para los dos mejores modelos.

Así encontramos que el modelo RF en conjunto con la combinación de una porción de caracteres (4-4) llamado char-gramas; nos da el mejor resultado en la exactitud (accuracy) alcanzado una predicción del 47,8% (47,08451627). El objetivo de este análisis de sentimiento es trazar la predicción más aceptable sobre el corpus que contiene las clases negative, neutral y positive y que se correlacionan a los tuits de este contexto. En el segundo mejor modelo es SVM con la combinación de una porción de caracteres de (4-9) y que tiene una exactitud de 44,76% (44,76960035). En la presentación de los todos los resultados el modelo K(K-Nearest) en combinación con el conjunto de palabras (3-3) llamados n-gramas o en este caso por ser una selección de tres palabras se denominan trigramas y que dan la menor exactitud con 33,74% (33,74099148).

Ahora observemos como trabajan el modelo RF en comparación con el modelo K; es decir, el mejor y menor resultado en las predicciones de la exactitud. Mientras que, el algoritmo RF crea un bosque de árboles de decisión con los textos y clases que contienen los tuits de corpus COVID, el algoritmo K busca la distancia entre un texto con su clase y todos los tuits que están más cercanos. Lo que implica que, los textos de estos tuits contienen información que no es constante en el tiempo; es decir, es muy variable o visto de otra forma es como un RF (random forest) y trata de pensar que todo está dicho sobre el covid; nos haría tener una perspectiva no apropiada para poder encontrar tuits más cercanos uno contra otros(K-Nearest).

Ahora bien, si vemos todos nuestros procesamientos como una máquina de inteligencia artificial podemos indicar que esta tiene un grado de confianza y una variabilidad en la exactitud (accuracy o Acc) que va desde un 33,74% (33,74099148 por el menor modelo K) hasta una exactitud del 47,08% (47,08451627 por el mejor modelo RF) sobre los tuits que correlacionan al corpus COVID.

Para terminar de ver nuestros resultados generales, el modelo SVM que por medio de una línea o hiperplano agrupa los tuits en las clases de este corpus como negative, neutral o positive; alcanza una exactitud del 44,79% (44,76960035); es decir que, el 55,24% (100% - 44,79%) son tuits con textos que encontramos tanto en clases negative, como en la clase neutral o en la clase positive.

Tabla 16 Resultados de las evaluaciones en modelos de machine learning para el corpus COVID

Conjunto	Modelo	Precision	Recall	F1	Acc
Palabra 1-1	SVM	41,5102029539	42,9788163354	41,3425620985	42,97881634
	RF	42,4902287228	46,0362524569	39,8786962253	46,03625246
	MNB	38,7073434937	37,9777243940	33,4756258721	37,97772439
	LR	40,2428993817	41,7121642280	40,3096817479	41,71216423
	K	35,8310751991	35,5317754968	34,2846775355	35,53177550
	J48	39,9728021427	40,0087355318	39,9655558319	40,00873553
Palabras 1-2	SVM	42,0590070328	43,8742083424	41,7434757746	43,87420834
	RF	40,7704896888	44,3546625901	38,3327624193	44,35466259
	MNB	40,3538268263	37,6283031230	31,3594139317	37,62830312
	LR	42,2091788978	43,5247870714	41,6745449843	43,52478707
	K	37,4928968504	36,1651015506	32,9604097514	36,16510155
	J48	39,0173154853	39,0259882070	38,9436072643	39,02598821
Palabras 1-3	SVM	41,8594110436	42,6949115527	39,8535113571	42,69491155
	RF	42,5898159096	45,4684428915	40,1948371801	45,46844289
	MNB	39,9868978181	38,1961126884	34,5409038504	38,19611269
	LR	41,9921921991	42,8696221883	40,1195370613	42,86962219
	K	38,2958668838	36,8421052632	33,9933378279	36,84210526
	J48	37,4168220242	36,9949770692	36,9864404500	36,99497707
Palabras 2-2	SVM	40,2367035036	41,1880323215	40,0231720790	41,18803232
	RF	40,2414714915	42,7822668705	38,3955387020	42,78226687
	MNB	40,1557711117	39,0915046953	36,6595337913	39,09150470
	LR	39,9480333128	40,5328674383	39,9027174027	40,53286744

Conjunto	Modelo	Precision	Recall	F1	Acc
Palabras 2-3	K	36,7278512999	33,8938632889	20,5450181726	33,89386329
	J48	38,0774145730	37,6501419524	36,6241968101	37,65014195
	SVM	39,3291277139	40,4673509500	39,1860769302	40,46735095
	RF	39,9885680594	40,6420615855	36,7736189063	40,64206159
	MNB	40,0670836524	39,1788600131	36,8469534994	39,17886001
	LR	39,6999057097	40,4236732911	39,5052313684	40,42367329
	K	37,1356269853	33,9157021184	20,2290673179	33,91570212
Palabras 3-3	J48	36,1257428876	35,6409696440	34,3882176621	35,64096964
	SVM	38,2217369650	38,5673727888	38,2881139164	38,56737279
	RF	39,0115800023	38,4145009827	33,9362613408	38,41450098
	MNB	37,9573896658	38,3489844944	37,5549476382	38,34898449
	LR	38,2365501065	38,6328892771	38,3069767800	38,63288928
	K	35,9058505837	33,7409914829	22,0450657523	33,74099148
	J48	35,8828001964	35,5317754968	33,4657796343	35,53177550
Caracteres 4-4	SVM	42,7148215206	44,4638567373	42,8637176779	44,46385674
	RF	44,5135249999	47,0845162699	43,6518949403	47,08451627
	MNB	39,3969493346	39,2225376720	34,3097690913	39,22253767
	LR	42,5660858505	43,8960471719	42,8372114686	43,89604717
	K	38,2566221646	37,2133653636	35,0462213258	37,21336536
	J48	39,8210929821	39,9213802140	39,8468009245	39,92138021
	SVM	42,7463035955	44,5512120550	42,9925203512	44,55121206
Caracteres 4-5	RF	44,2955699855	46,9316444639	42,5227753762	46,93164446
	MNB	40,1654405816	38,8075999126	32,0424725984	38,80759991
	LR	42,1162957872	43,5684647303	42,2247239066	43,56846473
	K	36,3815660102	35,7283249618	34,3270297971	35,72832496
	J48	37,9424291558	37,9777243940	37,9512062445	37,97772439
	SVM	42,5053178963	44,2017907840	42,7106337560	44,20179078
	RF	43,9195694313	46,6695785106	43,2833159525	46,66957851
Caracteres 4-6	MNB	40,8549373737	38,7639222538	31,7570010789	38,76392225
	LR	42,6262095217	43,9397248307	42,8835360171	43,93972483

Conjunto	Modelo	Precision	Recall	F1	Acc
Caracteres 4-7	K	36,2700444914	35,9903909150	34,2844540369	35,99039092
	J48	39,1507507365	39,2880541603	39,1868499361	39,28805416
	SVM	42,6987018317	44,5730508845	42,8146248756	44,57305088
	RF	43,6070453614	46,5822231928	42,0145162078	46,58222319
	MNB	40,8325562602	38,8731164010	31,8762832220	38,87311640
	LR	42,8049927091	44,0925966368	43,050830059	44,09259664
	K	35,4939335476	35,5099366674	33,9112925795	35,50993667
Caracteres 4-8	J48	39,7125866235	39,7248307491	39,6837166904	39,72483075
	SVM	42,5053962030	44,3983402490	42,6075163782	44,39834025
	RF	42,8021691924	46,1672854335	41,6182876467	46,16728543
	MNB	40,5740098616	38,7420834243	31,9602637228	38,74208342
	LR	42,7138443178	43,9834024896	42,9530398727	43,98340249
	K	35,6976368997	35,4880978380	34,0248246608	35,48809784
	J48	38,2607858903	38,2616291767	38,2297331519	38,26162918
Caracteres 4-9	SVM	42,9326194535	44,7696003494	42,9576653948	44,76960035
	RF	42,1845830812	45,8397029919	41,2387642705	45,83970299
	MNB	40,5426841985	38,8294387421	32,1616878042	38,82943874
	LR	42,7666343095	44,0270801485	42,9841697192	44,02708015
	K	35,2318456554	35,0731600786	33,5941675606	35,07316008
	J48	38,1127780670	38,1742738589	38,0718773900	38,17427386
	SVM	42,6895154661	44,5730508845	42,7129471093	44,57305088
Caracteres 4-10	RF	42,8424114292	46,0799301157	41,5294589236	46,07993012
	MNB	39,9574831393	38,7639222538	32,1422684981	38,76392225
	LR	42,7591984873	44,0270801485	42,9807060893	44,02708015
	K	35,0647666411	34,8547717842	33,5212587351	34,85477178
	J48	39,6313243115	39,6811530902	39,5828098355	39,68115309

En la Tabla 17, presentamos la comparación de las métricas resultante en los dos mejores modelos para el corpus COVID. Aquí los dos mejores modelos funcionan muy bien sobre estas combinaciones de conjuntos para char-gramas donde se toman proporción de caracteres

como; por ejemplo, en el modelo RF con caracteres (4-4) y el modelo SVM con caracteres (4-9).

Ahora hagamos una comparación cruzada a nivel de clase. En la clase negative para el modelo RF la precisión es de 51,22% (51,227496) y para la misma clase en el modelo SVM la precisión es del 47,90% (47,909138); esto implica que, predecir tuits con un sentimiento (negative) sobre el corpus COVID dan mejores resultados para los verdaderos negativos (TF) sobre la suma de los verdaderos negativos (TF) más los falsos negativos (FN); Pero, si vemos ahora la predicción en la clase positive el resultado de la métrica precisión en el modelo RF tiene un valor menor de 46,96% (46,966825) y para el modelo SVM tiene una precisión mayor de 47,10% (47,103569); es decir, que el análisis de sentimiento para la case positive tiene mejores resultados para los verdaderos positivos (TP) sobre la suma de los verdaderos positivos (TP) más los falsos positivos (FP).

En resumen, predecir tuits con análisis de sentimiento negativo y que correlacionan en el contexto de covid debemos utilizar el modelo RF con la combinación de char-gramas (4-4) y para predecir tuits con análisis de sentimiento positivo para este mismo contexto debemos utilizar el modelo SVM con la combinación de char-gramas (4-9) pero sin tener presente la exactitud en ambos modelos.

Tabla 17 Comparación de los dos mejores modelos de machine learning en el corpus COVID.

	Primer mejor modelo Acc 47,08451627				Segundo mejor modelo Acc 44,76960035			
Modelo	RF				SVM			
Conjunto	Caracteres 4-4				Caracteres 4-9			
	precision	recall	f1-score	support	precision	recall	f1-score	support
negative	51,227496	61,898484	56,059701	1517	47,909138	61,173368	53,734800	1517
neutral	35,534591	14,599483	20,695971	1548	33,976420	20,478036	25,554212	1548
positive	46,966825	65,455746	54,690949	1514	47,103569	53,170410	49,953460	1514
macro avg	44,576304	47,317904	43,815540	4579	42,996376	44,940605	43,080824	4579
weighted avg	44,513525	47,084516	43,651895	4579	42,932619	44,769600	42,957665	4579

En la **Figura 75** presentamos la matriz de confusión para comparar las predicciones resultantes de los mejores modelos en el corpus COVID. Ahora realizamos un análisis para ver desde la perspectiva de los falsos negativos (FN) y los falsos positivos (FP) para verificar que el nivel de error de nuestra máquina de inteligencia artificial para el contexto de esta pandemia. Para el modelo RF con Acc (47,08451627) los falsos negativos son el 20% y los falsos positivos son el 26%; Pero para el modelo SVM con Acc (44,76960035) tiene en los falsos negativos un 25% y en los falsos positivos el 21%. En resumen, el primer modelo RF se equivoca menos en los falsos negativos de un 20% vs los falsos negativos del modelo SVM con un 25%, para el modelo SVM en los falsos positivos de un 21% es menor el error en comparación con el modelo RF en los falsos positivos de un 26%.

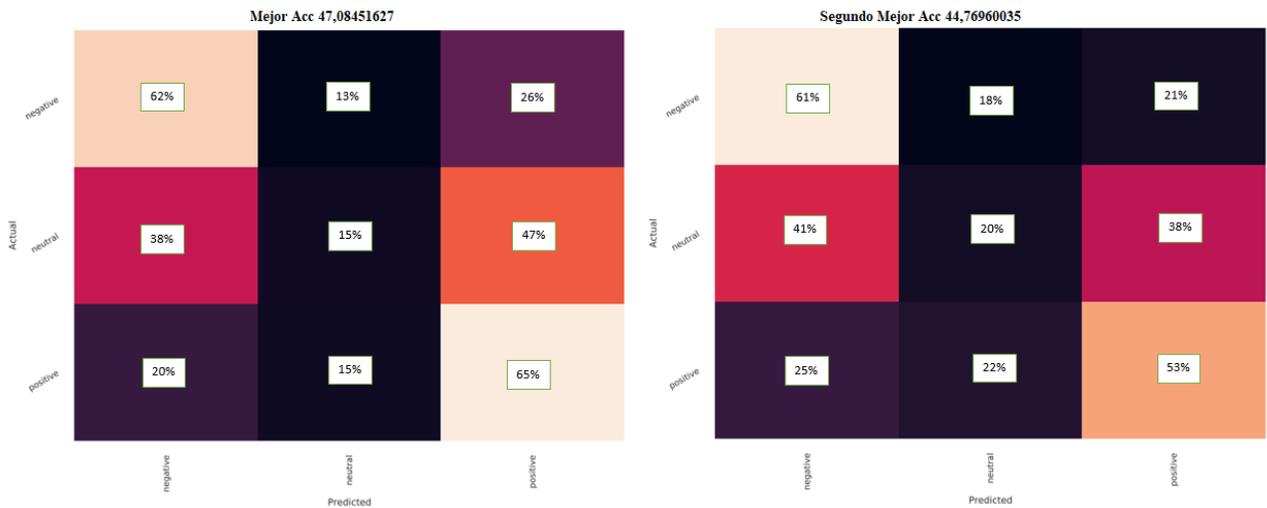


Figura 75 Matriz de confusión de los dos mejores modelos de machine learning en el corpus COVID

En la Tabla 18 , realizamos la comparación entre los dos mejores modelos para los n-gramas y los dos mejores modelos para los char-gramas. El modelo RF está presente en la mejor exactitud para los char-gramas con la combinación de caracteres (4-4) tiene un 47,08% (47,08451627) y para la combinación de caracteres (4-9) con un 46,03% (46,03625246). Esto implica que los usuarios que escriben los tuits del corpus COVID expresan sus sentimientos en los primeros caracteres dentro del texto de cada tuit. En el modelo RF que crea un bosque de árboles de decisiones nos permite tener una buena predicción a través de la proporción de caracteres como “covi”, “muer”, “tos ”, “dolo” o “vida”, etc. Aun cuando la maquina da buenos resultados las personas que expresan sus opiniones por medio del lenguaje natural

requieren ampliar un poco más el contenido de estos texto y por ende, aparece el modelo SVM que presenta la segunda mejor exactitud con la combinación de caracteres como “coronavir”, “hospital ”, “tos seca ”, “pulmones ”, entre otras. Esto permite que el modelo divida las clases(sentimientos) para ayudar a la predicción.

