

Análisis de tweets y su influencia en los seguros de vida en el ámbito colombiano

Trabajo Fin de Máster

Maria Paula Ávila Rodríguez

Universidad Politécnica de Valencia
Escuela Técnica Superior de Ingeniería Informática
Máster Universitario en Gestión de la Información
Valencia, España

Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería de Sistemas
Maestría Gestión de Información
Bogotá D.C., Colombia
2020

Análisis de tweets y su influencia en los seguros de vida en el ámbito colombiano

Trabajo Fin de Máster

Mat. María Paula Ávila Rodríguez

Trabajo de investigación para optar al título de
Magíster en Gestión de Información

Tutores:
PhD Daniel Vila
Didier Mauricio Calderón
PhD Victoria Eugenia Ospina Becerra
D. Antonia Ferrer Sapena

**Universidad Politécnica de Valencia
Escuela Técnica Superior de Ingeniería Informática
Máster Universitario en Gestión de la Información
Valencia, España**

**Escuela Colombiana de Ingeniería Julio Garavito
Decanatura de Ingeniería de Sistemas
Maestría en Gestión de Información
Bogotá D.C., Colombia
2020**

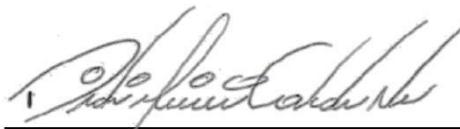
PÁGINA DE ACEPTACIÓN

El trabajo de grado de maestría titulado “Análisis de tweets y su influencia en los seguros de vida en el ámbito colombiano”, presentado por Maria Paula Ávila Rodríguez, cumple con los requisitos establecidos y recibe nota aprobatoria para optar al título de Magíster en Gestión de información.



Victoria Eugenia Ospina

Director del Trabajo de Grado



Didier Mauricio Calderón Novoa

Director del Trabajo de Grado



Dante Conti

Jurado



Ana María Gómez Lamus

Jurado

Bogotá, D.C., 01 de septiembre de 2020

© Únicamente se puede usar el contenido de las publicaciones para propósitos de información. No se debe copiar, enviar, recortar, transmitir o redistribuir este material para propósitos comerciales sin la autorización de la Escuela Colombiana de Ingeniería. Cuando se use el material de la Escuela se debe incluir la siguiente nota “Derechos reservados a Escuela Colombiana de Ingeniería” en cualquier copia en un lugar visible. Y el material no se debe notificar sin el permiso de la Escuela.

Publicado en 2020 por la Escuela Colombiana de Ingeniería Julio Garavito. Avenida 13 No 205-59 Bogotá. Colombia
TEL: +57 – 1 668 36 00

Reconocimiento o Agradecimientos

Agradezco a mi familia y a mi novio por apoyarme incondicionalmente durante toda mi formación académica y profesional. A Daniel Vila por su guía y dedicación en el desarrollo de esta tesis. También, a la Escuela Colombiana de Ingeniería Julio Garavito por su programa de becas para posgrado, del cual fui beneficiaria para adelantar los estudios en el programa de Maestría en Gestión de Información. Por último, agradezco al Sistema Movilidad Estudiantil y a las directoras del máster, Victoria Eugenia Ospina y Antonia Ferrer, quienes construyeron el convenio entre la Universidad Politécnica de Valencia y la Escuela Colombiana de Ingeniería Julio Garavito, permitiendo la experiencia de un intercambio académico en España.

Resumen

En los últimos años los seguros de vida han aumentado su participación dentro del mercado financiero colombiano. Para fortalecer este posicionamiento las aseguradoras tienen el reto de reinventar sus productos ajustándolos a las demandas de sus clientes y de este modo incrementar las ventas. Así pues, con el fin de incentivar integralmente el desarrollo de seguros de vida y dinamizar su oferta en el país, en este trabajo de máster se diseña, desarrolla y evalúa un modelo de procesamiento del lenguaje natural entrenado en la tarea de perfilado, a partir de datos no estructurados provenientes de Twitter. Con estos datos, se genera una segmentación de los usuarios evaluando los hábitos alimenticios, la actividad física y el sentimiento de las publicaciones, permitiendo ofrecer una tarifa preferencial para los fragmentos identificados con menor riesgo que genere un estímulo en la venta y cree valor económico para las aseguradoras. Se valida el modelo al relacionar los resultados con la tasa de mortalidad a nivel departamental, obteniendo una correlación de Pearson moderada del 0,56 para los departamentos urbanos de Colombia.

Palabras Clave: Tarificación, seguros de vida, Twitter, minería de datos, machine learning, procesamiento del lenguaje natural, Colombia.

Abstract

In recent years, life insurance has increased its participation in the Colombian financial market. To strengthen this position, insurers have the challenge of reinventing their products adjusting to the demands of their clients and thus increasing sales. Thus, to fully encourage the development of life insurance and streamline its offer in the country, in this master's work a natural language processing model trained in the profiling task is designed, developed and evaluated. unstructured data from Twitter. With these data, a segmentation of users is generated by analyzing eating habits, physical activity and the sentiment of the publications, allowing to offer a preferential rate for the fragments identified with less risk that generates a stimulus in the sale and creates economic value for insurers. The model is validated by relating the results to the mortality rate at the departmental level, obtaining a moderate Pearson's correlation of 0.56 for the urban departments of Colombia.

Keywords: Pricing, life insurance, Twitter, data mining, machine learning, natural language processing, Colombia.

Tabla de contenido

Lista de Ilustraciones

Lista de Tablas

Lista de Gráficas

1	INTRODUCCIÓN.....	1
1.1	PROBLEMÁTICA (JUSTIFICACIÓN).....	1
1.2	OBJETIVOS Y PREGUNTA DE INVESTIGACIÓN.....	2
1.3	ALCANCE Y LIMITACIONES.....	3
1.4	IMPACTO ESPERADO.....	4
1.5	METODOLOGÍA.....	5
2	MARCO TEÓRICO	7
2.1	ESTADO DEL ARTE.....	7
2.2	CRÍTICA AL ESTADO DEL ARTE.....	12
2.3	PROPUESTA.....	12
2.4	MARCO REFERENCIAL.....	13
2.4.1	TARIFICACIÓN DE SEGUROS DE VIDA.....	13
2.4.2	PROCESAMIENTO DE LENGUAJE NATURAL.....	16
2.4.3	DEPARTAMENTOS DE COLOMBIA.....	17
2.5	MARCO LEGAL.....	18
2.6	MARCO ÉTICO.....	20
3	BÚSQUEDA Y DEFINICIÓN DE FUENTES DE DATOS.....	23
3.1	TWITTER.....	23
3.2	HERRAMIENTAS DE EXTRACCIÓN DE DATOS DE MINERÍA SOCIAL.....	24
3.2.1	TWEET ARCHIVER	24
3.2.2	PYTHON.....	25
3.2.3	KNIME.....	26
3.3	SELECCIÓN DE LA HERRAMIENTA	26
3.4	EXTRACCIÓN DE DATOS.....	27
3.5	VARIABLES SELECCIONADAS	27
3.6	ANÁLISIS DESCRIPTIVO DE TWEETS	28
4	LIMPIEZA DE DATOS.....	30
4.1	HERRAMIENTA	30
4.2	LIMPIEZA.....	32
4.2.1	ANÁLISIS DE DATOS.....	32
4.2.2	DEFINICIÓN DE REGLAS DE MAPEO.....	33
4.2.3	TRANSFORMACIÓN	34
4.2.4	VERIFICACIÓN	39
5	DISEÑO DEL MODELO	43
5.1	PLAN DE TRABAJO.....	43

5.2	PRESUPUESTO.....	43
5.3	HERRAMIENTA	43
5.4	ARQUITECTURA DE RED NEURONAL	44
5.5	DESARROLLO DEL MODELO	46
5.5.1	CARGA Y PREPARACIÓN DE DATOS	46
5.5.2	ANOTACIÓN.....	51
5.5.3	ENTRENAMIENTO.....	58
5.5.4	VALIDACIÓN DEL MODELO	63
5.5.5	IMPLEMENTACIÓN: CASOS DE USO.....	72
6	VALIDACIÓN.....	75
6.1	TASA DE MORTALIDAD	75
6.1.1	PROYECCIONES DE POBLACIÓN.....	75
6.1.2	DEFUNCIONES.....	76
6.1.3	TASA DE MORTALIDAD POR DEPARTAMENTO.....	77
6.2	MÉTRICA DEL RESULTADO DEL MODELO POR DEPARTAMENTO	78
6.2.1	VALOR DEL MODELO POR DEPARTAMENTO.....	78
6.3	VALIDACIÓN DEL MODELO	79
7	CONCLUSIONES	83
7.1	CONCLUSIONES.....	83
7.2	RECOMENDACIONES.....	83
7.3	RELACIÓN DEL TRABAJO DESARROLLADO CON LOS ESTUDIOS CURSADOS.....	84
7.3.1	ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO.....	84
7.3.2	UNIVERSIDAD POLITÉCNICA DE VALENCIA	85
7.4	TRABAJOS FUTUROS.....	86
8	REFERENCIAS.....	87
ABREVIACIONES.....		92
ANEXOS		93

Lista de Ilustraciones

Ilustración 1 Fuentes de datos tradicionales y nuevas habilitadas por la digitalización.....	8
Ilustración 2 Funcionamiento de una neurona artificial.....	17
Ilustración 3 Perfil de audiencia de Twitter	24
Ilustración 4 Resultado Twitter Archiver	25
Ilustración 5 Librería Python para minería de datos.....	25
Ilustración 6 Nodos Knime.....	26
Ilustración 7 Nube de palabras del corpus	28
Ilustración 8 Nube de palabras del corpus reducido	29
Ilustración 9 Etapas limpieza de datos	30
Ilustración 10 OpenRefine 3.3	31
Ilustración 11 Aumentar de memoria OpenRefine.....	31
Ilustración 12 Text facet columna Location- OpenRefine	32
Ilustración 13 Add column base on this column- OpenRefine	34
Ilustración 14 Duplicar columnas- OpenRefine.....	35
Ilustración 15 Eliminar emoticones- OpenRefine	36
Ilustración 16 Text facet de la columna Location_Clean- OpenRefine.....	36
Ilustración 17 Función Cluster de OpenRefine.....	37
Ilustración 18 Keying Functions- OpenRefine.....	38
Ilustración 19 Opción “Split into several columns”- OpenRefine	40
Ilustración 20 Hallazgos de la verificación inicial- OpenRefine	40
Ilustración 21 Resultados del proceso de limpieza- OpenRefine	41
Ilustración 22 Exportar archivo- OpenRefine.....	41
Ilustración 23 Esquema plan de trabajo	43
Ilustración 24 Embedding	44
Ilustración 25 Encoding.....	45
Ilustración 26 Flujo de la arquitectura de Red Neuronal	46
Ilustración 27 Carga de datasets en biome	46
Ilustración 28 Output df.head()	47
Ilustración 29 Etiquetas biome	48

Ilustración 30 Proyectos biome	50
Ilustración 31 Anotaciones biome.....	50
Ilustración 32 Panel de anotación en biome	51
Ilustración 33 Opción Show labelled records.....	52
Ilustración 34 Ejemplos de anotaciones 1.....	52
Ilustración 35 Ejemplos de anotaciones 2.....	53
Ilustración 36 Ejemplos de anotaciones 3.....	54
Ilustración 37 Ejemplo de tweet con varias características.....	55
Ilustración 38 Ejemplo de tweet descartado.....	55
Ilustración 39 Botón “View metadata” de biome	56
Ilustración 40 Ejemplos anotaciones con sesgo por cuarentena	56
Ilustración 41 Ejemplos de biografías.....	57
Ilustración 42 Distribución de las anotaciones.....	57
Ilustración 43 Nube de palabras de las anotaciones	57
Ilustración 44 Predicted as- biome	60
Ilustración 45 Ejemplo de predicciones efectuadas por el modelo inicial 1	60
Ilustración 46 Ejemplo de predicciones efectuadas por el modelo inicial 2	61
Ilustración 47 Ejemplo correcciones a las predicciones efectuadas	61
Ilustración 48 Distribución de las anotaciones finales	62
Ilustración 49 Output pre-prosesamiento de tweets.....	63
Ilustración 50 Métricas para la categoría Non_healthy por epoch	64
Ilustración 51 Métricas para la categoría Non_Smoker por epoch	65
Ilustración 52 Métricas promedio por epoch	65
Ilustración 53 Métricas promedio por epoch de todas las categorías, sin Non_Smoker	66
Ilustración 54 Métricas por epoch de los cuatro experimentos	67
Ilustración 55 Métricas promedio del dataset de train y validation por epoch	69
Ilustración 56 Explorar predicciones en Biome	70
Ilustración 57 Ejemplos de falsas predicciones	71
Ilustración 58 Opciones matriz de confusión e intervalo de confianza biome	71
Ilustración 59 Nube de palabras de los casos de uso.....	73

Ilustración 60 Publicación Practice_sports del usuario Dos	73
Ilustración 61 Publicación Healthy del usuario Cinco.....	73
Ilustración 62 Publicación Negative del usuario Diez.....	74
Ilustración 63 Publicación Positive del usuario Doce	74
Ilustración 64 Proyecciones de Población 2018.....	76
Ilustración 65 Suite Marketplace.....	94
Ilustración 66 Aceptación de instalación de Tweet Archiver	94
Ilustración 67 Selección de cuenta de Google.....	94
Ilustración 68 Confirmación de confianza Archiver Tweets.....	95
Ilustración 69 Autorización de acceso a Twitter	95
Ilustración 70 Instalación del complemento finalizada	96
Ilustración 71 Complementos de la hoja de cálculo de Google.....	96
Ilustración 72 Creación de una regla de Twitter.....	97
Ilustración 73 Elegir archivos- OpenRefine	98
Ilustración 74 Configuración análisis de datos- OpenRefine	98
Ilustración 75 Vista previa- OpenRefine	99

Lista de Tablas

Tabla 1 Descuento según la categoría de riesgo.....	15
Tabla 2 Departamentos de Colombia con mayor población.....	17
Tabla 3 Categoría de los departamentos de Colombia	18
Tabla 4 Análisis de las necesidades y del beneficio público.....	21
Tabla 5 Información que Twitter recopila y usa	23
Tabla 6 Comparativa de herramientas de extracción de datos.....	26
Tabla 7 Palabras clave de búsqueda	27
Tabla 8 Entradas del modelo	28
Tabla 9 Errores o inconsistencias del campo ubicación	33
Tabla 10 Reglas de mapeo limpieza de datos.....	34
Tabla 11 Etiquetas de clasificación.....	48
Tabla 12 Ejemplo de preprocesador de texto.....	62

Tabla 13 Matriz de Confusión	63
Tabla 14 Resultados de los experimentos por clase.....	66
Tabla 15 Resultados F1 por intervalo de confianza y clase, del dataset de validación	71
Tabla 16 Variaciones de F1 por intervalo de confianza y clase, del dataset de validación.....	72
Tabla 17 Variaciones de F1 por intervalo de confianza y clase, del dataset de prueba	72
Tabla 18 Predicciones del modelo de los casos uso	74
Tabla 19 Correlación del valor del modelo con la tasa de mortalidad departamental	81

Lista de Gráficas

Gráfica 1 Índice de penetración en seguros por región	11
Gráfica 2 Número de tweets por departamentos.....	42
Gráfica 5 Resultado de las métricas por clase del dataset de validación	68
Gráfica 6 Resultado de las métricas por clase del dataset de prueba.....	69
Gráfica 7 Tasa de mortalidad por departamentos (ambos sexos, de 20 a 49 años)	77
Gráfica 8 Predicciones del modelo para el corpus total.....	78
Gráfica 9 Valor del modelo por departamento	79
Gráfica 10 Valor del modelo de clasificación de tweets versus tasa de mortalidad	80

1 Introducción

1.1 Problemática (Justificación)

Según el primer estudio de demanda de seguros realizado por la Federación de Aseguradores Colombianos, Fasecolda, y la Banca de las Oportunidades con el apoyo de la Superintendencia Financiera de Colombia (2018), el seguro más demandado en este país es el seguro de vida, con una participación del 19,2%. Mantener este crecimiento implica el desarrollo de una estrategia enfocada en las demandas de los consumidores de seguros. Dicha estrategia debe estar determinada por un trato personalizado que les permita percibir políticas de descuento que representen realmente sus condiciones de salud.

En efecto, Alejandro Romero señala en el *Insurance World Challenges 2017* que se requiere “entender al nuevo cliente y no solamente fijarse en un determinado rango de edad”, como tradicionalmente se hace. Para conseguirlo, es preciso involucrar datos internos y externos del negocio, así como nuevas tecnologías digitales que permitan la identificación de perfiles de riesgo y el desarrollo de un proceso de tarificación más exacto.

De hecho, las nuevas tecnologías digitales, permiten acceder a una nueva y valiosa fuente de información: el comportamiento online (Keller, 2018). Hoy en día, Twitter cuenta con más 340 millones de usuarios (We are social; Hootsuite, 2020), 3,2 millones ubicados en Colombia (Statista, 2020), quienes diariamente almacenan información sobre sus pensamientos, emociones, comportamientos, hábitos y otra información psicológica relevante. Estos datos han permitido la realización de varios estudios contundentes que ayudan a las empresas a conocer a sus clientes, medir la aceptación de un producto o servicio y de este modo, hacer crecer sus ganancias. De tal forma que el uso de manera adecuada de esos datos ha cobrado gran importancia en la estructuración de los planes de éxito de los diferentes negocios, incluyendo las compañías aseguradoras.

Así pues, alineada con la necesidad de conocer a los clientes y la oportunidad del aprovechamiento de los datos de redes sociales, este trabajo resuelve hacer uso de datos no estructurados provenientes de Twitter, para efectuar una caracterización de clientes conforme a su comportamiento, hábitos y estilo de vida y ofrecer un plan de descuentos aplicable a la tarifa de seguros de vida, de acuerdo con los segmentos de riesgo identificados.

1.2 Objetivos y Pregunta de Investigación

El propósito de este trabajo fin de máster es diseñar un modelo de aprendizaje automático que utilice la información disponible en Twitter como elemento de personalización de la tarifa de los seguros de vida riesgo, con el fin de innovar en el mercado asegurador colombiano. Para ello, se proponen tres objetivos específicos:

- Analizar y establecer los criterios necesarios para la segmentación de los usuarios de Twitter, a partir de sus publicaciones.
- Diseñar, desarrollar y evaluar un modelo de procesamiento del lenguaje natural, de modo que sea posible categorizar de acuerdo con los criterios establecidos la información capturada de Twitter.
- Identificar la relación existente entre los resultados del modelo y la tasa de mortalidad observada.

En consecuencia, este trabajo pretende responder la siguiente pregunta: ¿En qué medida los datos no estructurados -tweets en particular- aportan en la información para mejorar la segmentación de los consumidores de seguros de vida en Colombia?

1.3 Alcance y Limitaciones

En el presente estudio se desarrolla un modelo de clasificación de tweets utilizando aprendizaje automático supervisado con redes neuronales artificiales, que permite segmentar en cinco categorías de riesgo a los propietarios de una cuenta de Twitter, conforme a las publicaciones relacionadas con la práctica de deportes, niveles de estrés y consumo de tabaco y alimentos. Los resultados del clasificador se aplican para calcular el descuento sobre la tarifa del seguro de vida del usuario. El modelo abarca el procesamiento de lenguaje natural de aproximadamente 23,400 publicaciones escritas en español y se dirige a la población colombiana.

Por otro lado, debido a la complejidad del modelo y los costes de extracción de datos de las diferentes redes sociales, el modelo se limita al uso de datos de Twitter, sin integrar datos provenientes de otras plataformas, como Facebook o Instagram.

1.4 Impacto Esperado

La ejecución de este trabajo fin de máster, en cumplimiento de los objetivos, permitirá conseguir impactos en tres niveles: generación de nuevo conocimiento, transferencia de tecnología e impactos industriales.

Los principales avances en cuanto a generación de nuevo conocimiento son: la ampliación de las técnicas actuales para el conocimiento de los consumidores de seguros de vida; identificación del índice de fiabilidad entre la tasa de mortalidad y las publicaciones en redes sociales; la publicación de un artículo científico; y el presente documento, el cual contiene la descripción detallada del modelo construido y la estrategia aplicada para su desarrollo.

Por su parte, la transferencia de tecnología tendrá como resultado un informe técnico donde se describan las diferentes bases de datos, herramientas, entornos y algoritmos empleados para la construcción del modelo, siendo publicado en la plataforma de desarrollo colaborativo, GitHub, como un proyecto open source (código abierto).

Por último, los impactos industriales que se buscan generar son: la innovación en el cálculo de tarifas de seguros de vida en Colombia, el aporte de una nueva metodología para identificar las características particulares de cada asegurado, la oportunidad de involucrar fuentes externas al negocio asegurador y las técnicas de manejo y explotación de datos emergentes. A su vez, se derivan mejoras en la exactitud del desarrollo de la tarifa de los seguros de vida y aportaciones para el marketing personalizado, puesto que se identifica y trata a cada asegurado conforme a sus características propias.

1.5 Metodología

La metodología de este trabajo se desarrolla en cinco partes: la primera hace referencia al establecimiento del marco teórico de la investigación; la segunda representa la búsqueda y definición de las fuentes de datos, seguida por la limpieza de estos; en la cuarta etapa se realiza el diseño e implementación del modelo de clasificación de tweets; y en la etapa final se determina la viabilidad de la venta del producto.

En el primer componente, se investigan los avances realizados en esta temática, los principales casos de éxito a nivel internacional y los desafíos para el sector asegurador colombiano, buscando entender las oportunidades que tiene el desarrollo de una política de descuento para un seguro de vida con datos no estructurados provenientes de redes sociales. Así mismo, se realiza un marco referencial sobre los principales conceptos de la tarificación de los seguros de vida, el procesamiento de lenguaje natural y la división geopolítica del país. Por último, se estudian las diferentes normativas aplicables para el tratamiento de datos personales, los términos y condiciones del uso de Twitter y el marco ético en el que se desarrolla el proyecto.

Para la búsqueda y definición de fuentes de datos se procede a identificar las características de Twitter, seleccionar las variables a considerar dentro del modelo y finalmente, realizar la extracción de datos mediante el complemento de Google, Tweet Archiver. En la siguiente etapa -limpieza de datos- se emplea OpenRefine, para analizar e identificar los datos incompletos, incorrectos o inexactos del conjunto de datos; plantear y ejecutar los criterios para eliminar o agrupar la información y por último, verificar los resultados.

En la etapa de diseño del modelo, se utiliza la librería biome-text para realizar el proceso de anotación, entrenamiento y validación, basada en métricas derivadas de la matriz de confusión. Para ello se clasifican manualmente un aproximado de 3000 tweets para otorgar un entrenamiento inicial al modelo y medir el rendimiento particionando los datos en entrenamiento y prueba, para calcular las métricas de validación.

Finalmente, se realiza la validación del modelo, mediante una comparativa de la tasa de mortalidad identificada en las principales regiones de Colombia y el resultado del modelo tras analizar el contenido de Twitter, determinando así, si las publicaciones de la red social son un buen predictor de la mortalidad, teniendo en cuenta los factores no genéticos elegidos.

2 Marco teórico

2.1 Estado del arte

Actualmente, muchas aseguradoras centran sus esfuerzos en el reconocimiento de sus clientes con dos objetivos principales: cumplir las demandas de los consumidores de seguros al tener un trato personalizado que les permita percibir políticas de descuento que representen realmente sus condiciones particulares (Soto, 2019) y realizar una adecuada valoración del riesgo asumido en el contrato de seguros. Para conseguirlo, las compañías están adoptando soluciones tecnológicas que faciliten el manejo del gran volumen de información que recolectan diariamente en sus procesos, así como el proveniente las nuevas fuentes de información externas al negocio.

Estas soluciones las está proporcionando el big data, la inteligencia artificial y el internet de las cosas (*IoT- Internet of Things*). En los últimos años, el sector asegurador ha empezado a transformar el modelo de negocio, adaptándose al nuevo entorno digital y maximizando el potencial que los datos ofrecen para mejorar su proceso de segmentación de perfiles de riesgo, aplicado principalmente a la fijación de precios y suscripción. Esto bajo el supuesto de que “mejores datos hacen posible mejorar la alineación entre primas y riesgos y reducir el costo general del seguro” (Keller, 2018).

En efecto, conforme al estudio desarrollado en el 2019 por la Autoridad Europea de Seguros y Pensiones de Jubilación (EIOPA por sus siglas en inglés) se concluye que “las compañías de seguros consideran que el análisis del big data les permite desarrollar evaluaciones de riesgo más granulares y una mejor segmentación de los consumidores mediante la evaluación de los riesgos en áreas y segmentos que no eran posibles en el pasado”. En este sentido, resulta interesante revisar un comparativo entre las fuentes tradicionales de información y las nuevas que la implementación de esta reciente tecnología ha generado. En la ilustración 1 se muestra el resumen las transformaciones:

Ilustración 1 Fuentes de datos tradicionales y nuevas habilitadas por la digitalización

Traditional data sources	New data sources enabled by digitalisation
Medical data (e.g. medical history, medical condition, condition of family members)	IoT data (e.g. driving behaviour (car telematics), physical activity and medical condition (wearables).
Demographic data (e.g. age, gender, civil and family status, profession, address)	Online media data (e.g. web searches, online purchases, social media activities, job career information)
Exposure data (e.g. type of car, value of contents inside the car)	Insurance firms' own digital data (e.g. interaction with insurance firms (call centre data, users' digital account information, digital claim reports, online behaviour while logging in to insurance firms' websites or using insurance firms' app)
Behavioural data (except IoT data) (e.g. Smoking, drinking behaviour, distance driven in a year)	Geocoding data (i.e. latitude and longitude coordinates of a physical address)
Loss data (e.g. claim reports from car accidents, liability cases)	Genetics data (e.g. results of predictive analysis of a person's genes and chromosomes)
Population data (e.g. mortality rates, morbidity rates, car accidents)	Bank account / credit card data (e.g. consumer's shopping habits, income and wealth data)
Hazard data (e.g. frequency and severity of natural hazards)	Other digital data (e.g. selfie to estimate biological age of the consumer)
Other traditional data (e.g. credit scoring, claim adjustment reports, information from the auto repair shops)	

Fuente: EIOPA (2019)

A nivel internacional, se presentan múltiples ejemplos del uso de las nuevas fuentes de información. Compañías, como *Progressive* o *State Farm*, pertenecientes al sector asegurador americano, utilizan soluciones de IoT para segmentar a los asegurados. Monitoreando los hábitos de manejo a través de “cajas negras” instaladas en los vehículos, aplicaciones en el teléfono del tomador o dispositivos instalados en los carros, crean un perfil de conducción del contratante de la póliza de seguros de autos. Los datos recopilados incluyen posición geográfica, distancia recorrida, velocidad, aceleración, número de viajes, vibraciones de vehículo, entre otros. Datos con los que, en conjunto con otros datos de fuentes tradicionales (género, tipo de coche, etc.), les permite determinar una prima personalizada para cada asegurado. Estas iniciativas han tenido gran acogida en el mercado, aumentando su uso hasta en un 20% entre el 2015 y el 2016 (LexisNexis, 2016).

Por su parte, varias aseguradoras han optado por tomar datos provenientes de distintas redes sociales y explotar el poder de la minería de datos para segmentar a sus clientes de seguros generales. Este es el caso de la aseguradora inglesa *Admiral Insurance*, entidad que, en el 2016, innovó generando un plan de descuentos para el seguro de auto, cuyo objetivo era crear un perfil de riesgo de los jóvenes que no contaban con historial de conducción, usando un algoritmo de minería de texto que analizaba los ‘me gusta’, publicaciones y ubicación GPS del perfil de Facebook de los asegurados y de este modo, identificaba los rasgos en su personalidad que permitían emitir un juicio sobre el nivel de riesgo de conducción y otorgar una reducción en la prima del seguro (Zappa & Borrelli, 2017).

No obstante, esta estrategia fue detenida por Facebook, puesto que a pesar de tener consentimiento de los asegurados para el tratamiento de sus datos personales, la red social lo calificó como “intrusivo” e “inapropiado” (Peachey, 2016), basándose en el numeral 3.15 de la Política de la plataforma, donde se estipula que no es posible emplear los “datos obtenidos de Facebook para tomar decisiones sobre requisitos de participación, como decidir si se debe aprobar o rechazar una solicitud o cuánto interés se debe cobrar en un préstamo”.

En cuanto al uso de información interna, se destaca el artículo de Diego Zappa, titulado “*Text Mining In Insurance: From Unstructured Data To Meaning*” donde se presenta un caso de estudio en el que se analizan, con minería de texto, los reportes de accidentes escritos por los investigadores de la Administración Nacional de Seguridad del Tráfico en Carreteras (NHTSA) del Departamento de Transporte de EE. UU, con el fin de obtener un perfil de riesgo e incorporar las covariantes identificadas en la información necesaria para ajustar las primas de las pólizas de seguros de autos.

De igual manera, empresas como *John Hancock*, una de las aseguradoras de vida más grandes de Estados Unidos, cambió la manera de fijar el precio de los seguros de vida. Desde el 2019 exige a las personas interesadas en contratar un seguro, el uso de dispositivos telemáticos tales como manillas de actividad, relojes inteligentes o aplicaciones de teléfono. Estos instrumentos monitorean, almacenan y categorizan de manera constante la actividad física, el ritmo cardiaco, los pasos, horarios de sueño y demás hábitos diarios para otorgar descuentos a las personas con los perfiles de menor riesgo. Adicionalmente, tras adquirir el seguro ofrecen un sistema de recompensas por adoptar hábitos de saludables, así como asistencia médica en caso de emergencias reportadas por el dispositivo.

Otras iniciativas más actuales consideran el uso de información genética para fijar los criterios de tarificación de seguros de vida y salud, proponiendo remplazar la unidad de medida de la esperanza de vida actual (la edad cronológica) por la edad biológica, es decir, la derivada del estado funcional del organismo a partir del estudio de las células. Se alude que es un indicador que aporta mayor exactitud sobre el estado de salud de las personas y que refina la estimación de la siniestralidad. Sin embargo, se genera un debate en torno a la privacidad y datos personales, al referirse a un dato sensible que involucra, no solo al asegurado, sino a terceros (parientes consanguíneos), que no tienen relación con el contrato de seguros ni dan consentimiento para el tratamiento de sus datos sanitarios (Rodríguez C., 2011).

En pocas palabras, se evidencia que el sector asegurador (incluyendo seguros vida y seguros generales) se interesa con mayor frecuencia en la innovación y emprendimiento con datos para mejorar el proceso de fijación de tarifas, aprovechando los avances tecnológicos y científicos. Mientras que los usuarios se preocupan por obtener primas más exactas, sin desproteger la privacidad de sus datos.

En cuanto a la usabilidad de la información que se puede extraer de Twitter, se mencionan a continuación cuatro estudios dedicados a identificar las relaciones entre las publicaciones y las

condiciones de salud de los usuarios. Puesto que se considera importante soportar que, efectivamente, la información disponible en redes sociales permite realizar una segmentación de acuerdo con el estilo de vida y hábitos de los usuarios.

El primero es el estudio efectuado por la APS (*Association for Psychological Science*) en el 2015, en el que se realiza un análisis de los tweets de la población estadounidense utilizando una regresión lineal regularizada (*ridge regression*) para ajustar el modelo. Dicho estudio concluye que la información contenida en Twitter predice mejor la mortalidad por enfermedades cardíacas, que los factores de riesgo tradicionales, como los demográficos, sociales y sanitarios (tabaquismo, hipertensión, diabetes, obesidad). Afirma además que “los patrones de lenguaje que reflejan las relaciones sociales negativas, desconexión y emociones negativas, especialmente la ira, surgieron como factores de riesgo; las emociones positivas y el compromiso psicológico surgieron como factores protectores” (Eichstaedt, et al., 2015).

El segundo, es el desarrollado por el *Qatar Computing Research Institute* durante el 2015 en Estados Unidos, en el que se demuestra la posibilidad de realizar un monitoreo de la salud pública en el ámbito de los hábitos alimenticios utilizando el contenido de los tweets. Para efectuar este estudio se tomó información las publicaciones donde los usuarios relataban sus experiencias gastronómicas, para luego entrenar a un clasificador Naive Bayes y validar sus resultados bajo la relación de las calorías de los alimentos mencionados con las tasas de obesidad y diabetes de cada uno de los estados del país, encontrando una correlación de Pearson de 0,77 (Abbar, Mejova, & Weber, 2015).

El siguiente estudio, publicado por “Journal of Medical Internet Research” en 2019, integró y relacionó publicaciones de Twitter referentes a actividad física y ejercicios, con datos de reportes del nivel de actividad física del *Behavioral Risk Factor Surveillance System*, utilizando modelos de regresión lineal múltiple. Lo que permitió identificar que efectivamente existe una relación entre la proporción de personas que tocan este tema en redes sociales y las que lo practican. Sin embargo, en cuanto al análisis de sentimiento hacia el deporte (clasificado como positivo, negativo y neutro), no funcionó como predictor para el nivel de ejercicio de los usuarios (Liu, 2019).

