Universidad CENFOTEC

Maestría en Tecnologías de Base de Datos

Documento final de Proyecto de Investigación Aplicada 2

Evaluación de la precisión de distintos modelos de minería de datos para la posible aplicación en el contexto de atención de emergencias del sistema 9-1-1 de Costa Rica.

Durán Cerdas Johan Stiff

Julio, 2021

Declaratoria de derecho de autor

Dedicatoria

A mi esposa, Vanessa Meza Espinoza, por su paciencia, comprensión y por ser mi apoyo incondicional en todo momento.

Agradecimientos

Gracias a mi tutor, el profesor Diego Alfaro Bergueiro, por aceptar ser parte de este proceso y por su disposición, soporte y comprensión a lo largo del mismo.

universidad cenfotec_
tecnologías digitales

Universidad Cenfotec
Carrera de Postgrado
Maestría en Tecnología de Bases de Datos

TRIBUNAL EXAMINADOR

Este proyecto fue aprobado por el Tribunal Examinador de la carrera: **Maestría en Tecnología de Bases de Datos,** requisito para optar por el título de grado de **Maestría,** para el estudiante: **Durán Cerdas Johan.**

DIEGO ALONSO ALFARO BERGUEIRO (FIRMA)
Firmado digitalmente por DIEGO ALONSO ALFARO BERGUEIRO (FIRMA)
Fecha: 2021.08.27 18:32:41 -06'00'

MBD. Diego Alfaro Bergueiro
Tutor

LUIS CARLOS NARANJO ZELEDON (FIRMA)
Firmado digitalmente por LUIS CARLOS NARANJO ZELEDON (FIRMA)
Fecha: 2021.09.01 17:12:47 -06'00'

MBA. Luis C. Naranjo Zeledón
Lector 1

IGNACIO TREJOS ZELAYA (FIRMA)
Firmado digitalmente por IGNACIO TREJOS ZELAYA (FIRMA)
Fecha: 2021.09.06 22:40:14 -06'00'

M. Sc. Ignacio Trejos Zelaya
Lector 2

San José, Costa Rica, 17 de agosto de 2021

Firmada digitalmente, de conformidad con la Ley de Certificados, Firmas Digitales y Documentos Electrónicos N° 8454, destacando el artículo 9°-

Tabla de contenido

Resumen

Los sistemas de atención de emergencias son un ente fundamental en la sociedad mundial. Costa Rica no es la excepción con el sistema de atención de emergencias 9-1-1. Este ofrece un punto centralizado para atender emergencias con diferentes requerimientos, desde violencia hasta problemas de salud e incendios. El sistema 9-1-1 de Costa Rica se basa en llamadas telefónicas, sin costo, restricción ni necesidad de operadora telefónica (chip). Esto significa que, desde cualquier teléfono, se puede llamar al 9-1-1 sin importar las condiciones.

El surgimiento de nuevas tecnologías es creciente y acelerado, el área de minería de datos e inteligencia artificial avanza a grandes pasos, existen herramientas poderosas para convertir voz a texto, acceso a videoconferencia, el almacenamiento y procesamiento se ha vuelto barato, el GPS (Sistema de Posicionamiento Global, del inglés *Global Positioning System*) es un elemento presente en prácticamente todos los móviles. Sin embargo, ninguno de estos elementos es utilizado en la atención de emergencias.

En la presente investigación, se realiza la evaluación de distintos algoritmos de minería de datos para una posible aplicación en la clasificación de emergencias. Esto mediante aprendizaje supervisado y un conjunto de datos alternativo. Los datos del sistema 9-1-1 costarricense son de carácter confidencial y es por dicha razón que se hace uso de datos alternativos.

La posible aplicación de técnicas de minería de datos sobre una llamada puede traer beneficios desde distintos puntos. Desde un asistente para los operadores telefónicos, el cual brinde sugerencias basadas en datos; procesamiento en tiempo real para redireccionar unidades de forma rápida, hasta mensajes de voz y texto para solicitar ayuda, incluyendo metadatos como la ubicación.

El abanico de posibilidades de una migración a 9-1-1 basado en la web es muy grande y, a pesar de que puede traer consigo cambios necesarios en infraestructura y cultura, los beneficios para la oportuna atención de emergencias vitales pueden ser muy grande.