Por otro lado, en los n-gramas encontramos que los modelos RF y SVM también nos dan los mejores resultados en la exactitud. Para el modelo RF que genera un bosque de árboles de decisiones tenemos que la exactitud tiene el 46,03% (46,03625246) con la combinación de unigramas (una sola palabra); lo que indica que los usuarios que generan textos en los tuits para el contexto del covid, han centrado en una sola palabra el sentimiento o clase (negative, neutral y positive). Para el modelo SVM que separa las clases (negative, neutral y positive) por medio de un hiperplano y tiene una exactitud de 43,87% (43,87420834). Esto implica que la combinación de palabras (1-2) o unigramas + bigramas; como, por ejemplo, “tos seca”, “dolor cabeza”, “flema pulmones”, etc permite a la máquina de predicción producir mejores resultados.

Tabla 18 Comparación por modelo de los dos mejores n-gramas y los dos mejores char-gramas en el corpus COVID

Modelo	Conjunto	Precision	Recall	F1	Acc
RF	Caracteres 4-4	44,513525	47,084516	43,651895	47,08451627
SVM	Caracteres 4-9	42,932619	44,7696	42,957665	44,76960035
RF	Palabra1-1	42,490229	46,036252	39,878696	46,03625246
SVM	Palabras1-2	42,059007	43,874208	41,743476	43,87420834

En la Figura 76, nos da la comparación en las matrices de confusión para los n-gramas y char-gramas. Para el modelo RF con la combinación en caracteres (4-4) tenemos que los verdaderos negativos (TN) tiene un 62% y para la combinación de palabras (1-1) tiene un 61% en los mismos verdaderos negativos (TN). Esto implica que sentimientos negativos o clase negative tiene una predicción más alta con la combinación de los primeros caracteres del texto que contienen los tuits en el corpus COVID. Por otro lado, los verdaderos positivos (TP) en la combinación de caracteres (4-4) tiene 65% en comparación al 72% de los

verdaderos positivos en la combinación de palabras (1-1) o unigramas. Esto indica que predecir sentimiento positivos o clase positive tiene una mejor predicción en los tuits cuando solo se analiza los unigramas (una sola palabra).

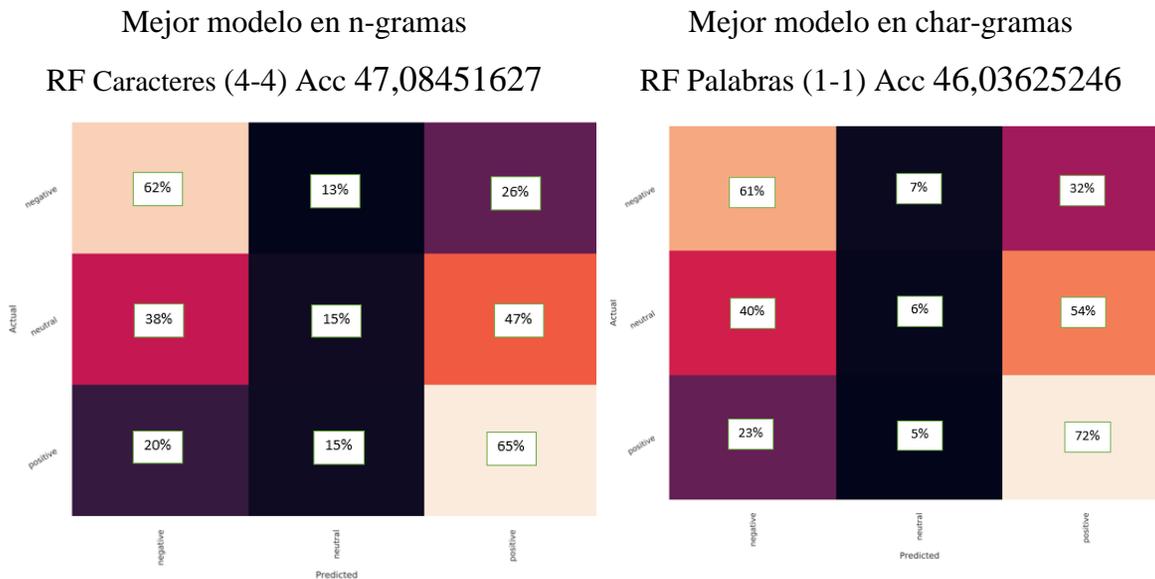


Figura 76 Matriz de confusión del mejor modelo para n-gramas y el mejor modelo para char-gramas para el corpus COVID

4.4.2 Evaluación para modelos de deep learning del corpus COVID.

Entre las técnicas de predicción tenemos también los modelos de deep learning o aprendizaje profundo que nos darán a conocer como estos algoritmos nos pueden ayudar en el análisis de sentimiento sobre los tuits que componen el corpus del COVID. Estos modelos ya se describen en la sección 4.3.2

En la Tabla 19 se muestran los resultados de las ejecuciones para los modelos de deep learning y sus métricas. Por lo cual, los modelos LSTM con un Acc (accuracy o exactitud) de 47,56% (47,56497052) y el modelo BILSTM con una exactitud de 47, 93% (47,93623062) son los mejores modelos para la predicción en el corpus del COVID. Y que el modelo con menor resultado en la exactitud es GRU con una exactitud del 45,18% (45,18453811); por lo cual podemos indicar que nuestra tabla de resultado o máquina de deep learning tiene un

grado de variabilidad en la exactitud que está en un rango que oscila desde el 45,18% hasta un 47,93%. Y si estos rangos lo comparamos con los resultados del rango obtenido en la sección 0 (desde un 33,74% hasta un 47,08%), los resultados de deep learning en este ámbito de los tuits generados por una pandemia; indican que, los modelos que usan una red neuronal artificial compuesta de un número en niveles jerárquicos nos permiten realizar mejores predicciones de las clases negative, neutral y positive.

Por consiguiente, las características contextuales que acompañan a este rango de efectividad para predecir en un rango del 45,18% hasta 47,93% con técnicas de deep learning sobre el corpus del COVID son: la información generada está en el ámbito de la infodemiología, seguidos de pandemia, variantes del virus con síntomas con diferentes síntomas, colapso el sistema sanitario, recesión en el comercio, restricciones de movilidad o reunión, la transformación digital como alternativa de actividades, vacunas que han detenido la tasa de mortalidad y el índice de camas UCI ocupadas. Por lo cual, podemos puntualizar que la información sobre el covid sigue siendo una información en constante cambio sobre una línea de tiempo; es decir, los usuarios que crean tuits que se refieren al covid generan textos sobre comentarios, problemas o soluciones que tienden a tener nuevos contextos.

Tabla 19 Resultados de las evaluaciones en modelos de deep learning para el corpus COVID

Modelo	Precision	Recall	F1	Acc
CNN	43,217821	45,293732	42,908807	45,29373226
LSTM	42,158870	47,564971	38,474346	47,56497052
GRU	44,367666	45,184538	44,415896	45,18453811
BI-GRU	41,467005	43,983402	40,778609	43,98340249
BILSTM	46,399217	47,936231	40,579154	47,93623062
DENSE	44,678303	46,953483	44,140776	46,95348329

En la Tabla 20, realizamos la comparación de los dos mejores modelos de deep learning ampliando en cada clase objetivo (negative, neutral y positive) y las métricas de evaluación como las medidas de macro avg (macro promedio) y weighted avg (promedio ponderado).

Primero hagamos un acercamiento al modelo LSTM que en la comparación tiene una menor exactitud que el modelo de BILSTM. El modelo LSTM son un tipo de redes recurrentes cuya característica principal es que la información puede persistir introduciendo bucles en el diagrama de la red, por lo que, básicamente, pueden recordar estados previos y utilizar esta información para decidir cuál será el siguiente; Esto quiere decir, que cuando el modelo LSTM toma un tuit para predecir su clase (negative, neutral o positive); mantienen o recuerda las palabras más representativas en la predicción y toman esta información para ayudar a predecir el siguiente tuit. Así la exactitud en la predicción de este modelo llego a un 47,56% (47,56497052).

Ahora hagamos el acercamiento al modelo BILSTM que en la comparación tiene una menor exactitud que el modelo de LSTM. El modelo BILSTM aplica el algoritmo de LSTM que toma un tuit lo predice y pasa la información representativa para ayudar a predecir el siguiente tuit. BILSTM hace los mismo, pero adicionando que cuando el último tuit se ha terminado de predecir, este es tomado como el principio y lo predice con un LSTM para que la información importante la pase a un inmediato anterior; es decir hace la predicción en dos direcciones. Este algoritmo ayuda a entender mejor la semántica del procesamiento del lenguaje natural para las predicciones en el corpus del COVID. Por lo cual, este modelo tiene una exactitud del 47,93% (47,93623062)

Tabla 20 Comparación de los dos mejores modelos de deep learning en el corpus COVID

	BILSTM Acc 47,93623062				LSTM Acc 47,56497052			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
negative	51,690821	63,480554	56,982249	1517	49,210037	69,808833	57,726901	1517
neutral	42,011834	45,865633	82,702388	1548	31,111111	0,904393	17,576899	1548
positive	45,583039	76,684280	57,178035	1514	46,389589	72,985469	56,724846	1514
macro avg	46,428565	48,250466	40,810174	4579	42,236912	47,899565	38,736479	4579
weighted avg	46,399217	47,936231	40,579154	4579	42,158870	47,564971	38,474346	4579

En la **Figura 77** presentamos la matriz de confusión comparando los dos mejores resultados que han arrojado las evaluaciones; en primer lugar, el modelo BILSTM (**Acc 47,93623062**) y en segundo lugar el modelo LSTM (**Acc 47,56497052**). Los verdaderos negativos, es decir

tuit que hay tenido un análisis de sentimiento de clase negative. El modelo BILSTM indica que los TN (verdaderos negativos) son un 63%, pero que los falsos negativos (FN) llegan a un 20%; es decir tenemos una tasa de error importante. En el modelo LSTM los verdaderos negativos llegan a un 70%, pero que los falsos negativos (FN) llegan a un 26%, esta variable de error es mucho mayor al primer modelo; lo que implica que tratar de alcanzar más predicciones hace que la variable de error aumente entre un modelo y otro para la clase negative.

Para el caso práctico, si se desea tener más predicciones negativas aun cuando existan más falsos positivos, tomaríamos LSTM, pero si lo que necesitamos es que las predicciones para tuit con sentimiento negativo tengan mejor exactitud, minimizando la variable de error, para este caso se prefiere el modelo BILSTM

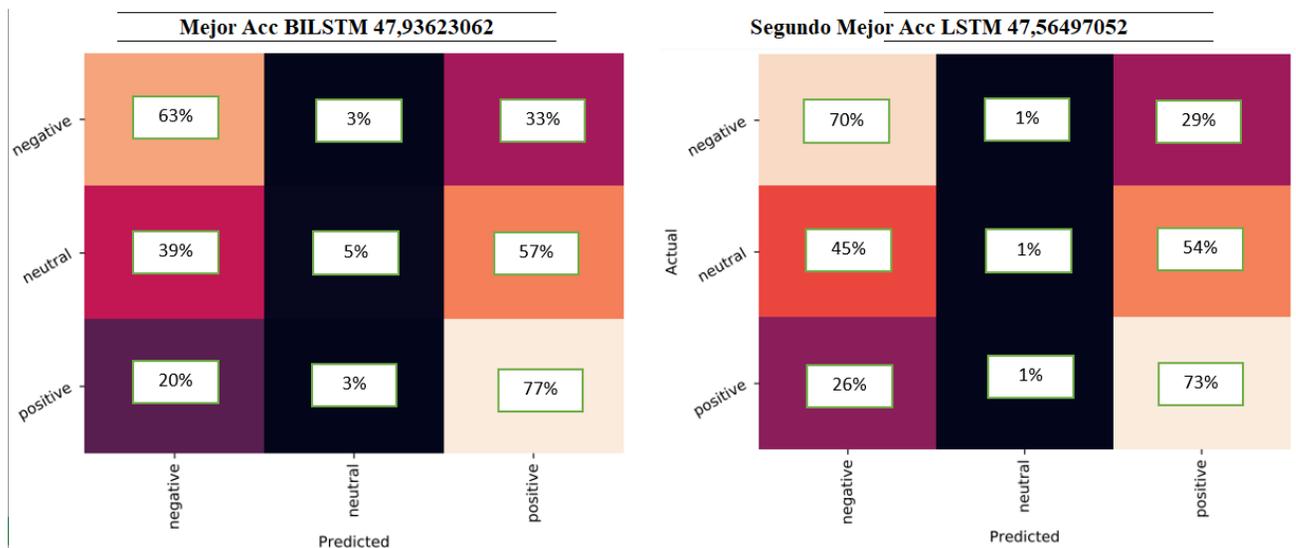


Figura 77 Matriz de confusión de los dos mejores modelos de deep learning en el corpus COVID

4.4.3 Discusión de los resultados del corpus COVID.

En la Tabla 21 presentamos un resumen de los mejores modelos con técnicas de machine learning y con modelos de deep learning que dieron como resultado la evaluación de tuits

para el corpus COVID. Los resultados tienen un grado de exactitud que va desde 43,87% (43,87420834) hasta el 47,93% (47,93623062) y los modelos que mejores resultados dieron son los de deep learning en los algoritmos de BILSTM y LSTM, las técnicas de machine learning son los siguientes mejores en resultados con los modelos RF y SVM en las combinaciones de char-gramas, para luego finalizar con las combinaciones de n-gramas.

Tabla 21 Mejores accuracy de los modelos en machine learning y Deep learning en el corpus COVID

Clasificador	Conjunto	Accuracy
BILSTM	FastText	47,93623062
LSTM	FastText	47,56497052
RF	Caracteres 4-4	47,08451627
SVM	Caracteres 4-9	44,76960035
RF	Palabra1-1	46,03625246
SVM	Palabras1-2	43,87420834

Ahora pasemos a realizar una comparación del resultado de aplicar técnicas de machine y deep learning en el corpus COVID versus los resultados que se dieron en la evaluación en el corpus del ZIKA y que se resume en la sección 4.3.3 con un rango de efectividad que va desde un 59,70% (59,70361348) hasta un 61,20% (61,20584653) y este rango de efectividad es mayor a los resultados de la evaluación en el corpus COVID (43,87% a 47,93%); por lo cual, es necesario hacer las siguientes evaluaciones.