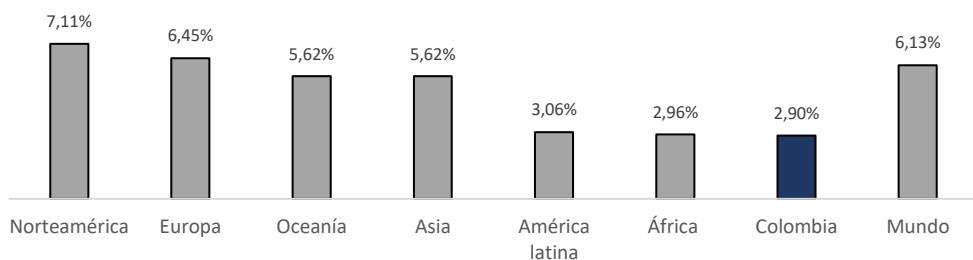
La cuarta investigación resuelve cuantificar y predecir el estado de salud de las personas basados en su actividad en Twitter. Para ello utilizan un clasificador de texto robusto que etiqueta los tweets como “enfermos”, si el contenido del texto indica que el usuario está enfermo, u “otros” si se relaciona con otra temática, mediante un procedimiento de aprendizaje en cascada SVM (Support Vector Machine). Adicionalmente, relaciona factores ambientales como la pobreza y la polución (a partir de la ubicación), y aspectos sociales como encuentros físicos o amigos enfermos, para tener un perfil de la salud del individuo. Concluyen que las publicaciones de Twitter son una fuente de información extremadamente rica para ejecutar estos estudios (Sadilek A. K., 2013).

A su vez, otros estudios que son muy usuales en redes sociales son los de mercados, orientados a conocer la forma en la que los usuarios interactúan con el nombre de la marca. Uno de los más

destacados es el elaborado para la aseguradora estadounidense *Allstate*, donde se detalla la forma en la que se aplican técnicas de minería de datos y analítica de texto a los datos extraídos de Twitter, para generar patrones que ayudan a identificar palabras clave y conceptos. Esto con el objetivo de aplicar los resultados ante diferentes temáticas como: área de atención al cliente, comprensión del sentimiento del cliente sobre la empresa, preferencias de los asegurados e identificación de tendencias más amplias en el mercado que la empresa puede aprovechar (Roosevelt, 2012).

Así pues, como se ha revisado a lo largo del capítulo, la mayoría de estas iniciativas se están planteando en el marco estadounidense y europeo. Esto se debe a que son las regiones con mayor índice de penetración, es decir, con mayor volumen de primas de seguros sobre el Producto Interno Bruto (PIB) y la inversión en innovación cobra relevancia. Según las cifras publicadas por la Federación de Aseguradores Colombianos, Fasecolda, durante el 2017 el índice de penetración norteamericano y el europeo era equivalente a 7.11% y 6.45%, respectivamente, mientras que en Colombia apenas alcanzaba el 2.90% (Durán, 2018), demostrando la baja cultura de seguros en el país. A continuación, se muestran las cifras mundiales:

Gráfica 1 Índice de penetración en seguros por región



Fuente: (Durán, 2018)

A pesar de la baja penetración del sector seguros en Colombia, hoy en día se ve como una nueva oportunidad de mercado, llevando a este sector a diversificar y flexibilizar su oferta para captar un mayor público. Fasecolda afirma que la baja penetración de los seguros en el país podría aumentar con el auge de las nuevas tendencias que se ajustan a los riesgos del siglo XXI y Juan Camilo Zarruk, Gerente de Negocio de la Agencia de Seguros Falabella, confirma la necesidad de innovar en el sector: “más allá de los productos tradicionalmente adquiridos por los colombianos, como el seguro de auto o de moto, hemos venido proponiendo alternativas de protección que suplen necesidades cada vez más específicas y que mejoran la calidad de vida de cada cliente” (Rodríguez Alvarez, 2017).

Dichas necesidades más específicas, se han venido enmarcando dentro de las, también, nuevas tendencias del mercadeo personalizado. Cada vez más común escuchar el término personalización, el cual hace referencia al “conjunto de estrategias y acciones que nos permiten ofrecer una oferta de productos y servicios diferenciados para cada cliente” (De Esteban), puesto que hoy en día se vuelve más relevante conocer los gustos e intereses de cada uno de los

consumidores y reinventar el proceso de compra para obtener el éxito en todos los sectores, incluyendo el asegurador.

2.2 Crítica al estado del arte

Tras el análisis de las distintas fuentes de información empleada para conocer a los asegurados, se percibe que los avances tecnológicos han potenciado la innovación en el sector asegurador y es posible explorar innumerables posibilidades para acceder a información más detallada sobre los clientes, siempre y cuando esto no invada su privacidad o esté en contra de las normativas aplicables.

El caso de la aseguradora *Admiral*, deja una lección invaluable en lo que respecta al emprendimiento del uso de datos de redes sociales: revisar los términos y condiciones de las plataformas con las que se desea acceder a la información, así como la normativa aplicable para el uso y tratamiento de los datos. Dado que se conoce este acontecimiento (Peachey, 2016), en este trabajo se opta por no usar datos de Facebook y evaluar con detenimiento las políticas de Twitter.

Por otro lado, se percibe que las fuentes de información seleccionadas con mayor frecuencia por las aseguradoras son los dispositivos telemáticos. Sin embargo, es necesario precisar que es posible que estos artefactos (pulseras, cajas negras, aplicaciones) sean saboteados con el fin de obtener mejores tarifas, por lo que resultaría importante realizar una validación de que la información que se está empleando efectivamente es real y pertenece al asegurado. Este es un punto que también se debe considerar dentro del desarrollo del trabajo y se contempla realizarlo mediante la opción de cuenta certificada, lo cual permite declarar que la cuenta efectivamente pertenece a quién se evalúa.

En cuanto a los estudios que se han realizado con la información extraída de Twitter (APS, *Qatar Computing Research Institute*, Liu S, Chen B, Kuo) se evidencia que en todos los casos los resultados obtenidos con el análisis de las publicaciones son contrastados con cifras derivadas de mediciones tradicionales. Por lo cual, se identifica que este tipo de validaciones son de gran importancia para conseguir fiabilidad en el modelo que se generará y resulta imprescindible realizarlas.

A su vez, como se mencionó, es pertinente innovar en el sector asegurador colombiano teniendo en cuenta las oportunidades y amenazas que ha enfrentado el mercado global y de la misma forma aumentar el índice de penetración.

2.3 Propuesta

Teniendo en cuenta la necesidad de segmentación de los clientes de las aseguradoras y la búsqueda de innovación que sugiere el sector colombiano, este trabajo analiza el uso potencial de la información disponible en Twitter como elemento de personalización de los seguros de vida, lo que pretende fomentar la venta de seguros de vida en el país.

Así pues, conforme a las aplicaciones que existen actualmente se identifican tres contribuciones para las aseguradoras:

- (1) El uso de datos no estructurados para el cálculo de tarifas de seguros en Colombia.
- (2) La aplicación de una segmentación basada en minería de texto de las publicaciones de redes sociales, en el ámbito de seguros de vida, para generar una tarifa más exacta.
- (3) Análisis de la correlación entre las publicaciones de Twitter evaluadas y la tasa de mortalidad observada.

2.4 Marco referencial

Con el fin de contextualizar al lector con la terminología que se usará a continuación, resulta imprescindible explicar algunos conceptos básicos sobre tarificación de seguros de vida, procesamiento de lenguaje natural y la división territorial de Colombia (departamentos).

2.4.1 Tarificación de seguros de vida

La actividad aseguradora se basa en el “reparto del riesgo entre un gran número de personas con las mismas o similares necesidades de protección” (Fundación Mapfre). Es decir, las compañías de seguros se encargan de recaudar las primas de un gran volumen de individuos, administrar este dinero y garantizar el pago de una indemnización a los asegurados o beneficiarios, en caso de materializarse el riesgo asumido. Para que esto funcione, las entidades deben calcular adecuadamente la siniestralidad esperada y basadas en esto, cobrar la tarifa que les permita cumplir con las obligaciones pactadas en el contrato de seguros.

Con el fin de identificar la siniestralidad es importante definir la cobertura de cada seguro. En particular, en los seguros de vida es posible resguardar dos aspectos: la muerte o la sobrevivencia. Los que cubren el primer aspecto, llamados comúnmente como vida riesgo, son aquellos que ofrecen al tomador el pago de la suma asegurada si el asegurado fallece dentro de la vigencia de la póliza, la cual puede ser vitalicia o temporal. Mientras que los que aseguran la sobrevivencia, denominados de ahorro, indemnizan si se sobrevive al finalizar la temporalidad contratada. Adicionalmente, se ofrecen seguros con las dos coberturas, nombrados dotales o mixtos. Este trabajo, se enfoca en los seguros de vida riesgo con una vigencia temporal.

Ahora bien, el valor de la prima depende del riesgo y del valor asegurado que se desee contratar. Como trabajaremos sobre un seguro de vida riesgo, para cuantificar del riesgo que cada asegurado representa, se debe evaluar la probabilidad de que el asegurado fallezca dentro de la vigencia de la póliza. Para ello, tradicionalmente, se emplean las tablas de mortalidad, las cuales son el instrumento de análisis demográfico que permite medir la incidencia de la mortalidad para cada edad y sexo específico (Instituto Nacional De Estadística, 2009), teniendo en cuenta aproximaciones estadísticas de las defunciones observadas sobre la población en estudio en un periodo de referencia.

Adicionalmente, para realizar una medición granular del riesgo también se emplean los cuestionarios de asegurabilidad, donde los clientes reportan su condición de salud, hábitos y

preexistencias respondiendo a preguntas específicas, con el fin de delimitar la cobertura del seguro. Por otro lado, en caso de que se considere necesario (por sumas aseguradas muy altas o condiciones de salud específicas) se somete al cliente a exámenes médicos rigurosos para evaluar su estado de salud. En ambos escenarios, se determina si el cliente es un riesgo estándar o si, por lo contrario, es necesario que asuma una extra-prima dadas sus características sanitarias. Para el presente trabajo, se excluirán estos valores adicionales, puesto que se considera que todos los asegurados son riesgo estándar y contratan sumas aseguradas promedio.

Así pues, para realizar el cálculo del riesgo estándar las áreas de actuaria de las aseguradoras emplean las tablas de mortalidad por edad y género adoptadas por la Superintendencia Financiera de Colombia. En la fecha de realización de este trabajo, las tablas vigentes son las reglamentadas en la Resolución 1112 de 2007, denominadas las “Tablas Colombianas de Mortalidad de los Asegurados por Sexos. Experiencia 1998 – 2003”. Sin embargo, estas tablas no revelan la mortalidad actual de los colombianos, según lo expone Fasecolda en la edición 166 de su revista, donde sugieren que la probabilidad de muerte se ha reducido hasta un 40%, en promedio (Torres & Mayorga, 2017). Por esto resulta necesario combinar las fuentes de información tradicionales (tablas de mortalidad), con las nuevas (redes sociales) y de este modo, ajustar esta desactualización.

En consecuencia, en este trabajo se busca otorgar una tarifa personalizada equivalente a la prima habitual (calculada con la tabla de mortalidad vigente), con una reducción porcentual derivada de los factores de riesgo identificados en las publicaciones de Twitter del asegurado. De este modo, en nomenclatura actuarial, el valor de prima pura del seguro corresponderá a:

$$P = (1 - PD_j) \cdot (VA \sum_{i=0}^{T-1} v^{i+1} \cdot ip_x \cdot q_{x+i})$$

Donde,

P : Prima pura de riesgo del seguro de vida

PD_j : Porcentaje de descuento aplicable para el tipo de riesgo j

VA : Valor asegurado

T : Temporalidad contratada para el seguro

v^i : Factor de actualización del periodo i

ip_x : Probabilidad que una persona de edad x sobreviva i años, teniendo en cuenta las tablas de mortalidad de asegurados Experiencia 1998 – 2003

q_{x+i} : Probabilidad que una persona de edad $x + i$ muera antes de cumplir $x + i + 1$ años, teniendo en cuenta las tablas de mortalidad de asegurados Experiencia 1998 – 2003

Para construir los grupos de riesgos j , se evalúan los factores no genéticos que suponen una vida saludable y más longeva sugeridos en el estudio publicado por la Fundación MAPFRE, donde

se menciona que “los hábitos de vida saludables, como no fumar, no comer carne roja, hacer ejercicio físico o evitar el estrés aumentan la esperanza de vida de 5 a 10 años” (Rodríguez J. M., 2011), cifras que concuerdan con el estilo de vida de varias poblaciones que se caracterizan por su longevidad.

Así pues, la política de descuento aplicable al modelo contempla los factores: alimentación saludable, práctica de actividad física, no fumar y nivelación del estrés, donde este último se evalúa conforme al sentimiento identificado en las publicaciones bien sea, positivo o negativo. También, se emplea la categoría “otros” para los textos que no se relacionen con estas temáticas

A partir de los factores descritos anteriormente, se definen cinco categorías de riesgo determinadas por la evaluación y segmentación de los cien últimos tweets escritos por el consumidor de seguros de vida. A continuación se describen sus características:

- Bajo: dentro de los contenidos evaluados se encuentran publicaciones etiquetadas como alimentación saludable, práctica de actividad física y no fumador.
- Medio bajo: se evidencia al menos una publicación referente a la alimentación saludable, práctica de deporte o abstinencia al tabaco. Adicionalmente, se observan textos clasificados como positivos.
- Medio: se verifica que contenga al menos una publicación relacionada con alimentación saludable, práctica de deporte o abstinencia al tabaco. También, se evidencian tweets clasificados como negativos.
- Medio alto: únicamente se presentan tweets etiquetados como positivos, negativos y otros.
- Alto: contiene únicamente tweets etiquetados como negativos y/u otros.

De estas categorías de riesgo, se establecen los porcentajes de descuento de la prima del seguro que se efectuarán:

Tabla 1 Descuento según la categoría de riesgo

Categoría de riesgo	Descuento de la prima
Bajo	40%
Medio Bajo	30%
Medio	15%
Medio Alto	5%
Alto	0%

De este modo, un usuario que tenga al menos una publicación sobre la práctica ejercicio y además sugiera tener bajos niveles de estrés (con textos positivos), obtendrá un descuento del 30% en el valor habitual del seguro. Mientras que un sujeto que no realice publicaciones relacionadas con las temáticas descritas pagará la prima normal del seguro.

Cabe aclarar que la siniestralidad no se ve afectada por esta reducción en la tarifa puesto que, por un lado, las mejoras de mortalidad por la selección de riesgo y la desactualización de las tablas actuales soportan la disminución y por el otro, al tratarse de un seguro temporal, no se

genera riesgo de longevidad, el cual hace referencia al aumento de la esperanza de vida a edades superiores a las estimadas en el cálculo.

2.4.2 Procesamiento de lenguaje natural

Para efectuar el análisis de las publicaciones, se implementarán técnicas de procesamiento de lenguaje natural. Por “Procesamiento de Lenguaje Natural (PLN, denominado también NLP por sus siglas en inglés) se entiende la habilidad de la máquina para procesar la información comunicada” (Gelbukh A. , 2010), es decir, no simplemente entender bytes y dígitos, sino comprender el lenguaje humano. Los procesos de PLN comprenden la construcción de modelos matemáticos con herramientas de lingüística, procesamiento léxico, morfosintáctico y semántico, entre otras.

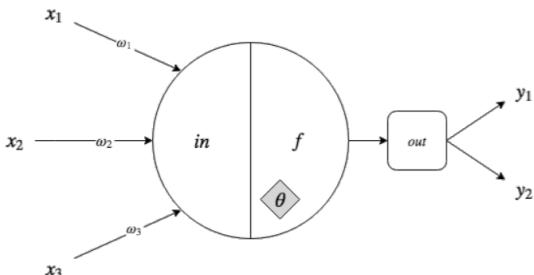
Dada la amplia variedad de estos procesos y líneas de investigación en PLN, en la literatura se habla frecuentemente de tareas, para las cuáles se desarrollan diferentes métodos y técnicas, así como conjuntos de datos de evaluación. Algunos ejemplos de tareas son: clasificación de textos, análisis de sentimiento, etiquetado de partes del discurso, procesado de dependencias sintácticas, implementación de chatbots o inferencia del lenguaje natural.

El desarrollo de técnicas, métodos y modelos de este trabajo está enmarcado dentro de la tarea denominada clasificación de textos y a la tarea particular de clasificación de tweets. Específicamente, en este trabajo se desarrollan y evalúan diferentes métodos, técnicas y modelos de clasificación de tweets utilizando aprendizaje automático supervisado con redes neuronales artificiales.

A saber, el aprendizaje automático es la “disciplina que estudia la construcción de sistemas computacionales que mejoran automáticamente con la ayuda de valoración de experiencias” (García, Martínez, & García, 2016). Este proceso se realiza empleando redes neuronales artificiales, las cuales son unidades de procesamiento capaces de “emular ciertas características propias de los humanos, como la capacidad de memorizar y de asociar hechos” (Matich, 2001), en otras palabras, son sistemas dinámicos que aprenden a partir de una experiencia inicial.

Las redes neuronales se componen de nodos interconectados (llamados neuronas artificiales) encargados de procesar las entradas (con un peso asignado) y generar un valor resultado. Para conseguirlo se aplica una función de entrada (i_n) que corresponde a una combinación lineal de las entradas y los pesos. A partir de este resultado, se utiliza la función de activación (f) que calcula el estado de actividad de la neuronal. En la Ilustración 2 se ejemplifica su funcionamiento:

Ilustración 2 Funcionamiento de una neurona artificial



Fuente: (Garrigues, 2019)

Así pues, en este trabajo se utiliza esta técnica para entrenar un modelo en la tarea de clasificación de tweets en categorías específicas. Para su desarrollo y evaluación se utiliza la librería de código abierto biome-text¹, que permite la configuración y evaluación de diferentes arquitecturas de redes neuronales y tipos de preprocesamiento para la clasificación de texto, como los que se exploran a continuación:

- Tokenización: hace referencia al proceso de separar las piezas del texto en entidades llamadas tokens, permitiendo agrupar las palabras como una unidad semántica útil para el procesamiento.
- Normalización: se refiere a la tarea de homogenizar el texto, convirtiéndolo en mayúsculas o minúsculas (según se defina), modificando los números por palabras, omitiendo los signos de puntuación y espacios.

2.4.3 Departamentos de Colombia

La validación del modelo se realiza desagregada por las unidades territoriales de Colombia, denominadas departamentos, por lo que resulta importante familiarizar al lector con esta temática. En la Constitución Política de Colombia 1991, se estipula que el territorio nacional tiene una división político-administrativa en departamentos, distritos, municipios y territorios indígenas. Como resultado, el país se fracciona en 32 departamentos y un distrito capital, Bogotá. A continuación, se presentan los diez departamentos (y sus capitales) donde se acumula el 68% de la población. El cuadro con la totalidad de los departamentos, capitales y población se muestra en el Anexo A. Población total de Colombia, con desagregación departamental.

Tabla 2 Departamentos de Colombia con mayor población

Departamento	Cod_Dpto	Capital	Población total (2018)
Bogotá (D.C.)	11	Bogotá	7.412.566
Antioquia	5	Medellín	6.407.102
Valle del Cauca	76	Cali	4.475.886
Cundinamarca	25		2.919.060
Atlántico	8	Barranquilla	2.535.517
Santander	68	Bucaramanga	2.184.837

¹ <https://github.com/recognai/biome-text>

Departamento	Cod_Dpto	Capital	Población total (2018)
Bolívar	13	Cartagena de Indias	2.070.110
Córdoba	23	Montería	1.784.783
Nariño	52	San Juan de Pasto	1.630.592
Norte de Santander	54	San José de Cúcuta	1.491.689

Fuente: (DANE, 2020)

Según la Comisión Económica para América Latina y el Caribe (CEPAL), estos departamentos se clasifican en urbanos, intermedios y rurales, conforme al Índice Territorial de Aglomeración (ITA). Los urbanos presentan alta densidad poblacional y gran tamaño territorial, los intermedios poseen un nivel de urbanización menor al de las urbanas, a las cuales se encuentran vinculadas, en su mayoría. Y las rurales presentan menor concentración demográfica. A continuación se muestran la categoría de cada departamento:

Tabla 3 Categoría de los departamentos de Colombia

Urbana	Intermedia	Rural
Quindío	Cesar	Amazonas
Risaralda	Caquetá	Guaviare
Valle del Cauca	Boyacá	Vichada
Atlántico	Huila	Vaupés
Magdalena	Arauca	Guainía
Caldas	La Guajira	
Norte de Santander	Córdoba	
Tolima	Nariño	
Sucre	Chocó	
Bolívar	Cauca	
Meta	Casanare	
Santander	Putumayo	
Antioquia	Cundinamarca	
Bogotá	San Andrés	

Fuente: (Ramírez & De Aguas, 2017)

2.5 Marco legal

Como todos los medios de la Web 2.0, en Twitter se pueden encontrar una variedad de publicaciones que contienen puntos de vista, opiniones, gustos y otros datos que revelan fácilmente información personal sobre los usuarios como el género, la orientación sexual, la nacionalidad, entre otros. Lo cual está dentro de la definición de datos personales, por tanto, resulta imprescindible abordar la normativa que lo reglamenta, es decir, la Ley de Protección de Datos Personales (Ley Orgánica 15/1999, de 13 de diciembre), cuyo objetivo es “garantizar y proteger, en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su honor, intimidad y privacidad personal y familiar”.

Para iniciar, debido a que Twitter es responsable de la recopilación de datos personales de los usuarios, la entidad está obligada a cumplir con las normativas aplicables dentro de su marco de actuación. Para el caso de los usuarios que residen en Colombia, la sede (*Twitter International Company*) encargada del tratamiento de los datos está ubicada en Dublín (Irlanda), por lo cual se reglamenta bajo el Reglamento General de Datos Personales de la Unión Europea (GDPR por sus siglas en inglés).

El reglamento (adaptado a la era digital) exige, entre otros aspectos, contar con un consentimiento por parte del interesado para acceder y tratar sus datos. Procedimiento que se satisface mediante la aceptación de los términos de servicio, que se realiza en el momento de registrar un perfil en la aplicación, en la que, además de conceder los permisos para el tratamiento, se dan los de reutilización con diversos fines (Orduña, 2020).

En efecto, al darse de alta en la red “otorga a Twitter una licencia a nivel mundial no exclusiva, libre de pago de derechos (con derecho a sublicencia) para usar, copiar, reproducir, procesar, adaptar, modificar, publicar, transmitir, mostrar o distribuir” el contenido publicado “en todos y cada uno de los medios de comunicación o métodos de distribución posibles, conocidos ahora o desarrollados con posterioridad”².

Así que Twitter manifiesta de manera explícita que la actividad de sus usuarios es pública y, a no ser que se trate de una publicación protegida, cualquier persona del mundo puede acceder al perfil y las publicaciones, con la posibilidad de reutilizar este contenido por terceros ajenos a la compañía con distintos fines y con restricciones únicamente referentes al volumen de información que se deseé descargar.

Para extraer los datos, Twitter dispone de la aplicación en sí misma y de las interfaces de programación de aplicaciones (APIs), las cuales, en términos generales “son la forma en que los programas informáticos “hablan” entre sí para solicitarse y enviarse información”³. Su uso se encuentra regulado por los términos del desarrollador (Developer terms)⁴, donde se restringe el uso inadecuado de los datos extraídos para conocer información sensible de sus usuarios como datos de salud, estados financieros o condiciones negativas, aficiones políticas, creencias, entre otras. Por lo tanto, en el presente trabajo se estudian las condiciones positivas que hace que los usuarios sean identificados como una categoría de riesgo bajo.

Por otro lado, mientras que a nivel europeo las actividades de elaboración de perfiles (“profiling”) (definida en el Artículo 4 del RGPD como “toda forma de tratamiento automatizado de datos personales consistente en utilizar datos personales para evaluar determinados aspectos personales de una persona física, en particular para analizar o predecir aspectos relativos al

² <https://twitter.com/es/tos>

³ <https://help.twitter.com/es/rules-and-policies/twitter-api>

⁴ <https://developer.twitter.com/en/developer-terms>

rendimiento profesional, situación económica, salud, preferencias personales, intereses, fiabilidad, comportamiento, ubicación o movimientos de dicha persona física") se encuentran reguladas en el Artículo 22 del reglamento, a nivel nacional, se evidencia la ausencia de normativas específicas sobre el uso de datos de redes sociales y perfilamiento. Sin embargo, existen algunas leyes que regulan los derechos a la intimidad, el buen nombre y al habeas data (Rendón, 2014).

La principal es la Ley 1581 de 2012, la cual constituye las disposiciones generales en lo que respecta a la protección de los datos personales en el país. Dicha ley salvaguarda el derecho del propietario de los datos a conocer, actualizar y rectificar sus datos personales y es aplicable a todas las bases de datos que contengan esta información de personas naturales. Por tanto, la realización de este trabajo se debe regir bajo esta normativa y seguir los principios para el tratamiento de datos que se consignan en el Artículo 4:

- Legalidad: sujeto a lo que diga la ley
- Finalidad: sujeto a una finalidad legítima de acuerdo con la Constitución y la Ley
- Libertad: sujeto a consentimiento previo, expreso e informado del Titular
- Veracidad: información veraz, completa, exacta y actualizada
- Transparencia: informar al Titular acerca de la existencia de datos que le conciernen
- Acceso y circulación restringida: sujeto a los límites de los datos personales
- Seguridad: los datos deben ser resguardados de forma efectiva
- Confidencialidad: garantizar la reserva de la información

Adicionalmente, la normativa aborda otro punto importante referente a los niños, niñas y adolescentes, quienes, a pesar de contar con un perfil registrado, no están habilitados para dar su consentimiento del tratamiento de sus datos y son protegidos por la normativa en defensa de sus derechos fundamentales. Por esta razón, si durante este trabajo se identifica que se está tratando una publicación de un menor de edad, esta se extraerá de la muestra.

2.6 Marco ético

Dado que el contrato de los términos y condiciones de uso de redes sociales es elaborado unilateralmente y para acceder a sus servicios, el usuario debe aceptarlo en su totalidad (sin tener opción de modificar aquellas con las que no esté de acuerdo), es posible asumir que se encuentren cláusulas abusivas. Lo cual pone en peligro el derecho a la privacidad de las personas. Por tanto, resulta imprescindible revisar la actuación ética en el que se enmarca el uso de datos recopilados de redes sociales y tratamiento para realizar ‘profiling’.

Uno de los sucesos relacionados más recientes es el de Cambridge Analytics, entidad que, durante las elecciones presidenciales de Estados Unidos del año 2016, desarrolló un algoritmo de perfilamiento de usuarios de Facebook, con el fin de ofrecer una publicidad personalizada para manipular su intención de voto, sin consentimiento de las personas y saltándose todas las aprobaciones éticas.

A causa de esta situación enmarcada dentro de la economía digital emergente, se crean tendencias más rigurosas centradas en los derechos y control sobre la utilización de datos del Titular, promoviendo las prácticas éticas. Uno de los pioneros en este tema ha sido el gobierno del Reino Unido, quienes establecieron el Consejo de Ética en Ciencias de Datos (Council of Data Science Ethics), permitiendo que el Gabinete Ministerial publicara un “Marco Ético de Ciencia de Datos” (Cañón, et al., 2017) bajo el cual se establecen seis principios éticos que servirán como guía para el presente trabajo:

- (i) empezar con un análisis de las necesidades y del beneficio público,
- (ii) usar los datos y herramientas que tengan un mínimo nivel de intrusión,
- (iii) crear modelos robustos de ciencias en datos,
- (iv) estar alerta de percepciones públicas,
- (v) contar con la mayor apertura y rendición de cuentas que sea posible y
- (vi) mantener los datos lo más seguro posible

En el siguiente cuadro se resume el análisis de las necesidades del sector asegurador, el cual se ha desarrollado con la información presentada a lo largo de este capítulo:

Tabla 4 Análisis de las necesidades y del beneficio público

Necesidades sector asegurador colombiano	Beneficios públicos
Segmentación de los consumidores de seguros	Tarifas más exactas, conforme a la condición del riesgo
Aumento de ventas de seguros de vida	Disminución de la tarifa para incentivar la venta en el país
Incremento del índice de penetración	Alternativas innovadoras para reportar el estado de salud, que fomenten el interés de los consumidores

Fuente: Elaboración propia

Se percibe que el consumidor de seguros es principal beneficiario al obtener un descuento en la prima de seguros de vida. De modo que, con el uso de los datos se generan beneficios para el público que van más allá de intereses monetarios. Por otro lado, es importante resaltar que con el desarrollo de este producto no se pretende ocasionar prejuicios o discriminación por la segmentación realizada, por lo contrario, se busca incentivar a los usuarios a tener un estilo de vida y hábitos más saludables. También, se debe mencionar que la implementación no supone selección del riesgo por parte de las aseguradoras.

Para el caso de las herramientas de extracción de datos, se utilizará aquellas que permitan acceder únicamente a la información que el usuario marca como pública para garantizar un nivel mínimo de instrucción en los contenidos generados. En cuanto a los numerales *iii* y *v*, se tiene como premisa que “los algoritmos deben ser desarrollados por y para los ciudadanos de manera abierta,

legítima y monitoreada, asegurando la igualdad de las personas y el control sobre sus datos” (Buenadicha & Galdon, 2019) y por tanto estarán disponibles para los interesados. Por su parte, el punto *iv* no se tendrá en cuenta puesto que es un trabajo académico.

Para finalizar, se reitera la importancia de integrar los derechos humanos al desarrollo tecnológico, para generar confianza y transparencia entre los actores del ecosistema de Big Data. Así como contemplar las prácticas éticas para el uso de Big Data que proporciona el Departamento Nacional de Planeación para impulsar el “emprendimiento de datos”, asegurando que no se generen infracciones de seguridad ni filtración de datos personales (Cañón, et al., 2017). Y de este modo, establecer un balance entre la protección de privacidad y la innovación, que permita la confianza de los consumidores, sin impedir los beneficios que los datos recolectados aportan (Keller, 2018).

3 Búsqueda y definición de fuentes de datos

3.1 Twitter

Para entrar en contexto es conveniente identificar las principales características de Twitter. Esta plataforma, creada en el 2006, es una red social que actúa como un servicio online de comunicación, limitado a 280 caracteres. Dentro de sus funcionalidades se encuentra la creación de cuentas, publicación de tweets (mensajes instantáneos), mención de otros usuarios, retweets de publicaciones ajena, citaciones, inclusión de la opción me gusta, entre otros. Así, mediante las diferentes interacciones la aplicación recopila y usa la siguiente información para proporcionar, comprender y mejorar su servicio:

Tabla 5 Información que Twitter recopila y usa

Información	Descripción
Básica de cuenta	Nombre, nombre de usuario, contraseña
De contacto	Dirección de correo electrónico o número de teléfono
Adicional	Agenda de contactos
Tweets, gente que sigue, listas, perfil y otra información pública	Información publicada en la red social
Mensajes directos y comunicaciones no públicas	Conversaciones privadas con otros usuarios de Twitter
Ubicación	Ubicación en los Tweets y en el perfil de Twitter
Enlaces	Interacciones con enlaces
Cookies	Cookies de sesión y cookies persistentes
Uso del Servicios	Interacciones con los servicios, incluso si no ha creado una cuenta
Datos de widgets	Datos de registro: página web que usted ha visitado y cookie que identifica su navegador
Servicios de comercio	Información de pago: número de su tarjeta de crédito o débito, fecha de vencimiento, código CVV
Publicidad	Mediciones de la efectividad de la publicidad
Terceros y afiliados	Información sobre usted de terceros, como otros usuarios de Twitter, socios

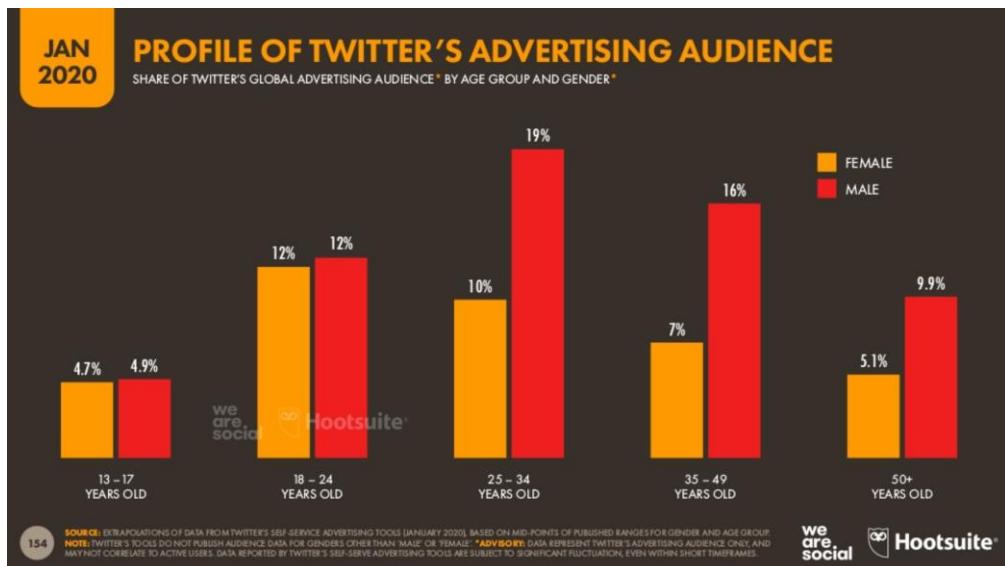
Fuente: Twitter⁵

A nivel mundial, diariamente se generan alrededor de 700 millones de tweets (a junio de 2019) y mensualmente ingresan a la plataforma 4.000 millones de usuarios (a mayo de 2019) (Orduña, 2020). Con respecto a su posicionamiento en el país, el Ministerio de Tecnologías de la Información y Comunicaciones (MinTIC) dio a conocer en el 2019 que Colombia se destaca por ser uno de los países con más usuarios en redes sociales en la región. Siendo Twitter uno de los más populares entre los colombianos, con aproximadamente 3,2 millones de usuarios (Statista, 2020).