Palabras clave: emergencia, clasificación, minería de datos, algoritmos, aprendizaje supervisado, 9-1-1.

Capítulo 1. Introducción

1.1 Generalidades

Para poder realizar una clasificación, es necesario tener un punto de referencia y un objetivo. En Costa Rica, la base de datos del 9-1-1 es de carácter confidencial inclusive para investigación. Por ende, no es posible obtener acceso a dichos datos para poder realizar análisis sobre estos. Dada la situación, en la presente investigación, se hace uso de un conjunto de datos alternativo, el cual contiene noticias clasificadas según cuatro categorías, en el idioma inglés. A pesar de que no es texto de emergencias, funciona para evaluar el posible desempeño de distintos modelos. Para su elección, se toman en consideración ciertas características con las que debe contar.

1.2 Antecedentes del problema

Durante los últimos años, el tema de atención de emergencias ha sido solucionado mediante el uso de llamada telefónica al 9-1-1. No se tiene registro de aplicaciones, sistemas o iniciativas para automatizar la atención, clasificación o soporte en la toma de decisiones relacionadas con situaciones de emergencias en Costa Rica.

Desde un punto internacional, con respecto a las tecnologías utilizadas en otras partes del mundo, se puede mencionar el 9-1-1 mejorado (E-911, del inglés *Enhanced* 9-1-1); en Europa, existe un sistema similar denominado E-112. Tal como se menciona en el sitio web del Gobierno de Estados Unidos (911 and E911 Services, s.f.), el sistema mejorado de 9-1-1 ofrece un elemento de gran valor al proveer la ubicación de la persona llamando. Para ello, la red telefónica transmite dicha información. De esta forma, se optimiza el tiempo de respuesta.

Por otra parte, nuevos sistemas surgen con los avances de la inteligencia artificial y el aprendizaje máquina. Sin embargo, dichos sistemas tienden a brindar soporte más que toma de decisiones. Un ejemplo de ello es Corti (Corti.ai), el cual ofrece un servicio de soporte al personal que atiende la emergencia. Actualmente, el sistema oficial de atención de emergencias de Copenhague lo utiliza para brindar soporte al personal de atención de emergencias. El sistema ofrece detección de distintas situaciones, incluyendo ataques al corazón, además, brinda información basada en datos durante y posterior a la emergencia.

1.3 Definición y descripción del problema

A pesar de que la fuente del reporte de la emergencia es generalmente una llamada, en la actualidad, la lista de herramientas y API (Interfaz de programación de aplicaciones, del inglés *Application Programming Interface*) que existen para convertir de voz a texto es muy grande, entre ellas se pueden mencionar:

- *Google speech-to-text.*
- *Amazon Transcribe.*
- *Microsoft Azure Speech to Text.*

Por ende, convertir de voz a texto no aporta a la investigación y el punto de partida es el texto como tal. El objetivo de este trabajo es brindar una visión del estado actual, en cuanto a alternativas y posibles candidatos a mejores algoritmos para el tema tratado específicamente. A pesar de que la presente investigación hace referencia a la evaluación de modelos de minería de datos, se muestra una serie de posibilidades, en términos de visión, que se pueden abrir basados en un correcto algoritmo de minería de datos.

La oportuna atención de una emergencia depende de varios factores, entre ellos se pueden mencionar:

- Disponibilidad de un operador para atender la llamada.
- Estado de la persona que está llamando. Por ejemplo: tranquilidad para brindar la información solicitada.
- Conocimiento del lugar para poder brindar la dirección.
- Distancia hasta el ente necesario: Fuerza Pública, Cruz Roja, Bomberos, entre otros.
- Disponibilidad de las unidades.

En caso de no haber un operador disponible, la persona debe esperar. Para el sistema 9-1-1, esto significa que debe tener una considerable cantidad de personas dedicadas a la atención de emergencias. Caso contrario, se pueden perder vidas. Por otra parte, la información que se puede proporcionar mediante una llamada se reduce a lo transmitido por la persona y no se hace uso de la gran cantidad disponible de herramientas en los dispositivos actuales.

1.4 Justificación

El sistema de atención de emergencias 9-1-1 es un ente fundamental en la sociedad costarricense. Una correcta atención y redirección de unidades basados

en la llamada puede marcar la diferencia entre la vida y la muerte. Sin embargo, son muchos los factores que influyen en la correcta atención. Uno de ellos es la capacidad del 9-1-1 de atender la llamada de forma rápida y oportuna. Para una correcta atención, es necesario contar con recurso humano y tecnológico suficiente.