De lo detallado en la sección 3.6 el corpus del COVID contiene 188.618 tuits en comparación a los 39.816 que contiene el corpus del ZIKA. Al tener una mayor cada de tuits que con llevan a que el texto de cada tuit lleva consigo una cantidad de hasta 280 caracteres, hará que los algoritmos de machine o deep learning deban realizar más evaluaciones sobre un corpus de información que está cambiando constantemente.

Cada corpus maneja 3 clases objetivos como son el negative, neutral y positive; dividiendo para tres las posibles alternativas que tienes algoritmos en predecir. Como la predicción que se realizar sobre el procesamiento del lenguaje natural existen muchas palabras que se

correlacionan para la clase negative y que también la encontramos en la positive o en su defecto en la clase neutral.

En la **Figura 42**, indica que la cantidad de texto en los tuis que se encuentran en el corpus del COVID es mayor a la cantidad de texto que contienen los tuits del corpus ZIKA indicado en la Figura 24. Por lo cual, diversifica la semántica del lenguaje natural haciendo caer la efectividad de los modelos de machine learning y haciendo más necesario el uso de los algoritmos de deep learning.

Además, en la Figura 53 indica que la línea de tendencia neutral del corpus COVID está por encima de las líneas en tendencia de la clase negative y mucho más de la clase positive; pero en la Figura 35 que presenta las líneas de tendencia del corpus ZIKA indica que las líneas de tendencia para la neutral, negative y positive se traslapan, indicando que los tuits del covid contienen información importante que no es ni positiva ni negativa que ayuden a la predicción tanto para los modelos de machine o deep learning.

El corpus del ZIKA que contiene tuits de las enfermedades infectocontagiosas como zika, dengue o chikungunya tiene un factor común del cual el vector trasmisor es el mosquito y que de una u otra manera se relacionan sus síntomas eso hace que el corpus contenga información que tiene una clara definición en su contexto a comparación con el corpus del COVID que la información aún está por definirse; ya que este virus al principio se lo relacionaba como una neumonía pero aún no termina en definir su contexto.

La calidad del corpus del COVID es mejorable si procedemos a ejecutar una actividad para Re etiquetar los tuits. Esto permitirá aprovechar el conocimiento ya adquirido de las personas, producto de enfrentar la pandemia y las acciones que se han seguido para poder aplanar la curva de mortalidad; así como, los efectos que han tenido las vacunas en la sociedad como, por ejemplo, minimizando los casos de hospitalización y la apertura del comercio como otras actividades.

4.5 Evaluación del corpus SATIRA

El corpus SATIRA, contiene dos clases objetivos; por lo cual, realizaremos una evaluación para una clasificación binaria. Esto nos dará una perspectiva diferente en los resultados si cambiamos el procesamiento de los modelos, pasando de clasificaciones ternaria (tres clases en los corpus del ZIKA y COVID) a una clasificación binaria.

En la sección 0 se realiza la evaluación para los modelos tradicionales de machine learning, en la sección 4.5.2 la evaluación para modelos deep learning y en la sección 4.5.3 la discusión de los mejores dos modelos de cada sección sobre el corpus de la SATIRA

4.5.1 Evaluación para modelos tradicionales de machine learning del corpus SATIRA.

En esta sección se presentan los resultados para cada evaluación; producto de aplicar técnicas tradicionales de machine learning en el corpus SATIRA. La estructuración de estos resultados, modelos y métricas; ya fue descrita en las secciones 4.1 y 4.2 de este capítulo. Además de los resultados detallados en la sección 0 que permitió descubrir que modelos y combinaciones dan mejores predicciones para el corpus que se analizó.

El corpus de la SATIRA se incorpora a esta tesis, producto del contenido en su texto que se orienta a uno de los géneros literarios del lenguaje natural. La sátira es un género que con el pasar del tiempo ahora se hace presente en el internet y más específicamente en las redes sociales. Este tipo literario se caracteriza como un guion irónico o cómico que llega al punto de ridiculizar una idea o por un constante sarcasmo y que también es vista como una parodia. Este corpus tiene en sus tuits dos clases que son; satírico o no satírico.

En la Tabla 22 presentamos los resultados obtenidos por los modelos de machine learning en combinación con el conjunto de palabras o porción de caracteres más la visualización de los resultados en las métricas, que nos permite evaluar el rendimiento de la predicción en el procesamiento de los textos que componen el corpus SATIRA. Así en la columna Acc(acurracy) presentamos el valor resaltado en negritas para los dos mejores modelos. Cabe recordar que la clasificación de estos tuits es binaria y se clasifican en satírico o no satírico por eso se espera que se obtenga mucho mejores resultados que con los corpus anteriores.

Así encontramos que el modelo SVM en conjunto con la combinación palabras (1-2) o unigramas + bigramas nos da el mejor resultado en a exactitud (accuracy) alcanzando una predicción del 83,73% (83,73056995). El objetivo de este análisis de clases (satírico o no satírico) es trazar la predicción más aceptable en los tuits de este contexto. En el segundo mejor modelo es el MNB con la combinación palabras (1-3) o unigramas + bigramas + trigramas y que tiene una exactitud de 83,52% (83,52331606). En la presentación de todos

los resultados el modelo K(K-Nearest) en combinación con el conjunto de palabras (3-3) llamados n-gramas o en este caso por ser una selección de tres palabras se denominan trigramas y que dan la menor exactitud con 55,95% (55,95854922).

Ahora observemos como trabaja el modelo SVM en comparación con el modelo K; es decir, el mejor y menor resultado en las predicciones de la exactitud. Mientras que, el algoritmo SVM crea un hiperplano para poder dividir las clases del corpus SATIRA y hacer una predicción sobre este entorno, el algoritmo K busca la distancia entre un texto con su clase y todos los tuits que están más cercanos. Lo que implica que, los textos de estos tuis contienen información que no es constante en el tiempo; es decir, es muy variable o visto de otra forma es como un SVM (support vector machine forest) y trata de pensar que todo está dicho sobre la sátira; lo cual nos haría tener una perspectiva no apropiada para poder encontrar tuits más cercanos uno contra otros(K-Nearest).

Ahora bien, si vemos todos nuestros procesamientos como una máquina de inteligencia artificial podemos indicar que esta tiene un grado de confianza y una variabilidad en la exactitud (accuracy o Acc) que va desde un 55,95% (55,95854922 por el menor modelo K) hasta una exactitud del 83,73% (83,73056995 por el mejor modelo SVM) sobre los tuits que correlacionan al corpus SATIRA.

Para terminar de ver nuestros resultados generales, el modelo MNB (Multinomial Naive Bayes) asume que las variables a predecir son independientes entre sí. Lo cual implica que, la presencia de una cierta característica en una clase no está en absoluto relacionada con la presencia de cualquier otra característica para las clases de este corpus. Las clases non-satire o satire alcanzan una exactitud del 83,52% (83,52331606); es decir que, el 16,48% (100% - 83,52%) son tuits con textos que encontramos tanto en clases non-satire como en la clase satire.

Tabla 22 Resultados de las evaluaciones en modelos de machine learning para el corpus SATIRA

Conjunto	Modelo	Precision	Recall	F1	Acc
Palabra	SVM	82,9708134752	82,9015544041	82,8377981118	82,90155440
1-1	RF	78,0328723596	77,8238341969	77,8531783527	77,82383420

Conjunto	Modelo	Precision	Recall	F1	Acc
	MNB	83,3485760567	82,9015544041	82,7359518441	82,90155440
	LR	80,0929778875	79,7927461140	79,6146889568	79,79274611
	K	79,8420354935	78,6528497409	78,1938968361	78,65284974
	J48	64,5020892529	64,0414507772	64,0783015873	64,04145078
Palabras	SVM	84,1675412791	83,7305699482	83,5801799443	83,73056995
1-2	RF	77,4398949739	76,6839378238	76,6906481377	76,68393782
	MNB	83,6214907833	83,0051813472	82,8048359240	83,00518135
	LR	82,2358887226	81,5544041451	81,3091191263	81,55440415
	K	79,9361226849	78,2383419689	77,6330548879	78,23834197
	J48	68,5729336396	68,0829015544	68,1143751427	68,08290155
Palabras	SVM	84,3078523458	83,6269430052	83,4258473860	83,62694301
1-3	RF	75,7970423558	74,9222797927	74,9163013153	74,92227979
	MNB	84,0988568020	83,5233160622	83,3409814792	83,52331606
	LR	81,1521909009	80,3108808290	79,9957432213	80,31088083
	K	79,9688710874	78,4455958549	77,8934525122	78,44559585
	J48	67,2743189028	66,3212435233	66,2781968296	66,32124352
Palabras	SVM	79,7026139470	79,6891191710	79,6280161808	79,68911917
2-2	RF	69,5104254690	69,5336787565	69,3325829011	69,53367876
	MNB	80,8642782541	80,3108808290	80,0690569832	80,31088083
	LR	78,5857412026	78,5492227979	78,4627191730	78,54922280
	K	76,9022056825	76,3730569948	76,0339666091	76,37305699
	J48	68,2738223964	65,9067357513	65,4951626142	65,90673575
Palabras	SVM	80,6629963510	80,5181347150	80,4038597449	80,51813472
2-3	RF	68,7962049758	68,7046632124	68,3357344454	68,70466321
	MNB	81,0795016348	80,5181347150	80,2788563833	80,51813472
	LR	79,4296893812	79,1709844560	78,9962560506	79,17098446
	K	76,4826963883	75,8549222798	75,4616027120	75,85492228
	J48	69,1078401255	66,4248704663	65,9430455555	66,42487047
Palabras	SVM	72,0991511782	72,0207253886	72,0447857750	72,02072539
3-3	RF	68,9538200277	63,4196891192	58,8062781938	63,41968912

Conjunto	Modelo	Precision	Recall	F1	Acc
	MNB	74,4261772832	72,3316062176	71,1796030746	72,33160622
	LR	74,2197399257	72,3316062176	71,2553310142	72,33160622
	K	74,8247197167	55,9585492228	48,2239661334	55,95854922
	J48	69,3807255746	61,0362694301	58,0894326643	61,03626943
Caracteres	SVM	82,3710950273	82,3834196891	82,3644372910	82,38341969
4-4	RF	80,3785667812	80,3108808290	80,3278122120	80,31088083
	MNB	80,3306929756	79,6891191710	79,4084231810	79,68911917
	LR	80,4048018199	80,4145077720	80,3782275373	80,41450777
	K	80,0711938342	78,6528497409	78,1361866632	78,65284974
	J48	68,7535280041	68,6010362694	68,6383531736	68,60103627
Caracteres	SVM	82,0905828802	82,0725388601	82,0258734732	82,07253886
4-5	RF	81,0965229436	81,0362694301	81,0516797694	81,03626943
	MNB	81,4518817810	80,3108808290	79,9253979278	80,31088083
	LR	81,6820499935	81,6580310881	81,6053707757	81,65803109
	K	77,0384492533	76,0621761658	75,5553401319	76,06217617
	J48	69,7161169574	69,7409326425	69,7258242997	69,74093264
Caracteres	SVM	81,3949480447	81,3471502591	81,2803691117	81,34715026
4-6	RF	81,0440555089	80,9326424870	80,9535860602	80,93264249
	MNB	81,3463404627	80,3108808290	79,9496265952	80,31088083
	LR	81,4355505980	81,3471502591	81,2631119626	81,34715026
	K	77,8311202541	76,8911917098	76,4315030236	76,89119171
	J48	70,5393921404	70,3626943005	70,4000120445	70,36269430
Caracteres	SVM	83,0444481014	83,0051813472	82,9540543553	83,00518135
4-7	RF	80,5634716139	80,3108808290	80,3375181921	80,31088083
	MNB	80,5896896670	79,6891191710	79,3406437709	79,68911917
	LR	81,1445303951	81,0362694301	80,9414797797	81,03626943
	K	77,7404282046	76,8911917098	76,4601704664	76,89119171
	J48	68,4475367396	68,0829015544	68,1236513405	68,08290155
Caracteres	SVM	83,0444481014	83,0051813472	82,9540543553	83,00518135
4-8	RF	80,3562631306	80,1036269430	80,1305446994	80,10362694

Conjunto	Modelo	Precision	Recall	F1	Acc
	MNB	80,4569284438	79,5854922280	79,2411904611	79,58549223
	LR	80,8213379532	80,7253886010	80,6322567019	80,72538860
	K	76,8600016969	76,0621761658	75,6156922768	76,06217617
	J48	70,5393921404	70,3626943005	70,4000120445	70,36269430
Caracteres	SVM	83,3335078640	83,3160621762	83,2769564696	83,31606218
4-9	RF	80,2899996104	80,1036269430	80,1294025528	80,10362694
	MNB	80,2366613904	79,4818652850	79,1649696039	79,48186528
	LR	80,6807259872	80,6217616580	80,5436158712	80,62176166
	K	75,7099608311	74,9222797927	74,4312963081	74,92227979
	J48	68,9805868739	68,7046632124	68,7470633461	68,70466321
Caracteres	SVM	83,1336923053	83,1088082902	83,0648403244	83,10880829
4-10	RF	80,5402904416	80,3108808290	80,3373107688	80,31088083
	MNB	80,1492735179	79,3782383420	79,0539828433	79,37823834
	LR	80,4844348962	80,4145077720	80,3293980001	80,41450777
	K	75,4407413422	74,6113989637	74,0902305457	74,61139896
	J48	69,3807255746	61,0362694301	58,0894326643	61,03626943

En la Tabla 23, presentamos la comparación de las métricas resultante en los dos mejores modelos para el corpus SATIRA. Aquí los dos mejores modelos funcionan muy bien sobre estas combinaciones de conjuntos para n-gramas donde se unigramas + bigramas + bigramas como; por ejemplo, en el modelo SVM con Palabras (1-2) y el modelo MNB con caracteres (1-3).