⁵https://twitter.com/es/privacy/previous/version_12

En cuanto a los datos demográficos de sus usuarios, se conoce que, el rango de edad de los usuarios donde se concentra más de la mitad de la población total de Twitter es de los 18 hasta los 49 años. Esto conforme a lo expuesto por las empresas We are Social y Hootsuite, en su estudio Digital 2020 Global Digital Overview (January 2020) v01. En la Ilustración 3 se muestra la distribución por edades y género observada en el estudio:

Ilustración 3 Perfil de audiencia de Twitter



Fuente: (We are social; Hootsuite, 2020)

Teniendo en cuenta estas características y cifras, Twitter se convierte en una fuente de información valiosa para encontrar los datos apropiados para realizar el análisis propuesto. Esencialmente por tres razones: recopila la información sobre las interacciones de los usuarios, cuenta con una gran presencia en el país y las edades de su mayor segmento coinciden con las de los consumidores de seguros de vida.

3.2 Herramientas de extracción de datos de minería social

Actualmente, existen una variedad de herramientas para la extracción de datos de Twitter. Por efectos de tiempo, para el trabajo se evalúan tres de las más utilizadas: Twitter Archiver, Python y Knime. A continuación, se exponen las principales características de cada una, con el fin de seleccionar la más conveniente para realizar el minado de información.

3.2.1 Tweet Archiver

Tweet Archiver es el complemento de Google que permite configurar y exportar una búsqueda de tweets, en una hoja de cálculo de Google Drive. Para hacerlo es necesario contar con una cuenta en Google y otra de Twitter. La información que se puede obtener mediante su uso es muy completa ya que detalla las interacciones de los usuarios. A continuación, se listan las columnas más relevantes que se pueden obtener mediante su uso:

- Fecha y hora de publicación
- Nombre completo de la persona que tuiteó
- Texto del Tweet
- Identificación del Tweet
- Tipo de sistema por el cual había subido los datos
- Número de seguidores que tiene el usuario
- Número de retweets
- Ubicación del usuario
- Biografía del usuario, donde se define a sí mismo

Ilustración 4 Resultado Twitter Archiver

Twitter Query: lang:es -filter:retweets -filter:replies								
Date	Screen Name	Full Name	Tweet Text	Tweet ID	Link(s)	Media	Location	Retweets
03/03/2020 13:11	@PostDataNoticia	PostData Noticias	Pide Violeta Lagunes auditoría al SOPAMA de Atlixco	1234949479821824000	http://www.postdatacomunicac https://pbs.twimgimg.com/media/ESNsAdOUQ			
03/03/2020 13:11	@jessica_carep	Bonnie	Juro solemnemente que mis intenciones no son buen	1234949477640757248				
03/03/2020 13:11	@LeettyG	Letty 🎉	Cúñate de los serios porque son los que salen con m	1234949476738981888				
03/03/2020 13:11	@ana_pea3	A 🌸	Realmente te extraño mucho perdón	1234949475782672384				
03/03/2020 13:11	@83742sanchez	Esquivel 83742sanchez	Este 2020https://www.instagram.com/p/B9Ng4lRbqP	1234949475715529904				
03/03/2020 13:11	@inxsmirla	anxiety prime	12 días no names	1234949474524397569				
03/03/2020 13:11	@Moar_AB	Adrien	Flume para mí ya está muerto desde que presencié a	1234949474478247936	https://twitter.com/fuemusic https://flu.me/TheDifferencePres			
03/03/2020 13:11	@priscylacr	Priscyla Castro	Llevas mi vida en tu bolsillo @natsnisino	1234949473924567041				
03/03/2020 13:11	@emmanue12832	emmanuel	No sé porque trato de ignorarte siempre si realmente	1234949473899401216				
03/03/2020 13:11	@dalumMG	Dalu Mtz	Estoy harta, harta, hartísima	1234949473198956544				
03/03/2020 13:11	@ramoncota	entenie	a todos nos vale verga la cantidad de droga que te me	1234949472972505088				
03/03/2020 13:11	@lilek097	Bandman4eva 🎵	yo imagine	1234949472880226305				
03/03/2020 13:11	@virkat5	virginkat	VtLww.JAJAJAJA puta ascoooo RT @auronplay: si yo he	1234949472775327744	https://twitter.com/auronplay/status/1234941538976878593 https://twitter.com/fernando8550123/status/1234937527259344901			
03/03/2020 13:11	@KimberlyOsuna	Kim pero no Kardashian	Estoy triste y me río, el concierto está lleno pero yo es	1234949471844220928				
03/03/2020 13:11	@lessjustelune	Jessi Pink 🌸	Un hombre real es aquél que tiene el abdomen marcí	1234949470988558336				
03/03/2020 13:11	@Jacquelineval2	आशाकृतिन	Mi dia estuvo pésimo, solo quiero llegar, acostarme y	1234949469814181888				

La herramienta tiene dos versiones: una gratuita y otra de pago. La primera, muestra actualizaciones cada hora y permite hacer búsqueda de únicamente dos hashtags, mientras que la segunda, se actualiza cada 5 minutos y múltiples entradas (De-Gracia).

3.2.2 Python

Python es el lenguaje de programación que se destaca por tener una sintaxis de código legible y capacidad para manejar grandes volúmenes de datos. Mediante el uso de la librería *tweepy* y la interfaz OAuth, es posible realizar un minado de datos haciendo uso del API de Twitter, la cual funciona como un punto de entrada a la aplicación y permite extraer e imprimir los tweets, seguidores y demás atributos haciendo uso del formato JSON (Bonzanini, 2016) .

Ilustración 5 Librería Python para minería de datos

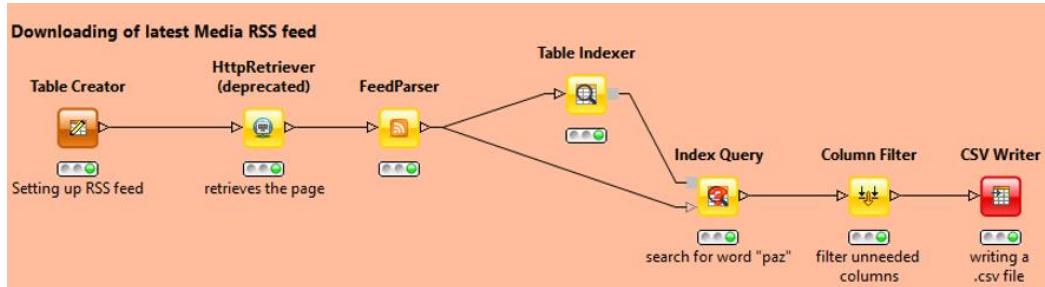
```
import tweepy
from tweepy import OAuthHandler
```

Funciona de forma gratuita y para realizar la extracción es necesario instalar el software Anaconda y contar con conocimientos de programación para lograr manipular la información descargada.

3.2.3 Knime

Knime (Konstanz Information Miner)⁶ es la plataforma de minería de datos desarrollada en Java, que cuenta con un entorno visual intuitivo, el cual funciona por medio de un modelo de nodos que secuencian las actividades de procesamiento de datos. Para realizar la extracción, utiliza la API de Twitter para configurar la conexión a la aplicación y ejecuta la búsqueda por medio queries que permiten configurar los filtros requeridos. Es un software de uso gratuito.

Ilustración 6 Nodos Knime



Fuente: (Torres, L. 2016)

3.3 Selección de la herramienta

Tras evaluar las diferentes herramientas, se realiza una comparativa teniendo en cuenta las características funcionales y de usabilidad, las cuales se resumen en la Tabla 6:

Tabla 6 Comparativa de herramientas de extracción de datos

Herramienta	Precio	Conectividad con Twitter	Requiere instalación de programas	Conocimientos de programación	Acceso a los datos
Tweet Archiver	Gratis	Integrada	No	No	Hoja de cálculo
Python	Gratis	API	Sí	Sí	JSON
Knime	Gratis	API/Nodo	Sí	Sí	CSV

Fuente: Elaboración propia

⁶ <https://www.knime.com/>

Por lo tanto, se determina que la herramienta empleada para el minado de datos en el presente trabajo es Tweet Archiver. Siendo los factores que mayormente inciden en la decisión la ausencia de instalaciones de software adicional, la conectividad integrada con Twitter y la interfaz amigable al usuario.

3.4 Extracción de datos

Para realizar la extracción de datos sin que estos presenten algún tipo de sesgo por un tema de tendencia en un momento concreto, se utiliza la técnica de la semana reconstruida. La cual consiste en extraer datos del lunes en la semana 1, del día martes en la 2 y así sucesivamente hasta reconstruir la semana completa. De este modo, los únicos filtros aplicables a los datos son idioma (español) y localización (cada uno de los departamentos de Colombia). Los tweets utilizados para este análisis se descargaron desde el jueves de la primera semana de marzo, hasta el miércoles de la tercera de abril del 2020. Se obtuvo un volumen de aproximadamente 19,000 registros.

Adicionalmente, para facilitar el entrenamiento del clasificador se ejecutan búsquedas con palabras relacionadas de manera directa con cada uno de los factores no genéticos que se consideran para realizar la segmentación del riesgo. En la Tabla 7 se muestran las palabras clave usadas en la búsqueda:

Tabla 7 Palabras clave de búsqueda

Factores no genéticos	Palabras clave
Tabaquismo	Cigarrillo, fumar, tabaco, nicotina, fumador
Hábitos alimenticios	Ensalada, comida, comer, vegano, cocinar, hamburguesa, pizza
Actividad física	Ejercicio, deporte, entrenamiento, crossfit, gimnasio, gym, fit
Estrés	Miedo, tensión, ansiedad, estrés, alegre, feliz, tranquilidad, depresión

Fuente: Elaboración propia

La extracción de los tweets con estos filtros se efectuó en la segunda semana de marzo y se adquirió un aproximado de 4,400 registros. El proceso para realizar la extracción de tweets con la herramienta Tweet Archiver se detalla en el Anexo B. Extracción de datos.

3.5 Variables seleccionadas

Para aplicar la política de descuento de la prima del seguro de vida, es necesario identificar los elementos que posibilitan la determinación de la categoría de riesgo al que el usuario pertenece, es decir, se precisa reconocer los factores no genéticos (alimentación saludable, práctica de actividad física, no fumar y niveles del estrés) entre sus publicaciones. Así pues, tras analizar los contenidos de las variables, se eligen las siguientes entradas para el modelo:

Tabla 8 Entradas del modelo

Variable	Tipo de dato	Descripción
Full Name	No estructurado	Nombre completo del usuario que tuiteó
Tweet Text	No estructurado	Texto de la publicación
Tweet ID	Estructurado	Identificador del Tweet
Location	No estructurado	Ubicación desde donde se realizó la publicación
Bio	No estructurado	Biografía del usuario

3.6 Análisis descriptivo de tweets

Con el fin de realizar un análisis descriptivo del corpus, se procede a realizar una representación visual de la frecuencia de las palabras de las publicaciones de Twitter, por medio de una nube de palabras generada a partir de la librería `wordcloud` de Python. Omitiendo las palabras monosílabas que no contribuyen a la visualización, se obtiene la Ilustración 7:

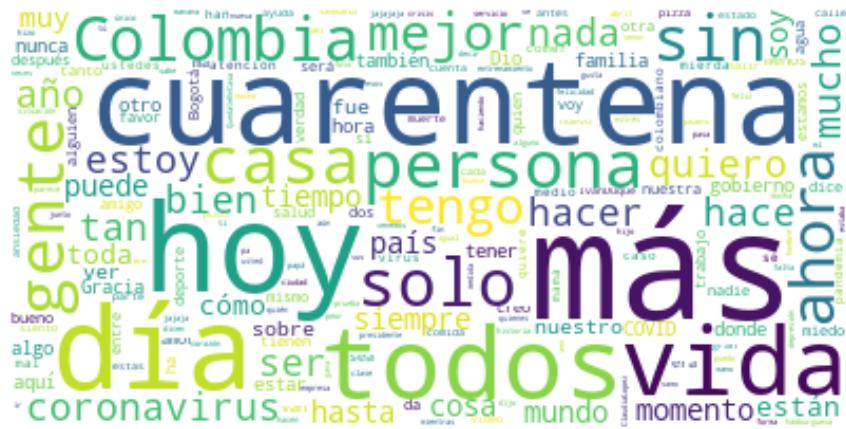
Ilustración 7 Nube de palabras del corpus



Se evidencia que, a pesar de utilizar el método de la semana reconstruida, existe un sesgo en los tweets debido a la pandemia del coronavirus que ocasionó que las personas estuvieran en cuarentena durante el periodo de extracción, siendo esta temática la más representativa dentro de las publicaciones. Por otro lado, se muestra que la información no fue sesgada por las palabras clave empleadas.

Adicionalmente, para evaluar la consistencia de la información se realiza el mismo procedimiento reduciendo aleatoriamente el 40% de los registros. Los resultados se muestran en la Ilustración 8:

Ilustración 8 Nube de palabras del corpus reducido



Se observa que los resultados son muy similares a los obtenidos con el dataset completo, permitiendo concluir que los datos son una buena representación que se mantiene equilibrada, aunque se efectúe una reducción de los datos.

4 Limpieza de datos

Los campos en los que el usuario de Twitter interviene de forma no estandarizada son: el texto del tweet, la ubicación y la biografía. Ahora bien, dadas las características del modelo las variables texto del tweet y biografía son tomadas de la fuente primaria, sin requerir limpieza de los datos (siempre que tengan contenido, de lo contrario se excluyen para la elaboración del modelo). Sin embargo, la ubicación cobra importancia para realizar la validación del modelo por departamentos.

Por lo tanto, este capítulo se dedica a realizar la limpieza y preparación de datos referentes a la ubicación de la publicación. Así pues, con el fin de asegurar la calidad de dichos datos, se llevan a cabo las etapas descritas en la Ilustración 9:



4.1 Herramienta

En la actualidad se han desarrollado diversas herramientas de limpieza y depuración de datos como: Trifacta, OpenRefine, Astera Centerprise, Paxata, así como los paquetes (bibliotecas) Dplyr para R o Pandas para Python. Analizando las distintas funcionalidades y el coste de aprendizaje de los diferentes aplicativos mencionados, se opta por emplear OpenRefine.

OpenRefine⁷ (antiguamente Google Refine) es un software de uso libre y abierto basado en Java, que permite la visualización y tratamiento de grandes volúmenes de datos. Esta herramienta se destaca por facilitar el proceso de limpieza, normalización y transformación de datos, permitiendo su exploración y vinculación. El aplicativo funciona como una base de datos relacional, es decir, con columnas y campos en lugar de celdas individuales, sin embargo, su apariencia es similar a una hoja de cálculo (Piña, 2018).

Otra característica destacable es que, a pesar de ser un software libre y utilizar un buscador web como interfaz, tiene políticas de datos privados y no reutiliza la información subida en el aplicativo, a no ser que otorguen los permisos para ello. Adicionalmente, es posible replicar las acciones que se realizan en una base, en otro conjunto de datos, gracias a su funcionalidad de almacenamiento.

⁷ <http://openrefine.org/>

La descarga se realiza desde la página web: <http://openrefine.org/download.html>, donde se encuentran las versiones más actualizadas del aplicativo. A su vez, como OpenRefine se basa en un entorno Java, es necesario contar con la una versión de Java en el equipo, la cual se encuentra en la página <http://java.com/download>. Para la ejecución del trabajo, se emplea la versión 3.3, lanzada en enero del 2020:

Ilustración 10 OpenRefine 3.3

OpenRefine 3.3

The final release of OpenRefine 3.3, released on January 31, 2020. Please backup your workspace directory before installing and report any problems that you encounter. A change log is provided on [the release page](#).

- **Windows kit**, Download, unzip, and double-click on *openrefine.exe*. If you're having issues with the above, try double-clicking on *refine.bat* instead.
- **Mac kit**, Download, open, drag icon into the Applications folder and double click on it.
- **Linux kit**, Download, extract, then type *./refine* to start.

Fuente: (OpenRefine. 2020)

Al ingresar a la aplicación, se abre automáticamente tanto la interfaz web como la consola de Java, donde se reflejan las acciones que se ejecutan en el aplicativo. Esta debe permanecer activa durante su uso. Para aumentar la memoria se requiere editar la línea Xmx (lenguaje de Java para "maximum heap size") del archivo *openrefine.l4j.ini*. Dicha línea define la cantidad de memoria de OpenRefine en Megabytes, en este caso se modifica de *Xmx1024M* (valor por defecto de 1GB) a *Xmx7168M*, debido al tamaño del archivo de tweets que se manejará.

Ilustración 11 Aumentar de memoria OpenRefine



```
openrefine.l4j: Bloc de notas
Archivo Edición Formato Ver Ayuda
# Launch4j runtime config

# initial memory heap size
-Xms256M

# max memory memory heap size
-Xmx7168M

# Use system defined HTTP proxies
-Djava.net.useSystemProxies=true

##-XX:+UseLargePages
##-Dsomevar="%SOMEVAR%"
```

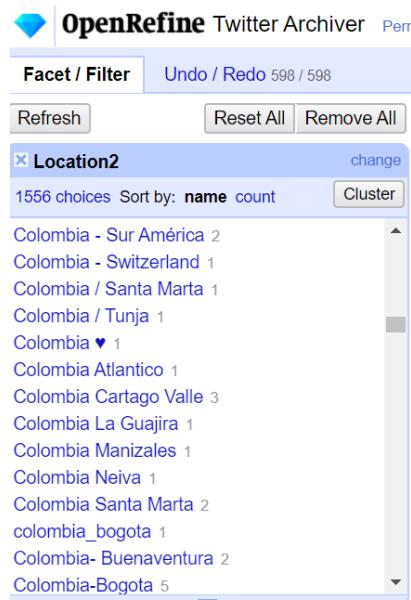
4.2 Limpieza

Para iniciar con el proceso de limpieza y transformación de datos es necesario crear un nuevo proyecto en OpenRefine. Para ello, se selecciona el archivo, se explora y valida el documento mediante una vista previa y por último se permite que el aplicativo cree un nuevo proyecto. Los pasos ejecutados se detallan en el Anexo C. Creación del proyecto en OpenRefine.

4.2.1 Identificación de tipo datos

A pesar de realizar la búsqueda de datos efectuando el filtro por cada uno de los departamentos, como el usuario puede agregar texto libre en el campo ubicación (sin ser verificado ni normalizado) las entradas que recibe la ubicación pueden presentar heterogeneidad y problemas de calidad. Para identificar el tipo de contenido alojado en esta variable se realiza un análisis exploratorio de los datos, utilizando la función “Text facet” de OpenRefine, (presente en la opción “Facet” del desplegable de la columna) permitiendo evidenciar el contenido y la frecuencia de cada registro:

Ilustración 12 Text facet columna Location- OpenRefine



Tras una revisión del listado presentado por el aplicativo se identifican los siguientes errores (o inconsistencias) y algunos ejemplos de la casuística:

Tabla 9 Errores o inconsistencias del campo ubicación

Errores o inconsistencias del campo ubicación	Descripción	Ejemplo (Tweet ID: Location)
Campo en blanco	Registros con la variable vacía	1245024960340537344 1245024811337818119 1245024524636168201
Emoticones	Se presentan emojis dentro del campo	1235652992386334720: La Calera, Colombia ☀️☀️☀️☀️☀️ 1238481739414110208: Bogota-Colombia.
Información indeterminada	Ubicaciones indeterminadas o no desagregadas por departamento	1245024860029489158: En mi casita. 1235651350014373890: el mundo 1244801188761145347: Colombia♥
Países diferentes	Se incluyen países diferentes a Colombia	1238371732639096832: Bélgica 1238406294010290182: Venezuela 1235652983834021888: Panamá, Panamá
Puntuación o variaciones ortográficas	Textos con representaciones alternativas del mismo lugar	1245023385593303042: Atlántico, Colombia 1245023333013508098: Colombia Atlantico
Subdivisiones de departamentos	Se registran ciudades, municipios, pueblos, corregimientos del departamento	1245025284635664384: Zipaquirá, Colombia 1245025115353624583: Cajicá, Colombia 1245027699225214976: Chía, Colombia
Ubicaciones homónimas	Ubicaciones con el mismo nombre, pero que pertenecen a distintos departamentos	1245028207717420000: Sabanalarga, Colombia. (Municipio de Antioquia, Atlántico, y Casanare)
Varios departamentos	Se reporta más de un departamento en la misma celda	1238481402569592835: Medellín, Bogotá 1237580373514579970: Cali-Bogotá 1238421311602470913: Bogotá - Cartagena.

4.2.2 Definición de reglas de mapeo

Para solucionar estas inconsistencias o errores, se plantea tres tipos de decisiones: la primera, aplicada para aquellas filas en las que no es posible identificar el departamento (bien sea porque no pertenece al país, no es claro o se reportan varios lugares) es excluir el tweet en la etapa de validación del modelo, sin embargo, esta fila se tendrá en cuenta en el diseño del modelo, es decir, se evaluará el tweet. La segunda, consiste en eliminar la inconsistencia y la tercera radica en agrupar por departamentos la información. En la siguiente tabla se expone el detalle de las decisiones y la regla que se aplica:

Tabla 10 Reglas de mapeo limpieza de datos

Errores o inconsistencias del campo ubicación	Decisión	Regla
Campo en blanco	Excluir tweet en la etapa de la validación	Sustituir por “No aplica”
Emoticones	Eliminar	Sustituir por “”
Información indeterminada	Excluir tweet en la etapa de la validación	Sustituir por “No aplica”
Países diferentes	Excluir tweet en la etapa de la validación	Sustituir por “No aplica”
Puntuación o variaciones ortográficas	Agrupar	Sustituir por nombre del departamento, con la ortografía adecuada
Subdivisiones de departamentos	Agrupar	Sustituir por nombre del departamento al que pertenece ⁸
Ubicaciones homónimas	Excluir tweet en la etapa de la validación	Sustituir por “No aplica”
Varios departamentos	Excluir tweet en la etapa de la validación	Sustituir por “No aplica”

4.2.3 Transformación

En la fase de transformación se aplican las reglas de mapeo definidas. El primer paso es duplicar la columna “Location” para almacenar la información original y lograr contrastar los resultados de la limpieza. En OpenRefine este procedimiento se realiza usando la opción “Add column base on this column”, disponible en el menú desplegable de la columna:

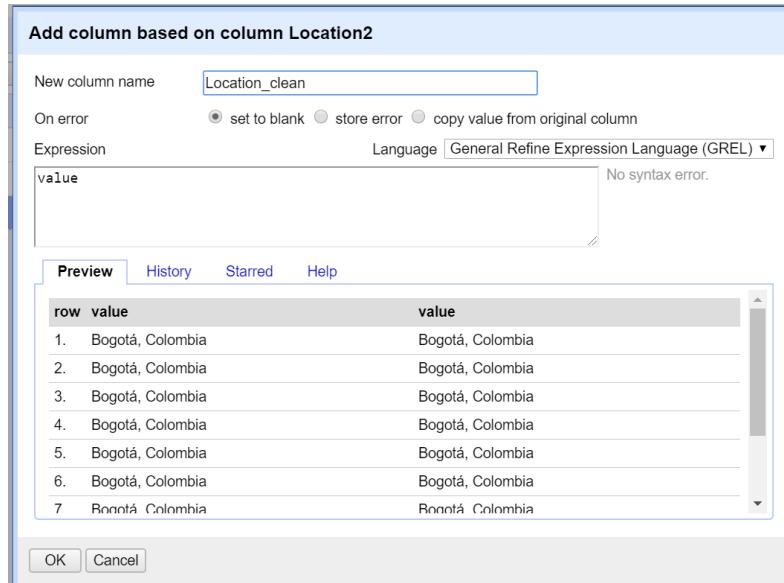
Ilustración 13 Add column base on this column- OpenRefine



⁸ La agrupación por departamento se realiza conforme a la información reportada en el documento oficial de la División Político Administrativa (DIPOLA) disponible en el siguiente enlace: <https://www.minsalud.gov.co/Documentos%20y%20Publicaciones/RUAF%20DIVIPOLA.pdf>

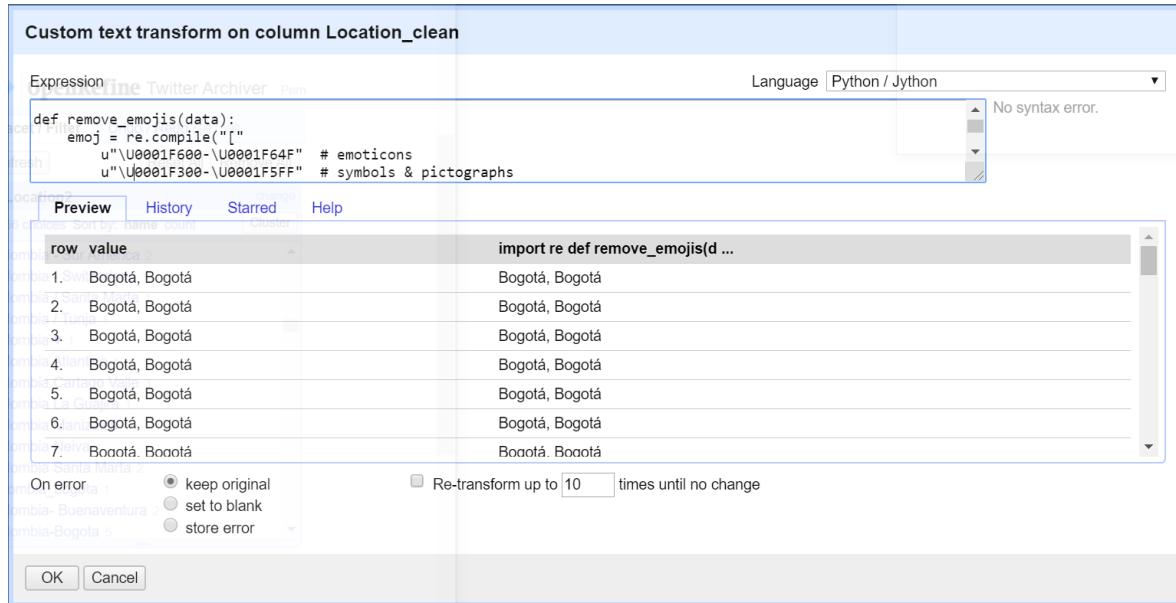
En la nueva columna, nombrada “Location_Clean”, se incluyen los mismos valores de la columna “Location” usando la expresión **value** y posteriormente se da click en ok:

Ilustración 14 Duplicar columnas- OpenRefine



La primera modificación de la nueva columna es eliminar los emojis, para ello se elige la opción “Transform…”, que se encuentra desplegando el menú de la columna y eligiendo el botón “Edit Cells”. A partir de la ventana emergente, se emplea la expresión en lenguaje Jython/Python (declarada en el Anexo D. Script de Jython/Python para eliminar emoticones en OpenRefine) para reemplazar los emoticones encontrados por vacío, permitiendo así manejar con mayor facilidad los datos.

Ilustración 15 Eliminar emoticones- OpenRefine



Ahora bien, con el fin de combinar y estandarizar las redundancias evidenciadas, se utiliza nuevamente la función Text facet (esta vez aplicada a la columna construida) y se da click en el botón Cluster ubicado en la esquina superior del cuadro de facetas:

Ilustración 16 Text facet de la columna Location_Clean- OpenRefine

OpenRefine Twitter Archiver [Permalink](#)

Facet / Filter **Undo / Redo** 3 / 3

140 matching rows (13554 total)

Show as: **rows** **records** Show: 5 10 25 50 rows

Location_clean change [Import from URL](#)

1556 choices Sort by: name count Cluster

	Full Name	Tweet Text	Tweet ID	Link(s)	User Since	Location2	Location_clean
Jorge Cuervo	Teniendo en cuenta datos del @DANE_Colombia solo deberian entrar en cuarentena obligatoria y de responsabilidad los hogares y sitios donde hayan mayores de 70 años. No impactaria tanto la economia.	1242874417602080000				Arauca, Colombia	Arauca, Colombia
Llanera.com	Señora, José Angel, Música llanera un solo llano!	1242873912813330000	https://musica.llanera.com/se%C3%B1ora-jose-angel/62/727			Arauca, Colombia	Arauca, Colombia
alex devia cruz	Izemos Nuestra Bandera en Honor a la Vida. #UNIDOSOMOSMAS	1242873103576950000				Arauca, Colombia	Arauca, Colombia

Mediante esta opción se abre una nueva ventana, donde se muestran las agrupaciones que el programa efectúa de acuerdo con el algoritmo de cluster seleccionado. El algoritmo que se aplica por defecto es el que emplea el método key collision (colisión de teclas) y la función fingerprint, el cual asocia términos que resultan muy similares, salvo la puntuación o variaciones ortográficas:

Ilustración 17 Función Cluster de OpenRefine

Cluster & Edit column "Location_clean"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method key collision ▾ Keying Function fingerprint ▾ 226 clusters found

Cluster Size	Row Count	Values in Cluster	Merge? New Cell Value
15	328	<ul style="list-style-type: none"> • Medellín, Colombia (271 rows) • Medellin, Colombia (18 rows) • Medellin, Colombia. (13 rows) • Medellin - Colombia (6 rows) • Medellin - Colombia (5 rows) • Medellin, colombia (3 rows) • Medellin colombia (2 rows) • medellin - colombia (2 rows) • medellin colombia (2 rows) • COLOMBIA - MEDELLÍN (1 rows) • MEDELLIN COLOMBIA (1 rows) • MEDELLÍN, COLOMBIA (1 rows) • Medellin Colombia (1 rows) • Medellin (Colombia) (1 rows) • Medellín(Colombia) (1 rows) 	<input type="checkbox"/> Medellín, Colombia
11	108	<ul style="list-style-type: none"> • Cúcuta, Colombia (80 rows) • Cucuta, Colombia (8 rows) • Cúcuta - Colombia (8 rows) • Cucuta - Colombia (4 rows) • Cucuta Colombia (2 rows) • CUCUTA COLOMBIA (1 rows) • Colombia - Cúcuta (1 rows) • Colombia - cúculta (1 rows) • Cucuta colombia (1 rows) 	<input type="checkbox"/> Cúcuta, Colombia

Choices in Cluster
Rows in Cluster
Average Length of Choices
Length Variance of Choices

Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Para realizar los cambios propuestos por OpenRefine, se debe marcar con un check la columna “Merge?” y dar click en el botón “Merge Selectec & Re-Cluster”. Este algoritmo resulta útil para asociar y homogenizar los campos, pero también se puede aprovechar para incluir el departamento del cluster identificado, agregando la estructura “Municipio, Departamento” en el campo de texto “New Cell Value”, como se muestra a continuación:

Cluster & Edit column "Location_clean"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method key collision ▾ Keying Function fingerprint ▾ 226 clusters found

Cluster Size	Row Count	Values in Cluster	Merge? New Cell Value
9	263	<ul style="list-style-type: none"> • Cartagena, Colombia (242 rows) • Cartagena - Colombia (9 rows) • Cartagena Colombia (3 rows) • Cartagena, Colombia. (3 rows) • Cartagena (COLOMBIA) (2 rows) • Cartagena (Colombia) (1 rows) • Cartagena - Colombia (1 rows) • cartagena colombia (1 rows) • cartagena, colombia (1 rows) 	<input checked="" type="checkbox"/> Cartagena, Bolívar
9	111	<ul style="list-style-type: none"> • Barranquilla, Colombia (97 rows) • Barranquilla - Colombia (7 rows) • Barranquilla - colombia (1 rows) • Barranquilla Colombia (1 rows) • Barranquilla, Colombia. (1 rows) • Colombia - Barranquilla (1 rows) • Colombia _ Barranquilla (1 rows) • barranquilla (colombia) (1 rows) • barranquilla colombia (1 rows) 	<input checked="" type="checkbox"/> Barranquilla, Atlántico

Choices in Cluster
Rows in Cluster
Average Length of Choices

Para los casos en los que solo se reporte el departamento, de igual forma, se mantiene la estructura “Departamento, Departamento” para lograr trabajar con esta columna más adelante:

6	100	<ul style="list-style-type: none"> • La Guajira, Colombia (91 rows) • La Guajira - Colombia (5 rows) • Colombia La Guajira (1 rows) • Colombia, La Guajira (1 rows) • La Guajira - colombia (1 rows) • La Guajira Colombia (1 rows) 	<input checked="" type="checkbox"/>	La Guajira, La Guajira
---	-----	---	-------------------------------------	------------------------

Así pues, de conformidad con las reglas establecidas para los campos que reportan países diferentes o presenten una ubicación indeterminadas o estén vacíos, se reemplazan por No aplica de la siguiente forma:

7	69	<ul style="list-style-type: none"> • San Cristóbal, Venezuela (58 rows) • San Cristóbal - Venezuela (6 rows) • SAN CRISTOBAL, VENEZUELA (1 rows) • San Cristobal, Venezuela (1 rows) • San Cristóbal - VENEZUELA (1 rows) • San Cristóbal-Venezuela (1 rows) • San cristóbal - Venezuela (1 rows) 	<input type="checkbox"/>	No aplica
6	14	<ul style="list-style-type: none"> • La Vida (5 rows) • La vida (3 rows) • LA VIDA (2 rows) • La vida. (2 rows) • La VIDA. (1 rows) • la vida (1 rows) 	<input type="checkbox"/>	No aplica
3	387	<ul style="list-style-type: none"> • (385 rows) • . (1 rows) • ... (1 rows) 	<input checked="" type="checkbox"/>	No aplica

Después de evaluar las sugerencias de cluster dadas por el algoritmo por defecto, es posible explorar distintas funciones como (n-gram o metaphone3) para continuar agrupando por municipios y departamentos:

Ilustración 18 Keying Functions- OpenRefine

Method	key collision	Keying Function
Cluster Size	Row Count	Values in Cluster
15	328	<ul style="list-style-type: none"> • Medellín, Colombia (271 rows) • Medellin, Colombia (18 rows) • Medellín. Colombia. (13 rows)

fingerprint

fingerprint

ngram-fingerprint

metaphone3

cologne-phonetic

Daitch-Mokotoff

Beider-Morse

Es importante resaltar, que si bien, el software es muy potente para realizar los clusters, se debe revisar con detenimiento las agrupaciones que sugiere, puesto que se pueden encontrar situaciones como la que se expone en la siguiente imagen:

The screenshot shows the OpenRefine interface with the following settings:

- Method:** key collision
- Keying Function:** cologne-phonetic

A cluster for the value "Cucuta, Colombia" is selected, indicated by a blue border. The cluster contains 25 rows, with 8 distinct suggestions listed:

- Popayan Colombia (1 rows)
- Popayán Colombia (1 rows)
- 8 25 • Cucuta, Colombia (8 rows)
 - Caquetá, Colombia (5 rows)
 - Cucuta - Colombia (4 rows)
 - Gachetá, Colombia (3 rows)
 - Cucuta Colombia (2 rows)
 - CUCUTA COLOMBIA (1 rows)
 - Cucuta colombia (1 rows)
 - cucuta-colombia (1 rows)

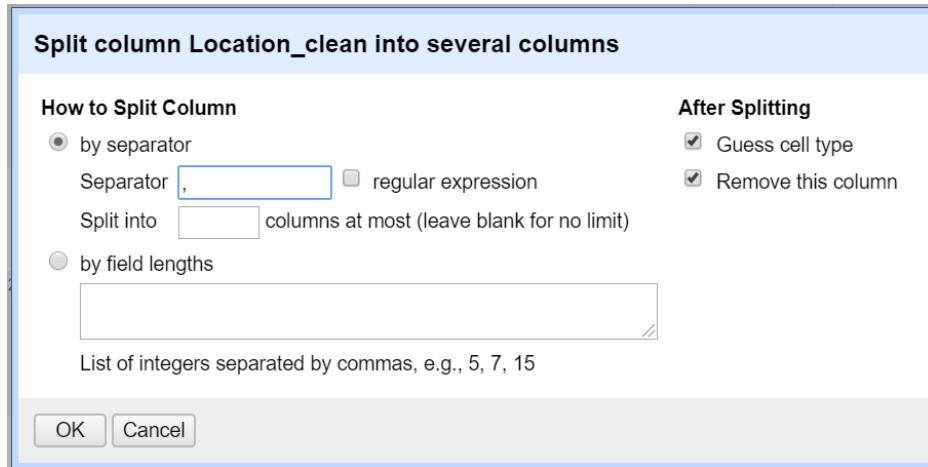
A su vez, se utilizó el método Neighbor (el vecino más cercano), técnica en la que se aplica una comparación de los valores por medio de una función de distancia. Adicionalmente, en los casos con condiciones muy particulares fue necesario realizar las transformaciones de forma manual mediante la función edit, de la faceta.