En la actualidad, una debilidad del sistema 9-1-1 es la dependencia a llamadas telefónicas comunes, en una época donde el Internet gobierna con una variedad muy grande de sensores y utilidades que se podrían integrar para mejorar la atención. Además, tal como se puede observar en la Figura 1, el número de llamadas reales por incidente es prácticamente igual al número de llamadas realizadas de forma errónea.

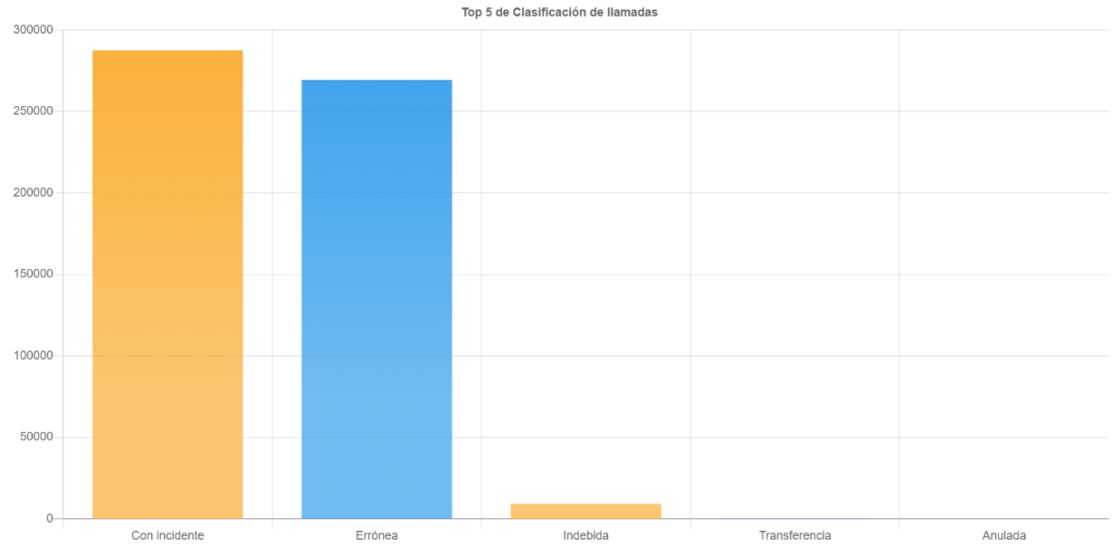


Figura 1: Clasificación de llamadas.

Fuente: (9-1-1 Costa Rica, 2019).

Basado en la información presentada anteriormente, es claro que el sistema 9-1-1 podría verse beneficiado, en gran medida, gracias a los avances en temas de minería de datos, inteligencia artificial, aprendizaje máquina, entre otros. Además, la capacidad computacional ha dejado de ser un problema con los grandes avances en computación en la nube y servicios por los cuales se paga lo que se utiliza.

Si se unen los puntos, para obtener un producto altamente efectivo, se podrían mencionar sensores como el GPS para obtener la ubicación exacta de la persona y atender la emergencia de forma directa. Además, para un manejo automático o supervisado de toma de decisiones, se debe iniciar con la creación y

entrenamiento de algoritmos altamente efectivos que logren diferenciar situaciones de emergencia y clasificarlas, según la necesidad, de forma rápida, efectiva y oportuna. Este también puede funcionar tanto para atención directa como para soporte en la toma de decisiones.

A modo de resumen, un sistema capaz de clasificar las llamadas puede traer consigo beneficios tanto a nivel de procesos como a nivel de atención de emergencia y tiempo de respuesta. Es por dichas razones que el presente trabajo se enfoca en el análisis y evaluación de distintos modelos de minería de datos para la posible aplicación en el contexto de atención de emergencias del sistema 9-1-1, de esta forma, se podría tener una base para una gran variedad de futuros trabajos.