Ahora hagamos una comparación cruzada a nivel de clase. En la clase non-satire para el modelo SVM la precisión es de 81,06% (81,067126) y para la misma clase en el modelo MNB la precisión es del 80,47% (80,475382); esto implica que, predecir tuits con un sentimiento (non-satire) sobre el corpus SATIRA dan mejores resultados para los verdaderos negativos (TF) sobre la suma de los verdaderos negativos (TF) más los falsos negativos (FN); Pero, si vemos ahora la predicción en la clase satire el resultado de la métrica precisión en el modelo SVM tiene un valor menor de 87,76% (87,760417) y para el modelo SVM tiene una

precisión mayor de 88,29% (88,297872); es decir, que el análisis de sentimiento para la case positive tiene mejores resultados para los verdaderos positivos (TP) sobre la suma de los verdaderos positivos (TP) más los falsos positivos (FP).

En resumen, predecir tuits con análisis de sentimiento negativo y que correlacionan en el contexto de la sátira debemos utilizar el modelo SVM con la combinación de n-gramas Palabras (1-2) y para predecir tuits con análisis de sentimiento satire para este mismo contexto debemos utilizar el modelo MNB con la combinación de n-gramas Palabras (1-3) pero sin tener presente la exactitud en ambos modelos.

Tabla 23 Comparación de los dos mejores modelos de machine learning en el corpus SATIRA

		Primer mejor modelo Acc 83,73056995				Segundo mejor modelo Acc 83,52331606			
Modelo		SVM				MNB			
Conjunto		Palabras 1-2				Palabras 1-3			
		precision	recall	f1-score	support	precision	recall	f1-score	support
non_satire		81,067126	90,926641	85,714286	518	80,475382	91,505792	85,636856	518
satire		87,760417	75,391499	81,107100	447	88,297872	74,272931	80,680437	447
macro avg		84,413771	83,159070	83,410693	965	84,386627	82,889361	83,158647	965
weighted avg		84,167541	83,730570	83,580180	965	84,098857	83,523316	83,340981	965

En la **Figura 78** presentamos la matriz de confusión para comparar las predicciones resultantes de los mejores modelos en el corpus SATIRA. Ahora realizamos un análisis para ver desde la perspectiva de los falsos negativos (FN) y los falsos positivos (FP) para verificar que el nivel de error de nuestra máquina de inteligencia artificial para el contexto de esta pandemia. Para el modelo SVM con Acc (83,73056995) los falsos negativos son el 25% y los falsos positivos son el 9%; Pero para el modelo MNB con Acc (83,52331606) tiene en los falsos negativos un 26% y en los falsos positivos el 8%.

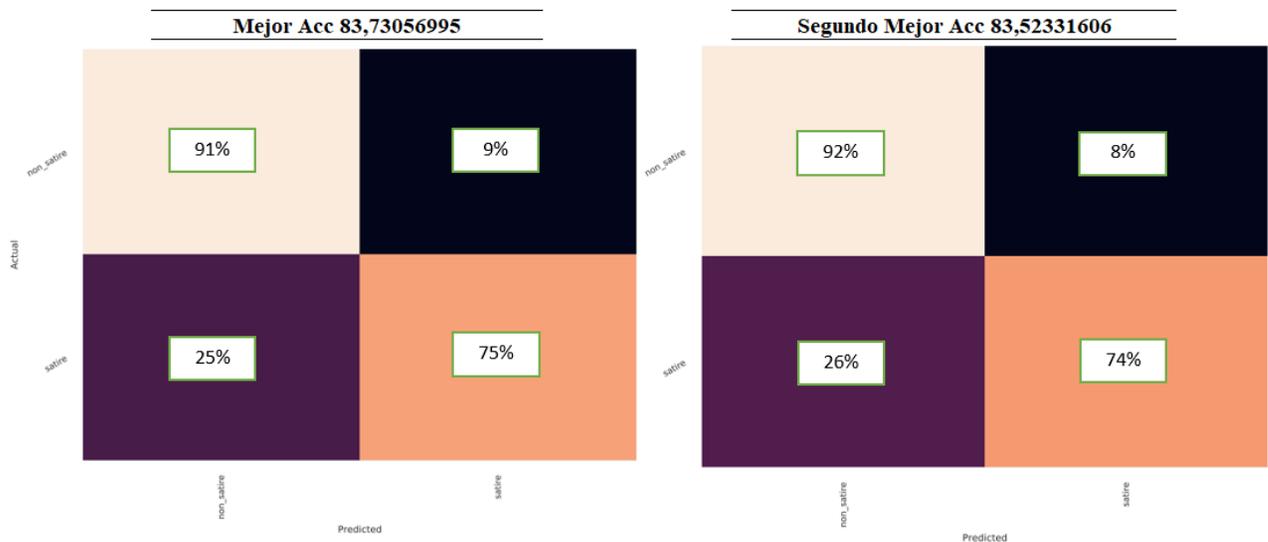


Figura 78 Matriz de confusión de los dos mejores modelos de machine learning en el corpus SATIRA

En la Tabla 24 , realizamos la comparación entre los dos mejores modelos para los n-gramas y los dos mejores modelos para los char-gramas. El modelo SVM está presente en la mejor exactitud para los n-gramas con la combinación de palabras (1-2) tiene un 83,73% (83,73056995) y para la combinación de caracteres (4-9) con un 83,31% (83,31606218). Esto implica que los usuarios que escriben los tuits del corpus SATIRA expresan sus sentimientos en n-gramas o char-gramas que están dentro del texto de cada tuit. En el modelo MNB se tiene en cuenta el número de ocurrencia de un término (n-gramas) sobre el corpus y se asume que los términos son independientes entre sí. Por lo cual, el modelo es capaz de predecir si un tuit es non-satire o satire. Tanto los modelos de SVM y LR utilizan un hiperplano para dividir ambas clases non-satire o satire, así podemos concluir que SVM es determinista (n-gramas o char-gramas pertenecen a una clase u otra) y LR es probabilístico. (la probabilidad de que n-grama o char-grama sea de una clase u otra)

Tabla 24 Comparación por modelo de los dos mejores n-gramas y los dos mejores char-gramas en el corpus SATIRA

Modelo	Conjunto	Precision	Recall	F1	Acc
SVM	Palabras1-2	84,16754128	83,73056995	83,58017994	83,73056995
MNB	Palabras1-3	84,0988568	83,52331606	83,34098148	83,52331606
SVM	Caracteres 4-9	83,33350786	83,31606218	83,27695647	83,31606218
LR	Caracteres 4-5	81,68204999	81,65803109	81,60537078	81,65803109

En la **Figura 79**, nos da la comparación en las matrices de confusión para los n-gramas y char-gramas. Para el modelo SVM con la combinación en Palabras (1-2) tenemos que los verdaderos non-satire (TN) tiene un 91% y para la combinación de caracteres (4-9) tiene un 87% en los mismos verdaderos non-satire (TN). Esto implica que sentimientos negativos o clase non-satire tiene una predicción más alta con la combinación de los primeros unigramas + bigramas del texto que contienen los tuits en el corpus SATIRAA. Por otro lado, los verdaderos satire (TP) en la combinación de palabras (1-2) tiene 75% en comparación al 79% de los verdaderos positivos en la combinación de Caracteres (4-9) o char-gramas. Esto indica que predecir sentimiento positivos o clase satire tiene una mejor predicción en los tuits cuando solo se analiza una proporción de caracteres en el texto.

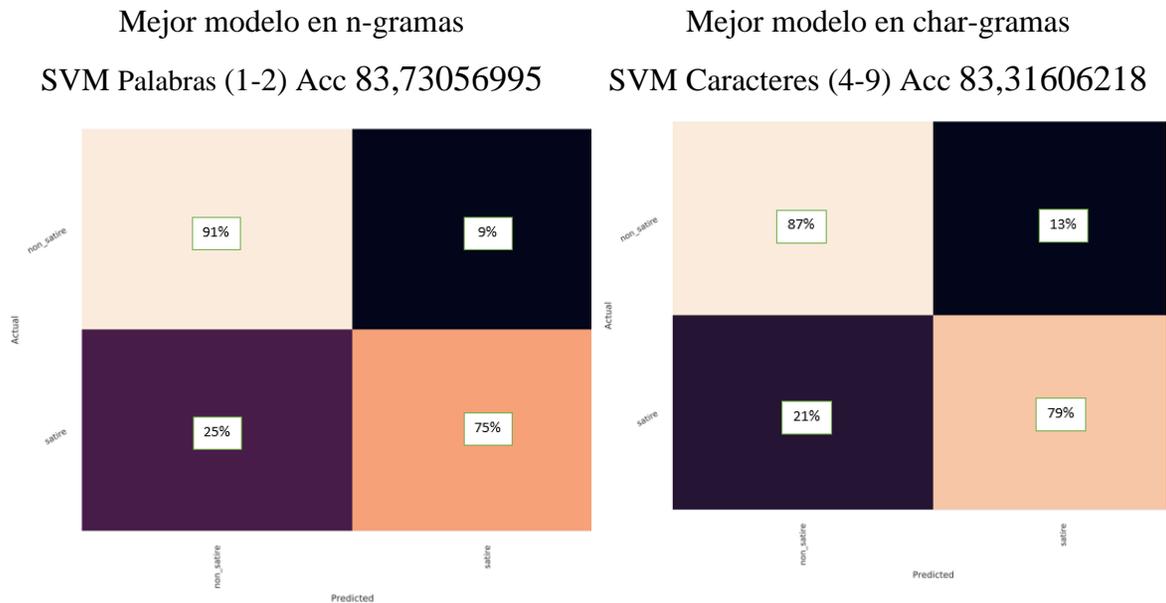


Figura 79 Matriz de confusión del mejor modelo para n-gramas y el mejor modelo para char-gramas para el corpus SATIRA

4.5.2 Evaluación para modelos de deep learning del corpus SATIRA.

En la Tabla 25 se muestran los resultados de las ejecuciones para los modelos de deep learning y sus métricas. En este caso, los modelos DENSE con un Acc (accuracy o exactitud) de 82,79% (82,79792746) y el modelo CNN con una exactitud de 82,38% (82,38341969) son los mejores modelos para la predicción en el corpus SATIRA. Y que el modelo con menor resultado en la exactitud es LSTM con una exactitud del 80,82% (80,82901554); por lo cual podemos indicar que nuestra tabla de resultado o máquina de deep learning tiene un grado de variabilidad en la exactitud que está en un rango que oscila desde el 80,82% hasta un 82,79%. Y si estos rangos lo comparamos con los resultados del rango obtenido en la sección 0 (desde un 83,52% hasta un 83,73%) correspondiente en SATIRA. Los resultados de deep learning en este ámbito de los tuits generados por un género literario; indican que, los modelos que usan una red neuronal artificial compuesta de un número en niveles jerárquicos nos permiten realizar mejores predicciones de las clases non-satire y satire.

Por consiguiente, las características contextuales que acompañan a este rango de efectividad para predecir en un rango del 45,18% hasta 47,93% con técnicas de deep learning sobre el corpus SATIRA son: la información obtenida está en el ámbito del lenguaje natural;

específicamente, en el género literario de la sátira con los siguientes parámetros que definen el texto en los tuits: el idioma que fueron escritos, las características que identifican al texto como satirico o no satírico y el dominio donde se indica entidad que genero el texto.

Tabla 25 Resultados de las evaluaciones en modelos de deep learning para el corpus SATIRA

Modelo	Precision	Recall	F1	Acc
CNN	82,5280949694	82,3834196891	82,4048095280	82,38341969
LSTM	81,1290127575	80,8290155440	80,6681958675	80,82901554
GRU	81,3366761696	81,1398963731	81,0154084626	81,13989637
BI-GRU	82,0589790870	82,0725388601	82,0585818715	82,07253886
BILSTM	82,2679787152	82,2797927461	82,2588545126	82,27979275
DENSE	82,9596467297	82,7979274611	82,7031099371	82,79792746

En la Tabla 26, realizamos la comparación de los dos mejores modelos de deep learning ampliando en cada clase objetivo (non-satire y satire) y las métricas de evaluación como las medidas de macro avg (macro promedio) y weighted avg (promedio ponderado).

Primero hagamos un acercamiento al modelo DENSE que en la comparación tiene una menor exactitud que el modelo de CNN. El modelo DENSE también llamada capa completamente conectada, se refiere a la capa cuyas neuronas internas se conectan a cada neurona en la capa anterior; Es decir, es una estructura o topología de red que la arquitectura del modelo toma información de las capas anteriores y luego pasa información a la siguiente capa. Este modelo consiguió una exactitud del 82,79% (82,79792746).

Ahora hagamos el acercamiento al modelo CNN que en la comparación tiene una menor exactitud que el modelo de DENSE. El modelo CNN es una red convolucional que procesa sus capas y transforma los datos de entrada utilizando la operación matemática de convolución; así, permite abordar problemas asociados con tareas de procesamiento del lenguaje natural como clasificación de oraciones, clasificación de texto, análisis de sentimientos, resumen de texto, traducción automática y relaciones de respuesta en tuits con texto satíricos o no satíricos. Este algoritmo alcanzo una exactitud sobre el corpus SATIRA de 82,38% (82,38341969)

Tabla 26 Comparación de los dos mejores modelos de deep learning en el corpus SATIRA

	DENSE Acc 82,79792746				CNN Acc 82,38341969			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
non_satire	81,316726	88,223938	84,62963	518	85,222672	81,274131	83,201581	518
satire	84,863524	76,510067	80,47059	447	79,405520	83,668904	81,481481	447
macro avg	83,090125	82,367003	82,550109	965	82,314096	82,471518	82,341531	965
weighted avg	82,959647	88,797928	84,703110	965	82,528095	82,383420	82,404810	965

En la **Figura 80** Matriz de confusión de los dos mejores modelos de deep learning en el corpus SATIRA presentamos la matriz de confusión comparando los dos mejores resultados que han arrojado las evaluaciones; en primer lugar, el modelo DENSE (**Acc 82,79792746**) y en segundo lugar el modelo CNN (**Acc 82,38341969**). Los verdaderos negativos, es decir tuit que hay tenido un análisis de sentimiento de clase negative. El modelo DENSE indica que los TN (verdaderos non-satire) son un 88%, pero que los falsos non-satire (FN) llegan a un 23%; es decir tenemos una tasa de error importante. En el modelo CNN los verdaderos non-satire llegan a un 81%, pero que los falsos non-satire (FN) llegan a un 16%, esta variable de error es mucho menor al primer modelo; lo que implica que tratar de alcanzar más predicciones hace que la variable de error disminuye entre un modelo y otro para la clase non-satire.