Por último, cabe destacar que OpenRefine guarda automáticamente todos los cambios realizados y almacena la trazabilidad de las modificaciones efectuadas en el proyecto. Esto con el fin de aplicar este procedimiento a distintas bases de datos. Funcionalidad que ha sido muy útil para conseguir limpiar las extracciones efectuadas a lo largo de las 7 semanas de recolección de datos.

4.2.4 Verificación

Con el fin de comprobar que la limpieza ha sido exitosa, se verifica que la columna generada a partir de las reglas de mapeo definidas en las etapas anteriores cuente con 34 opciones: los 32 departamentos, el distrito capital y la etiqueta “No aplica”. Para conseguirlo, inicialmente, se divide la columna “Location_clean” en dos por el separador “,” empleando la opción “Split into several columns”, ubicada en la función “Edit column” del desplegable de la columna:

Ilustración 19 Opción “Split into several columns”- OpenRefine



Al realizar este procedimiento y generar las facetas en la columna resultante, se identifica que se presentan 39 opciones, las 5 adicionales corresponden a varios departamentos duplicados (diferenciados por un espacio adicional) y datos sin clasificar con la estructura “Municipio, Colombia”. Adicionalmente se encuentran registros en blanco, correspondientes a los clasificados como No aplica:

Ilustración 20 Hallazgos de la verificación inicial- OpenRefine

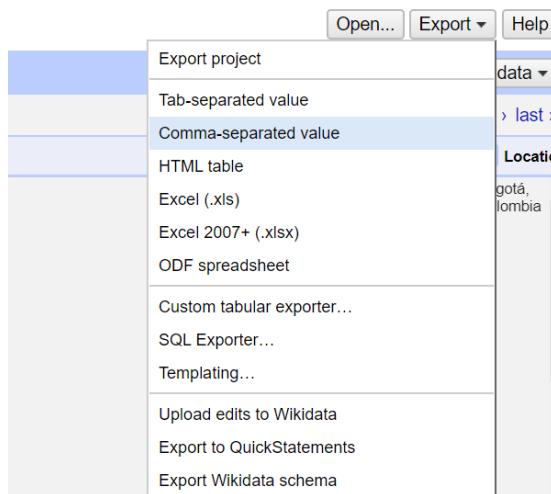
Para corregir los primeros casos se utiliza nuevamente la función cluster y de este modo se asocian los departamentos con diferencias por los espacios. Posteriormente se procede a clasificar los datos faltantes y a incluir la etiqueta No aplica, para los registros en blanco (verificando previamente que no se trate de ningún caso diferente a los clasificados con este texto en la fase anterior). Tras realizar estos procedimientos se comprueba que la columna tiene 34 opciones, de acuerdo con lo esperado:

Ilustración 21 Resultados del proceso de limpieza- OpenRefine

Location_clean 2	
34 choices	Sort by: name count
	Cluster
Atlántico	395
Bogotá	4513
Bolívar	416
Boyacá	364
Caldas	69
Caquetá	88
Casanare	194
Cauca	260
Cesar	279
Chocó	223
Córdoba	262

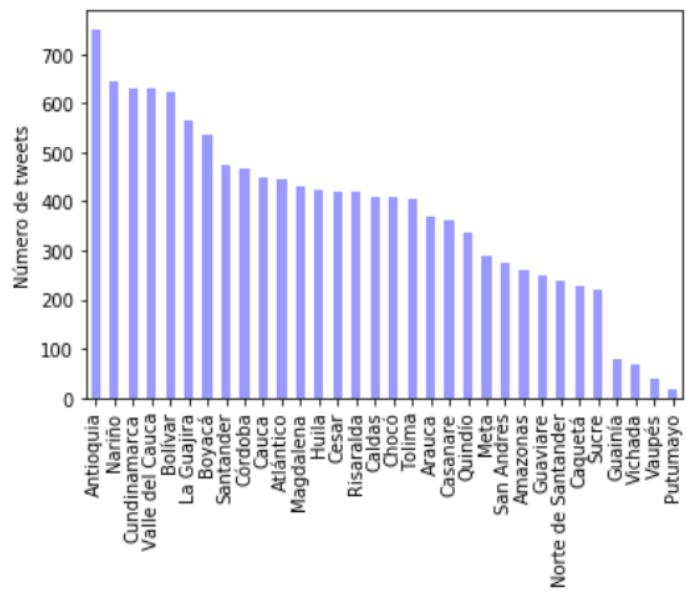
Al finalizar, se exporta el archivo usando el botón “Export” de la esquina superior derecha de la pantalla y se elige el formato deseado como se muestra a continuación:

Ilustración 22 Exportar archivo- OpenRefine



Tras ejecutar la limpieza de los datos, se evalúa la cantidad de tweets resultantes por departamento, identificando que el 27% están etiquetados como “No aplica”, el 21% pertenecen a Bogotá y el restante presenta la siguiente distribución:

Gráfica 2 Número de tweets por departamentos



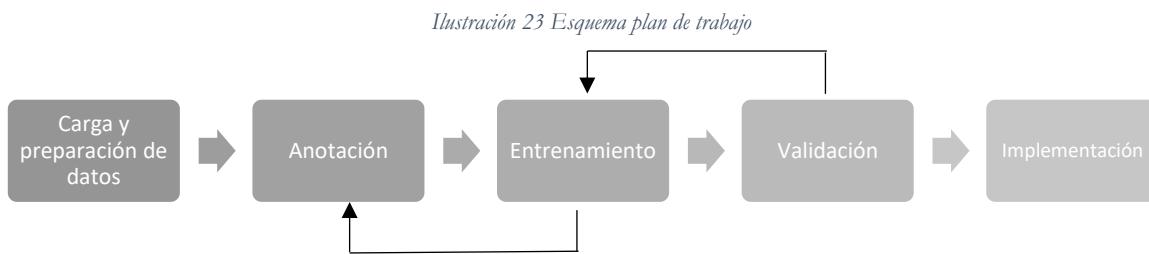
Se evidencia que, después de Bogotá, los departamentos con mayor actividad en Twitter son Antioquia, Nariño, Cundinamarca y Valle de Cauca; mientras que Guainía, Vichada, Vaupés y Putumayo tienen una baja densidad de tweets (inferior a 100), conservando una proporción similar al número de habitantes por población.

5 Diseño del modelo

5.1 Plan de trabajo

El desarrollo del modelo de clasificación de tweets, utilizando aprendizaje automático supervisado con redes neuronales artificiales, se lleva a cabo mediante cinco fases principales: carga de datos, anotación, entrenamiento, validación e implementación. En la primera etapa se cargan y preparan los datos recolectados y depurados en el entorno de clasificación. Posteriormente, se clasifica manualmente una muestra de los datos, otorgando una única categoría a cada tweet con un enfoque supervisado y luego comienza el entrenamiento.

Tras un entrenamiento inicial, se retorna a la etapa de anotación teniendo en cuenta las predicciones del modelo, donde se corrigen o asignan nuevas etiquetas que sirven para entrenar el modelo nuevamente. Seguidamente, se evalúan los resultados aplicando métricas de exactitud, sensibilidad y puntaje F1. A partir de estos datos, se analizan las modificaciones en la configuración del entrenamiento que optimizan su rendimiento y se capacita una vez más. Finalmente, se implementa el modelo mediante la valoración de distintos casos de uso.



5.2 Presupuesto

El presupuesto necesario para el diseño y la implementación del modelo de segmentación de usuarios contempla tres aspectos:

- (1) Honorarios del estudiante: correspondientes al total de horas dedicadas en la ejecución del modelo.
- (2) Honorarios de los tutores: referentes al total de horas dedicadas en el acompañamiento y monitoreo de las actividades de la construcción del modelo.
- (3) Costos de la herramienta: La suscripción al entorno Biome facilitado por Recognai, tiene un precio aproximado de 100 euros al mes. Sin embargo, para la realización de este proyecto el costo no será asumido por tratarse de un uso académico y un acceso gratuito al entorno cloud.

5.3 Herramienta

Para elaborar el modelo se emplea la herramienta de procesamiento de lenguaje natural Biome-text, desarrollada por la empresa española Recognai. Este instrumento de código abierto está construido con AllenNLP y soportado sobre Python 3.6. Mediante su uso es posible: (1) entrenar

un modelo a través de comandos de entrenamiento básico para modelos iniciales o ajustar los entrenados previamente, (2) hacer predicciones sobre un conjunto de datos usando un modelo entrenado y (3) explorar los resultados realizando distintas validaciones del modelo⁹.

Biome-text fue elegida para el desarrollo del modelo debido a la experiencia y dominio de la plataforma por parte de los autores. Adicionalmente, por su funcionalidad para entrenar modelos de clasificación supervisada de manera sencilla y ágil, destacándose en el mercado de herramientas de PLN por (1) los tiempos reducidos de desarrollo, (2) proveer un entorno integrado de diseño de modelos, evaluación y anotación de datos y (3) su especialización en textos cortos, como tweets o registros de cliente. Otras alternativas open source como Keras o FastAI requieren un mayor tiempo de desarrollo e integración, así como la necesidad de combinar dichas librerías con herramientas de anotación propietarias como Prodi.gy¹⁰ o herramientas open source como Doccano¹¹.

5.4 Arquitectura de red neuronal

Para construir el modelo de clasificación se define la arquitectura de red neuronal con una estrategia de cinco pasos. El primero es el uso de vectores de palabras, proceso conocido como *embedding*, que permite “tratar palabras individuales como unidades de significado relacionadas, en lugar de identificaciones completamente distintas” (Honnibal, 2016). Este procedimiento se realiza a nivel de carácter y de palabra, con el fin de otorgar un sistema robusto ante errores ortográficos.

Mediante el *embedding* se genera un ID para cada letra y cada palabra, representado por un vector binario de dimensión 200 y 300 respectivamente:



Fuente: (Honnibal, 2016)

Una práctica habitual para ejecutar este proceso es contar con vectores de palabras pre-entrenados con información lingüística. Para este trabajo, se empleó el modelo disponible para el idioma español, en <https://fasttext.cc>. Este algoritmo, con alto rendimiento y calidad, entrena los vectores de palabras con un corpus compuesto por datos de Wikipedia (enciclopedia gratuita en línea) y Common Crawl (organización sin ánimo de lucro que rastrea la web y publica los datos recolectados) (Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018).

⁹ Biome-text. <https://github.com/recognai/biome-text>

¹⁰ <https://prodi.gy/>

¹¹ <https://github.com/doccano/doccano>

Dicho modelo fue capacitado utilizando una extensión del modelo de CBOW (modelo continuo Skip-gram que realiza un submuestreo de palabras frecuentes para aprender representaciones de palabras más regulares (Mikolov, 2013), que adiciona pesos dependientes de la posición e información de subpalabras teniendo en cuenta la morfología, representando así, las palabras como una bolsa de caracteres n-grams (Bojanowski, Grave, Joulin, & Mikolov, 2017).

El segundo paso es la codificación (*encode*), proceso mediante el cual se facilita la comprensión de oraciones. A partir de la secuencia de vectores de palabras que conforman el tweet, se calcula una representación llamada matriz de oración, donde cada fila representa el significado de cada token en el contexto del resto de la oración (Honnibal, 2016). Para ello se utiliza una Red Neuronal Recurrente (RNN, por sus siglas en inglés) tipo GRU (*Gated Recurrent Unit*) (Chung, Gulcehre, Cho, & Bengio, 2014) de 3 capas, que transforma los vectores de palabras de dimensión 500 (200 por carácter + 300 por palabra), en vectores de tweets de dimensión 300.



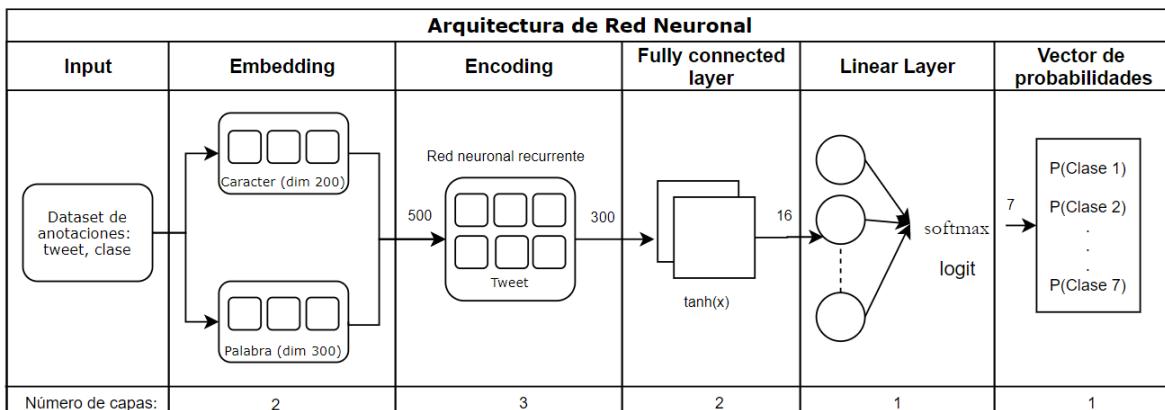
Fuente: (Honnibal, 2016)

Posteriormente, para agregar estas representaciones ricas en contexto a la representación de la oración, se conecta la salida del *encoding* (RNN) con una capa completamente conectada, conocida como *fully connected layer*, que emplea la función de activación de tangente hiperbólica (tanh), para transformar las entradas de la capa anterior en un vector de dimensión 16.

El resultado se conecta con una capa lineal, cuyo tamaño de salida obedece al número de clases, en este caso, 7. Esas activaciones, llamadas *logits*, son un vector de predicciones sin procesar (no normalizadas). Con el fin de normalizar dichos logits, se pasan a una función softmax y de este modo, se pueden interpretar como una distribución de probabilidad, con un valor para cada categoría. La red optimiza la entropía cruzada (*cross entropy*) entre la clase verdadera y esa distribución de probabilidad.

En la ilustración 37 se resume gráficamente el flujo de los tweets dentro de la arquitectura de red neuronal, que determina la categoría a la que pertenece:

Ilustración 26 Flujo de la arquitectura de Red Neuronal



Fuente: Elaboración propia

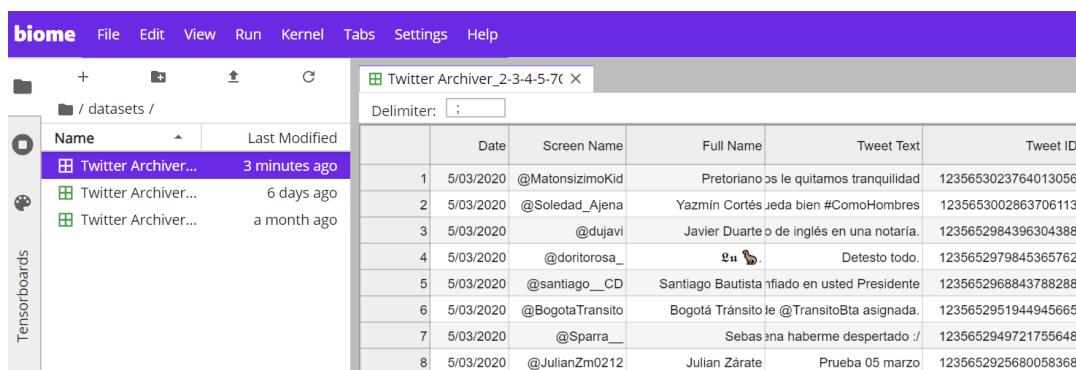
5.5 Desarrollo del modelo

Con el fin de aplicar la arquitectura de red neuronal descrita, se ejecutan las siguientes fases:

5.5.1 Carga y preparación de datos

En la etapa inicial de la construcción del modelo, el archivo CSV resultado del proceso de limpieza de OpenRefine, se carga en el entorno de biome mediante la opción “Upload File”. Para mantener un orden dentro del proyecto global se crea una carpeta denominada “datasets” para incluir los archivos:

Ilustración 27 Carga de datasets en biome



	Date	Screen Name	Full Name	Tweet Text	Tweet ID
1	5/03/2020	@MatonsizmoKid	Pretoriano	os le quitamos tranquilidad	1235653023764013056
2	5/03/2020	@Soledad_Ajena	Yazmin Cortés	jeda bien #ComoHombres	1235653002863706113
3	5/03/2020	@dujaví	Javier Duarte	o de inglés en una notaría.	1235652984396304388
4	5/03/2020	@doritorosa_	guita.	Detesto todo.	1235652979845365762
5	5/03/2020	@santiago_CD	Santiago Bautista	rifiado en usted Presidente	1235652968843788288
6	5/03/2020	@BogotaTransito	Bogotá Transito	le @TransitoBta asignada.	1235652951944945665
7	5/03/2020	@Sparra_	Sebas	una haberme despertado :/	1235652949721755648
8	5/03/2020	@JulianZm0212	Julian Zárate	Prueba 05 marzo	1235652925680058368

Con el fin de preparar los datos de entrada al modelo, se crea un notebook (entorno interactivo de Jupyter Notebook, diseñado para trabajar con Python) y se importa la librería “pandas” usando el siguiente comando:

```
import pandas as pd
```

Dicha librería proporciona a Python una estructura de datos tabular (DataFrame) compuesta por filas y columnas ordenadas¹², además contiene una variedad de funciones para el importe, manipulación y análisis de datos. Posteriormente se crea el data frame ‘df’, el cual contiene la información del archivo cargado en el entorno:

```
df = pd.read_csv('datasets/Twitter Archiver_1-2-3-4-5-6-7ColombiaEspañol.csv', sep=';')
```

A continuación, se muestran las primeras filas del documento:

```
df.head()
```

Ilustración 28 Output df.head()

	Date	Screen Name	Full Name	Tweet Text	Tweet ID	Link(s)	Media	Location	Retwe...
0	5/03/2020	@MatonsizimoKid	Pretoriano	En un país en manos del crimen organizado, don...	1235653023764013056	NaN	NaN	NaN	NaN
1	5/03/2020	@Soledad_Ajena	Yazmín Cortés	Un hombre tan lindo como tú y con tatuajes, es...	1235653002863706113	NaN	NaN	NaN	NaN
2	5/03/2020	@dujavi	Javier Duarte	Otro logro de nuestro subpresidente, ya no nec...	1235652984396304388	https://twitter.com/elespectador/status/123565... https://pbs.twimg.com/media/ESWb1x-XKAASg0Q.jpg			NaN
3	5/03/2020	@doritorosa_	Guacamole	Detesto todo.	1235652979845365762	NaN	NaN	NaN	NaN
4	5/03/2020	@santiago_CD	Santiago Bautista	#YoVotePorDuque yo también voté por @IvanDuque...	1235652968843788288	NaN	NaN	NaN	NaN

5 rows × 21 columns

Seguidamente, se procede a construir y guardar una base de datos (denominada Twitter_1-7_colombia) que contenga únicamente las variables seleccionadas en el numeral 3.5. A saber, Full Name, Tweet Text, Tweet ID, Bio y Location, para conseguirlo se usa el script:

```
df[['Full Name', 'Tweet Text', 'Tweet ID', 'Bio', 'Location_clean']].to_csv('datasets/Twitter_1-7_colombia.csv')
```

Ahora bien, para realizar la creación del proyecto de clasificación es necesario definir los grupos en los que se desea clasificar la información de los tweets y la biografía. Con este fin, se crean las siguientes etiquetas para cada uno de los factores elegidos:

¹²<https://bioinf.comav.upv.es/courses/linux/python/pandas.html>

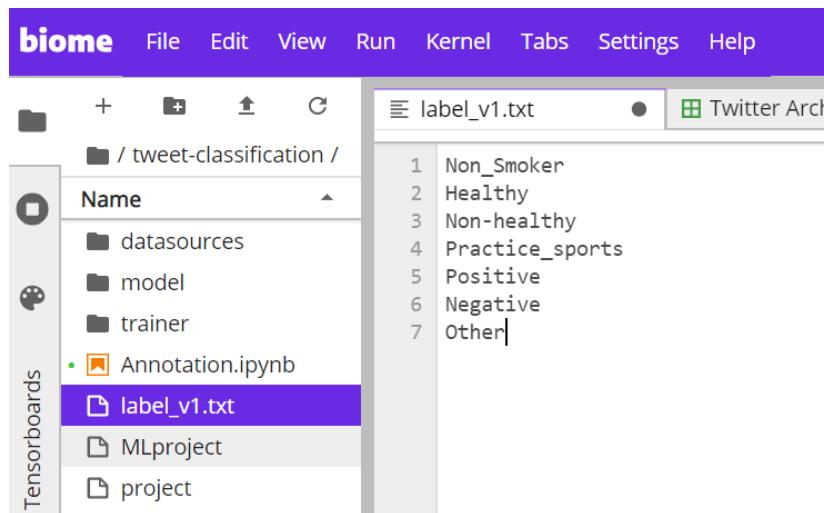
Tabla 11 Etiquetas de clasificación

Factor no genético	Etiqueta	Descripción
Tabaquismo	Non_smoker	La publicación contiene comentarios en contra del consumo de tabaco
Hábitos alimenticios	Healthy	Consumo de alimentos bajos en grasas y sales
	Non_healthy	Consumo de comidas rápidas
Actividad física	Practice_sports	Incluye información sobre sus prácticas de ejercicio
Estrés	Positive	Publicaciones con connotaciones positivas
	Negative	Publicaciones con connotaciones negativas
Otros	Other	Publicaciones que no se puedan clasificar dentro de las categorías previas

Fuente: Elaboración propia

En el entorno, se crea el listado de etiquetas en un txt para que puedan ser leídas y tenidas en cuenta por el clasificador:

Ilustración 29 Etiquetas biome



El proyecto se crea desde un Notebook incluyendo el siguiente comando:

```

!biome.classifier feedback new --help

Usage: biome.classifier feedback new [OPTIONS] NAME

Creates a new feedback session from data

Options:
--source PATH           Source file where the data is
--input TEXT            Mapping definition for fields from data to use as
                        classifier input data [required]
--format TEXT           The datasource format (csv, json,...)
--labels PATH           If provided, make a random annotation using the
                        labels described inside file
--predict-with TEXT     If provided, use the experiment output model for
                        make annotations
--include-package TEXT  additional packages to include
--help                  Show this message and exit.

```

Los atributos del modelo se incluyen mediante el siguiente script, donde se declara el nombre del proyecto, la dirección en la que se encuentra almacenado el archivo desde la que tomará los datos (source), el “token”, es decir, la variable que se desea clasificar y las etiquetas (labels) definidas para realizar las anotaciones:

```

!biome.classifier feedback new tweets-v3 --source ../datasets/Twitter_2-
7_colombia.csv --input tokens='Tweet Text' --labels label_v1.txt

```

A partir de esto, se crea un nuevo proyecto en una interfaz web de biome, desde donde es posible realizar las anotaciones conforme a las etiquetas. Para tener un mayor control de la clasificación, se decide crear dos modelos, uno destinado para la segmentación de tweets y otro para las biografías. Esto se realiza modificando la variable tokens, por ‘bio’. De esta forma se crean los siguientes proyectos:

Ilustración 30 Proyectos biome

Projects

Name ▾

My projects (2) Favorites

tweet-classification
20 days ago

bio-classification
20 days ago

Cada uno de los proyectos contiene las anotaciones creadas. En este caso, se crean varias para trabajar con diferentes archivos que incluyen diferentes días de la semana. Sin embargo, es posible guardar y consolidar los resultados de las diferentes versiones:

Ilustración 31 Anotaciones biome

Explorations Annotations (2)

Search 🔍

Delete Data source name ▾ Model ▾

tweets-v2 4 minutes ago	../datasets/Twitter_2-7_colombia.csv	none	F1: - Recall: - Precision: -	Delete
tweets-v1 20 days ago	../datasets/Twitter Archiver_juevesColombiaEsp...	none	F1: - Recall: - Precision: -	Delete

Dentro de las versiones de las anotaciones se visualizan los datos de entrada del modelo y las posibles etiquetas para asignar, así como un panel con el conteo de los tweets clasificados por cada categoría

Ilustración 32 Panel de anotación en biome

The screenshot shows the biome annotation interface. At the top, there's a purple header bar with the word "biome". Below it, a navigation bar shows "Projects / tweet-classification / tweets-v4". A file path "..../datasets/Twitter_1-7_colombia.csv (18445 Records)" is displayed, along with a "Search records" button and a magnifying glass icon.

Below the navigation, there are several filter and search options:

- "Show labelled records" toggle switch (on).
- Filter by "Labelled as": "Other, Non-healthy...".
- Filter by "Predicted as": "Non_Smoker".
- Filter by "Confidence": a dropdown menu.
- Sort by: "Text A - Z".

The main area displays a list of tweets. Each tweet has a checkbox, a "Label as..." dropdown, and a "Discard" button.

Tweet 1:

- Text: "tokens: ¿Por qué es importante implementar una política pública basada en la clasificación de enfermedades? #ForosSemana y @SiemensColombia, realizarán un conversatorio para debatir sobre este tema. Próximo 25 de marzo en Bogotá. Más información en 📅".
- Labels: Non_Smoker (0%), Healthy (0%), Non-healthy (0%).
- More labels dropdown.

Tweet 2:

- Text: "tokens: @intiasprilla @AlvaroUribeVel @EnriquePenalosa @petrogustavo Camarón, camarón, camarón: con pura 💩💩💩 en la cabeza. Solamente un pobre imbécil puede sacar semejantes conclusiones."
- Labels: Non_Smoker (0%), Healthy (0%), Non-healthy (0%).

To the right of the tweets, there's a summary table:

	4.36%
All	18445
Annotated	804
Other	400
Non-healthy	181
Practice_sports	89
Negative	44
Positive	37
Healthy	34
Non_Smoker	19

5.5.2 Anotación

En esta etapa se realiza la clasificación de un conjunto de tweets de manera manual, con el fin dar una muestra de entrenamiento al clasificador. Para comenzar con este proceso, se recomienda activar la opción “Show labelled records” (ubicada en la esquina superior izquierda), puesto que permite visualizar el contenido del token marcado, de lo contrario, este se desaparece al anotar y su rastreo se dificulta:

Ilustración 33 Opción Show labelled records

Projects / tweet-classification / tweets-v2

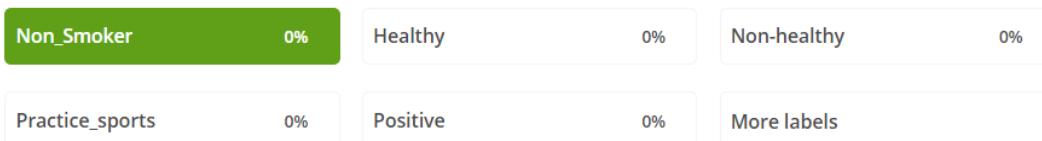
◇ ./datasets/Twitter_2-7_colombia.csv

Show labelled records

Como se puede suponer los datos contienen mucho ruido, así que las anotaciones iniciales se realizan efectuando la búsqueda por las palabras clave de cada uno de los factores no genéticos que se evaluarán, mediante la opción “Search records” de la esquina superior derecha, la cual permite hacer este tipo de filtros dentro de la totalidad del contenido. A continuación se presentan algunos ejemplos de los resultados y de la clasificación efectuada:

Ilustración 34 Ejemplos de anotaciones 1

- tokens: Este aislamiento me hace valorar aún más una de las mejores decisión que he tomado en mi vida: dejar el cigarrillo!



- tokens: Me voy a terminar tatuando las mismas preguntas con la respuesta que a uno le hacen cuando dice que es vegetariano



- tokens: El neurocirujano hablándole a mi mamá: Toca que le dé hamburguesas, pizza, malteada, papas, tocino... Mucha grasa. Mamá: ¿Solo comida chatarra? Yo: no cuestiones, toca hacerle caso al médico.

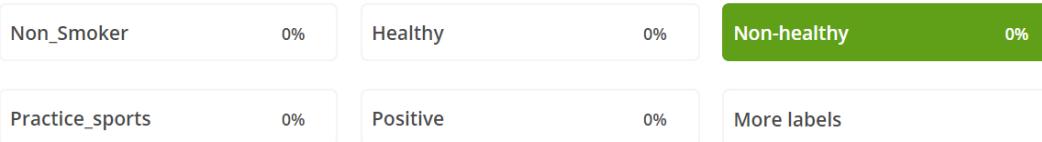
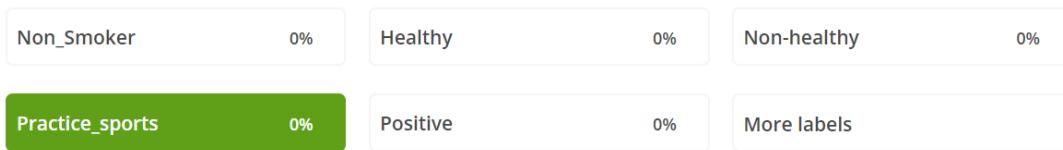


Ilustración 35 Ejemplos de anotaciones 2

- tokens: Entrena siempre para ser motivación #health #fitness #fit #fitnessmodel #workout #cardio #gym #training #healthy #motivation #determination #lifestyle #exercise #duberfitness #duberfit #lovefitness #core...



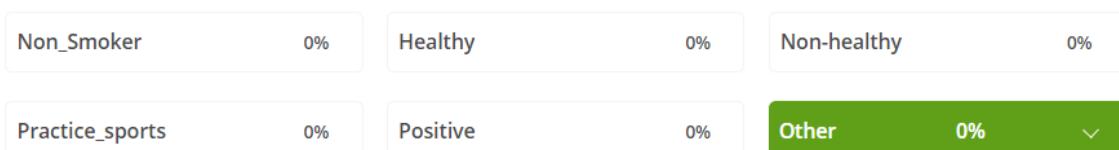
- tokens: #FelizViernesAtodos y tengan calma ante todo lo que nos rodea. ¡Sean felices y cuídense!