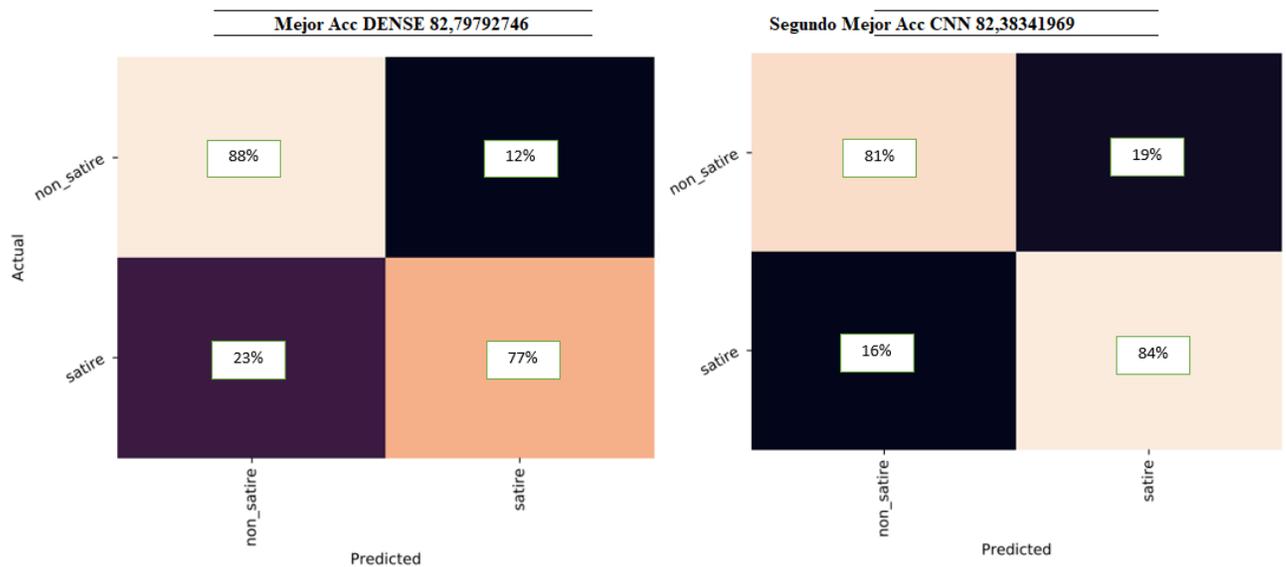


Figura 80 Matriz de confusión de los dos mejores modelos de deep learning en el corpus SATIRA

4.5.3 Discusión de los resultados del corpus SATIRA.

En la Tabla 27 vamos a revisar los resultados en conjunto de los dos mejores resultados aplicando técnicas tradicionales de machine learning y deep learning sobre el corpus SATIRA. Los resultados tienen un grado de exactitud que va desde 80,82% (80,82901554) hasta el 83,73% (83,73056995) y los modelos que mejores resultados dieron son los de machine learning en los algoritmos SVM y MNB, con la combinación de n-gramas de palabras (1-2), palabras (1-3) y char-gramas caracteres (4-9) respectivamente. Las técnicas de deep learning son las siguientes mejores en resultado con los modelos DENSE y LSTM; pero este último anticipado por la exactitud que tiene el modelo LR con la combinación de char-gramas de caracteres (4-5) que corresponde a machine learning.

Tabla 27 Mejores accuracy de los modelos en machine learning y Deep learning en el corpus SATIRA

Clasificador	Conjunto	Accuracy
SVM	Palabras 1-2	83,73056995
MNB	Palabras 1-3	83,52331606
SVM	Caracteres 4-9	83,31606218
DENSE	FastText	82,79792746
LR	Caracteres 4-5	81,65803109
LSTM	FastText	80,82901554

Ahora realicemos una comparación de los resultados de aplicar tecinas de machine y deep learning en el corpus SATIRA versus los resultados que se dieron en la evaluación en los corpus del ZIKA y COVID, detallados en las secciones 4.3.3 y 4.4.3 respectivamente. Las técnicas aplicadas en el corpus ZIKA tiene un grado de exactitud que va desde 59,79% hasta un 61,20%, el corpus de COVID tiene un grado de exactitud que va desde un 43,87% hasta un 47,93%; para terminar, indicando que el corpus SATIRA tiene un grado de exactitud mucho mayor desde 80,82% hasta un 83,73%. Aquí primero tenemos que tener en cuenta que la clasificación en los dos primeros corpus es ternaria y en el último corpus es binaria por lo que es normal que se obtengan muchos mejores resultados para la clasificación de la sátira.

Sigamos profundizando la comparación de los tres corpus y en la sección 3.6, indica que el corpus del ZIKA tiene una cantidad de 39.816 en tuits correspondientes para tres clases y la exactitud más alta conseguida es de 61,20%, el corpus COVID contiene 188.618 tuits con tres clases y una exactitud más alta de 47,93%; para terminar, indicando que el corpus SATIRA una cantidad de 9.688 tuit en dos clases y una exactitud más alta de 83,73%. En la Tabla 28, podemos ver un resumen de la comparación de resultados de los tres corpus.

Si utilizamos el 100% de la masa de tuits y teniendo como objetivo de predicción tres clases; ocasionara que nuestros modelos tengan un máximo de exactitud del 47,93%. Esto implica que, los modelos tienen un 100% de masa como presión, ocasionando mayor variabilidad en los cálculos y una pérdida de 42,76% en la exactitud.

Si solo utilizamos el 21% de la masa de tuits y teniendo como objetivo de predicción tres clases; ocasionara que nuestros modelos tengan un máximo de exactitud del 61,20%. Esto implica que, los modelos tienen un 79% menos de masa como presión, ocasionando más variabilidad en los cálculos y una pérdida de 26,91% en la exactitud.

Si solo utilizamos el 5% de la masa de tuits y teniendo como objetivo de predicción dos clases; ocasionara que nuestros modelos tengan un máximo de exactitud del 83,73%. Esto implica que, los modelos tienen un 95% menos de masa como presión, ocasionando menos variabilidad en los cálculos.

Tabla 28 Comparación de los resultados en la predicción para los corpus ZIKA, COVID y SATIRA

Corpus	Clases	Masa	% Masa	% Masa menos	Max. Acc	% Acc	% Acc Menos
COVID	3	188618	100%	0%	47,93%	57,24%	42,76%
ZIKA	3	39816	21%	79%	61,20%	73,09%	26,91%
SATIRA	2	9688	5%	95%	83,73%	100%	0%

Ahora revisemos la cantidad de caracteres que contienen los tuits de los tres corpus. En la **Figura 24** que tiene la información de la cantidad de texto en el corpus ZIKA (concentración de texto de 45 a 145 caracteres), en la **Figura 42** detalla la información de la cantidad de texto en el corpus COVID (concentración de texto de 45 a 280 caracteres), y en la **Figura 59** la información detallada del corpus SATIRA (concentración de texto de 45 a 110 caracteres). Por lo cual, concluimos que sobre el corpus COVID los usuarios de la red social escriben más caracteres para poder expresar el impacto que ha tenido la pandemia sobre las actividades diarias y su efecto en la salud acompañado de una alta tasa de mortalidad. En el corpus ZIKA se presenta menos cantidad de caracteres que por consiguiente es información relacionada sobre salud, pero su tasa de mortalidad se ha aplanado o contenida y terminamos por relacionar el corpus SATIRA que tiene la menor cantidad de concentración de caracteres pero que la información no relacionada a la salud o que haya implicado algún índice de

mortalidad y su aspecto del lenguaje natural se ha enfocado en expresiones lingüísticas y psicológicas.

Terminemos de comparar los mejores modelos en las técnicas tradicionales de machine learning y los mejores modelos en las técnicas de deep learning para los corpus ZIKA, COVID y SATIRA. En la Tabla 29 se puede concluir que, para el corpus SATIRA que contiene lenguaje natural con un contexto de tendencia psicológica y lingüística el mejor de modelo de predicción es SVM que corresponde a las técnicas de machine learning y de las técnicas de deep learning con el modelo DENSE; enmarcado en una cantidad de tuits de menos de 10 mil cargados en el corpus. Para el corpus ZIKA cuyo contexto es la salud con una tasa de mortalidad controlada; el mejor modelo a utiliza es SVM que corresponde a la técnica de machine learning con el conjunto de palabras (1-2) que correlacionan a los unigramas + bigramas y el mejor modelo de las tecnicas de deep learning es el modelo DENSE; enmarcado en una cantidad de tuits no mayor a 40 mil cargados en el corpus. Para Finalizar en el corpus COVID cuyo contexto es la salud con una tasa de mortalidad aún en proceso de controlar; el mejor modelo a utilizar es BILSTM que corresponde a las tecnia de deep learning y mejor modelo de las técnicas de machine learning es RF con el conjunto de char-gramas de caracteres (4-4). Esto enmarcado en una cantidad de hasta 180 mil tuits.

Tabla 29 Comparación de los mejores modelos de machine y deep learning para los corpus ZIKA, COVID y SATIRA

Corpus	Learning	Clasificador	Conjunto	Accuracy
SATIRA	Machine	SVM	Palabras 1-2	83,7305700
	Deep	DENSE	FastText	82,7979275
ZIKA	Machine	SVM	Palabras 1-3	61,2058465
	Deep	DENSE	FastText	60,6577345
COVID	Deep	BILSTM	FastText	47,9362306
	Machine	RF	Caracteres 4-4	47,0845163

5 Capítulo 5. Conclusiones y trabajos futuros.

5.1 Introducción

En este capítulo se detallan las conclusiones que se han generado de las actividades llevadas en esta tesis doctoral. Puntualizamos el trabajo de investigación que se enfocó en la inteligencia artificial y su ámbito en el procesamiento del lenguaje natural. También analizamos con las herramientas del aprendizaje automático y aprendizaje profundo, la validación de los dominios de ZIKA, COVID y SATIRA. Para concluir en la evaluación de las métricas resultantes; además, se proponen investigaciones futuras.

5.2 Conclusiones

La información que se genera en el internet es valiosa y voluminosa, pero requiere de un procesamiento previo o de validación para que se convierta en una información de calidad. Tomar los datos sin una verificación previa podría ocasionar tener decisiones no acertadas, estrategias equivocadas y planificaciones erróneas. Por lo cual, ejecutamos este trabajo doctoral para aportar con el rendimiento que tienen las ciencias computacionales en el ámbito del procesamiento del lenguaje natural.

En esta tesis se propone distintas técnicas para el análisis de emociones en tuits relacionados con las enfermedades infecciosas. Estas técnicas se basan en machine learning tradicional y tecnologías de word embeddings en Deep learning. Estas técnicas se valoran en 3 dominios distintos: ZIKA, COVID y SÁTIRA. En los dos primeros corpus se evaluó una clasificación de tuits con tres etiquetas distintas (positivo, negativo y neutro) y en el último corpus de la sátira se evaluó con una clasificación binaria (satírico y no satírico).

Para el desarrollo de esta tesis se ha estudiado en detalle el estado del arte en cada uno de los conceptos y tecnologías interviniente. Concretamente en los fundamentos de la inteligencia artificial que permite diseñar operaciones que se consideran propias de la inteligencia humana, el procesamiento del lenguaje natural que es el campo de conocimiento específico que permitió investigar la manera de máquinas con las personas mediante el uso de lenguajes naturales, las técnicas de aprendizaje automático o machine learning, que detallan los

algoritmos matemáticos para analizar datos, aprender de estos y tomar decisiones basadas en lo aprendido y las técnicas de aprendizaje profundo o deep learning, que estructuran algoritmo de capas para crear una red neuronal artificial, que puede aprender y tomar decisiones. Para terminar de estudiar las técnicas en análisis del lenguaje subjetivo y ampliando los conceptos de análisis de sentimientos como los modelos populares de la clasificación de textos.

El desarrollo de la tesis requirió construir dos corpus de tuits relacionados con las enfermedades infecciosas en español. Para ello, se recopiló un conjunto de tuits de las enfermedades relacionadas con el zika, dengue, chikungunya y malaria. Todo este conjunto de tuits fue etiquetado formando el corpus ZIKA que representa una de las primeras contribuciones de esta tesis ya que la escasez de recursos lingüísticos y corpus anotados en español hace que este tipo de corpus sea muy necesario actualmente.

El segundo corpus llamado COVID se basó en la enfermedad del COVID-19 que se refiere a la última pandemia que el mundo está viviendo; ambos corpus fueron enfocados en lo que se describe como la ciencia de la infodemiología. La compilación y etiquetado manual de este corpus es la segunda contribución importante en esta tesis por las mismas razones expuestas anteriormente.

Por último, para validar las mismas tecnologías en otro ámbito de clasificación de texto se utilizó el corpus de la SATIRA que permite identificar si un tuit es satírico o no.

Los dos primeros corpus ZIKA y COVID fueron etiquetados por un grupo de personas que validada cada tuit y lo clasificaba en positivo, negativo y neutro a través de una aplicación web llamada UMUCorpusClassifier. Esta actividad se coordinó con estudiantes de universidades que ingresaban a la aplicación por un periodo determinado de tiempo, para solo hacer una cantidad específica por día y por horas. Esto permitió que la minería de opiniones no tenga un sesgo equivocado producto del cansancio diario de las personas como también facilitar la incorporación de estas tareas a la investigación.

El siguiente paso fue el de crear distintos programas de computacionales por medio de un entorno integral de desarrollo que pueda ser utilizado a través de los navegadores web y sin dependencia de ningún proveedor específico. Los primeros programas permitieron hacer la

carga de los tuits accediendo a Twitter por medio sus las APIs de acceso y depositando el resultado obtenido en base de datos y archivos. Los siguientes programas permitieron describir la información en forma resumida de los tuits y que aportaron información de composición y comparación entre un corpus u otro.

El siguiente paso fue crear programas computacionales para la ejecución y visualización de las técnicas de aprendizaje automático o machine learning y programas para las del aprendizaje profundo que se han descrito a lo largo de la tesis doctoral. La ejecución de estos programas requirió de un alto consumo en la capacidad de procesamiento y tiempo de espera hasta que los programas terminasen de ejecutar los algoritmos que permitía tener datos resultantes como grafico de barras, grafico de tendencia, indicadores de evaluación, matrices de correlación como son la de confusión y detallando las métricas para evaluar cada modelo.

Esto permitió tener la información necesaria para crear primero las tablas que contienen los resultados de las ejecuciones en los modelos de técnicas tradicionales como machine learning ampliando el estudio a combinación de conjuntos como n-gramas o char-gamas que nos permiten tener varias alternativas que permitan ver el rendimiento de cada algoritmo. también Incorporamos las técnicas del aprendizaje profundo o deep learnng. Así pudimos ver los resultados que son mejores que otros y hacer comparaciones entre los dos mejores modelos que tuvieron la exactitud más alta dentro de cada corpus. También ampliamos las comparaciones entre los dos mejores modelos cada machine learning para n-gramas, los dos mejores modelos para char-gramas y los dos mejores modelos de deep learning.