- tokens: A mi me va a matar el estrés 😱😱



- tokens: Necesito un abrazo y un cigarrillo... Pero el cigarro no es urgente!



- tokens: Hoy hicimos berenjenas rellenas de arroz, tomate y champiñones, también ensalada de espinaca con fresas. Se me olvidó tomarle foto al plato 😞

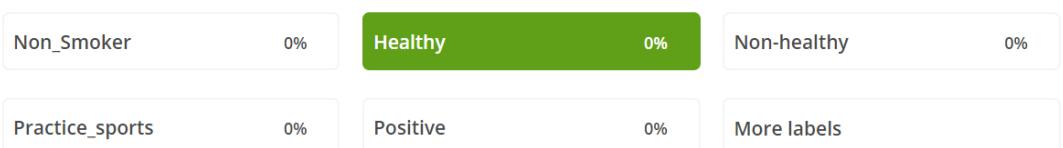


Ilustración 36 Ejemplos de anotaciones 3

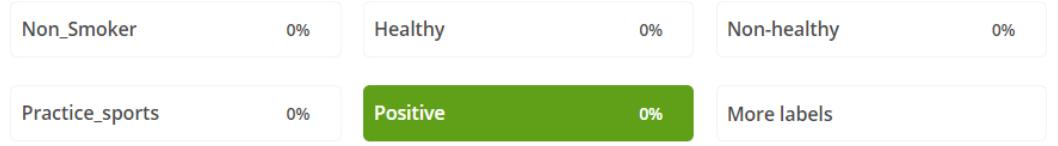
tokens: yo: yo soy una persona que se alimenta muy bien. yo tmnb: desayuna pizza y coca cola.



tokens: Hoy trabajamos una rutina para todas partes, no nos desmotivamos y cada vez vamos para adelante #entrenamiento #entrenamientofuncional #discipline #disciplina #xtremotivation #fitnessmotivation...



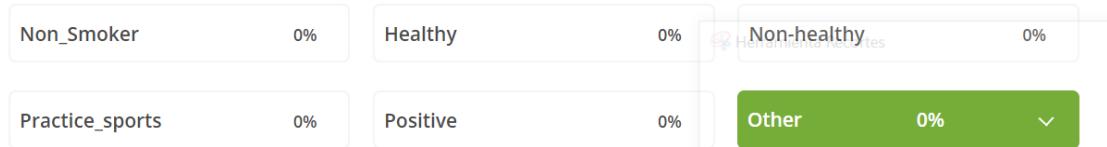
tokens: 2 años de felicidad



tokens: Ya que no quiero existir, me acabo de pedir una pizza para al menos morir sin hambre.



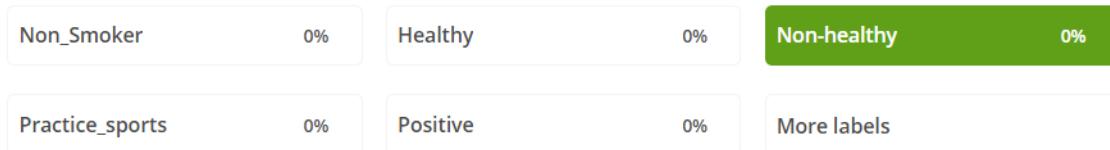
tokens: Contaminación del aire reduce esperanza de vida más que el cigarrillo o el sida



Para los casos en los que se reporten más de una característica, se opta por clasificarlo con la etiqueta que revele la condición negativa:

Ilustración 37 Ejemplo de tweet con varias características

- tokens: Ya no extraño el cigarrillo, pero no dejo de extrañar las empanadas.



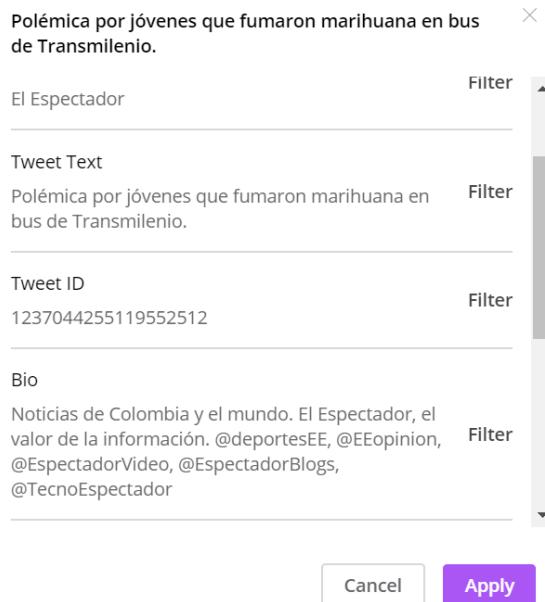
Adicional a las etiquetas designadas, es posible descartar las publicaciones que se consideren, mediante el botón Discard de la esquina superior derecha. Esta opción se emplea para los tweets sin texto, los cuales presentan debido a que el usuario sube contenido multimedia, sin agregar ningún escrito.

Ilustración 38 Ejemplo de tweet descartado



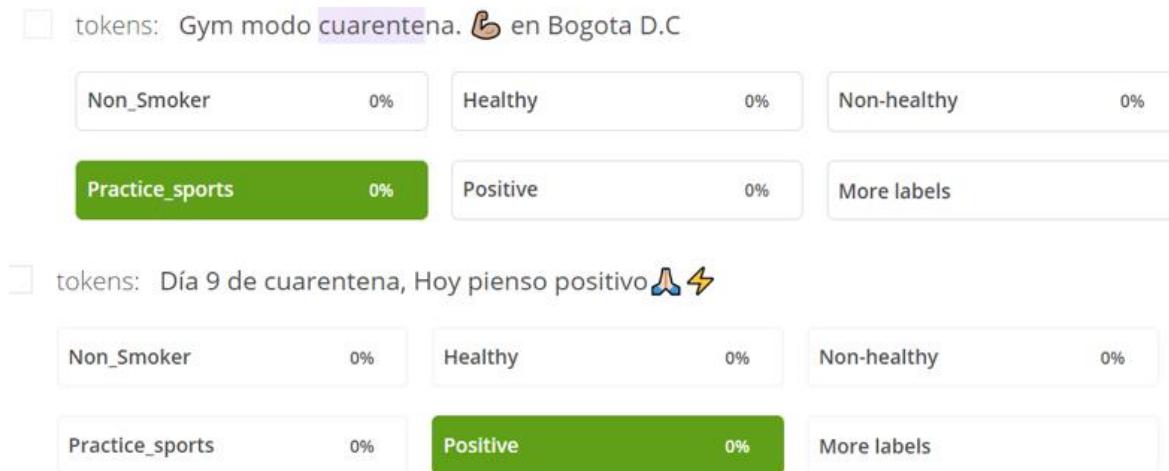
También es posible ver los metadatos de cada tweet y filtrarlos a través del botón “View metadata”, permitiendo observar los datos que acompañan al tweet y realizar un filtro por cada uno de ellos:

Ilustración 39 Botón “View metadata” de biome



Dentro de esta etapa se confirmó que la información presenta un sesgo debido a la cuarentena generada por la pandemia. A pesar de esto, fue posible asignar las etiquetas correspondientes. Se presentan algunos ejemplos:

Ilustración 40 Ejemplos anotaciones con sesgo por cuarentena



Por otro lado, tras una evaluación detallada del contenido de las biografías de los usuarios, se decide no construir un modelo para este dato. Se considera que el modelo resultante usaría datos no balanceados puesto que la mayoría de los datos capturados, reportan su profesión (o estudios) y esto no aporta la información suficiente para describir sus hábitos de consumo de cigarrillo o

alimento, ni prácticas de deportes o nivel de estrés y se clasificarían como “Other”. Seguidamente se señalan algunos ejemplos:

Ilustración 41 Ejemplos de biografías

- tokens: Contadora Pública.
 - tokens: Administrador de Empresas, perseverante y firme!!!
 - tokens: Estudio en la universidad autonoma de occidente COMUNICACION SOCIAL - PERIODISMO

Por último, se consolidan los resultados de las diferentes versiones de anotaciones de tweets y se evalúa la cantidad de anotaciones efectuadas, vigilando que las categorías con mayor número de casos no sesguen a las clases minoritarias y se genere un modelo no balanceado. A continuación se presenta la distribución de las anotaciones iniciales:

Ilustración 42 Distribución de las anotaciones

Other	374
Negative	295
Positive	286
Non-healthy	249
Practice_sports	186
Healthy	108
Non Smoker	26

Se evidencia que la categoría “Non_Smoker”, tiene pocos registros puesto que es poco usual tratar este tema en Twitter. Adicionalmente, se realiza una nube de palabras con los datos anotados para tener un análisis descriptivo de su contenido. Se observa que, efectivamente, se destacan las temáticas relacionadas con las categorías evaluadas, característica inherente de los modelos supervisados:

Ilustración 43 Nube de palabras de las anotaciones



5.5.3 Entrenamiento

Antes de iniciar esta etapa, es necesario dividir el conjunto los datos anotados en dos grupos: el de entrenamiento (train) y el de validación (validation). El primero, correspondiente al 80% de los datos, se emplea para construir el modelo y el segundo (el 20% restante) se utiliza en la fase de validación para comprobar su rendimiento. La división en estas proporciones es una práctica habitual en este tipo de modelos.

Para efectuar esta partición se realiza una división aleatoria utilizando la función `datasource new` de biome, mediante el siguiente comando, donde se señala los datos que se toman (en este caso, las anotaciones realizadas en las versiones 2 y 4) y las proporciones para aplicar el split del dataset:

```
!biome.classifier datasource new v2 --input tokens="Tweet Text" --label label  
--validation 0.2 --test 0.2 --source datasets/annotations-v2-v4-final.csv -  
overwrite  
  
Created train dataset with 1219 examples and 7 unique labels  
Created validation dataset with 305 examples and 7 unique labels
```

Ahora bien, por medio de la función `learn new` se realiza el entrenamiento del modelo con el conjunto de datos fraccionado (v2). La ejecución tarda unos pocos minutos:

```
!biome.classifier learn new v2 --ds v2
```

Esta función, emplea las siguientes características para su ejecución:

- `--model TEXT`: el modelo utiliza la arquitectura de red neuronal predeterminada, es decir, la configurada con la estructura descrita en el numeral 5.4.
- `--from-learn TEXT`: por medio de esta opción es posible que comience a aprender de los resultados de aprendizaje referenciados, sin embargo, no se requiere para este trabajo.
- `--ds TEXT`: el *datasource* que se maneja para el entrenamiento es el que contiene las anotaciones efectuadas.
- `--trainer TEXT`: la configuración *trainer* se realizó teniendo en cuenta los siguientes aspectos:
 - Número de *epochs* (épocas): “un *epoch* es cuando un conjunto de datos entero se pasa a través de la red neuronal” (Sharma, 2017). Se define un máximo de 30 *epochs* para este modelo.
 - Métricas de validación: la medida de rendimiento que se monitorea para finalizar el entrenamiento es el puntaje f1 (métrica que se define en el siguiente capítulo). Para ello, se utiliza el proceso de *early stopping* (detención temprana), el cual se basa en “detener el entrenamiento cuando aumenta el error en el conjunto de datos de validación, ya que esto es una señal de sobreajuste al conjunto de datos de entrenamiento” (Prechelt, Orr, & Müller, 2012). Este procedimiento orienta

el número de *epochs* necesarias para mejorar el rendimiento, conforme a los objetivos perseguidos. En particular, se busca aumentar el promedio de la métrica de validación f1, por lo que se introduce la siguiente instrucción: `+average/f1`

- *Patience* (pacienza): este argumento hace referencia a la cantidad de *epochs* que se está dispuesto a efectuar sin observar mejoras en el rendimiento, antes de terminar el entrenamiento del modelo. En el trabajo actual, se ajusta en 5.
- Optimizador: se utiliza el método de Adam, algoritmo para la optimización basada en gradiente de primer orden de funciones objetivo-estocásticas. El método es apropiado para objetivos no estacionarios y problemas con gradientes muy ruidosos (Kingma & Ba, 2014), como el estudiado.

Tras la finalización del entrenamiento, se realiza una retroalimentación del modelo resultante, en la que se validan y corrigen las predicciones calculadas, con el fin de otorgar una mayor experiencia de aprendizaje al modelo. Para ello, se emplea la versión final de la base de datos (`Twitter_1-7_colombia_final`), en la que se consolida la totalidad de los tweets extraídos:

```
!biome.classifier feedback new tweets-v5-with-model --source  
..../datasets/Twitter_1-7_colombia_final.csv --input tokens='Tweet Text' --  
predict-with v2
```

De este modo, se crea una nueva versión de anotaciones en la que se incluye la probabilidad de que el tweet pertenezca a cada clase y se determina la predicción con la etiqueta que mayor probabilidad posea. Dicha estimación se puede visualizar y filtrar mediante la opción “Predicted as” disponible en la interfaz web de biome:

Ilustración 44 Predicted as- biome

The screenshot shows the biome interface for a project titled "tweets-v5-with-model". The main area displays a CSV file named "Twitter_1-7_colombia_final.csv" with 23626 records. A toggle switch "Show labelled records" is turned off. The top navigation bar includes tabs for "Labelled as", "Predicted as" (which is currently selected), and "Metadata". A search bar "Search label..." is present. A dropdown menu lists several categories with counts: Other (9021), Positive (7429), Negative (4117), Healthy (1530), Non-healthy (851), Practice_sports (465), and Non_Smoker (1). Below the dropdown, a text snippet from a tweet is shown: "fumador crónico en los tiempos del coronavirus.". At the bottom of the dropdown are "Cancel" and "Apply" buttons, along with probability percentages: 6% for Practice_sports and 1% for Non-healthy. A "More labels" button is also visible.

En esta etapa, el objetivo es validar manualmente la forma en la que el modelo está realizando las predicciones y aportar más características para que pueda aprender. Así pues, se etiquetan los casos en los que se considere que el modelo no proporciona la clasificación esperada. Se muestran a continuación, algunos ejemplos de las probabilidades arrojadas y las modificaciones realizadas:

Ilustración 45 Ejemplo de predicciones efectuadas por el modelo inicial 1

- tokens: Hubo un tiempo en mi vida donde la nicotina era mi desayuno, almuerzo y cena; ahora pasa alguien fumando cigarrillo al lado mío y me dan ganas de estrellarle un ladrillo en la cara.



Ilustración 46 Ejemplo de predicciones efectuadas por el modelo inicial 2

- tokens: Sabiduría convencional para el fumador crónico en los tiempos del coronavirus. #COVID-19

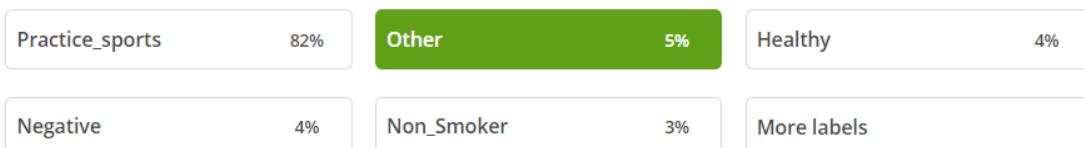


- tokens: Mi mama me dio la vida pero la salchipapa mixta de míster pizza me dio las ganas de vivirla

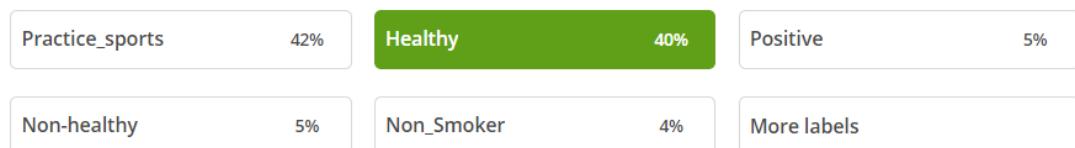


Ilustración 47 Ejemplo correcciones a las predicciones efectuadas

- tokens: Extraño el gimnasio al que nunca he ido :(



- tokens: Mi desayuno favorito, mucha nutrición, personalizado en 27gr de Proteína y solo 200 calorías. #SomosDorayAlex #HerbalifeNutrición #Fit #rendimientofisico #lamesa #Nutricióninteligente #Pequeñadisciplina...



- tokens: Ya me estoy volviendo loca sin ir al gym, con esta semana ya van 22 años.



Con la evaluación manual realizada se concluye que las categorías con mejor comportamiento fueron las relacionadas con la práctica de deporte y alimentación poco saludable. Mientras que

alimentación saludable presentó mayores desaciertos en la clasificación y fue necesario ajustar las anotaciones de una porción de tweets, conforme a las condiciones.

Después de realizar los ajustes, se procede a incorporar las anotaciones realizadas en la versión consolidada de las anotaciones efectuadas anteriormente. De esta manera, el nuevo archivo tiene la siguiente distribución:

Ilustración 48 Distribución de las anotaciones finales

Other	581
Negative	468
Positive	466
Practice_sports	302
Non-healthy	289
Healthy	200
Non_Smoker	38

Posteriormente, se procede a realizar el preprocesamiento de los tweets empleando la librería **ekphrasis**, herramienta open source orientada a las publicaciones de redes sociales, como Twitter, “que realiza tokenización, normalización de palabras, segmentación y corrección ortográfica” (Baziotis, Pelekis, & Doulkeridis, 2017), usando redes de memoria a corto y largo plazo (LSTM por sus siglas en inglés), entrenadas con estadísticas de una colección de 330M de tweets. Así pues, utilizando la función **text_processor** (disponible en el Anexo E. Script de Python para el pre-procesamiento de tweets), es posible definir fácilmente el preprocesamiento para el dataset de tweets final. La Tabla 12 muestra un ejemplo del texto efectuado:

Tabla 12 Ejemplo de preprocesador de texto

original	The *new* season of #TwinPeaks is coming on May 21, 2017. CANT WAIT \o/ !!! #tvseries #davidlynch :D
processed	the new <emphasis> season of <hashtag> twin peaks </hashtag> is coming on <date> . cant <allcaps> wait <allcaps> <happy> ! <repeated> <hashtag> tv series </hashtag> <hashtag> david lynch </hashtag> <laugh>

Fuente: (Baziotis, Pelekis, & Doulkeridis, 2017)

El resultado del preprocesamiento se agrega como una columna adicional del dataframe de anotaciones (denominada **preprocessed_tweet**), utilizando el siguiente script:

```
df['preprocessed_tweet'] = df['Tweet Text'].apply(lambda x: "
".join(text_processor.pre_process_doc(x)))
df.head()
```

El output generado se muestra a continuación:

Ilustración 49 Output pre-procesamiento de tweets

Unnamed: 0		Tweet Text	label	preprocessed_tweet
0	0	Hice la pizza casera más deliciosa do mundo.	Non-healthy	hice la pizza casera más deliciosa do mundo .
1	1	No amén sin ser amado, ese deporte mata .	Negative	no amén sin ser amado , ese deporte mata .
2	3	En esta cuarentena en vez de encontrarme conm...	Negative	en esta cuarentena en vez de encontrarme conm...
3	4	Yo feliz de que mientras estoy encerrada como ...	Positive	yo feliz de que mientras estoy encerrada como ...
4	5	Que asco las aceitunas, le quitan todo lo buen...	Non-healthy	que asco las aceitunas , le quitan todo lo bue...

5.5.4 Validación del modelo

Para realizar la evaluación del rendimiento del modelo de clasificación se emplean métricas definidas a partir de los elementos de la Matriz de Confusión (herramienta que permite identificar la “correlación entre la etiqueta y la clasificación del modelo” (Google Developers, 2019)). En la Tabla 13 se presenta su versión binaria:

Tabla 13 Matriz de Confusión

Predicción			
Positivos			
Real	Positivos	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Fuente: Elaboración propia

Existe una variedad de métricas que se desprenden de esta matriz, sin embargo, las más utilizadas para validar este tipo de modelos son exactitud, sensibilidad y puntaje F1, razón por la cual el modelo se valida teniendo en cuenta dichas mediciones:

- Exactitud o, en inglés, “precision”: permite medir la calidad del modelo, identificando la proporción de las predicciones positivas que son correctas:

$$p = \frac{VP}{VP + FP}$$

- Sensibilidad (Recall): esta métrica resuelve la proporción de positivos reales se identificó correctamente:

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

- Puntaje F1: este indicador es el promedio armónico de la exactitud y la sensibilidad. Alcanza su mejor valor en 1 (que representa *precision* y *recall* perfectas) y su peor valor en 0. Se define con el siguiente cociente:

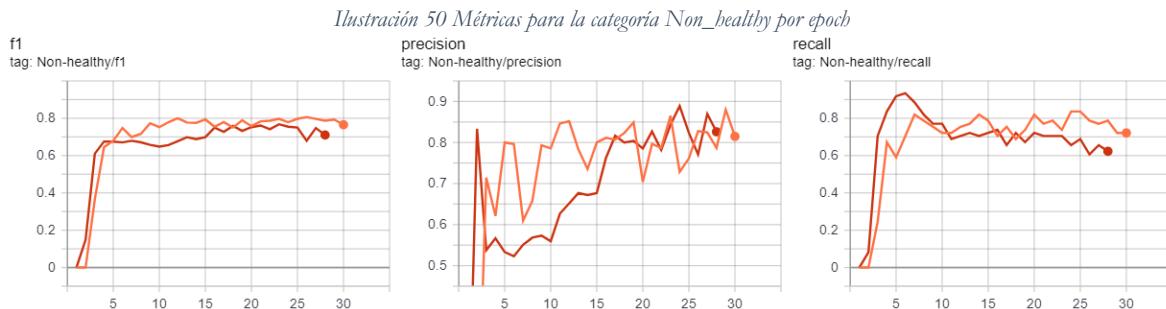
$$F1 = 2 \cdot \frac{p \cdot r}{p + r}$$

Ahora bien, estas métricas se evalúan en los modelos resultantes de dos experimentos de entrenamiento que tienen en cuenta la arquitectura de redes neuronales descrita y la inclusión o exclusión del preprocesamiento de tweets. Los experimentos se realizan con el fin de evaluar cuál es el modelo que mejor predice la clasificación. A continuación se definen los experimentos efectuados a partir del archivo final de anotaciones de la sección anterior:

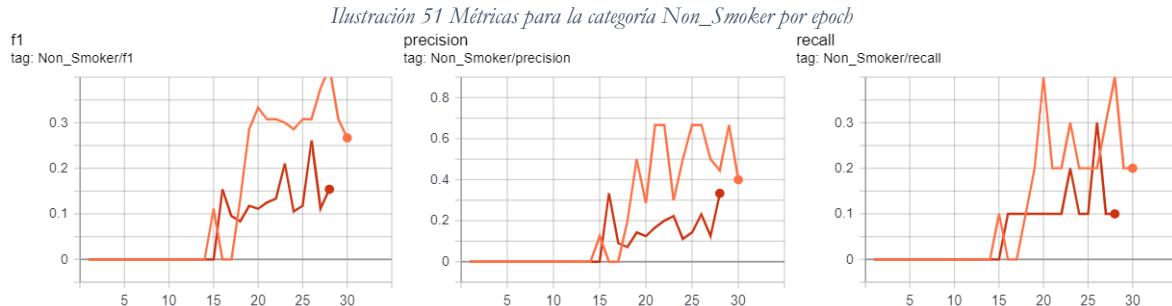
- Arquitectura de red neural base, sin preprocesamiento:
En este experimento (denominado *nopreprocessed*) se entrena el modelo aplicando la arquitectura de red neuronal descrita en el apartado anterior, en la que se manejan las 5 fases definidas. Adicionalmente, se tienen en cuenta las publicaciones sin preprocesamiento.
- Arquitectura de red neural base, con preprocesamiento:
Para este escenario, nombrado *preprocessed*, el modelo se entrena con la misma arquitectura del escenario anterior, pero usando la columna de textos con preprocesamiento realizado con la librería **ekphrasis**.

Los resultados de las métricas calculadas en cada epoch del entrenamiento se almacenan (por categoría) en un archivo con formato JSON, previamente configurado en biome. Para visualizar los resultados se emplea Tensorboard¹³, herramienta integrada al entorno cloud que provee seguimiento y visualización de métricas para experimentos de aprendizaje automático.

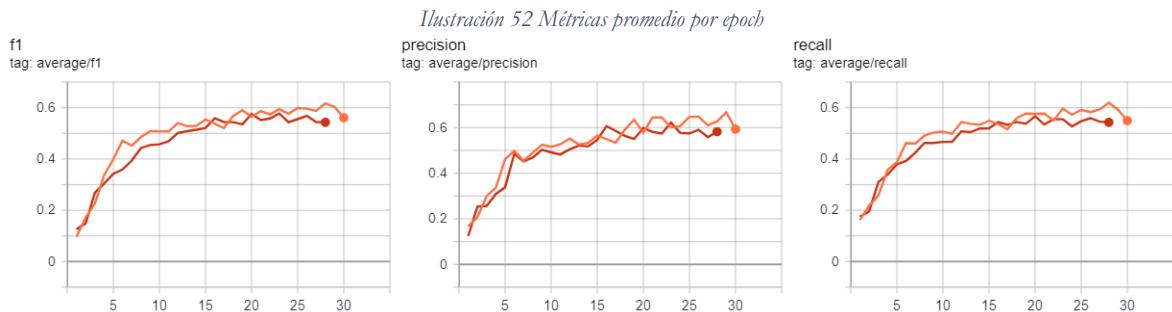
Así pues, se generan los gráficos de línea de la evolución obtenida en cada epoch de los experimentos descritos. A continuación, se presentan los resultados para las categorías con mayor y menor desempeño (*Non_healthy* y *Non_Smoker*, respectivamente), donde la línea naranja representa el primer escenario y la rojo el segundo:



¹³<https://www.tensorflow.org/tensorboard>



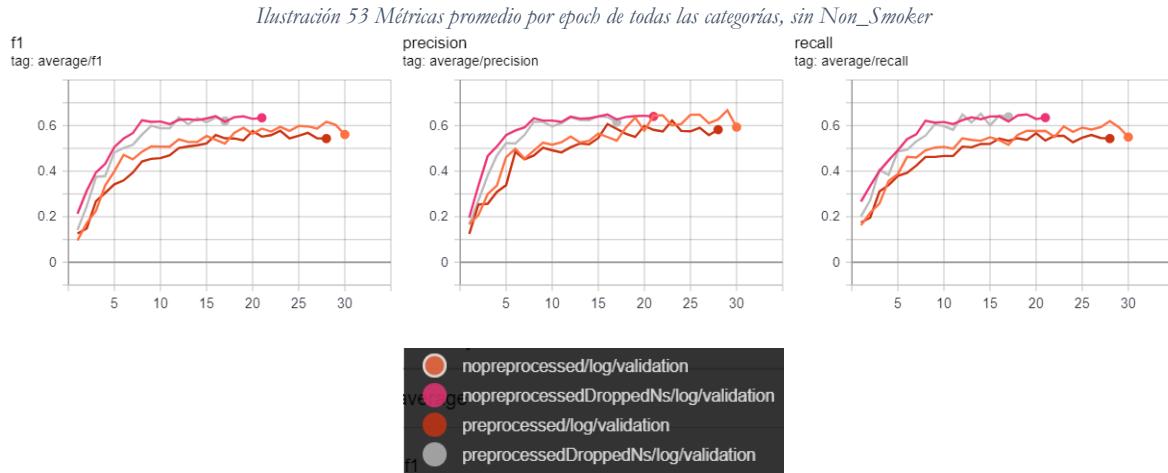
Para visualizar el resultado global del modelo, se estudian las métricas del promedio de todas las clases, donde se evidencia que para la mayoría de las etiquetas, los resultados sin preprocesar superan a los preprocesados:



Los valores más altos del puntaje f1 del promedio, se producen en las epochs 29 y 28, de cada experimento, con los siguientes valores:

Name	Smoothed	Value	Step
nonpreprocessed/log/validation	0.603	0.603	29
preprocessed/log/validation	0.543	0.543	28

Como se puede apreciar, los resultados son susceptibles de mejoras. En particular, se evidencia que la clase “Non_Smoker” presenta un desempeño bajo (a pesar de contar con una precisión alta), la métrica global f1 oscila entre 0 y 0.4, haciendo que la media de todas las etiquetas se vea afectada, por lo cual se decide eliminar esta categoría del modelo. Para ello, se excluyen las publicaciones anotadas como Non_Smoker del dataset (mediante la función `drop` de Pandas) y se procede a entrenar nuevamente el modelo bajo los dos escenarios descritos. Los resultados del promedio se muestran a continuación:



De acuerdo con lo esperado, se observa que el modelo sin la etiqueta referente a los no fumadores presenta mejor rendimiento que el que la contiene y este resultado se extiende en todas las categorías Sin embargo, se propone realizar más experimentos, por lo que se procede a modificar el tipo de red neural recurrente (initialmente configurado como GRU) por redes LSTM (*Long Short Term Memory*) dado que solventan los problemas de acumulación de errores y pueden ser más expresivas y proporcionar mejores resultados ante contextos complejos (Weiss, Goldberg, & Yahav, 2018), como el trabajado. De esta manera, se generan dos nuevos experimentos:

- Arquitectura de red neural modificada, sin preprocesamiento:
Se ejecuta el entrenamiento aplicando la arquitectura de red neuronal descrita, pero variando el tipo de red neural recurrente por LSTM. También, se tienen en cuenta las publicaciones sin preprocesamiento. Se nombra nopreprocessedDroppedNs-lstm.
- Arquitectura de red neural modificada, con preprocesamiento:
En este escenario, nombrado preprocessedDroppedNs-lstm, el modelo se entrena con la misma arquitectura del escenario anterior y empleando la columna de textos con preprocesamiento.

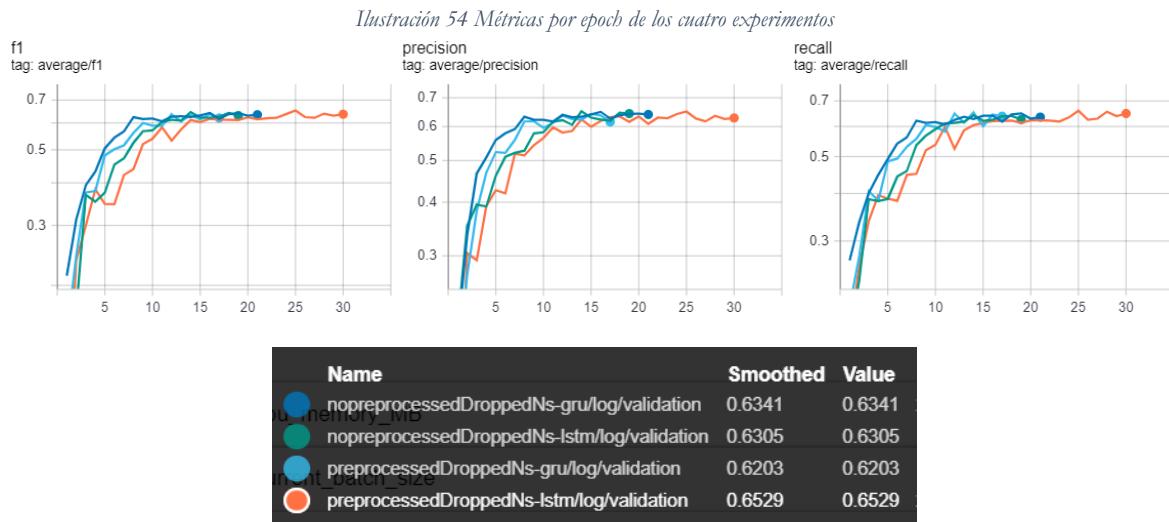
A continuación se presenta una tabla que resume los resultados (por categoría) del mejor epoch de cada experimento realizado:

Tabla 14 Resultados de los experimentos por clase

Etiqueta	Mejor modelo	Mayor F1	Epoch
Non_healthy	Prepo-lstm	0,8182	24
Practice_sports	Noprepo-lstm	0,7627	13
Positive	Prepo-gru	0,6452	11
Healthy	Prepo-lstm	0,6250	24
Other	Noprepo-lstm	0,6073	13

Negative	Noprepo-gru	0,5660	15
----------	-------------	--------	----

Se observa que la red neuronal recurrente LSTM tiene mejor rendimiento para la mayoría de las clases, mientras que la distribución del preprocesamiento es igualitaria entre las categorías, con una participación de 3 a 3. Así que para determinar el modelo con mayor rendimiento, se analizan las métricas del promedio. A continuación se muestran el puntaje F1 del promedio de las clases de los cuatro experimentos:



Se evidencia que, en promedio, todos los experimentos tienen un comportamiento similar, sin embargo, se destaca el entrenamiento ejecutado con la red neuronal recurrente LSTM, frente al rendimiento de la GRU. Adicionalmente, el experimento con preprocesamiento de tweets, supera al que no lo tiene. Así que, el mejor aprendizaje es el generado con solo 6 etiquetas, con preprocesamiento de tweets y utilizando una red recurrente LSTM presentando métricas consideradas altas, al ser provenientes de un dataset ruidoso derivado de redes sociales.

Ahora bien, con el objetivo de optimizar los resultados de este modelo, se procede a realizar nuevas anotaciones que permitan un mayor aprendizaje. Así que, se analizan y clasifican de manera manual aproximadamente 700 tweets adicionales. También, se modifican los siguientes aspectos en el Trainer:

1. Uso de un planificador de la ratio de aprendizaje (learning rate scheduler en inglés), *ReduceLROnPlateau* (PyTorch, 2019), que reduce la ratio de aprendizaje cuando la métrica de validación del modelo (validation/f1) deja de mejorar tras la validación del modelo al final de cada epoch, permitiendo al modelo ajustar sus pesos de manera más granular durante las iteraciones finales del proceso de entrenamiento.
2. Bloquear los pesos de la capa de *embedding* para reutilizar al máximo los vectores de palabra pre-entrenados.

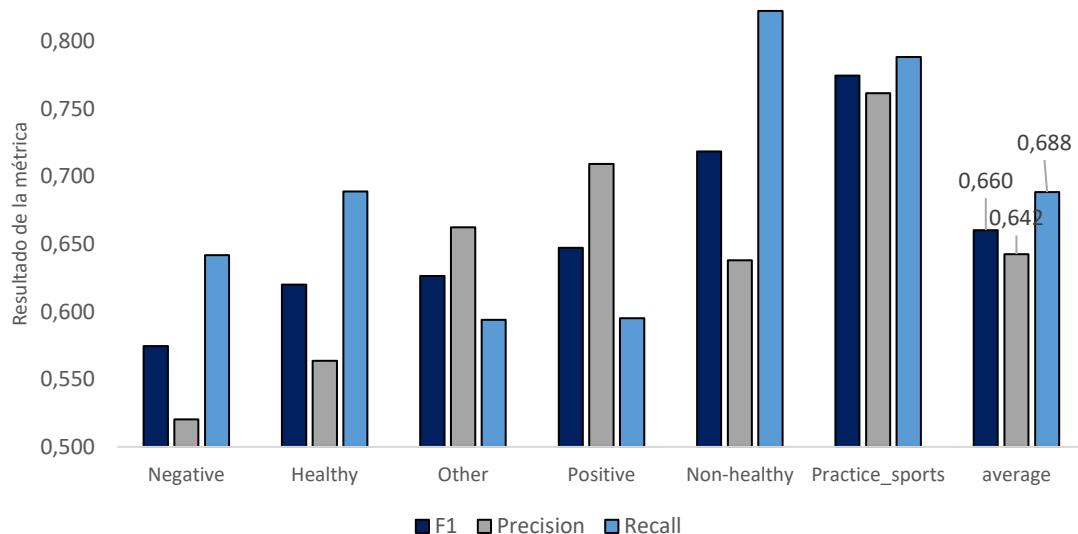
Al aumentar el tamaño del conjunto de datos de anotaciones, es posible generar una nueva partición e incorporar una porción aleatoria de los datos destinada únicamente para pruebas (*test*), sin que intervenga en el proceso de entrenamiento. Puesto que, aunque el conjunto de validación no se emplea en la capacitación, se utiliza durante el proceso para controlar el *early stopping* y el ratio de aprendizaje con el planificador mencionado anteriormente.

Esta división es una práctica habitual en los modelos predictivos, que permite medir el comportamiento del modelo y encontrar un mejor modelo de manera más eficiente. Por consiguiente, se procede a realizar la partición de la base de datos asignando un 20% para validación, 10% para pruebas y el 70% restante para entrenamiento:

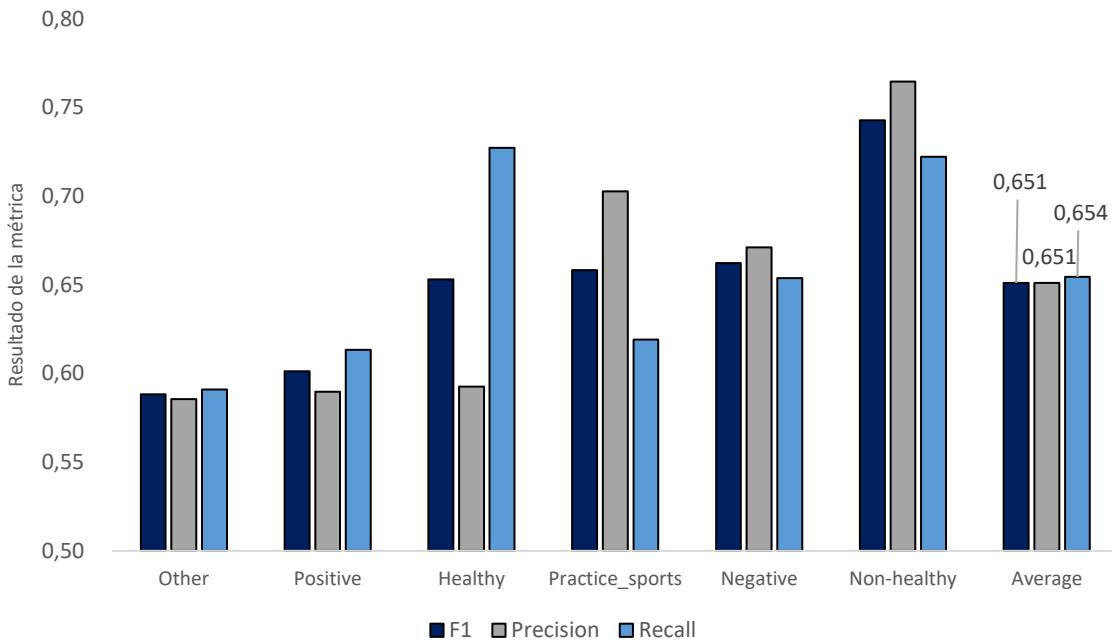
```
!biome.classifier datasource new preprocessed-final --input
tokens="preprocessed_tweet" --label label --validation 0.2 --test 0.1 --
source datasets/annotations_dropped-v2-v4-v5-usecasenew_prepo_str.csv
```

A partir de esta partición, se realiza un nuevo entrenamiento que incorpore las características del modelo con mejor aprendizaje y las trasformaciones del tariner descritas. Se denomina preprocessedDroppedNs-lstm-no-trainable-final. Así mismo, se procede a calcular las métricas para el conjunto de datos de validación y prueba por categoría. Los resultados se muestran a continuación:

Gráfica 3 Resultado de las métricas por clase del dataset de validación



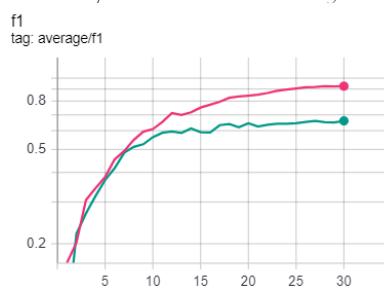
Gráfica 4 Resultado de las métricas por clase del dataset de prueba



En conjunto se identifica que, en promedio, las tres métricas de los datos de prueba son similares a las de validación. Evidenciando, además, que las que poseen mayor rendimiento en ambos conjuntos son Practice_sports y Non-healthy. No obstante, se observa una diferencia en los resultados entre las otras clases: para el conjunto de pruebas la categoría que tiene menor rendimiento es Other, mientras que para validación es Negative, la cual es una de las mejores en pruebas.

A pesar de las diferencias, se aprecia que el modelo tiene buen rendimiento en promedio, con métricas superiores a 0,64. Por otro lado, se resalta que el modelo final no genera *overfitting* puesto que al comparar los resultados de entrenamiento con los de validación se observa que la métrica F1 para el conjunto de validación crece de manera constante, hasta alcanzar una meseta en el *epoch* 29 (momento el que el sistema de *Early Stopping* detiene el proceso de entrenamiento registrando el modelo con la mejor métrica F1 en el conjunto de evaluación).

Ilustración 55 Métricas promedio del dataset de train y validation por epoch



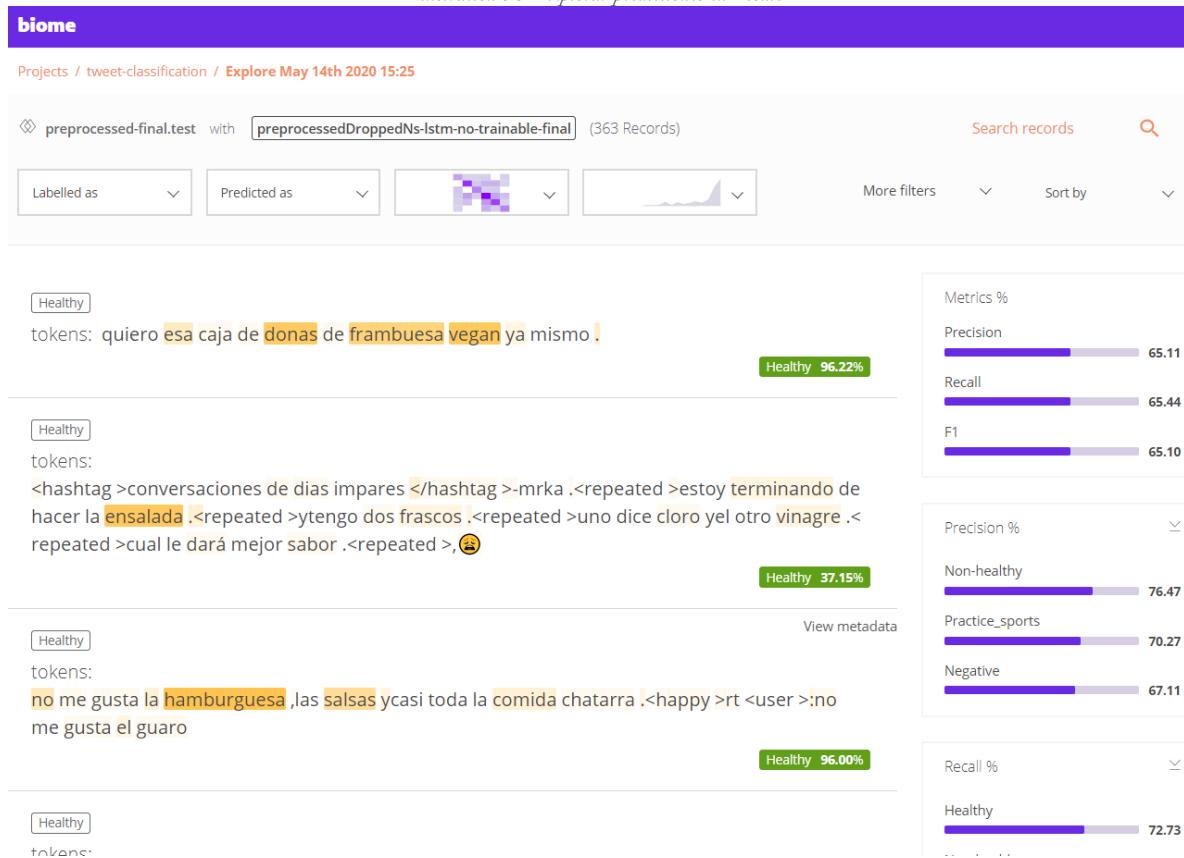
Name	Smoothed	Value
preprocessedDroppedNs-Lstm-no-trainable-final/log/train	0.9252	0.9252
preprocessedDroppedNs-Lstm-no-trainable-final/log/validation	0.6602	0.6602

5.5.4.1 Análisis de errores

Con el fin de mitigar las falsas predicciones del aprendizaje seleccionado, se procede a establecer el umbral de confianza aplicable, es decir, determinar la probabilidad mínima con la que se acepta la predicción del modelo para cada etiqueta. En los casos cuya probabilidad sea inferior a la fijada, el tweet se clasifica como “Other”.

Para definirlo, se realiza un análisis de los resultados de validación y prueba, empleando la función `explore new` de biome, por la cual se genera una interfaz web para explorar las predicciones del modelo, donde muestra la anotación efectuada, la probabilidad con la que se asigna la etiqueta predicha y los resultados de las 3 métricas. Adicionalmente, presenta el peso de las palabras más importantes dentro del texto (calculada a partir de la función integrada de biome, `interpret`):

Ilustración 56 Explorar predicciones en Biome



A continuación se presentan algunos ejemplos de falsas predicciones que se podrían evitar ajustando el intervalo de confianza:

Ilustración 57 Ejemplos de falsas predicciones

[view meta](#)

Other

tokens:
<hashtag>maltrato animal </hashtag>denuncian masacre de perros ygatos con comida
envenenada en guadalupe ,huila

Negative 35.58%

[View metad](#)

Non-healthy

tokens:
me comí unos frijoles con picante y me puso asudar .sigueme para más consejos de ejercicio
pasivo

Practice_sports 42.14%

[view metad](#)

Negative

tokens:
ami no me gusta que me den atención ,ami entre más me ignore más me gusta se convierte
como un reto personal .(me encanta sufrir)

Practice_sports 36.08%

Dentro de la interfaz web, también se encuentran otras funcionalidades importantes: representación de la matriz de confusión y la opción para filtrar los resultados de la muestra dependiendo del intervalo de probabilidad de la predicción:

Ilustración 58 Opciones matriz de confusión e intervalo de confianza biome



Usando esta última funcionalidad, se busca el intervalo que maximiza el resultado del puntaje F1, tanto para validación como para prueba. Para ello, se ajustan los rangos y se evalúan los distintos efectos. En la Tabla 15 se presentan los valores de la métrica (por clase) para los datos de validación:

Tabla 15 Resultados F1 por intervalo de confianza y clase, del dataset de validación

Rango de confianza	Practice_sports	Non_healthy	Positive	Other	Healthy	Negative
0%-100%	77,46	71,84	64,72	62,63	62,00	57,43
50%-100%	79,76	72,00	65,98	63,52	62,63	59,38
60%-100%	81,25	73,47	66,91	65,93	65,93	59,75
70%-100%	84,56	75,79	69,88	66,84	65,91	63,72
80%-100%	86,36	76,60	71,05	65,85	68,29	64,68

Teniendo en cuenta estos valores se calculan las variaciones entre las filas. Esto con el fin de identificar el rango que optimice la métrica. Los resultados se muestran en la Tabla 16:

Tabla 16 Variaciones de F1 por intervalo de confianza y clase, del dataset de validación

Variaciones	Practice_sports	Non_healthy	Positive	Other	Healthy	Negative	Total
0-50	2,97%	0,22%	1,95%	1,42%	-4,23%	3,40%	5,73%
50-60	1,87%	2,04%	1,41%	3,79%	11,03%	0,62%	20,77%
60-70	4,07%	3,16%	4,44%	1,38%	-0,03%	6,64%	19,66%
70-80	2,13%	1,07%	1,67%	-1,48%	3,61%	1,51%	8,51%

Considerando las variaciones totales mostradas, se evidencia que para el dataset de validación la mayor se produce al modificar la cota inferior de 50% por 60%, correspondiente a 20,77%. De igual modo, se realiza el análisis para los datos de prueba y se observan las siguientes variaciones:

Tabla 17 Variaciones de F1 por intervalo de confianza y clase, del dataset de prueba

Variaciones	Non_healthy	Negative	Practice_sports	Healthy	Positive	Other	Total
0-50	1,44%	2,72%	1,29%	-4,73%	-0,53%	1,68%	1,87%
50-60	0,53%	0,68%	0,73%	2,28%	4,66%	4,06%	12,95%
60-70	1,53%	-2,66%	9,96%	2,33%	3,90%	-0,27%	14,79%
70-80	2,30%	9,34%	8,33%	4,87%	10,24%	8,14%	43,22%
80-90	2,64%	0,78%	2,04%	1,00%	4,60%	-6,20%	4,86%

Teniendo en cuenta los datos de prueba, el variación máxima se produce al cambiar la cota inferior de 70% por 80%, representando un 43,22%. Por consiguiente, se define el umbral de confianza como el promedio entre los resultados de validación y prueba, es decir, el promedio entre 60% y 80%, equivalente al 70%. De modo que, todas las predicciones estimadas con una probabilidad inferior serán categorizadas como “Other”.

5.5.5 Implementación: Casos de uso

En esta etapa, se muestran los resultados del modelo de clasificación aplicado a las publicaciones de 12 cuentas públicas de Twitter, las cuales fueron elegidas aleatoriamente por la gran cantidad y variedad de contenido registrado en la aplicación. Cabe resaltar que por temas de protección de datos se anonimizan los perfiles empleando los números del uno al doce, para cada usuario.

Para realizar la extracción de tweets se emplea Keyhole¹⁴, herramienta de *Hashtag Tracking* (seguimiento de hashtag), que permite dar seguimiento a las cuentas de Twitter de perfiles seleccionados y a la vez da acceso al total de sus publicaciones públicas mediante un archivo

¹⁴ <https://keyhole.co/>

Excel. De esta forma se accede a las últimas publicaciones de cada usuario, las cuales son consolidadas y evaluadas descriptivamente mediante una nube de palabras:

Ilustración 59 Nube de palabras de los casos de uso



Se aprecia que las publicaciones no presentan sesgo sobre los temas evaluados, lo cual es un buen indicativo para observar el rendimiento del modelo con este tipo de información.

Ahora bien, para observar el desempeño del modelo se procede a cargar los datos en el entorno de biome y aplicar el entrenamiento a este nuevo dataset, mediante la función **feedback new**. De esta forma (al igual que en la etapa de entrenamiento), se genera una interfaz web en la que se observan las probabilidades arrojadas por el aprendizaje seleccionado. A continuación se presentan algunos ejemplos:

Ilustración 60 Publicación Practice_sports del usuario Dos

tokens: ayer hice la rutina tan contenta y cuando decían opción avanzada las hacía así todas y hoy me duele todo , más tarde me voy a morir cuando entrene .

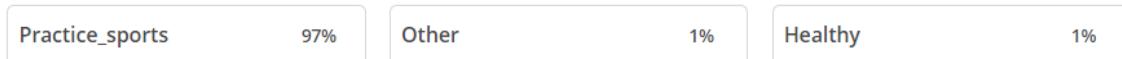


Ilustración 61 Publicación Healthy del usuario Cinco

tokens: hice una ensalada de papa con atún para mi solo pero mi loza dice que hice bandeja paisa para personas

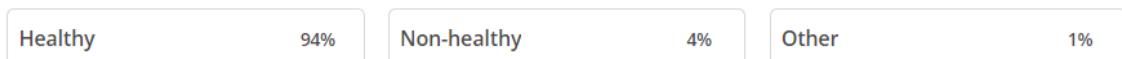


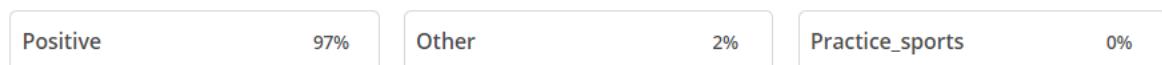
Ilustración 62 Publicación Negative del usuario Diez

tokens: cómo quisiera saber de ti , poder preguntar sí dormiste , sí comiste , sí todo está medianamente bien . pero solo fueron días , en los que nos volvimos mierda 😞



Ilustración 63 Publicación Positive del usuario Doce

tokens: bogotá , que lindo día nos regalas hoy . te disfrutamos desde casa ☀



Según un análisis exploratorio se evidencia que el modelo tiene buen rendimiento para clasificar los textos. Así que se procede a exportar y realizar un conteo de la totalidad de las predicciones efectuadas (usando el paquete Pipeline de biome). Tras evaluar los resultados por usuario y clase, se asigna la categoría de riesgo conforme a la política establecida, como se muestra en la Tabla 18:

Tabla 18 Predicciones del modelo de los casos uso

Usuario anonimizado	Predicción						Categoría de riesgo	Descuento prima
	Healthy	Practice_sports	Positive	Negative	Non-healthy	Other		
Uno	2	0	19	8	0	58	Medio bajo	30%
Dos	4	10	27	8	3	39	Bajo	40%
Tres	2	0	29	12	0	43	Medio bajo	30%
Cuatro	2	1	13	11	1	65	Bajo	40%
Cinco	2	0	12	2	0	79	Medio bajo	30%
Seis	7	0	16	12	2	53	Medio bajo	30%
Siete	1	2	11	5	1	37	Bajo	40%
Ocho	0	0	9	11	1	65	Medio alto	5%
Nueve	1	2	14	10	2	24	Bajo	40%
Diez	1	2	17	10	0	35	Bajo	40%
Once	0	1	6	10	0	32	Medio bajo	30%
Doce	1	0	24	7	0	49	Medio bajo	30%

De esta manera, como los usuarios Dos, Cuatro, Siete, Nueve y Diez tienen publicaciones categorizadas como Healthy, Practice_sports y Positive, se determina que la categoría de riesgo Bajo. En cambio, como dentro de los textos del usuario Uno solo se presentan Healthy y Positive, pertenece a la categoría Medio bajo. Por su parte, el usuario Ocho solo tiene publicaciones clasificadas como Positive, así que corresponde a riesgo Medio alto.

Así pues la prima de seguro de vida que cada usuario pagará será el resultado de aplicar la fórmula definida en el numeral 2.4.1., teniendo en cuenta las categorías de riesgo determinadas.

6 Validación

La validación se realiza calculando la correlación existente entre los valores arrojados por el modelo y la tasa de mortalidad observada por departamento. Para conseguirlo, se calcula la tasa de mortalidad por departamento, posteriormente, se construye una métrica para evaluar los resultados del modelo por departamentos, para luego identificar la correlación de las dos medidas, utilizando los métodos de Pearson, Spearman y Kendall.

6.1 Tasa de mortalidad

La tasa de mortalidad (TM) es un indicador demográfico equivalente a la proporción de defunciones de personas pertenecientes a un determinado segmento poblacional, respecto al total de habitantes de dicho colectivo, en un periodo de tiempo (Instituto Nacional de Estadísticas, 2019). Se define a continuación la tasa de mortalidad para el departamento d , de personas de edad x (de 20 a 49 años), en el periodo t como:

$$TM_d = \frac{D_{d,x}^t}{P_{d,x}^t}$$

donde:

$D_{d,x}^t$: Defunciones registradas durante el año t de personas residentes en el departamento d y edad x .

$P_{d,x}^t$: Población total durante el año t , residentes en el departamento d y edad x .

Cabe aclarar que, el rango de edad establecido obedece al segmento identificado con mayor participación en Twitter y las edades más frecuentes de los consumidores de seguros de vida.

Ahora bien, con el fin de realizar el análisis demográfico de la población desagregada por departamento para $t = 2018$, se emplean datos abiertos provenientes de fuentes estadísticas elaboradas por el organismo público colombiano (Departamento Administrativo Nacional de Estadística DANE). Entidad encargada de “planear, implementar y evaluar procesos rigurosos de producción y comunicación de información estadística a nivel nacional, que cumplan con estándares internacionales y se valgan de la innovación y la tecnología”¹⁵.

6.1.1 Proyecciones de población

En febrero de 2020, el DANE publicó las proyecciones de población calculadas teniendo en cuenta los resultados del Censo Nacional de Población y Vivienda -CNPV- 2018, ajustados por edad y sexo. Para la elaboración el DANE utilizó una proyección determinística, por

¹⁵ <https://www.dane.gov.co/index.php/acerca-del-dane/informacion-institucional/generalidades>

componentes de cohortes a nivel total, departamental cabecera y centros poblados y rural disperso (DANE, 2020).

La base de datos empleada para este trabajo se denomina “Series de proyecciones de Población 2018-2023 con desagregación nacional, departamental y municipal, por grupos quinquenales de edad, edades simples (0 a 28 años) y sexo”. La información del 2018 necesaria para el análisis se presenta en la hoja “MPIO”, en las siguientes columnas:

- Código: Código del departamento, según la División Político- Administrativa del DANE
- Grupos de edad: Agrupación de edades, por quinquenios
- Ambos Sexos: población total
- Hombres: población total registrada con género masculino
- Mujeres: población total registrada con género femenino

Ilustración 64 Proyecciones de Población 2018



DANE
INFORMACIÓN PARA TODOS

Proyecciones de Población

Proyecciones de Población 2018-2023, total nacional, departamental y municipal por grupos quinquenales de edad y sexo
A Junio 30

Código	Grupos de edad	TOTAL 2018		
		Ambos Sexos	Hombres	Mujeres
05	Antioquia			
	Total	6.407.102	3.094.159	3.312.943
	00-04	451.139	230.491	220.648
	05-09	459.957	235.055	224.902
	10-14	483.185	246.682	236.503
	15-19	527.055	267.830	259.225
	20-24	566.674	284.886	281.788
	25-29	551.560	274.746	276.814

Fuente: (DANE, 2020)

6.1.2 Defunciones

En enero del 2020 la Dirección de Censos y Demografía (DCD), en colaboración con el DANE, publica las Estadísticas Vitales - EEVV - 2017 - 2018, calculadas a partir de la información registrada en el Sistema de Registro Civil y Estadísticas Vitales. Dentro del estudio se reportan los registros de las defunciones no fetales, los cuales se obtienen de los certificados de defunción, y se asocia con las diferentes subdivisiones geográficas (departamentos, municipios y áreas).

El archivo de datos que contiene la información de defunciones durante el 2018 es “nofetal2018” y las variables empleadas para realizar el análisis son:

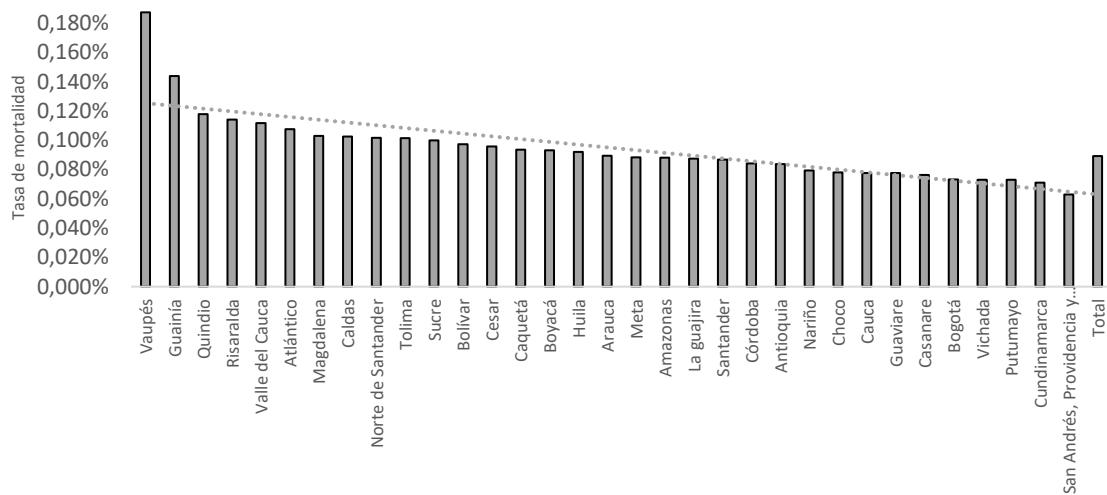
- GRU_ED1: Agrupación de edades, según la edad del fallecido, por quinquenios
- SEXO: Sexo del fallecido
- CODPTORE: Código del departamento de residencia habitual del fallecido, según la División Político- Administrativa del DANE
- PMAN_MUER: Probable manera de muerte (Natural, Violenta, En estudio)
- MAN_MUER: Probable manera de muerte violenta (Suicidio, Homicidio, Accidente de tránsito, Otro accidente, En estudio, Sin información)

De este modo, se obtiene la información de defunciones desagregada por departamento género, edad y causa de muerte. Ahora bien, dado que el modelo desarrollado no tiene en cuenta las maneras de muerte violenta, este tipo de características se debe excluir, al igual que las que están en estudio. Sin embargo, puesto que se mide los niveles de estrés con publicaciones positivas o negativas, se mantienen los registros relacionados con el suicidio.

6.1.3 Tasa de mortalidad por departamento

En el Anexo G se presenta un resumen de la información consolidada de los dos estudios para el grupo de edad de 20 a 49 años y ambos sexos, desagregada por departamento. En la Gráfica 7 se muestra la tasa de mortalidad calculada por departamento:

Gráfica 5 Tasa de mortalidad por departamentos (ambos sexos, de 20 a 49 años)



Como se aprecia en la gráfica, la tasa de mortalidad presenta baja variación por departamento, exceptuando los valores de Vaupés y Guainía, que en promedio superan al dato más próximo (Quindío) por 0,05%. Debido a esto, se consideran atípicos y se decide excluir estos datos de la muestra para la validación. De modo que, el departamento con la tasa de mortalidad más alta es Quindío, donde (conforme a la información prevista de defunciones) la causa de muerte más frecuente para este grupo etario se presenta por Choque séptico (hipotensión prolongada).

Adicionalmente, se observa que San Andrés, Providencia y Santa Catalina, Cundinamarca y Putumayo presentan las menores tasas de mortalidad del país.

6.2 Métrica del resultado del modelo por departamento

El valor del modelo por departamento (vm_d) corresponde al cociente entre las publicaciones identificadas como saludables (etiquetadas como healthy, practice sports, positive) y el total de los tweets efectuados del departamento.

$$vm_d = \frac{h_d + ps_d + p_d}{ns_d + h_d + nh_d + ps_d + p_d + n_d + o_d}$$

Donde:

h_d : tweets etiquetados como healthy del departamento d

nh_d : tweets etiquetados como non_healthy del departamento d

ps_d : tweets etiquetados como practice_sports del departamento d

p_d : tweets etiquetados como positive del departamento d

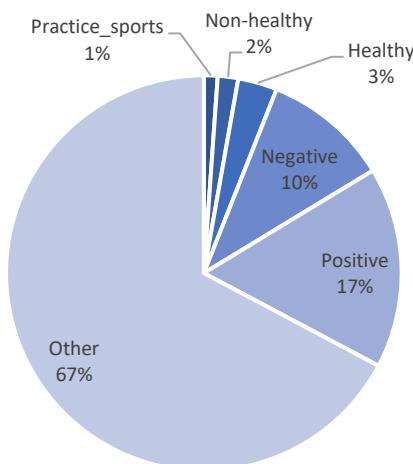
n_d : tweets etiquetados como negative del departamento d

o_d : tweets etiquetados como other del departamento d

6.2.1 Valor del modelo por departamento

Con el fin de realizar el cálculo de la métrica para los 31 departamentos definidos en el numeral 6.1.3, se extraen y analizan las predicciones de clasificación de tweets generadas por el modelo. Los resultados por departamento y clase de todos los tweets evaluados se incluyen en el Anexo H. Valor del modelo con desagregación departamental. En la Gráfica 6 se resumen la participación de la cada una de las categorías dentro del corpus evaluado:

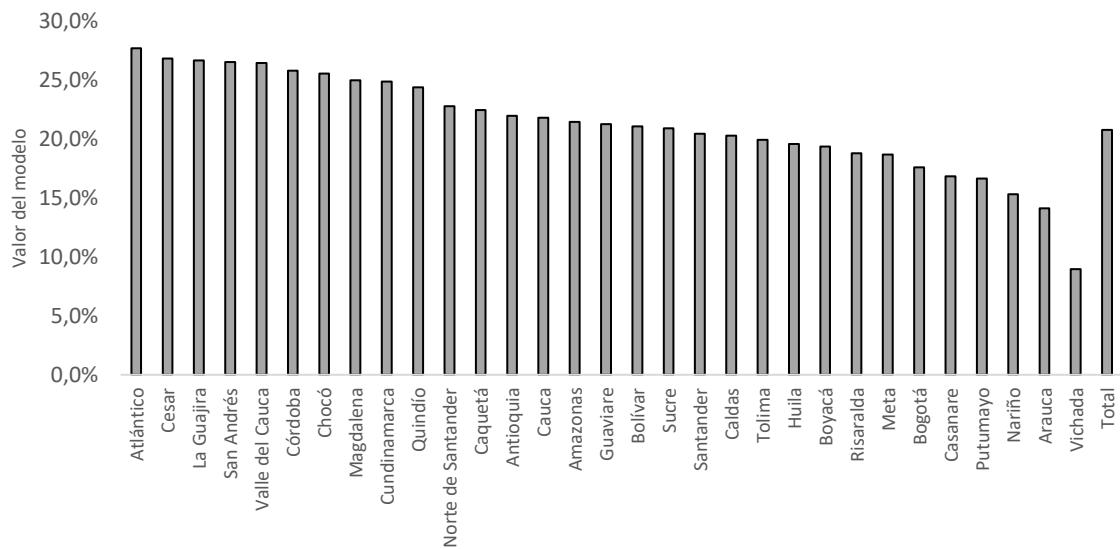
Gráfica 6 Predicciones del modelo para el corpus total



El ruido, las diversas temáticas que se comparten en Twitter y las predicciones con baja probabilidad, generan que la mayor parte de los tweets extraídos se clasifiquen como “Other”. Las porciones más representativas de los datos restantes son los etiquetados como “Positive” y “Negative”, debido a la densidad de publicaciones que revelan el sentimiento del usuario. También, se evidencia que es poco frecuente que las personas compartan textos afines a la práctica de deporte.

A partir de esta información, se calcula el valor del modelo por departamento. En la Gráfica 7 se muestran los valores obtenidos:

Gráfica 7 *Valor del modelo por departamento*



Debido a la alta densidad de tweets etiquetados como “Other” los valores del modelo oscilan entre 10% y 28%. El departamento donde se concentran la mayor proporción de tweets saludables es Atlántico, mientras que Vichada presenta la menor, debido a que ningún tweet fue clasificado como “Healthy” ni como “Practice_sports”. En ambos casos, las publicaciones saludables con mayor participación son las clasificadas como “Positive”.