El resultado obtenido de los experimentos permitió ver el grado de eficiencia que tienen los modelos sobre la información contextual de los corpus en este estudio. Hacer una comparación de las características orgánicas que componen el corpus, para luego hacer aseveraciones de porque este u el otro modelo funcionan mejor. La estrategia de validación consistió crear combinaciones en n-gramas como son unigramas que es la representación de una sola palabra, los bigramas que corresponden a dos palabras y trigramas que indican tres palabras. También se diseñaron conjuntos de porción de caracteres sobre el texto, que van en grupo de 1 a 4 caracteres y de ahí pasamos en combinaciones de (4, 5), (4, 6), (4, 7), (4, 8), (4, 9) y de (4, 10). Lo anterior se pudo aplicar en las técnicas de machine learning. Por otro

lado, las técnicas de aprendizaje profundo tomaron cada corpus por completo para evaluar el resultado de cada algoritmo creando redes neuronales sobre texto.

Ahora veamos el resultado en las evaluaciones sobre el corpus ZIKA arrojó que las técnicas de aprendizaje automático con el modelo SVM y combinación de palabras (1-3) unigramas + bigramas + trigramas dio la mejor exactitud con el 61,20% para la eficiencia en la predicción. La menor exactitud lo produjo el modelo K con la combinación de palabras (3-3) trigramas con un valor en su ejecución de 44,33%. Esto implica que, hacer arboles de decisiones con el modelo K minimiza la efectividad en la predicción y que por lo contrario hacer predicción con el modelo SVM a través de un hiperplano nos dará el mejor resultado. Por lo cual, concluimos que para un corpus que contiene información de salud, que los efectos en la sociedad ya han sido controlados y con tres objetivos de predicción.

Por otra parte, del resultado en las evaluaciones sobre el corpus COVID arrojó que las técnicas de aprendizaje profundo con el modelo BILSTM dio la mejor exactitud con el 47,93% para la eficiencia en la predicción. La menor exactitud lo produjo el modelo K con la combinación de palabras (3-3) trigramas con un valor en su ejecución de 33,74%. Esto implica que, hacer arboles de decisiones con el modelo K minimiza la efectividad en la predicción y que por lo contrario hacer predicción con el modelo BILSTM a través de una red neuronal en ambos sentidos nos dará el mejor resultado. Por lo cual, concluimos que para un corpus que contiene información de salud, que los efectos en la sociedad no han sido controlados y con tres objetivos de predicción.

Por consiguiente, resultado en las evaluaciones sobre el corpus SATIRA arrojó que las técnicas de aprendizaje automático con el modelo SVM y el conjunto palabras (1-2) unigramas + bigramas dio la mejor exactitud con el 83,73% para la eficiencia en la predicción. La menor exactitud lo produjo el modelo K con la combinación de palabras (3-3) trigramas con un valor en su ejecución de 55,95%. Esto implica que, hacer arboles de decisiones con el modelo K minimiza la efectividad en la predicción y que por lo contrario hacer predicción con el modelo SVM a través de un hiperplano nos dará el mejor resultado. Por lo cual, concluimos que para un corpus que contiene información de lingüística, psicológica y con dos objetivos de predicción.

Las evaluaciones dejaron ver que el resultado obtenido sobre el corpus SATIRA arrojó los mejores resultados en las predicciones y el contexto que permite alcanzar esta eficiencia al ser una tarea más sencilla de clasificación binaria. El siguiente mejor resultado lo obtuvo el corpus del ZIKA con un grado de eficiencia menor, pero pudiendo describir las razones técnicas de peso del porque produce este efecto en el ámbito de la salud. Y para finalizar la eficiencia obtenida en el corpus del COVID fue la menor de entre los tres corpus, pero permitiendo ver que es una información muy dinámica y que aun la sociedad no termina de superar.

En los modelos de machine learning que se evaluaron, el algoritmo SVM con n-gramas tuvo el mejor resultado en los corpus de SATIRA y ZIKA. Para el corpus COVID el mejor resultado lo obtuvo en el modelo RF usando char-gramas. Se concluye que para los dos primeros corpus existen palabras que ya definen una clase objetivo, permitiendo al modelo hacer mejores predicciones y para el corpus COVID hacer un análisis e los primeros caracteres en un boque de árboles de decisiones nos dan una alternativa válida para esta información que aún tiene un grado alto de variabilidad

En los modelos de deep learning que se evaluaron, el algoritmo DENSE tuvo el mejor resultado en los corpus de SATIRA y ZIKA. Para el corpus COVID el mejor resultado lo obtuvo en el modelo BILSTM. Se concluye que para los dos primeros corpus tener una red neuronal completamente conectada permite al modelo hacer mejores predicciones y para el corpus COVID hacer un análisis bidireccional nos dan una alternativa válida para esta información que aún tiene un grado alto de variabilidad

Este estudio de distintos algoritmos de machine y Deep learning ha permitido poder analizar el comportamiento de estas técnicas para tanto el análisis de sentimientos en el dominio de la salud como para la detección de sátira. Aunque hubiese sido deseable probar otras tecnologías para la clasificación de texto no fue posible y en el apartado siguiente se presentan algunas vías futuras que surgen a partir de esta tesis doctoral.

5.3 Trabajos futuros

En este trabajo hemos presentado una evaluación de las características del recuento de términos y las word embeddings previamente entrenados para la clasificación automática

de tuits. Esta evaluación involucró varias incorporaciones de palabras previamente entrenadas, diferentes arquitecturas de aprendizaje profundo y varias características de conteo de términos. Esta evaluación se realizó sobre tres conjuntos de datos compuestos por tweets extraídos de Twitter de dos variantes: salud y sátira. Sin embargo, estos resultados deben considerarse con cautela porque encontramos un sesgo importante con respecto al conjunto de datos salud y sátira. Es más, para superar las dificultades e inconvenientes encontrados en este trabajo, sugerimos tres direcciones de investigación:

Mejora de la calidad de los conjuntos de datos

La primera limitación del trabajo está en la calidad de los conjuntos de datos. Se puede observar en los resultados que el corpus del ZIKA da muchos mejores resultados que el corpus del COVID. Esto hace presuponer que el etiquetado manual de este tipo de tuits relacionados con la salud no es trivial y es necesario realizar un etiquetado manual cuidadoso para tener un conjunto de datos de calidad. Además, el número reducido de cuentas de Twitter diferente sesgó los resultados entre la sátira y la salud ocasionando un efecto directo en los modelos. Para superar estos inconvenientes, proponemos compilar más tweets de diferentes países de América Latina y anotarlos manualmente por varios anotadores para filtrar esos tweets.

Otra limitación que se nos presenta en el trabajo está en la cantidad de clases objetivos a predecir. Se puede crear subconjuntos de tuits con clases binarias de objetivos como son positivo y negativo solamente. Esto permitirá evaluar los nuevos resultados para los corpus ZIKA y COVID. A partir de, esta presunción las personas desean conocer en el ámbito de la salud cosas que afecten de forma positiva o negativa de lo que contiene la información. Así las personas tendrán una base de conocimiento que emita información validada y de calidad para no caer en el sesgo neutral que no infiere objetivo alguno para la predicción.

En los modelos también descubrimos que la información está tomada en un tiempo determinado. Donde, en ese rango de fechas la información que se genera por los corpus ZIKA y COVID tienen entornos diferentes, con aspectos que pueden cambiar de un periodo a otro. Para superar estos inconvenientes, proponemos compilar más tweets con hitos que reflejen un antes, durante y después, de cada una de las enfermedades que este estudio cubrió. Con esto podemos saber el efecto que tiene la vigencia de la información en el uso de los modelos de aprendizaje automático y aprendizaje profundo.

Análisis de emociones

La dificultad de etiquetar los tuits manualmente en positivo, negativo y neutro en el dominio de la salud es elevada y sería interesante ampliar este estudio a un análisis de emociones basado en los seis tipos de emociones definidos por Ekman (Matsumoto & Ekman, 1989). El estudio fue realizado en 1987 proporcionando evidencia de desacuerdo cultural sobre las calificaciones en la intensidad de las expresiones faciales. El estudio examinó la base de estas diferencias culturales. Los sujetos japoneses y estadounidenses hicieron dos clasificaciones de intensidad separadas de poses japoneses y caucasianos que retratan ira, disgusto, miedo, felicidad, tristeza y sorpresa. Los estadounidenses tenían índices de intensidad media más altos que los japoneses para todas las emociones excepto el disgusto, independientemente de la cultura o el género. Los estadounidenses dieron a las fotos felices y enojadas las calificaciones más altas de intensidad, mientras que los japoneses dieron a las fotos de disgusto las calificaciones más altas. Pero hubo una consistencia intercultural considerable en las diferencias relativas a las fotos

Para ello sería necesario etiquetar de nuevo los dos corpus detectando este tipo de emociones. Como el siguiente trabajo realizado (c, Martín-Valdivia, Ureña-López, & Mitkov, 2020, pp. 114120) que indican que han experimentado con la integración de emociones en modelos computacionales que puede mejorar los sistemas de interacción humano-ordenador. Logrando una precisión aceptable sobre los modelos de aprendizaje automático.

Intrerepretabilidad de modelos

En tercer lugar, una línea de investigación prometedora es mejorar la interpretabilidad de los modelos. En este sentido, en trabajos anteriores ((García-Díaz et al., 2020, pp. 641-657), (García-Díaz et al., 2021, pp. 506-518)), hemos evaluado la combinación de word embeddings con características lingüísticas mediante la concatenación de dos arquitecturas de aprendizaje automático, una para la capa de inserción y otra para las características lingüísticas que lograron buenos resultados. Específicamente aplicaron tecnologías de análisis de sentimiento y computación social para la detección de mensajes misóginos en Twitter. También, recopilaron un corpus equilibrado sobre la misoginia en español, y lo han clasificado en tres subconjuntos relacionados con violencia hacia las mujeres, mensajes

de acoso a mujeres y rasgos generales relacionados con la misoginia. La investigación combina una clasificación basada en Word-embedding y características lingüísticas para entender los fenómenos lingüísticos relacionados.

Arquitecturas basadas en Transformers

Finalmente, consideramos que es interesante explorar nuevas arquitecturas de aprendizaje profundo para combinar arquitecturas de aprendizaje profundo para cada conjunto de características. Por ejemplo, algunos de los modelos basados en transformers más prometedores como BERT y ELMO deberían evaluarse en estos dominios porque han logrado muy buenos resultados en otras tareas de clasificación de texto como la detección de discursos de odio (Plaza-del-Arco, Molina-González, Ureña-López, & Martín-Valdivia, 2021, pp. 114120). ELMO (Embeddings from Language Model) es un modelo un word-embedding para representar una secuencia de palabras en una secuencia correspondiente de vectores. Los tokens a nivel de caracteres se toman como entradas a un LSTM bidireccional que produce Word-level embedding. .ELMo Embeddings son sensibles al contexto, produciendo diferentes representaciones para palabras que comparten la misma ortografía, pero tienen diferentes significados. En cambio, el modelo BERT (Bidirectional Encoder Representations from Transformers), es un algoritmo basado en aprendizaje profundo que fue desarrollado por Google para comprender mejor las búsquedas que realizan los usuarios en su navegador web. Este modelo juzga las palabras que están ubicadas a la izquierda y a la derecha de cada sentencia

6 Capítulo 6. Contribuciones científicas derivadas de la tesis doctoral

6.1 Publicaciones en revistas

1. Apolinario-Arzube O., García-Díaz J.A., Medina-Moreira J., Luna-Aveiga, H. & Valencia-García R. (2019) Evaluating Information-Retrieval Models and Machine-Learning Classifiers for Measuring the Social Perception towards Infectious Diseases. *Appl. Sci.* 2019, 9(14), 2858; <https://doi.org/10.3390/app9142858> (factor de impacto 2020: 2,678. Categoría: Engineering, Multidisciplinary, Q2)
2. Apolinario-Arzube O., García-Díaz J.A., Medina-Moreira J., Luna-Aveiga, H. & Valencia-García R. (2020) Comparing Deep-Learning Architectures and Traditional Machine-Learning Approaches for Satire Identification in Spanish Tweets. *Mathematics* 2020, 8(11), 2075; <https://doi.org/10.3390/math8112075>(factor de impacto 2020: 1,747. Categoría: Mathematics, Q1)
3. Apolinario, Ó., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., García-Díaz, J. A., & Valencia-García, R. (2018). Tecnologías inteligentes para la autogestión de la salud. *Procesamiento del Lenguaje Natural*, 61, 159-162.
4. Apolinario, O., Medina-Moreira, J., Luna-Aveiga, H., García-Díaz, J.A., Valencia-García, R., Estrade-Cabrera, J.I. (2019) Prevención de enfermedades infecciosas basada en el análisis inteligente en RRSS y participación ciudadana. *Procesamiento del Lenguaje Natural*, 63, 163-166.