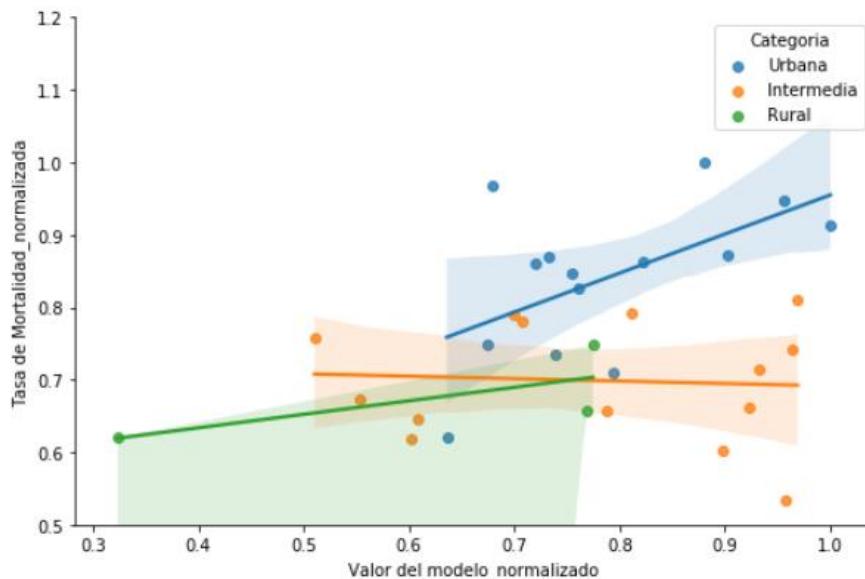
6.3 Validación del modelo

Inicialmente, se normalizan los datos para que los rangos de la tasa de mortalidad y del valor del modelo sean consistentes. De este modo, se facilitan los cálculos estadísticos y se permite una comparación justa de las dos variables, asegurando que tengan el mismo impacto. Para ello, se utiliza el método de escalado de características simples (*simple feature scaling*), el cual consiste en dividir cada valor inicial por el valor máximo de la columna, haciendo que los nuevos valores oscilen entre cero y uno:

$$x_{final} = \frac{x_{initial}}{x_{max}}$$

Tras realizar la normalización, se explora la relación entre las variables en la Gráfica 8, donde se asigna un color a cada departamento de acuerdo con su categoría (urbana, intermedia, rural):

Gráfica 8 *Valor del modelo de clasificación de tweets versus tasa de mortalidad*



Se evidencia que los departamentos urbanos presentan una mayor tasa de mortalidad, mientras que los rurales la menor. Estos hallazgos se deben a la exclusión de las muertes por causa violenta (exceptuando el suicidio), puesto que este motivo de fallecimiento representa un 10% para los urbanos, 14% para los intermedios y 20% para los rurales, según la información sobre defunciones reportada por la DCD y el DANE. Esto sugiere que la valoración de la mortalidad obedece diferentes patrones dependiendo de la situación del territorio y por ende debe contemplarse un tratamiento distinto para estos datos, en contraste con el manejo actual donde se emplea la misma tabla de mortalidad sin distinguir la región.

A su vez, se observa que los departamentos urbanos tienen un mayor valor del modelo que los rurales, aunque no se produce una tendencia clara sobre las agrupaciones de departamentos puesto que la mayoría de los puntos se concentran en el rango de 0,6 a 1. En cuanto a la relación entre la tasa de mortalidad y el valor del modelo se identifica que para la categoría urbana es creciente, es decir, entre mayor valor del modelo, se genera una mayor tasa de mortalidad; mientras que los intermedios presentan una relación inversa, con una ligera pendiente negativa. Respecto a los datos de categoría rural, se observa que tienen una pendiente positiva, sin embargo, no representan una porción de datos significativa.

Ahora bien, para evaluar la correlación entre la tasa de mortalidad y los resultados del modelo se usan los métodos de Pearson, Spearman y Kendall. El primero es la medida estadística paramétrica más utilizada para investigar la relación lineal entre dos variables continuas X, Y. El coeficiente de correlación de Pearson ($\rho_{pearson_{X,Y}}$) es equivalente al cociente entre la

covarianza de las dos variables aleatorias $cov(X, Y)$ y el producto de su desviación estándar (σ) (Arizmendi Echecopar & Condor Espinoza, 2019):

$$\rho_{pearson_{X,Y}} = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

La correlación de Spearman es la medida robusta asociación que mide la monotonidad, es decir, evalúa las similitudes en el orden de las variables. El coeficiente de correlación de Spearman se calcula de la misma forma que la correlación de Pearson, pero empleando rangos en lugar de las observaciones reales (Rousseau, Egghe, & Guns, 2018). La correlación de rango de Kendall es una técnica no paramétrica, que mide la asociación ordinal entre variables. Al igual que Spearman, este indicador se basa en rangos para analizar si se mantiene o no el orden entre las variables. El coeficiente τ de Kendall se deriva de pares de observaciones y establece la fuerza de asociación a partir del patrón de concordancia y discordancia entre los pares (Magiya, 2019).

Estos métodos aportan tanto el coeficiente de correlación, como el p-valor para evaluar la asociación que define la relación entre las dos variables, bajo la hipótesis global $H_0: \beta_1 = 0$. El resultado del coeficiente de correlación se encuentra en el rango de -1 a 1, donde 1 representa una correlación perfecta y positiva; -1 una correlación perfecta y negativa; y 0 indica la ausencia de una relación lineal. Por otro lado, el p-valor determina la significancia estadística del coeficiente de correlación para aceptar o rechazar H_0 , si el valor es menor que 0,001, se tiene una fuerte certeza sobre el coeficiente de correlación que se calcula, si está entre 0,001 y 0,05 la certeza es moderada, si está entre 0,05 y 0,1 es débil y si resulta mayor que 1 no presenta ninguna certeza de correlación.

Así pues, usando la librería científica **SciPy**¹⁶ de Python se calculan los tres coeficientes de correlación y los p-valores para las categorías de los departamentos. En la Tabla 19 muestra los resultados:

Tabla 19 Correlación del valor del modelo con la tasa de mortalidad departamental

	Pearson		Spearman		Kendall	
	p-pearson	p-value	p-spearman	p-value	p-kendall	p-value
Total	0,27212	0,13862	0,23266	0,20784	0,17850	0,16470
Urbana	0,55632	0,03882	0,50769	0,06384	0,36264	0,07946
Intermedia	-0,06628	0,82188	0,07692	0,79381	0,07692	0,74719
Rural	0,73548	0,47392	1,00000	0,00000	1,00000	0,33333

Las correlaciones más altas se obtienen con el método de Pearson, esto implica que la relación entre los resultados del modelo y la mortalidad tiene un comportamiento lineal. Motivo por el

¹⁶ <https://www.scipy.org/>

cual el coeficiente de Kendall reporta valores más bajos. En los tres métodos la correlación es mayor cuando consideramos los departamentos rurales, sin embargo, el p-valor indica que no hay significancia estadística. Por su parte, los departamentos urbanos presentan un coeficiente de correlación de Pearson positivo y moderado, con un p-valor inferior a 0,05. Mientras que el p-valor es débil tanto para Spearman como para Kendall. La categoría intermedia tiene una correlación débil, con poca significancia en todos los métodos, de la misma manera que la totalidad de los departamentos.

En resumen, la única correlación se produce con el método de Pearson para los departamentos urbanos. Esto se puede presentar por razones culturales, puesto que en este contexto se fomenta con mayor frecuencia el uso de redes para publicar la vida cotidiana, incluyendo sus hábitos alimenticios, actividad física y sentimientos, lo que permite contar con más información y robustecer los resultados del modelo. Así mismo, se conoce que la mayoría de los usuarios de Twitter en Colombia, se concentran en zonas urbanas, lo que sugiere que existe una comunidad de usuarios en estas poblaciones, generando una mayor visibilidad de la plataforma y permitiendo que se compartan diariamente las actividades y pensamientos entre todos los miembros.

Otro factor que incide en los resultados es la accesibilidad a Internet. En el 2018, “mientras que en Bogotá solo el 9% no tenían acceso, en el área rural es más del 30% y lo mismo sucedía en las pequeñas ciudades” (Lemoine, 2018). Al no existir esta acercamiento por parte de la población se inhabilita cualquier tipo de interacción en la red social y los resultados del modelo se pueden ver sesgados por publicaciones de tipo informativas, en lugar de recreativas como las que se evalúan.

En conclusión, los datos de Twitter son un potencial predictor de la mortalidad en los departamentos urbanos de Colombia en una medida moderada, debido a las características de usabilidad y accesibilidad presentes en estos territorios. Lo que implica que los datos provenientes de este red social aportan información para personalizar la tarifa de seguros de vida en estos sectores del país y permiten ofrecer primas más exactas que fomenten el consumo de este tipo de productos.

7 Conclusiones

7.1 Conclusiones

Se proporciona un caso de estudio que demuestra que las publicaciones de Twitter son un elemento potencial para personalizar las tarifas de los seguros de vida riesgo, en los departamentos urbanos de Colombia. Los criterios que permiten establecer la segmentación de perfiles de riesgo son los hábitos de alimentación, la práctica de actividad física y los niveles de estrés, considerados factores no genéticos que inciden directamente en la mortalidad de la población.

El modelo de aprendizaje automático supervisado que se propone se entrena en la tarea de clasificación de tweets, con procesamiento de lenguaje natural, para detectar las temáticas establecidas. El algoritmo se basa en una arquitectura de red neuronal de nueve capas, que utiliza procesos de *embedding* y *encoding*, redes neuronales recurrentes LSTM, una capa completamente conectada y una lineal, obteniendo un puntaje F1 de 0,66, en el conjunto de datos de validación. Es factible mejorar el rendimiento del modelo al emplear únicamente las categorías con mayor puntaje F1 (Practice_sports y non_healty), las cuales superan el 0,7

Los resultados muestran que el contenido de las publicaciones está vinculado con la tasa de mortalidad a nivel departamental, con una correlación de Pearson moderada del 0,56 en los territorios urbanos. Sin embargo, para las zonas intermedias y rurales la correlación es débil, motivo por el que para caracterizar el riesgo de los usuarios de estas poblaciones se requiere desarrollar herramientas más precisas y granulares, con una mayor cantidad de datos y enfoques específicos de acuerdo con su condición.

En consecuencia, el análisis efectuado expone que los datos no estructurados provenientes de Twitter aportan en una proporción moderada en la segmentación de los consumidores de seguros de vida. Esto representa que es posible refinar el enfoque tradicional de las aseguradoras para valorar el perfil de riesgo de sus clientes, abriendo el camino hacia la innovación.

7.2 Recomendaciones

Se recomienda detallar los términos de aceptación del riesgo puesto que el algoritmo de clasificación no es un sistema infalible. Los usuarios podrían incluir información ficticia en su perfil con el fin de recibir un beneficio económico en la prima de su seguro de vida.

Por otro lado, se sugiere contemplar el debate respecto al uso de información privada para la tarificación de seguros, donde las aseguradoras deben garantizar el adecuado tratamiento, actualización y corrección de los datos, así como la transparencia sobre su manejo, para que el cliente mantenga la confianza en la empresa y autorice el uso de estos.

7.3 Relación del trabajo desarrollado con los estudios cursados

El contenido del presente trabajo coordina los conocimientos adquiridos durante los estudios del Máster Universitario en Gestión de la Información en dos universidades: la Escuela Colombiana de Ingeniería Julio Garavito y la Universidad Politécnica de Valencia, entidades que cuentan con un convenio de doble titulación para dicho programa y aportan a la generación de competencias curriculares y transversales.

Las competencias transversales trabajadas en la elaboración del trabajo fin de máster son:

- Aplicación y pensamiento práctico: utilizar el conocimiento previo sobre el estudio de la información disponible en las diferentes redes sociales para adaptarse a los nuevos modelos de negocio del entorno asegurador, con el fin de ser altamente competitivo y encontrando las soluciones apropiadas frente a las normativas y tecnologías actuales. El nivel del dominio es el tercero puesto que se diseña un plan coherente en acciones concretas para abordar situaciones complejas de forma individual, estableciendo objetivos, consiguiendo información relevante y llevando a cabo el plan para trazar la forma en la que se realiza una segmentación de clientes de seguros de vida.
- Análisis y resolución de problemas: ante la problemática de identificar las características de los usuarios para generar tarifas personalizadas, se genera una solución empleando las tecnologías y tendencias actuales para el manejo de la información, siguiendo una metodología estructurada.
- Innovación creatividad y emprendimiento: como se ha revisado en el estado del arte, se genera un producto innovador y creativo, que permite tener una perspectiva diferente ante la forma en la que se diseñan y calculan las tarifas de los seguros de vida. Adicionalmente, es posible analizar el valor que genera dentro del sector asegurador colombiano.
- Diseño y proyecto: al elaborar este trabajo la idea de generar una segmentación de acuerdo con el riesgo se convierte en realidad al diseñar el modelo de clasificación y la metodología para su desarrollo incluyendo planificación de acciones, recursos y tiempos.
- Planificación del tiempo: se desarrolla un cronograma plasmando las actividades necesarias para cumplir los objetivos y evaluando la planificación y resultados.
- Conocimiento de problemas contemporáneos: al realizar el estado del arte es posible analizar cada uno de los aspectos actuales de la implementación de nuevas fuentes de información incorporadas al sector asegurador, de modo que permita identificar los problemas actuales y la manera de solucionarlos.

Por otro lado, las competencias curriculares (específicas para cada universidad) que permitieron la realización de este trabajo se describen a continuación:

7.3.1 Escuela Colombiana de Ingeniería Julio Garavito

Las materias que permitieron la realización de este trabajo fueron impartidas en las materias Gestión Estratégica, Gestión del conocimiento y técnicas actuales de extracción y análisis de

información mediante Minería Social. La primera aborda temas sobre el diseño de una estrategia empresarial que identifique integralmente la organización, con el fin de construir una visión alineada con las necesidades de la empresa. En este caso, la necesidad de las aseguradoras colombianas es segmentar a sus clientes y otorgar tarifas personalizadas para aumentar sus ganancias y el índice de penetración, adaptándose a las nuevas tecnologías.

Así que, mediante la formulación de la estrategia óptima se identifica que la visión es atender las necesidades de los consumidores de seguros de vida de forma personalizada. Para cumplirla, inicialmente se realiza la caracterización de las necesidades de las aseguradoras y se evalúan las nuevas tecnologías digitales y de información emergentes., considerando las oportunidades y amenazas actuales. Posteriormente, se toma la decisión de modificar el modelo de negocio de venta de seguros, agregando a los cuestionarios de asegurabilidad, un sistema de medición del riesgo, basado en la actividad de los usuarios en Twitter. Lo que fomenta el uso de información externa y activa la venta de seguros en el país.

Gestión del conocimiento contribuye al entendimiento de conceptos básicos como la extracción, la capitalización, la estructuración y gestión de los conocimientos, para sistemas basados en la valoración del capital intangible, lo que permite desarrollar habilidades para la documentación y transferencia del material producido.

Por su parte, la tercera asignatura aporta las herramientas para realizar el proceso de análisis de datos mediante los siguientes pasos: definición del problema, preparación de datos, exploración de datos, desarrollo de modelo predictivo y por último visualización de los resultados. A su vez, aporta un panorama de las distintas aplicaciones de la minería de texto y la utilidad en distintos sectores como el policial, académico, bancario, entre otros, aportando una visión del potencial de los datos provenientes de redes sociales para generar conocimiento en todos los ámbitos.

7.3.2 Universidad Politécnica de Valencia

Los cursos que habilitan la realización del presente trabajo son Explotación de Datos Masivos, Analítica Digital y Servicios de Datos y Contenidos. Durante el primero se discutió sobre el aumento de datos en los últimos años y la oportunidad de transformarlos en conocimiento, mediante la implementación de distintas herramientas para su manejo como R y Phyton. Por consiguiente, en el desarrollo del trabajo se han empleado los datos disponibles para generar conocimiento acerca de los usuarios de Twitter, permitiendo segmentarlos y de esta forma contar con conocimiento valioso para la toma de decisiones de las asegurados.

En la asignatura de analítica digital se dieron a conocer los métodos de recolección de datos de las plataformas de redes sociales, incluyendo Twitter. A su vez, se despertó un pensamiento analítico y crítico para comparar y elegir la mejor herramienta conforme a las necesidades del proyecto, habilitando el análisis para seleccionar cada una de las herramientas empleadas para la construcción del modelo.

Por último, la asignatura de servicios de datos y contenidos aportó las habilidades para realizar la búsqueda de información de calidad, contribuyendo con la divulgación de las principales fuentes y recursos de información disponibles para la elaboración de trabajos de investigación. A su vez, fomenta la creación de servicios innovadores basados en datos, contenidos y fuentes web reutilizables, lo cual se aplica en la construcción tanto del modelo, como de la tasa de mortalidad empleada para su validación.

7.4 Trabajos futuros

Con base en los análisis efectuados en la realización de este trabajo, se puede expandir el número de factores no genéticos que captura el modelo, para generar una segmentación más granular. Además, es posible incorporar elementos externos como: factores ambientales evaluando la población o niveles de pobreza derivados de la ubicación del usuario, y las relaciones sociales, teniendo en cuenta las interacciones con los seguidores.

De igual forma, es factible mejorar el rendimiento del modelo al continuar lanzando nuevos experimentos de aprendizaje que incluyan una mayor cantidad de anotaciones, puesto que ello supone un aumento de experiencia en el entrenamiento del clasificador y aporta más características para el aprendizaje. También es posible incluir un factor de corrección teniendo en cuenta la volumetría de los tweets presentes en cada perfil.

Adicionalmente, existe la posibilidad de utilizar redes sociales como Instagram y algoritmos de reconocimiento de imágenes, para realizar un perfilamiento más efectivo, teniendo en cuenta los términos de servicios de las plataformas.

Por último, cabe señalar que la misma arquitectura de red neuronal puede reproducirse en diferentes contextos, permitiendo obtener clasificadores de textos para distintas temáticas y enfoques.

Los datos y el material se encuentran disponibles en GitHub. Se puede acceder mediante el siguiente enlace: <https://github.com/mpavila/tweets-classification-for-insurance-pricing>

8 Referencias

- Abbar, S., Mejova, Y., & Weber, I. (2015). You Tweet What You Eat: Studying Food Consumption Through Twitter.
- Arizmendi Echecopar, L. F., & Cóndor Espinoza, I. (2019). *Fundamentos de Estadística y Probabilidades con aplicaciones: (en R, Python y otros softwares de tipo GNU/GPL)*. Lima.
- Autoridad Europea de Seguros y Pensiones de Jubilación (EIOPA). (2019). *Big Data Analytics in motor and health insurance: a thematic review*.
- Banca de las Oportunidades; Federación de Aseguradoras Colombianas (Fasecolda); Superintendencia Financiera de Colombia. (2018). *Estudio de demanda de seguros 2018*.
- Baziotis, C., Pelekis, N., & Doulkeridis, C. (2017). DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. En *Proceedings of the 11th International Workshop on Semantic Evaluation* (págs. 747-754). Vancouver: Association for Computational Linguistics.
- Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. En *Transactions of the Association for Computational Linguistics*, 5 (págs. 135-146).
- Bonzanini, M. (2016). *Mastering Social Media Mining with Python*. United States of America: Packt Publishing.
- Borreli, M., & Zappa, D. (2018). From unstructured data and word vectorization to meaning: text mining in insurance.
- Buenadicha, C., & Galdon, G. (2019). *La gestión ética de los datos*. Banco Interamericano de Desarrollo (BID).
- Cañón, V., Clavijo, A., Godoy, L., Letouzé, E., Pestre, G., & Ricard, J. (2017). *Definición de la estrategia de Big Data para el Estado Colombiano y para el desarrollo de la industria de Big Data en Colombia*.
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.
- DANE. (17 de Febrero de 2020). *Proyecciones de población*. Obtenido de <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion>

De Esteban, F. (s.f.). *El futuro del marketing está en la personalización*. Obtenido de Deloitte: <https://www2.deloitte.com/es/es/pages/consumer-business/articles/El-futuro-del-marketing-esta-en-la-personalizacion.html>

De-Gracia, P. (s.f.). *La monitorización de Twitter como técnica para el estudio de la sociología del tiempo*.

Dirección de Censos y Demografía, Departamento Administrativo Nacional de Estadística. (24 de Enero de 2020). *COLOMBIA - Estadísticas Vitales - EEVV - 2017 - 2018*. Obtenido de Archivo Nacional de Datos: http://microdatos.dane.gov.co/index.php/catalog/652/get_microdata

Durán, V. A. (2018). Mercado mundial de seguros 2017. *Revista Fasecolda, (171)*, 21-26.

Eichstaedt, J., Schwartz, H., Kern, M., Park, G., Labarthe, D., Merchant, R., . . . Seligman, M. (2015). Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science 2015, Vol. 26(2)*, 159 –169.

Facebook. (s.f.). *Política de la plataforma de Facebook*. Obtenido de Facebook: <https://developers.facebook.com/policy/>

Fundación Mapfre. (s.f.). *Los seguros, ¿qué son y cómo funcionan?* Obtenido de <https://segurosypensionesparatodos.fundacionmapfre.org/syp/es/seguros/definicion-seguro-asegurar/el-seguro/que-son-como-funcionan/>

García, J., Martínez, J., & García, E. (2016). Revisión de técnicas de pre-procesamiento de textos para la clasificación automática de tweets en español. *Revista de Cómputo Aplicado Vol.1 No.2*, 1-11.

Garrigues, P. (2019). *Diseño e implementación de un clasificador mediante redes neuronales para un sistema de inspección industrial 3D*. Valencia.

Gelbukh, A. (2010). Procesamiento de Lenguaje Natural y sus Aplicaciones. *Komputer Sapiens*, 6-32.

Gelbukh, A. (2010). Procesamiento de Lenguaje Natural y sus Aplicaciones. *Komputer Sapiens*, 6-32.

Google Developers. (2019). *Glosario sobre aprendizaje automático*. Obtenido de Google Developers: https://developers.google.com/machine-learning/crash-course/glossary?hl=es-419#confusion_matrix

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning Word Vectors for 157 Languages. En *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

- Honnibal, M. (10 de Noviembre de 2016). *Embed, encode, attend, predict: The new deep learning formula for state-of-the-art NLP models*. Obtenido de Explosion: <https://explosion.ai/blog/deep-learning-formula-nlp>
- Instituto Nacional De Estadística. (2009). *Tablas de mortalidad Metodología*. Obtenido de https://www.ine.es/daco/daco42/mortalidad/metodo_9107.pdf
- Instituto Nacional de Estadísticas. (2019). *Indicadores Demográficos Básicos*.
- Keller, B. (2018). *Big Data and Insurance: Implications for Innovation, Competition and Privacy*. The Geneva Association.
- Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.
- Lemoine, C. (2018). ACIEMTELECOM 2018. (ACIEM, Entrevistador)
- LexisNexis. (Agosto de 2016). *2016 Usage-based insurance (UBI) research results for the U.S. consumer market*. Obtenido de White paper: <https://www.lexisnexis.com/risk/downloads/whitepaper/2016-ubi-study-white-paper.pdf>
- Liu, S. C. (2019). A Monitoring Physical Activity Levels Using Twitter Data: Infodemiology Study. *J Med Internet Res* 2019;21(6):e12394.
- Magiya, J. (17 de Junio de 2019). *Kendall Rank Correlation Explained*. Obtenido de Towards data science: <https://towardsdatascience.com/kendall-rank-correlation-explained-dee01d99c535>
- Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. Universidad Tecnológica Nacional – Facultad Regional Rosario.
- Mikolov, T. S. (2013). Distributed representations of words and phrases and their compositionality. En *Advances in neural information processing systems* (págs. 3111-3119).
- Ministerio de Tecnologías de la Información y Comunicaciones. (2019). *Colombia es uno de los países con más usuarios en redes sociales en la región*.
- Orduña, E. (2020). Investigando con Twitter: una mirada según el Reglamento General de Protección de Datos. En *Marco Jurídico de la Ciencia de Datos*. Valencia: Tirant Lo Blanch.
- Peachey, K. (2 de Noviembre de 2016). Facebook blocks Admiral's car insurance discount plan. *BBC News*.
- Piña, C. (2018). *Aplicaciones Avanzadas de Bases de Datos (OpenRefine)*.

- Prechelt, L., Orr, G., & Müller, K.-R. (2012). Early Stopping — But When? En G. Montavon, & K.-R. Müller, *Neural Networks: Tricks of the Trade* (págs. 53–67). Springer Berlin Heidelberg.
- PyTorch. (2019). *How to adjust learning rate.* Obtenido de PyTorch: https://pytorch.org/docs/stable/optim.html#torch.optim.lr_scheduler.ReduceLROnPlateau
- PyTorch. (2019). *How to adjust learning rate.* Obtenido de PyTorch: https://pytorch.org/docs/stable/optim.html#torch.optim.lr_scheduler.ReduceLROnPlateau
- Ramírez, J. C., & De Aguas, J. M. (2017). *Configuración territorial de las provincias de Colombia: ruralidad y redes.* Bogotá: Naciones Unidas.
- Rendón, W. (2014). *Redes sociales: el valor de la información y la protección de la privacidad.* Bogotá.
- Rodríguez Alvarez, Y. A. (2017). Innovación, la estrategia de los seguros en Colombia.
- Rodríguez, C. (2011). Genética, seguros y derechos de terceras personas. *Revista de Bioética y Derecho, num 23.*
- Rodríguez, J. M. (2011). *La incertidumbre bioactuarial en el riesgo de longevidad: Reflexiones bioéticas.* Fundación MAPFRE.
- Romero, A. (2017). Protagonistas de Insurance World Challenges 17. (C. o. Insurance, Entrevistador)
- Roosevelt, C. M. (2012). Social Media Analytics: Data Mining Applied to Insurance Twitter Posts. *Casualty Actuarial Society EForum.*
- Rousseau, R., Egghe, L., & Guns, R. (2018). Statistics. En *Becoming Metric-Wise* (págs. 67-97).
- Sadilek, A. K. (2013). Modeling the impact of lifestyle on health at scale. *In WSDM.*
- Sadilek, A., & Kautz, H. (2013). Modeling the Impact of Lifestyle on Health at Scale. *WSDM'13.*
- Sharma, S. (2017). *Epoch vs Batch Size vs Iterations.* Obtenido de Towards data science: <https://towardsdatascience.com/epoch-vs-iterations-vs-batch-size-4dfb9c7ce9c9>
- Soto, L. (30 de Octubre de 2019). *3 formas en que las compañías de seguros pueden mejorar la experiencia del cliente.* Obtenido de Signaturit: <https://blog.signaturit.com/es/3-formas-en-que-las-companias-de-seguros-pueden-mejorar-la-experiencia-del-cliente>

Statista. (Febrero de 2020). *Latin America: Twitter users 2020, by country*. Obtenido de Statista: <https://www.statista.com/statistics/977791/number-twitter-users-in-latin-american-countries/>

Torres, D., & Mayorga, W. (2017). Graduación de una nueva tabla de mortalidad de asegurados de vida individual. *Fasecolda*, (166), 44-53.

Twitter. (s.f.). *Información sobre las API de Twitter*. Obtenido de Twitter: <https://help.twitter.com/es/rules-and-policies/twitter-api>

Twitter. (s.f.). *Twitter Términos de servicio*. Obtenido de <https://twitter.com/es/tos>

Twitter. (s.f.). *Understand our developer policies and agreements*. Obtenido de Developer Twitter: <https://developer.twitter.com/en/developer-terms>

We are social; Hootsuite. (30 de Enero de 2020). *Digital 2020: Global digital overview*. Obtenido de Datareportal: <https://datareportal.com/reports/digital-2020-global-digital-overview>

Weiss, G., Goldberg, Y., & Yahav, E. (2018). On the Practical Computational Power of Finite Precision RNNs for Language Recognition.

Zappa, D., & Borrelli, M. (2017). From unstructured data and word vectorization to meaning: text mining in insurance.

Zappa, D., Borrelli, M., Clemente, G. P., & Nino, S. (2019). Text Mining In Insurance: From Unstructured Data To Meaning.

Abreviaciones

ACIEM	Asociación Colombiana de Ingenieros
API	Interfaz de programación de aplicaciones (Application Programming Interface)
APS	Asociación para la ciencia psicológica (Association for Psychological Science)
DANE	Departamento Administrativo Nacional de Estadística
EIOPA	Autoridad europea de seguros y pensiones de jubilación (European Insurance and Occupational Pensions Authority)
GRU	Unidades recurrentes cerradas (Gated Recurrent Unit)
IoT	Internet de las cosas (Internet of Things)
LSTM	Memoria a largo plazo (Long Short-Term Memory)
MinTIC	Ministerio de Tecnologías de la Información y Comunicaciones
PLN	Procesamiento de Lenguaje Natural
RGPD	Reglamento general de protección de datos (General Data Protection Regulation)
RNN	Red neuronal recurrente (Recurrent Neural Network)
TM	Tasa de mortalidad

Anexos

Anexo A. Población total de Colombia, con desagregación departamental

Departamento	Cod_Dpto	Capital	Población total (2018)
Bogotá	11	Bogotá	7.412.566
Antioquia	5	Medellín	6.407.102
Valle del Cauca	76	Cali	4.475.886
Cundinamarca	25	Bogotá	2.919.060
Atlántico	8	Barranquilla	2.535.517
Santander	68	Bucaramanga	2.184.837
Bolívar	13	Cartagena de Indias	2.070.110
Córdoba	23	Montería	1.784.783
Nariño	52	San Juan de Pasto	1.630.592
Norte de Santander	54	San José de Cúcuta	1.491.689
Cauca	19	Popayán	1.464.488
Magdalena	47	Santa Marta	1.341.746
Tolima	73	Ibagué	1.330.187
Boyacá	15	Tunja	1.217.376
Cesar	20	Valledupar	1.200.574
Huila	41	Neiva	1.100.386
Meta	50	Villavicencio	1.039.722
Caldas	17	Manizales	998.255
Risaralda	66	Pereira	943.401
Sucre	70	Sincelejo	904.863
La guajira	44	Riohacha	880.560
Quindío	63	Armenia	539.904
Chocó	27	Quibdó	534.826
Casanare	85	Yopal	420.504
Caquetá	18	Florencia	401.849
Putumayo	86	Mocoa	348.182
Arauca	81	Arauca	262.174
Vichada	99	Puerto Carreño	107.808
Guaviare	95	San José del Guaviare	82.767
Amazonas	91	Leticia	76.589
San Andrés, Providencia y Santa Catalina	88	San Andrés	61.280
Guainía	94	Inírida	48.114
Vaupés	97	Mitú	40.797
Colombia	00	Bogotá	48.258.494

*Fuente: DANE (2020)¹⁷

¹⁷ DANE. (17 de Febrero de 2020). *Proyecciones de población*. Obtenido de <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion>

Anexo B. Extracción de datos

Para realizar la extracción de datos se instala el complemento Tweet Archiver desde la Suite Marketplace de Google (categoría herramientas para empresas), usando el botón “Instalar”:

Ilustración 65 Suite Marketplace



Para continuar la instalación se debe aceptar las condiciones de servicio y políticas de privacidad de Google, haciendo click en continuar:

Ilustración 66 Aceptación de instalación de Tweet Archiver

Prepárate para la instalación

Para iniciar la instalación de **Tweet Archiver**,
antes debes dar permiso.

Al hacer clic en Continuar, aceptas que tu información se
utilice de conformidad con las [condiciones de servicio](#) y la
[política de privacidad](#) de esta aplicación.

[CANCELAR](#) [CONTINUAR](#)

Seguidamente, se selecciona la cuenta de Google con la que se efectúa la extracción de datos:

Ilustración 67 Selección de cuenta de Google



Después, se confirma que se confía en el complemento y se aceptan los términos del servicio y las políticas de privacidad, pulsando el botón “Permitir”

Ilustración 68 Confirmación de confianza Archiver Tweets

Confirma que confías en Archive Tweets in Spreadsheet

Puede que estés compartiendo información sensible con este sitio web o esta aplicación. Consulta los **términos del servicio** y las **políticas de privacidad** de Archive Tweets in Spreadsheet para saber cómo se tratarán tus datos. Puedes ver o retirar el acceso en cualquier momento en tu [cuenta de Google](#).

[Más información sobre los riesgos](#)

[Cancelar](#)

[Permitir](#)

Por último, se realiza la conexión con la cuenta de Twitter introduciendo el usuario y la contraseña para autorizar la aplicación a acceder a todos los tweets, información de perfil y de seguidores (cuentas seguidas, silenciadas o bloqueadas). Así como a la API de Twitter, por medio de la cual es posible extraer la información requerida.

Ilustración 69 Autorización de acceso a Twitter



El proceso de instalación del complemento y la conexión con Twitter termina con la siguiente pantalla:

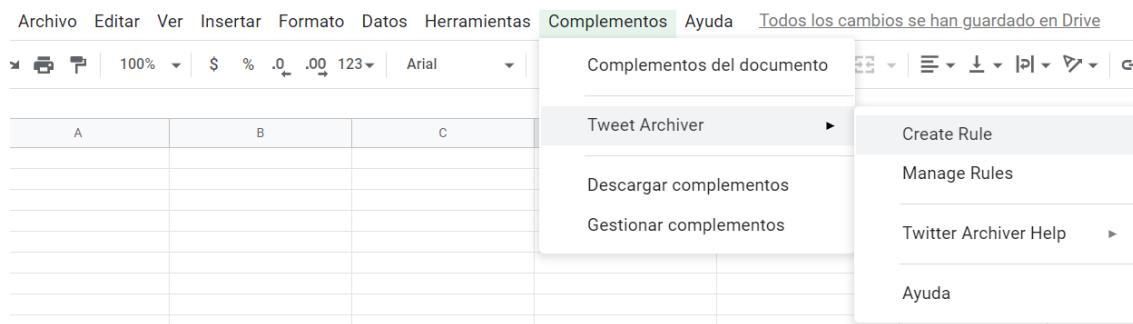
Ilustración 70 Instalación del complemento finalizada

Success!