6.2 Publicaciones en Congresos

1. Luna-Aveiga, H., Medina-Moreira, J., Lagos-Ortiz, K., Apolinario, O., Paredes-Valverde, M.A., del Pilar Salas-Zárate, M. & Valencia-García, R. (2017). Sentiment Polarity Detection in Social Networks: An Approach for Asthma Disease Management. In Proceedings of ICCSAMA 2017 (AISC, volume 629 pp 141-152)
2. García-Díaz, J. A., Apolinario-Arzube, Ó., Medina-Moreira, J., Luna-Aveiga, H., Lagos-Ortiz, K., & Valencia-García, R. (2018). Sentiment Analysis on Tweets related to infectious diseases in South America. In Proceedings of the Euro American Conference on Telematics and Information Systems (pp. 1-5).
3. Apolinario-Arzube O., García-Díaz J.A., Luna-Aveiga, H., Medina-Moreira J. & Valencia-García R. (2020) Knowledge Extraction from Twitter Towards Infectious Diseases in Spanish. In Sixth International Conference on Technologies and Innovation (pp. 43-57). Guayaquil. Ecuador
4. Apolinario-Arzube O., García-Díaz J.A., Pinto S., Luna-Aveiga, H., Medina-Moreira J., Gomez-Berbís J.M. & Valencia-García R. (2020) CollaborativeHealth: Smart Technologies to Surveil Outbreaks of Infectious Diseases Through Direct and Indirect Citizen Participation. In 9th CSOC (177-190)

Referencias

- Abdi, A., Shamsuddin, S. M., Hasan, S., & Piran, J. (2018). Machine learning-based multi-documents sentiment-oriented summarization using linguistic treatment. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2018.05.010>
- Adams, O., Makarucha, A., Neubig, G., Bird, S., & Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*. <https://doi.org/10.18653/v1/e17-1088>
- Agrawal, A., Gans, J. S., & Goldfarb, A. (2017). What to expect from artificial intelligence. *MIT Sloan Management Review*.
- Ali, N., Neagu, D., & Trundle, P. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences*. Ali N, Neagu D, Trundle P (2019) Evaluation of k-Nearest Neighbour Classifier Performance for Heterogeneous Data Sets. *SN Appl Sci*. Doi: 10.1007/s42452-019-1356-9. <https://doi.org/10.1007/s42452-019-1356-9>
- Aljawarneh, S., Yassein, M. B., & Aljundi, M. (2019). An enhanced J48 classification algorithm for the anomaly intrusion detection systems. *Cluster Computing*. <https://doi.org/10.1007/s10586-017-1109-8>
- Andrew Burgess. Palgrave Macmillan. (2018). The Executive Guide to Artificial Intelligence: How to identify and implement applications for AI in your organization. Chapter 1. In *The Executive Guide to Artificial Intelligence: How to identify and implement applications for AI in your organization*.
- Angarita, R., Lefèvre, B., Ahvar, S., Ahvar, E., Georgantas, N., & Issarny, V. (2019). Universal social network bus: Toward the federation of heterogeneous online social network services. *ACM Transactions on Internet Technology*. <https://doi.org/10.1145/3323333>
- Apolinario-Arzuabe, Ó., García-Díaz, J. A., Luna-Aveiga, H., Medina-Moreira, J., &

- Valencia-García, R. (2020). Knowledge Extraction from Twitter Towards Infectious Diseases in Spanish. In *Communications in Computer and Information Science*. https://doi.org/10.1007/978-3-030-62015-8_4
- Apolinario-Arzuabe, Ó., García-Díaz, J. A., Medina-Moreira, J., Luna-Aveiga, H., & Valencia-García, R. (2020). Comparing deep-learning architectures and traditional machine-learning approaches for satire identification in spanish tweets. *Mathematics*. <https://doi.org/10.3390/math8112075>
- Arbib, M. A. (2018). From cybernetics to brain theory, and more: A memoir. *Cognitive Systems Research*. <https://doi.org/10.1016/j.cogsys.2018.04.001>
- Barbieri, F., Ronzano, F., & Saggion, H. (2015a). Do we criticise (and Laugh) in the same way? Automatic detection of multi-lingual satirical news in twitter. In *IJCAI International Joint Conference on Artificial Intelligence*.
- Barbieri, F., Ronzano, F., & Saggion, H. (2015b). Is this tweet satirical? A computational approach for satire detection in Spanish. *Procesamiento de Lenguaje Natural*.
- Barkalov, A., Titarenko, L., & Mazurkiewicz, M. (2019). Programmable Logic Controllers. In *Studies in Systems, Decision and Control*. https://doi.org/10.1007/978-3-030-11961-4_6
- Barker, J., Marxer, R., Vincent, E., & Watanabe, S. (2016). The third “CHiME” speech separation and recognition challenge: Dataset, task and baselines. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015 - Proceedings*. <https://doi.org/10.1109/ASRU.2015.7404837>
- Bem, D. J. (2017). An experimental analysis of self-persuasion. In *Attitude Change: The Competing Views*. <https://doi.org/10.4324/9781351315364-7>
- Bhandare, A., Bhide, M., Gokhale, P., & Chandavarkar, R. (2016). Applications of Convolutional Neural Networks. *International Journal of Computer Science and Information Technologies*.
- Bunker, A., Magarkar, A., & Viitala, T. (2016). Rational design of liposomal drug delivery systems, a review: Combined experimental and computational studies of lipid

- membranes, liposomes and their PEGylation. *Biochimica et Biophysica Acta - Biomembranes*. <https://doi.org/10.1016/j.bbamem.2016.02.025>
- Bury, R. (2017). Technology, fandom and community in the second media age. *Convergence*. <https://doi.org/10.1177/1354856516648084>
- Buzzelli, A., Holdan, E. G., Rota, D., & McCarthy, J. (2016). Utilizing twitter for concept learning. *International Journal of Information and Communication Technology Education*. <https://doi.org/10.4018/IJICTE.2016010106>
- Cabanelas Omil, J. (2019). Inteligencia artificial ¿Dr. Jekyll o Mr. Hyde? *Mercados y Negocios*.
- Cahan, D. (2015). Helmholtz, Hermann Ludwig Ferdinand von (1821-94). In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. <https://doi.org/10.1016/B978-0-08-097086-8.61049-8>
- Cao, K., & Rei, M. (2016). A Joint Model for Word Embedding and Word Morphology. <https://doi.org/10.18653/v1/w16-1603>
- Carbonell, J. G., Cullingford, R. E., & Gershman, A. V. (1981). Steps Toward Knowledge-Based Machine Translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. <https://doi.org/10.1109/TPAMI.1981.4767124>
- Castro, D., & Benavides, L. (2019). Uo-cerpamid at IroSvA: Impostor method adaptation for irony detection IroSvA@IberLEF 2019. In *CEUR Workshop Proceedings*.
- Chandankhede, C., Devle, P., Waskar, A., Chopdekar, N., & Patil, S. (2017). ISAR: Implicit sentiment analysis of user reviews. In *International Conference on Computing, Analytics and Security Trends, CAST 2016*. <https://doi.org/10.1109/CAST.2016.7914994>
- Chen, F., Ji, R., Su, J., Cao, D., & Gao, Y. (2018). Predicting Microblog Sentiments via Weakly Supervised Multimodal Deep Learning. *IEEE Transactions on Multimedia*. <https://doi.org/10.1109/TMM.2017.2757769>
- Cignarella, A. T., & Bosco, C. (2019). ATC at IroSva 2019: Shallow syntactic dependency-

- based features for irony detection in Spanish variants. In *CEUR Workshop Proceedings*.
- Cowan, N. (2015). George Miller's magical number of immediate memory in retrospect: Observations on the faltering progression of science. *Psychological Review*. <https://doi.org/10.1037/a0039035>
- Dascalu, M., Gutu, G., Ruseti, S., Paraschiv, I. C., Dessus, P., McNamara, D. S., ... Trausan-Matu, S. (2017). ReaderBench: A multi-lingual framework for analyzing text complexity. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-66610-5_48
- De Asis, K., Fernando Hernandez-Garcia, J., Zacharias Holland, G., & Sutton, R. S. (2018). Multi-step reinforcement learning: A unifying algorithm. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*.
- Deloitte. (2018). Artificial Intelligence Innovation Report. *Innovation Report 2018 Deloitte*.
- Deon, D. J., & De Freitas, L. A. (2019). UFPelRules to irony detection in Spanish variants. In *CEUR Workshop Proceedings*.
- Dev, S., Li, T., Phillips, J. M., & Srikumar, V. (2020). On Measuring and Mitigating Biased Inferences of Word Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*. <https://doi.org/10.1609/aaai.v34i05.6267>
- Dragoni, M., & Petrucci, G. (2017). A Neural Word Embeddings Approach for Multi-Domain Sentiment Analysis. *IEEE Transactions on Affective Computing*. <https://doi.org/10.1109/TAFFC.2017.2717879>
- Dudău, D. P., & Sava, F. A. (2020). The development and validation of the Romanian version of Linguistic Inquiry and Word Count 2015 (Ro-LIWC2015). *Current Psychology*. <https://doi.org/10.1007/s12144-020-00872-4>
- Duwairi, R. M., Ahmed, N. A., & Al-Rifai, S. Y. (2015). Detecting sentiment embedded in Arabic social media - A lexicon-based approach. *Journal of Intelligent and Fuzzy Systems*. <https://doi.org/10.3233/IFS-151574>

- Eisenbrand, F. (2011). The euclidean algorithm. In *Algorithms Unplugged*.
https://doi.org/10.1007/978-3-642-15328-0_12
- Eklund, P., Galán, M. A., & Gähler, W. (2009). Partially Ordered Monads for Monadic Topologies, Rough Sets and Kleene Algebras. *Electronic Notes in Theoretical Computer Science*. <https://doi.org/10.1016/j.entcs.2008.12.067>
- Ertel, W. (2017). *Introduction to Artificial Intelligence (Undergraduate Topics in Computer Science)*. Book.
- Esteves, M., Ganz, E., Sousa, N., & Leite-Almeida, H. (2020). Asymmetrical Brain Plasticity: Physiology and Pathology. *Neuroscience*.
<https://doi.org/10.1016/j.neuroscience.2020.01.022>
- Esuli, A., & Sebastiani, F. (2015). SentiWordNet: A High-Coverage Lexical Resource. *LRE.Tex*.
- Evren, B., Evren, C., Dalbudak, E., Topcu, M., & Kutlu, N. (2018). Relationship of internet addiction severity with probable ADHD and difficulties in emotion regulation among young adults. *Psychiatry Research*. <https://doi.org/10.1016/j.psychres.2018.08.112>
- Fischetti, M. A. (2013). A review of progress at MCC: Although its current objectives do not include “beating Japan,” this U.S. organization is positioned to help industry do so. *IEEE Spectrum*. <https://doi.org/10.1109/mspec.1986.6371032>
- Fragaki, H. (2012). Clocks and dials with automata: The mosaic of Qasr El-Lebya. In *History of Mechanism and Machine Science*. https://doi.org/10.1007/978-94-007-4132-4_16
- Franchi, S., & Bianchini, F. (2018). THE ONTOLOGY OF THE ENEMY: NORBERT WIENER AND THE CYBERNETIC VISION. In *The Search for a Theory of Cognition*.
https://doi.org/10.1163/9789401207157_003
- Franzese, M., & Iuliano, A. (2018). Hidden markov models. In *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*.
<https://doi.org/10.1016/B978-0-12-809633-8.20488-3>
- Frenda, S., & Patti, V. (2019). Computational models for irony detection in three Spanish

- variants. In *CEUR Workshop Proceedings*.
- García-Díaz, J. A., Almela, A., Alcaraz-Marmol, G., & Valencia-García, R. (2020). UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for Natural Language Processing tasks. *Procesamiento de Lenguaje Natural*. <https://doi.org/10.26342/2020-65-22>
- García-Díaz, J. A., Apolinario-Arzupe, O., Medina-Moreira, J., Salavarría-Melo, J. O., Lagos-Ortiz, K., Luna-Aveiga, H., & Valencia-García, R. (2018). Opinion mining for measuring the social perception of infectious diseases. an infodemiology approach. In *Communications in Computer and Information Science*. https://doi.org/10.1007/978-3-030-00940-3_17
- García-Díaz, J. A., Cánovas-García, M., Colomo-Palacios, R., & Valencia-García, R. (2021). Detecting misogyny in Spanish tweets. An approach based on linguistics features and word embeddings. *Future Generation Computer Systems*. <https://doi.org/10.1016/j.future.2020.08.032>
- García-Díaz, J. A., Cánovas-García, M., & Valencia-García, R. (2020). Ontology-driven aspect-based sentiment analysis classification: An infodemiological case study regarding infectious diseases in Latin America. *Future Generation Computer Systems*. <https://doi.org/10.1016/j.future.2020.06.019>
- García, L., Moctezuma, D., & Muñiz, V. (2019). A contextualized word representation approach for irony detection. In *CEUR Workshop Proceedings*.
- Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*. <https://doi.org/10.1613/jair.5714>
- Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society. Series A: Statistics in Society*. <https://doi.org/10.1111/rssa.12276>
- Ghosh, R., Ravi, K., & Ravi, V. (2016). A novel deep learning architecture for sentiment classification. In *2016 3rd International Conference on Recent Advances in Information*

- Technology, RAIT 2016*. <https://doi.org/10.1109/RAIT.2016.7507953>
- González, J. A., Hurtado, L. F., & Pla, F. (2019). ELIRF-UPV at IroSvA: Transformer encoders for Spanish irony detection. In *CEUR Workshop Proceedings*.
- Gotthelf, A. (2013). Aristotle's Conception of Final Causality. In *Teleology, First Principles, and Scientific Method in Aristotle's Biology*. <https://doi.org/10.1093/acprof:oso/9780199287956.003.0001>
- Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2019). Learning word vectors for 157 languages. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*.
- Graves, A. (2014). *Neural Turing Machines arXiv : 1410 . 5401v2 [cs . NE] 10 Dec 2014*. Google DeepMind, London, UK.
- Gregersen, A. (2014). Cognition. In *The Routledge Companion to Video Game Studies*. <https://doi.org/10.4324/9780203114261>
- Hafez, G., Ismail, R., & Karam, O. (2017). Temporal sentiment analysis and time tags for opinions. In *2017 IEEE 8th International Conference on Intelligent Computing and Information Systems, ICICIS 2017*. <https://doi.org/10.1109/INTELCIS.2017.8260065>
- Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics*. <https://doi.org/10.1093/qje/qjx045>
- Haro, P. A. de. (2018). INTRODUCCIÓN A LA SÁTIRA COMO PROBLEMA ESTÉTICO GENERAL. *Revista de Estudios Literarios*. https://doi.org/10.14195/2183-847x_7_2
- He, H., & Deng, Y. (2015). The mental lexicon and english vocabulary teaching. *English Language Teaching*. <https://doi.org/10.5539/elt.v8n7p40>
- Hellrich, J., & Hahn, U. (2017). Don't get fooled by word embeddings: better watch their neighborhood. *Digital Humanities 2017—Conference Abstracts of the 2017 Conference of the Alliance of Digital Humanities Organizations (ADHO)*.
- Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*.