Your Google Account is now connected with Twitter. Please close this browser tab, switch to your Google Spreadsheet and create a new rule from the Addons > Twitter Archiver menu.
Please visit our [web store](#) if you would like to upgrade to premium edition of Twitter Archiver.
For help, tweet the developer [@labnol](#).

Una vez se encuentre instalado el complemento, se crea una nueva hoja de cálculo de Google, en la que se puede observar que en la pestaña Complementos se ha incluido la opción Tweet Archiver. A través de este botón se despliega el listado de funcionalidades del complemento. Desde aquí es posible generar una regla de extracción de datos de Twitter eligiendo las palabras clave, hashtags o características que se deseen tener en cuenta para la extracción:

Ilustración 71 Complementos de la hoja de cálculo de Google



El complemento permite crear reglas de extracción por palabras, hashtags, ubicación, cuentas, idiomas o búsquedas avanzadas por código. Para el presente trabajo se ejecutan dos clases de búsquedas:

- (1) Sin sesgo: para la cual se aplican dos filtros. El primero es en campo “Written in” y se selecciona “Spanish (español)”. Mientras que el segundo se ejecuta en el campo “Near This Place” (cerca a este lugar) agregando cada uno de los departamentos de Colombia en búsquedas diferentes, de esta forma, se realiza la búsqueda por las coordenadas cercanas al lugar seleccionado.
- (2) Con palabras específicas: incluye el filtro anterior, pero adicionalmente, se agrega en el campo “All of these words” las palabras mencionadas en la Tabla 7 Tabla 7 Palabras clave de búsqueda.

Ilustración 72 Creación de una regla de Twitter

Create Twitter Rule X

All of these words	<input type="text"/>	This exact phrase	<input type="text"/>
Any of these words	<input type="text"/>	None of these words	<input type="text"/>
These #hashtags	<input type="text"/>	Written in	<input type="text"/> Spanish (español) ▼
Near This Place	<input type="text"/> Introduce una ubicación	Advanced Rules	<input type="text"/>

People

To these accounts	<input type="text"/>	Mentioning accounts	<input type="text"/>
From these accounts	<input type="text"/>		

Twitter Search Query:

Create Search Rule Upgrade to Premium Cancel

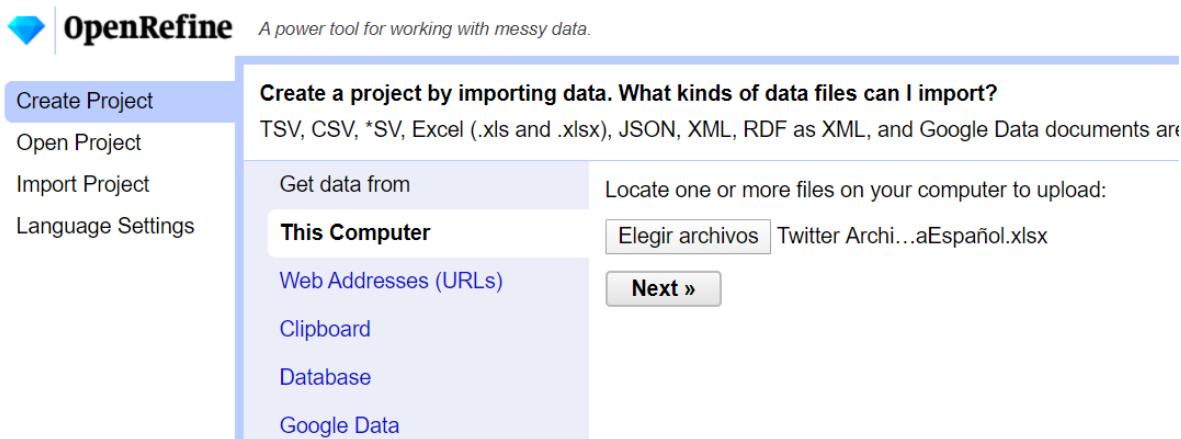
De este modo, al dar click en el botón “Create Search Rule”, se inicia el guardado de los datos en la hoja de cálculo teniendo en cuenta los filtros seleccionados. Este procedimiento tarda aproximadamente 15 segundos.

Cabe resaltar que la versión gratuita solo permite crear una regla de extracción, por lo que, para generar las consultas adicionales, fue requerido eliminar o editar la regla inicial. Siendo esta una limitación de la versión gratuita del complemento.

Anexo C. Creación del proyecto en OpenRefine

Primero se selecciona la opción Create Project, luego se elige el archivo previamente guardado en el equipo (es posible importarlo desde distintas extensiones como CSV, xls, JSON, entre otras) y posteriormente se da click en el botón Next:

Ilustración 73 Elegir archivos- OpenRefine



Una vez cargado el archivo, es posible configurar el análisis de datos que el aplicativo debe implementar para la lectura del documento. Por defecto, la primera fila se toma como encabezados de columna. Adicionalmente, OpenRefine analiza el tipo de dato que contiene cada columna (cadenas de texto, números, fechas y URL entre otros) para ajustar el formato y la vista. También es posible excluir las líneas o celdas en blanco, sin embargo, en el desarrollo de este trabajo se mantendrán para realizar el análisis, por lo que se marcan las opciones “Store blank rows” y “Store blank cells as nulls”:

Ilustración 74 Configuración análisis de datos- OpenRefine



Se comprueba que se haya cargado adecuadamente el archivo explorando las columnas y las filas (utilizando el scrollbar) en la vista previa que proporciona OpenRefine y posteriormente se da click en el botón “Create Project”.

Ilustración 75 Vista previa- OpenRefine

Configure Parsing Options					
	Date	Screen Name	Full Name	Tweet Text	Tweet ID
1.	Thu Mar 05 20:46:52 CET 2020	@MatonsizimoKid	Pretoriano	En un país en manos del crimen organizado, donde los narcotraficantes y paramilitares hacen lo que les da la gana es muy difícil pensar que los capos algún día estén presos, pero ver a Uribe desesperado dando explicaciones reconforta y mucho; por lo menos le quitamos tranquilidad	1235653023764013056
2.	Thu Mar 05 20:46:47	@Soledad_Ajena	Yazmin Cortés	Un hombre tan lindo como tú y con tatuajes, eso no les queda bien #ComoHombres	1235653002863706113

Después de unos pocos minutos se crea un nuevo proyecto, con una apariencia similar a un Excel. Desde esta pantalla es posible explorar los datos, realizar filtros, revisar la información y aplicar las demás funcionalidades de limpieza que ofrece el software:

13554 rows					
	Date	Screen Name	Full Name	Tweet Text	Tweet ID
1.	Thu Mar 05 20:46:52 CET 2020	@MatonsizimoKid	Pretoriano	En un país en manos del crimen organizado, donde los narcotraficantes y paramilitares hacen lo que les da la gana es muy difícil pensar que los capos algún día estén presos, pero ver a Uribe desesperado dando explicaciones reconforta y mucho; por lo menos le quitamos tranquilidad	1235653023764013056
2.	Thu Mar 05 20:46:47 CET 2020	@Soledad_Ajena	Yazmin Cortés	Un hombre tan lindo como tú y con tatuajes, eso no les queda bien #ComoHombres	1235653002863706113
3.	Thu Mar 05 20:46:43 CET 2020	@djaví	Javier Duarte	Otro logro de nuestro subpresidente, ya no necesitamos TOEFL o IELTS I RT @elEspectador: Denuncian que cónsul en Orlando certificó su conocimiento de inglés en una	1235652984396304388 https://twitter.com/elEspectador/status/123565048476635137 http://bit.ly/2vy8B74

Anexo D. Script de Jython/Python para eliminar emoticones en OpenRefine

```
import re

def remove_emojis(data):
    emoj = re.compile("["
        u"\U0001F600-\U0001F64F"  # emoticons
        u"\U0001F300-\U0001F5FF"  # symbols & pictographs
        u"\U0001F680-\U0001F6FF"  # transport & map symbols
        u"\U0001F1E0-\U0001F1FF"  # flags (iOS)
        u"\U00002500-\U00002BEF"  # chinese char
        u"\U00002702-\U000027B0"
        u"\U00002702-\U000027B0"
        u"\U000024C2-\U0001F251"
        u"\U0001f926-\U0001f937"
        u"\U00010000-\U0010ffff"
        u"\u2640-\u2642"
        u"\u2600-\u2B55"
        u"\u200d"
        u"\u23cf"
        u"\u23e9"
        u"\u231a"
        u"\ufe0f"  # dingbats
        u"\u3030"
    "]+", re.UNICODE)
    return re.sub(emoj, '', data)
return remove_emojis(value)
```

Anexo E. Script de Python para el pre-procesamiento de tweets

```
from ekphrasis.classes.preprocessor import TextPreProcessor
from ekphrasis.classes.tokenizer import SocialTokenizer
from ekphrasis.dicts.emoticons import emoticons

text_processor = TextPreProcessor(
    # terms that will be normalized
    normalize=['url', 'email', 'percent', 'money', 'phone', 'user',
               'time', 'url', 'date', 'number'],
    # terms that will be annotated
    annotate={"hashtag", "allcaps", "elongated", "repeated",
              'emphasis', 'censored'},

    fix_html=True, # fix HTML tokens

    # corpus from which the word statistics are going to be used for word
    segmentation
    segmenter="twitter",
    # corpus from which the word statistics are going to be used for spell
    correction
    corrector="twitter",

    unpack_hashtags=True, # perform word segmentation on hashtags
    unpack_contractions=True, # Unpack contractions (can't -> can not)
    spell_correct_elong=False, # spell correction for elongated words

    # select a tokenizer. You can use SocialTokenizer, or pass your own
    # the tokenizer, should take as input a string and return a list of
    tokens
    tokenizer=SocialTokenizer(lowercase=True).tokenize,

    # list of dictionaries, for replacing tokens extracted from the text,
    # with other expressions. You can pass more than one dictionaries.
    dicts=[emoticons]
)
```

Fuente: <https://github.com/cbaziotis/ekphrasis>

**Anexo F. Población y defunciones 2018, con desagregación departamental,
por sexos y grupos quinquenales de las edades de estudio (20 a 49 años)**

Departamento	Cod._ Dpto	Grupos de edad	Proyección población 2018*			Defunciones 2018**		
			Ambos Sexos	Hombres	Mujeres	Ambos Sexos	Hombres	Mujeres
Antioquia	5	20-49	2.861.677	3.094.159	3.312.943	2.394	1.373	1.021
Antioquia	5	20-24	566.674	284.886	281.788	229	149	80
Antioquia	5	25-29	551.560	274.746	276.814	284	172	112
Antioquia	5	30-34	502.813	247.457	255.356	343	200	143
Antioquia	5	35-39	465.107	224.267	240.840	407	241	166
Antioquia	5	40-44	392.383	184.017	208.366	464	249	215
Antioquia	5	45-49	383.140	176.109	207.031	667	362	305
Atlántico	8	20-49	1.100.761	1.234.444	1.301.073	1.184	628	556
Atlántico	8	20-24	218.507	109.227	109.280	118	60	58
Atlántico	8	25-29	209.118	103.612	105.506	129	73	56
Atlántico	8	30-34	191.854	94.047	97.807	170	107	63
Atlántico	8	35-39	179.966	86.925	93.041	205	107	98
Atlántico	8	40-44	155.504	74.010	81.494	227	114	113
Atlántico	8	45-49	145.812	68.512	77.300	335	167	168
Bogotá	11	20-49	3.582.806	3.544.078	3.868.488	2.626	1.452	1.174
Bogotá	11	20-24	711.698	354.265	357.433	269	169	100
Bogotá	11	25-29	695.382	344.761	350.621	314	181	133
Bogotá	11	30-34	629.954	309.841	320.113	328	198	130
Bogotá	11	35-39	580.516	279.587	300.929	452	252	200
Bogotá	11	40-44	500.214	234.000	266.214	469	243	226
Bogotá	11	45-49	465.042	212.319	252.723	794	409	385
Bolívar	13	20-49	868.415	1.027.323	1.042.787	845	459	386
Bolívar	13	20-24	177.390	89.105	88.285	86	46	40
Bolívar	13	25-29	165.669	82.467	83.202	107	66	41
Bolívar	13	30-34	150.117	73.820	76.297	99	52	47
Bolívar	13	35-39	137.478	66.879	70.599	138	72	66
Bolívar	13	40-44	122.515	59.362	63.153	159	87	72
Bolívar	13	45-49	115.246	55.710	59.536	256	136	120
Boyacá	15	20-49	500.317	599.293	618.083	466	270	196
Boyacá	15	20-24	94.288	48.062	46.226	48	28	20
Boyacá	15	25-29	85.015	42.188	42.827	37	21	16
Boyacá	15	30-34	83.245	40.623	42.622	50	30	20
Boyacá	15	35-39	83.445	40.630	42.815	66	40	26
Boyacá	15	40-44	78.893	38.317	40.576	109	61	48
Boyacá	15	45-49	75.431	36.095	39.336	156	90	66
Caldas	17	20-49	420.460	483.738	514.517	431	240	191
Caldas	17	20-24	78.710	39.545	39.165	39	25	14
Caldas	17	25-29	75.348	37.202	38.146	46	28	18
Caldas	17	30-34	69.400	33.790	35.610	55	30	25
Caldas	17	35-39	69.305	33.295	36.010	64	30	34
Caldas	17	40-44	62.833	29.938	32.895	82	51	31
Caldas	17	45-49	64.864	30.646	34.218	145	76	69
Caquetá	18	20-49	166.930	203.804	198.045	156	85	71
Caquetá	18	20-24	35.578	18.559	17.019	20	13	7
Caquetá	18	25-29	32.261	16.500	15.761	11	9	2
Caquetá	18	30-34	28.878	14.656	14.222	20	8	12
Caquetá	18	35-39	26.136	13.070	13.066	24	15	9
Caquetá	18	40-44	23.020	11.213	11.807	39	22	17
Caquetá	18	45-49	21.057	10.107	10.950	42	18	24
Cauca	19	20-49	638.155	725.274	739.214	495	250	245

Departamento	Cod_Dpto	Grupos de edad	Proyección población 2018*			Defunciones 2018**		
			Ambos Sexos	Hombres	Mujeres	Ambos Sexos	Hombres	Mujeres
Cauca	19	20-24	133.572	67.672	65.900	63	41	22
Cauca	19	25-29	120.501	60.158	60.343	58	32	26
Cauca	19	30-34	111.620	55.112	56.508	57	34	23
Cauca	19	35-39	104.778	51.339	53.439	82	40	42
Cauca	19	40-44	88.528	43.005	45.523	108	47	61
Cauca	19	45-49	79.156	38.019	41.137	127	56	71
Cesar	20	20-49	509.155	594.330	606.244	487	237	250
Cesar	20	20-24	105.642	52.729	52.913	49	28	21
Cesar	20	25-29	98.987	48.611	50.376	51	32	19
Cesar	20	30-34	88.710	43.112	45.598	80	42	38
Cesar	20	35-39	80.536	38.946	41.590	86	40	46
Cesar	20	40-44	71.286	34.130	37.156	81	35	46
Cesar	20	45-49	63.994	30.387	33.607	140	60	80
Córdoba	23	20-49	743.130	888.548	896.235	626	309	317
Córdoba	23	20-24	149.809	74.938	74.871	53	32	21
Córdoba	23	25-29	136.512	67.384	69.128	78	42	36
Córdoba	23	30-34	125.024	60.797	64.227	85	34	51
Córdoba	23	35-39	118.519	57.436	61.083	94	50	44
Córdoba	23	40-44	109.186	53.035	56.151	140	71	69
Córdoba	23	45-49	104.080	50.861	53.219	176	80	96
Cundinamarca	25	20-49	1.296.719	1.442.200	1.476.860	922	493	429
Cundinamarca	25	20-24	251.625	127.717	123.908	92	52	40
Cundinamarca	25	25-29	235.802	117.645	118.157	98	56	42
Cundinamarca	25	30-34	224.482	110.447	114.035	114	69	45
Cundinamarca	25	35-39	215.253	105.095	110.158	148	73	75
Cundinamarca	25	40-44	189.493	91.405	98.088	187	102	85
Cundinamarca	25	45-49	180.064	86.323	93.741	283	141	142
Chocó	27	20-49	212.853	264.228	270.598	166	72	94
Chocó	27	20-24	46.753	23.629	23.124	19	8	11
Chocó	27	25-29	43.935	21.330	22.605	18	8	10
Chocó	27	30-34	39.256	18.815	20.441	32	14	18
Chocó	27	35-39	31.809	15.221	16.588	33	20	13
Chocó	27	40-44	27.561	13.104	14.457	29	9	20
Chocó	27	45-49	23.539	11.177	12.362	35	13	22
Huila	41	20-49	454.546	549.094	551.292	418	204	214
Huila	41	20-24	93.104	47.341	45.763	42	30	12
Huila	41	25-29	85.194	42.781	42.413	42	25	17
Huila	41	30-34	77.431	38.527	38.904	46	22	24
Huila	41	35-39	72.916	35.981	36.935	78	34	44
Huila	41	40-44	64.962	31.594	33.368	81	30	51
Huila	41	45-49	60.939	29.307	31.632	129	63	66
La guajira	44	20-49	353.259	431.247	449.313	309	150	159
La guajira	44	20-24	82.098	40.559	41.539	31	21	10
La guajira	44	25-29	70.299	34.054	36.245	36	15	21
La guajira	44	30-34	62.993	30.213	32.780	47	18	29
La guajira	44	35-39	54.610	25.920	28.690	60	31	29
La guajira	44	40-44	45.075	21.261	23.814	64	27	37
La guajira	44	45-49	38.184	17.986	20.198	71	38	33
Magdalena	47	20-49	553.066	671.519	670.227	569	299	270
Magdalena	47	20-24	114.700	58.252	56.448	62	39	23
Magdalena	47	25-29	103.602	51.879	51.723	58	29	29
Magdalena	47	30-34	94.346	46.569	47.777	82	35	47
Magdalena	47	35-39	87.197	42.575	44.622	108	56	52

Departamento	Cod_Dpto	Grupos de edad	Proyección población 2018*			Defunciones 2018**		
			Ambos Sexos	Hombres	Mujeres	Ambos Sexos	Hombres	Mujeres
Magdalena	47	40-44	79.855	38.939	40.916	120	72	48
Magdalena	47	45-49	73.366	35.725	37.641	139	68	71
Meta	50	20-49	460.255	525.903	513.819	406	215	191
Meta	50	20-24	90.791	46.771	44.020	33	22	11
Meta	50	25-29	86.028	43.740	42.288	45	21	24
Meta	50	30-34	80.344	40.622	39.722	57	31	26
Meta	50	35-39	75.516	37.965	37.551	77	43	34
Meta	50	40-44	66.616	33.214	33.402	74	37	37
Meta	50	45-49	60.960	30.112	30.848	120	61	59
Nariño	52	20-49	719.422	798.195	832.397	571	320	251
Nariño	52	20-24	143.731	72.884	70.847	56	39	17
Nariño	52	25-29	131.997	65.948	66.049	54	36	18
Nariño	52	30-34	124.666	61.024	63.642	78	48	30
Nariño	52	35-39	117.569	56.460	61.109	117	71	46
Nariño	52	40-44	104.661	49.528	55.133	104	53	51
Nariño	52	45-49	96.798	45.319	51.479	162	73	89
Norte de Santander	54	20-49	650.847	735.493	756.196	662	392	270
Norte de Santander	54	20-24	133.375	67.629	65.746	56	34	22
Norte de Santander	54	25-29	122.848	61.935	60.913	72	37	35
Norte de Santander	54	30-34	112.841	56.160	56.681	80	47	33
Norte de Santander	54	35-39	104.314	51.092	53.222	129	81	48
Norte de Santander	54	40-44	91.286	43.842	47.444	123	76	47
Norte de Santander	54	45-49	86.183	40.675	45.508	202	117	85
Quindío	63	20-49	229.910	260.251	279.653	271	145	126
Quindío	63	20-24	44.324	22.121	22.203	30	18	12
Quindío	63	25-29	41.324	20.297	21.027	24	15	9
Quindío	63	30-34	37.708	18.206	19.502	33	28	5
Quindío	63	35-39	37.787	18.086	19.701	60	30	30
Quindío	63	40-44	34.075	16.079	17.996	45	19	26
Quindío	63	45-49	34.692	16.009	18.683	79	35	44
Risaralda	66	20-49	405.408	451.418	491.983	463	247	216
Risaralda	66	20-24	77.012	38.111	38.901	56	29	27
Risaralda	66	25-29	74.498	36.168	38.330	46	31	15
Risaralda	66	30-34	68.261	32.777	35.484	64	36	28
Risaralda	66	35-39	67.767	32.198	35.569	64	35	29
Risaralda	66	40-44	59.115	27.840	31.275	100	47	53
Risaralda	66	45-49	58.755	27.191	31.564	133	69	64
Santander	68	20-49	964.570	1.069.599	1.115.238	836	471	365
Santander	68	20-24	187.499	95.152	92.347	71	43	28
Santander	68	25-29	175.937	88.002	87.935	81	49	32
Santander	68	30-34	164.965	81.827	83.138	107	63	44
Santander	68	35-39	156.087	76.490	79.597	162	87	75
Santander	68	40-44	141.943	68.191	73.752	157	91	66
Santander	68	45-49	138.139	65.234	72.905	258	138	120
Sucre	70	20-49	372.552	454.988	449.875	372	208	164
Sucre	70	20-24	75.445	38.525	36.920	33	16	17
Sucre	70	25-29	69.621	34.638	34.983	49	34	15
Sucre	70	30-34	62.894	30.960	31.934	53	30	23
Sucre	70	35-39	58.828	28.899	29.929	54	33	21
Sucre	70	40-44	53.701	26.406	27.295	79	44	35
Sucre	70	45-49	52.063	25.696	26.367	104	51	53
Tolima	73	20-49	536.626	659.286	670.901	544	310	234
Tolima	73	20-24	104.751	53.428	51.323	60	38	22

Departamento	Cod_Dpto	Grupos de edad	Proyección población 2018*			Defunciones 2018**		
			Ambos Sexos	Hombres	Mujeres	Ambos Sexos	Hombres	Mujeres
Tolima	73	25-29	95.081	47.625	47.456	65	40	25
Tolima	73	30-34	87.687	43.305	44.382	70	41	29
Tolima	73	35-39	86.470	42.359	44.111	81	36	45
Tolima	73	40-44	81.107	39.216	41.891	112	66	46
Tolima	73	45-49	81.530	38.982	42.548	156	89	67
Valle del Cauca	76	20-49	1.868.737	2.126.546	2.349.340	2.086	1.143	943
Valle del Cauca	76	20-24	362.100	178.396	183.704	164	107	57
Valle del Cauca	76	25-29	341.703	165.592	176.111	246	151	95
Valle del Cauca	76	30-34	313.694	149.562	164.132	277	144	133
Valle del Cauca	76	35-39	302.982	142.121	160.861	371	207	164
Valle del Cauca	76	40-44	275.836	127.477	148.359	410	214	196
Valle del Cauca	76	45-49	272.422	124.369	148.053	618	320	298
Arauca	81	20-49	114.198	132.519	129.655	102	57	45
Arauca	81	20-24	24.259	12.389	11.870	14	7	7
Arauca	81	25-29	22.612	11.356	11.256	8	6	2
Arauca	81	30-34	19.983	9.981	10.002	13	6	7
Arauca	81	35-39	17.966	8.893	9.073	16	10	6
Arauca	81	40-44	15.405	7.483	7.922	17	9	8
Arauca	81	45-49	13.973	6.740	7.233	34	19	15
Casanare	85	20-49	192.906	212.548	207.956	147	80	67
Casanare	85	20-24	37.901	19.130	18.771	13	7	6
Casanare	85	25-29	36.082	17.886	18.196	20	12	8
Casanare	85	30-34	34.658	17.080	17.578	26	13	13
Casanare	85	35-39	31.347	15.473	15.874	25	15	10
Casanare	85	40-44	28.285	14.119	14.166	27	12	15
Casanare	85	45-49	24.633	12.455	12.178	36	21	15
Putumayo	86	20-49	153.486	175.691	172.491	112	65	47
Putumayo	86	20-24	32.643	16.897	15.746	18	8	10
Putumayo	86	25-29	28.838	14.490	14.348	9	4	5
Putumayo	86	30-34	27.097	13.384	13.713	16	9	7
Putumayo	86	35-39	24.492	12.033	12.459	8	8	-
Putumayo	86	40-44	21.412	10.424	10.988	25	17	8
Putumayo	86	45-49	19.004	9.196	9.808	36	19	17
San Andrés, Providencia y Santa Catalina	88	20-49	26.975	29.595	31.685	17	13	4
San Andrés, Providencia y Santa Catalina	88	20-24	4.706	2.428	2.278	1	1	-
San Andrés, Providencia y Santa Catalina	88	25-29	5.336	2.652	2.684	2	1	1
San Andrés, Providencia y Santa Catalina	88	30-34	4.747	2.277	2.470	1	1	-
San Andrés, Providencia y Santa Catalina	88	35-39	4.305	2.017	2.288	3	2	1
San Andrés, Providencia y Santa Catalina	88	40-44	3.872	1.792	2.080	6	4	2
San Andrés, Providencia y Santa Catalina	88	45-49	4.009	1.827	2.182	4	4	-
Amazonas	91	20-49	29.504	39.911	36.678	26	20	6
Amazonas	91	20-24	6.473	3.536	2.937	4	4	-
Amazonas	91	25-29	5.908	3.094	2.814	5	5	-
Amazonas	91	30-34	5.335	2.719	2.616	3	3	-
Amazonas	91	35-39	4.628	2.376	2.252	2	1	1
Amazonas	91	40-44	3.848	1.998	1.850	5	2	3
Amazonas	91	45-49	3.312	1.695	1.617	7	5	2

Departamento	Cod_Dpto	Grupos de edad	Proyección población 2018*			Defunciones 2018**		
			Ambos Sexos	Hombres	Mujeres	Ambos Sexos	Hombres	Mujeres
Guainía	94	20-49	16.685	25.138	22.976	24	13	11
Guainía	94	20-24	4.332	2.368	1.964	4	3	1
Guainía	94	25-29	3.205	1.648	1.557	3	1	2
Guainía	94	30-34	2.751	1.402	1.349	5	3	2
Guainía	94	35-39	2.465	1.244	1.221	1	-	1
Guainía	94	40-44	2.113	1.076	1.037	6	4	2
Guainía	94	45-49	1.819	939	880	5	2	3
Guaviare	95	20-49	33.536	44.260	38.507	26	10	16
Guaviare	95	20-24	7.413	4.165	3.248	1	1	-
Guaviare	95	25-29	6.003	3.234	2.769	2	-	2
Guaviare	95	30-34	5.614	2.928	2.686	2	1	1
Guaviare	95	35-39	5.114	2.625	2.489	7	3	4
Guaviare	95	40-44	5.004	2.574	2.430	6	3	3
Guaviare	95	45-49	4.388	2.320	2.068	8	2	6
Vaupés	97	20-49	11.746	21.425	19.372	22	18	4
Vaupés	97	20-24	3.062	1.735	1.327	7	7	-
Vaupés	97	25-29	2.243	1.219	1.024	2	2	-
Vaupés	97	30-34	1.952	1.052	900	4	3	1
Vaupés	97	35-39	1.661	907	754	3	1	2
Vaupés	97	40-44	1.482	824	658	3	2	1
Vaupés	97	45-49	1.346	742	604	3	3	-
Vichada	99	20-49	41.039	57.242	50.566	30	13	17
Vichada	99	20-24	9.958	5.410	4.548	5	3	2
Vichada	99	25-29	8.088	4.249	3.839	3	-	3
Vichada	99	30-34	7.004	3.668	3.336	3	-	3
Vichada	99	35-39	6.139	3.253	2.886	3	2	1
Vichada	99	40-44	5.400	2.900	2.500	4	4	-
Vichada	99	45-49	4.450	2.447	2.003	12	4	8

*Fuente: DANE (2020)¹⁸

**Fuente: Dirección de Censos y Demografía (2020)¹⁹

¹⁸ DANE. (17 de Febrero de 2020). *Proyecciones de población*. Obtenido de <https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion>

¹⁹ Dirección de Censos y Demografía, Departamento Administrativo Nacional de Estadística. (24 de Enero de 2020). *COLOMBIA - Estadísticas Vitales - EEVV - 2017 - 2018*. Obtenido de Archivo Nacional de Datos: http://microdatos.dane.gov.co/index.php/catalog/652/get_microdata

Anexo G. Tasa de mortalidad con desagregación departamental

Departamento	Cod_Dpto	Grupo de edad	Población Ambos Sexos 2018 (1)	Defunciones Ambos Sexos 2018 (2)	TM Ambos Sexos (2/1)
Bogotá	11	20-49	3.582.806	2.626	0,073%
Antioquia	5	20-49	2.861.677	2.394	0,084%
Valle del Cauca	76	20-49	1.868.737	2.086	0,112%
Cundinamarca	25	20-49	1.296.719	922	0,071%
Atlántico	8	20-49	1.100.761	1.184	0,108%
Santander	68	20-49	964.570	836	0,087%
Bolívar	13	20-49	868.415	845	0,097%
Córdoba	23	20-49	743.130	626	0,084%
Nariño	52	20-49	719.422	571	0,079%
Norte de Santander	54	20-49	650.847	662	0,102%
Cauca	19	20-49	638.155	495	0,078%
Magdalena	47	20-49	553.066	569	0,103%
Tolima	73	20-49	536.626	544	0,101%
Boyacá	15	20-49	500.317	466	0,093%
Cesar	20	20-49	509.155	487	0,096%
Huila	41	20-49	454.546	418	0,092%
Meta	50	20-49	460.255	406	0,088%
Caldas	17	20-49	420.460	431	0,103%
Risaralda	66	20-49	405.408	463	0,114%
Sucre	70	20-49	372.552	372	0,100%
La guajira	44	20-49	353.259	309	0,087%
Quindío	63	20-49	229.910	271	0,118%
Choco	27	20-49	212.853	166	0,078%
Casanare	85	20-49	192.906	147	0,076%
Caquetá	18	20-49	166.930	156	0,093%
Putumayo	86	20-49	153.486	112	0,073%
Arauca	81	20-49	114.198	102	0,089%
Vichada	99	20-49	41.039	30	0,073%
Guaviare	95	20-49	33.536	26	0,078%
Amazonas	91	20-49	29.504	26	0,088%
San Andrés, Providencia y Santa Catalina	88	20-49	26.975	17	0,063%
Guainía	94	20-49	16.685	24	0,144%
Vaupés	97	20-49	11.746	22	0,187%
Total	00	20-49	21.090.651	18.811	0,089%

Anexo H. Valor del modelo con desagregación departamental

Departamento	Predicción					Total Tweets (1)	Healthy Tweets (2)	vm _d (2/1)
	Healthy	Practice_sports	Positive	Negative	Non-healthy			
Atlántico	13	10	100	52	1	268	444	123 27,7%
Cesar	19	10	84	49	1	258	421	113 26,8%
La Guajira	18	5	128	72	2	341	566	151 26,7%
San Andrés	8	3	62	43	5	154	275	73 26,5%
Valle del Cauca	18	12	137	81	7	376	631	167 26,5%
Córdoba	22	6	92	57	7	281	465	120 25,8%
Chocó	16	1	87	54	4	245	407	104 25,6%
Magdalena	19	5	84	72	1	251	432	108 25,0%
Cundinamarca	20	5	132	76	13	385	631	157 24,9%
Quindío	12	4	66	51	2	201	336	82 24,4%
Norte de Santander	4	4	46	22	3	158	237	54 22,8%
Caquetá	8	5	38	19	1	156	227	51 22,5%
Antioquia	28	5	132	76	4	506	751	165 22,0%
Cauca	8	8	82	38	3	310	449	98 21,8%
Amazonas	14	1	41	10	1	194	261	56 21,5%
Guaviare	8	2	43	14	7	175	249	53 21,3%
Bolívar	26	11	94	62	6	422	621	131 21,1%
Sucre	4	4	38	20	4	150	220	46 20,9%
Santander	9	10	78	71	1	305	474	97 20,5%
Caldas	11	5	67	45	1	280	409	83 20,3%
Tolima	15	2	64	41	0	284	406	81 20,0%
Huila	6	5	72	54	3	284	424	83 19,6%
Boyacá	15	3	86	68	4	361	537	104 19,4%
Risaralda	10	7	62	54	6	281	420	79 18,8%
Meta	6	0	48	34	2	199	289	54 18,7%
Bogotá	161	44	642	395	184	3383	4809	847 17,6%
Casanare	15	4	42	15	2	284	362	61 16,9%
Putumayo	1	0	2	0	0	15	18	3 16,7%
Nariño	14	6	79	62	3	482	646	99 15,3%
Arauca	12	2	38	29	3	284	368	52 14,1%
Vichada	0	0	6	3	1	57	67	6 9,0%
Total	540	189	2.772	1.739	282	11.330	16.852	3.501 20,8%