<https://doi.org/10.1126/science.aaa8685>

- Huang, Q., Chen, R., Zheng, X., & Dong, Z. (2017). Deep sentiment representation based on CNN and LSTM. In *Proceedings - 2017 International Conference on Green Informatics, ICGI 2017*. <https://doi.org/10.1109/ICGI.2017.45>
- Iranzo-Sánchez, J., & Ruiz-Dolz, R. (2019). Vrain at IroSvA 2019: Exploring classical and transfer learning approaches to short message irony detection. In *CEUR Workshop Proceedings*.
- Kim, Y. (2014). Convolutional neural networks for sentence classification. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. <https://doi.org/10.3115/v1/d14-1181>
- Krasnowska-Kieras, K., & Wróblewska, A. (2020). Empirical linguistic study of sentence embeddings. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. <https://doi.org/10.18653/v1/p19-1573>
- Krawczak, K. (2016). Objectivity, subjectivity and intersubjectivity. *Functions of Language*. <https://doi.org/10.1075/fo1.23.2.02kra>
- Krejzl, P., Hourová, B., & Steinberger, J. (2017). Stance detection in online discussions. *ArXiv*.
- Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., ... Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In *33rd International Conference on Machine Learning, ICML 2016*.
- Larsson, G., Maire, M., & Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-46493-0_35
- Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*. <https://doi.org/10.1038/nature14539>
- Lederberg, J., Sutherland, G. L., Buchanan, B. G., Feigenbaum, E. A., Robertson, A. V.,

- Duffield, A. M., & Djerassi, C. (1969). Applications of Artificial Intelligence for Chemical Inference. I. The Number of Possible Organic Compounds. Acyclic Structures Containing C, H, O, and N. *Journal of the American Chemical Society*. <https://doi.org/10.1021/ja01039a025>
- Lexcellent, C. (2019). Artificial Intelligence. In *SpringerBriefs in Applied Sciences and Technology*. https://doi.org/10.1007/978-3-030-21445-6_2
- Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*.
- Liu, Y., Ji, L., Huang, R., Ming, T., Gao, C., & Zhang, J. (2019). An attention-gated convolutional neural network for sentence classification. *Intelligent Data Analysis*. <https://doi.org/10.3233/IDA-184311>
- Luna-Aveiga, H., Medina-Moreira, J., Lagos-Ortiz, K., Apolinario, O., Paredes-Valverde, M. A., Salas-Zárate, M. D. P., & Valencia-García, R. (2018). Sentiment polarity detection in social networks: An approach for asthma disease management. In *Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-3-319-61911-8_13
- Lysaker, J. T. (2018). *Philosophy, Writing, and the Character of Thought*. *Philosophy, Writing, and the Character of Thought*. <https://doi.org/10.7208/chicago/9780226569734.001.0001>
- Maiese, M. (2016). Affective scaffolds, expressive arts, and cognition. *Frontiers in Psychology*. <https://doi.org/10.3389/fpsyg.2016.00359>
- Matsumoto, D., & Ekman, P. (1989). American-Japanese cultural differences in intensity ratings of facial expressions of emotion. *Motivation and Emotion*. <https://doi.org/10.1007/BF00992959>
- McLachlan, C. J., & Garcia, R. J. (2015). Philosophy in practice? Doctoral struggles with ontology and subjectivity in qualitative interviewing. *Management Learning*. <https://doi.org/10.1177/1350507615574634>
- Mehri, A., & Jamaati, M. (2017). Variation of Zipf's exponent in one hundred live languages: A study of the Holy Bible translations. *Physics Letters, Section A: General, Atomic and Solid State Physics*. <https://doi.org/10.1016/j.physleta.2017.05.061>

- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality arXiv : 1310 . 4546v1 [cs . CL] 16 Oct 2013. *ArXiv Preprint ArXiv:1310.4546*.
- Mikolov, T., Grave, E., Bojanowski, P., Puhersch, C., & Joulin, A. (2017). Advances in pre-training distributed word representations. *ArXiv*.
- Miranda-Belmonte, H. U., & López-Monroy, A. P. (2019). Early fusion of traditional and deep features for irony detection in twitter. In *CEUR Workshop Proceedings*.
- Morales Harley, R. (2017). Risa y argumentación: algunos ejemplos en la literatura latina. *Pensamiento Actual*. <https://doi.org/10.15517/pa.v17i28.29554>
- Munro, E., & Hardie, J. (2019). Why We Should Stop Talking About Objectivity and Subjectivity in Social Work. *British Journal of Social Work*. <https://doi.org/10.1093/bjsw/bcy054>
- Murdock, K. K., Horissian, M., & Crichlow-Ball, C. (2017). Emerging Adults' Text Message Use and Sleep Characteristics: A Multimethod, Naturalistic Study. *Behavioral Sleep Medicine*. <https://doi.org/10.1080/15402002.2015.1120203>
- Nakamura, H., & Kawaguchi, J. (2016). People like logical truth: Testing the intuitive detection of logical value in basic propositions. *PLoS ONE*. <https://doi.org/10.1371/journal.pone.0169166>
- Núñez Cardenas, F. de J., Tomás Mariano, V. T., Crespo Hernandez, J. D., & Andrade Hernández, E. (2017). Reconocimiento facial mediante redes neuronales “hopfield”, “backpropagation” y algoritmo pca: un enfoque comparativo. *Ciencia Huasteca Boletín Científico de La Escuela Superior de Huejutla*. <https://doi.org/10.29057/esh.v5i9.2225>
- Nurmatova, G. (2019). Corpus-based Study of Engineering Terms: Linguistic and Pedagogical Implications. *Bulletin of Science and Practice*. <https://doi.org/10.33619/2414-2948/48/51>
- Organización Mundial de la Salud. (2020). Primera Conferencia de la OMS sobre Infodemiología.

- Ortega-Bueno, R., Rangel, F., Farías, D. I. H., Rosso, P., Montes-Y-Gómez, M., & Medina-Pagola, J. E. (2019). Overview of the task on irony detection in Spanish variants. In *CEUR Workshop Proceedings*.
- Peng, H., Ma, Y., Li, Y., & Cambria, E. (2018). Learning multi-grained aspect target sequence for Chinese sentiment analysis. *Knowledge-Based Systems*. <https://doi.org/10.1016/j.knosys.2018.02.034>
- Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. <https://doi.org/10.3115/v1/d14-1162>
- Plaza-del-Arco, F. M., Martín-Valdivia, M. T., Ureña-López, L. A., & Mitkov, R. (2020). Improved emotion recognition in Spanish social media through incorporation of lexical knowledge. *Future Generation Computer Systems*. <https://doi.org/10.1016/j.future.2019.09.034>
- Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*. <https://doi.org/10.1016/j.eswa.2020.114120>
- Poria, S., Cambria, E., Howard, N., Huang, G. Bin, & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*. <https://doi.org/10.1016/j.neucom.2015.01.095>
- Price, W. N., Gerke, S., & Cohen, I. G. (2019). Potential Liability for Physicians Using Artificial Intelligence. *JAMA - Journal of the American Medical Association*. <https://doi.org/10.1001/jama.2019.15064>
- Ramalingam, V. V., & Ragavendran, R. (2020). Prediction of liver disease using artificial neural network with adam optimizer. *Journal of Critical Reviews*. <https://doi.org/10.31838/jcr.07.17.164>
- Rao, G., Huang, W., Feng, Z., & Cong, Q. (2018). LSTM with sentence representations for document-level sentiment classification. *Neurocomputing*.

- <https://doi.org/10.1016/j.neucom.2018.04.045>
- Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. <https://doi.org/10.18653/v1/d17-1317>
- Ravi, K., & Ravi, V. (2017). A novel automatic satire and irony detection using ensembled feature selection and data mining. *Knowledge-Based Systems*. <https://doi.org/10.1016/j.knosys.2016.12.018>
- Reganti, A. N., Maheshwari, T., Kumar, U., Das, A., & Bajpai, R. (2016). Modeling Satire in English Text for Automatic Detection. In *IEEE International Conference on Data Mining Workshops, ICDMW*. <https://doi.org/10.1109/ICDMW.2016.0141>
- Rieger, C. (1976). An organization of knowledge for problem solving and language comprehension. *Artificial Intelligence*. [https://doi.org/10.1016/0004-3702\(76\)90001-1](https://doi.org/10.1016/0004-3702(76)90001-1)
- Rocklage, M. D., Rucker, D. D., & Nordgren, L. F. (2018). The Evaluative Lexicon 2.0: The measurement of emotionality, extremity, and valence in language. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-017-0975-6>
- Rubinstein, D., & Ami Podrebarac, A. (2019). Introduction to Natural Language Processing. In *Fragmentation of the Photographic Image in the Digital Age*. <https://doi.org/10.4324/9781351027946-15>
- Salas-Zárate, M. del P., Paredes-Valverde, M. A., Rodríguez-García, M. Á., Valencia-García, R., & Alor-Hernández, G. (2017). Automatic detection of satire in Twitter: A psycholinguistic-based approach. *Knowledge-Based Systems*. <https://doi.org/10.1016/j.knosys.2017.04.009>
- Salaz Zárate, M. del P. (2018). Detection of psycholinguistic patterns for the analysis of subjective language in Spanish. *Procesamiento de Lenguaje Natural*. <https://doi.org/10.26342/2018-60-10>
- Salur, M. U., & Aydin, I. (2020). A Novel Hybrid Deep Learning Model for Sentiment Classification. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2020.2982538>

- Sánchez-Rada, J. F., & Iglesias, C. A. (2019). Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison. *Information Fusion*. <https://doi.org/10.1016/j.inffus.2019.05.003>
- Sandmo, A. (2019). ADAM SMITH AND MODERN ECONOMICS. In *Adam Smith*. <https://doi.org/10.2307/j.ctvc77g77.19>
- Serna, A. A., Acevedo, E. M., & Serna, E. M. (2017). Principios de la Inteligencia Artificial en las Ciencias Computacionales. *Actas de Ingeniería*.
- Sharma, A. Sen, Mridul, M. A., & Islam, M. S. (2019). Automatic Detection of Satire in Bangla Documents: A CNN Approach Based on Hybrid Feature Extraction Model. In *2019 International Conference on Bangla Speech and Language Processing, ICBSLP 2019*. <https://doi.org/10.1109/ICBSLP47725.2019.201517>
- Slaby, J., & Gallagher, S. (2015). Critical Neuroscience and Socially Extended Minds. *Theory, Culture & Society*. <https://doi.org/10.1177/0263276414551996>
- Smithers, L. E., & Eaton, P. W. (2017). Nomadic Subjectivity: Movement in Contemporary Student Development Theory. *Thresholds in Education*.
- Spechler, J. W. (2017). Digital Equipment Corporation. In *Reasonable Accommodation*. <https://doi.org/10.1201/9780203741207-7>
- Stigler, G. J., Irwin, D., Stigler, G. J., & Irwin, D. (2018). Leon Walras. In *Production and Distribution Theories*. <https://doi.org/10.4324/9781315127477-9>
- Sung, W., Ahn, J., & Black, J. B. (2017). Introducing Computational Thinking to Young Learners: Practicing Computational Perspectives Through Embodiment in Mathematics Education. *Technology, Knowledge and Learning*. <https://doi.org/10.1007/s10758-017-9328-x>
- Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*. <https://doi.org/10.3115/v1/p14-1146>

- Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*. <https://doi.org/10.1177/0261927X09351676>
- Thompson, M. J. (2014). Axel Honneth and the neo-Idealist turn in critical theory. *Philosophy and Social Criticism*. <https://doi.org/10.1177/0191453714541583>
- Tian, Y., Lo, D., & Lawall, J. (2014). Automated construction of a software-specific word similarity database. In *2014 Software Evolution Week - IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering, CSMR-WCRE 2014 - Proceedings*. <https://doi.org/10.1109/CSMR-WCRE.2014.6747213>
- Toçoğlu, M. A., & Onan, A. (2019). Satire Detection in Turkish News Articles: A Machine Learning Approach. In *Communications in Computer and Information Science*. https://doi.org/10.1007/978-3-030-27355-2_8
- Tsonkov, T. V., & Koychev, I. (2015). Automatic detection of double meaning in texts from the social networks. In *CEUR Workshop Proceedings*.
- Valencia-Garcia, R. (2005). *Un entorno para la extracción incremental de conocimiento desde texto en lenguaje natural*. Departamento de Ingeniería de la Información y las Comunicaciones.
- Van der Aalst, W. (2016). *Process mining: Data science in action*. *Process Mining: Data Science in Action*. <https://doi.org/10.1007/978-3-662-49851-4>
- Van Long, N. (2015). Optimal Control Theory. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. <https://doi.org/10.1016/B978-0-08-097086-8.43071-0>
- Waldrop, M. M. (2016). The chips are down for Moore's law. *Nature*. <https://doi.org/10.1038/530144a>
- Wang, Y., Huang, M., Zhao, L., & Zhu, X. (2016). Attention-based LSTM for aspect-level sentiment classification. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. <https://doi.org/10.18653/v1/d16-1058>

- Weiss, S. M., Indurkha, N., Zhang, T., & Damerau, F. J. (2005). *Text mining: Predictive methods for analyzing unstructured information*. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. <https://doi.org/10.1007/978-0-387-34555-0>
- Wilensky, R. (1983). Story grammars versus story points. *Behavioral and Brain Sciences*. <https://doi.org/10.1017/S0140525X00017520>
- Wright, J. (2017). The Turing Bombe Victory and the first naval Enigma decrypts. *Cryptologia*. <https://doi.org/10.1080/01611194.2016.1219786>
- Wulff, D. U., Hills, T. T., Lachman, M., & Mata, R. (2016). The aging lexicon: Differences in the semantic networks of younger and older adults. *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.
- Xiong, S., Lv, H., Zhao, W., & Ji, D. (2018). Towards Twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing*. <https://doi.org/10.1016/j.neucom.2017.11.023>
- Yang, C., Zhang, H., Jiang, B., & Li, K. (2019). Aspect-based sentiment analysis with alternating coattention networks. *Information Processing and Management*. <https://doi.org/10.1016/j.ipm.2018.12.004>
- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016 - Proceedings of the Conference*. <https://doi.org/10.18653/v1/n16-1174>
- Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *ArXiv*.
- Yu, J., Jiang, J., & Xia, R. (2019). Global inference for aspect and opinion terms co-extraction based on multi-task neural networks. *IEEE/ACM Transactions on Audio Speech and Language Processing*. <https://doi.org/10.1109/TASLP.2018.2875170>
- Yu, L. C., Wang, J., Robert Lai, K., & Zhang, X. (2018). Refining Word Embeddings Using Intensity Scores for Sentiment Analysis. *IEEE/ACM Transactions on Audio Speech and*

- Language Processing*. <https://doi.org/10.1109/TASLP.2017.2788182>
- Yuan, Z., Wu, S., Wu, F., Liu, J., & Huang, Y. (2018). Domain attention model for multi-domain sentiment classification. *Knowledge-Based Systems*. <https://doi.org/10.1016/j.knosys.2018.05.004>
- Zhang, S., Zhang, X., Chan, J., & Rosso, P. (2019). Irony detection via sentiment-based transfer learning. *Information Processing and Management*. <https://doi.org/10.1016/j.ipm.2019.04.006>
- Zhang, Z., Wang, L., Zou, Y., & Gan, C. (2018). The optimally designed dynamic memory networks for targeted sentiment classification. *Neurocomputing*. <https://doi.org/10.1016/j.neucom.2018.04.068>
- Zhao, W., Guan, Z., Chen, L., He, X., Cai, D., Wang, B., & Wang, Q. (2018). Weakly-supervised deep embedding for product review sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*. <https://doi.org/10.1109/TKDE.2017.2756658>
- Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*.