1.5 Viabilidad

Durante los últimos años, la minería de datos ha tenido gran impacto en temas de análisis textual y también en algoritmos de clasificación. Por ende, a pesar de que en Costa Rica no se cuenta con un medio tecnológico para clasificar emergencias basado en la tecnología y técnicas actuales, es muy viable el desarrollo de este proyecto. A continuación, se desarrollan los motivos técnicos, operativos y económicos por los que se consideran que el desarrollo del proyecto de investigación es factible.

1.5.1 Punto de vista técnico.

Como profesional con experiencia en distintas áreas de las ciencias de la computación y, además, como futuro Máster en Tecnologías de Base de Datos, el autor cuenta con el conocimiento y experiencia necesario para poder desarrollar, aplicar y evaluar distintos modelos de minería de datos, con el fin de obtener resultados satisfactorios. Además, gracias al proceso de investigación, se espera recopilar y obtener conocimiento proveniente de los trabajos realizados por otros investigadores.

Por otra parte, desde el punto de vista tecnológico, hay tres puntos primordiales: poder computacional, bibliotecas (incluyendo las de minería de datos) y un lenguaje de programación o utilidad capaz de procesar. Con respecto al poder computacional, en los últimos años, tener acceso a poder computacional es muy sencillo y barato. De hecho, muchos de los algoritmos pueden ser ejecutados inclusive por una máquina convencional. En términos de bibliotecas, existe una gran variedad de ellas para distintas acciones, desde creación de los algoritmos, limpieza

de datos y vectorización, a lo largo del desarrollo de la investigación se explica cada uno de ellos. Finalmente, con respecto a el lenguaje de programación, hay una gran variedad. Sin embargo, para procesamiento de minería de datos, no todos son igual de funcionales; en el mercado, hay dos que lideran el área: *Python* y R.

1.5.2 Punto de vista operativo.

Tal como se mencionó anteriormente, la base de datos del 9-1-1 es de carácter confidencial. Por dicha razón no fue posible obtener datos de esta. Además, se contactó con distintos entes internacionales e investigadores de otras instituciones que publicaron al respecto. De todas las formas, no fue posible obtener una base de datos de situaciones de emergencia con las respectivas transcripciones para ser utilizada como entrada para los algoritmos. Por la anterior razón, se hace uso de una base de datos alternativa, la cual es de noticias con cuatro categorías. Además de la base de datos, se hace uso de una herramienta de terceros de pago, para obtener el poder de procesamiento necesario para la tarea.

La herramienta utilizada para escribir el código necesario se denomina *Colaboratory*, propiedad de *Google*. La razón de la elección es porque esta herramienta ofrece *la* posibilidad de programar en *Python* y, además, involucra *Jupyter Notebooks*, la cual es de mucha utilidad para mantener los resultados y que sean amigable con el usuario. Además, la versión Pro, de pago, tiene un precio muy accesible y, con ello, se obtiene CPU, Memoria y GPU extra. Lo anterior en conjunto con un mayor tiempo de procesamiento. Esto debido a que, en la versión estándar, luego de una cantidad pequeña de procesamiento, la máquina se cierra, lo que conlleva a que el algoritmo no complete su entrenamiento.

1.5.3 Punto de vista económico.

En cuanto al costo del proyecto, los gastos de investigación, *hardware*, *software*, licencias y demás gastos asociados serán incurridos por el autor. Se utilizan como puntos de referencia los datos de salarios de sitio Glassdoor, específicamente para el puesto de Científico de Datos. Esta información es de Estados Unidos. Tal como se puede observar en la Figura 2, el salario promedio es de alrededor de $113 000. Sin embargo, dado que dichos datos son para Estados Unidos, como punto de referencia para Costa Rica se utiliza el límite inferior, es decir, $83 000 anuales.

Figura 2: Rango salarial para Científico de Datos en U.S.A.

Fuente: (Data Scientist Salaries, 2021)

En la siguiente tabla, se puede observar el costo hora para consultor, tomando como referencia el valor anteriormente mencionado. Además, se considera la semana laboral de 40 horas y 52 semanas laborales al año.

Tabla 1: Desglose de salario.

Salario anual	**Salario mensual**	**Salario diario**	**Salario hora**
$83 000	$6916	$345	$43

Fuente: Elaboración propia basada en la información obtenida de Data Scientist Salaries (2021).

Posteriormente, los algoritmos a ejecutar son complejos y pesados para una computadora personal convencional. Por dicha razón, se hace uso de una licencia paga de terceros para obtener procesamiento en la nube. La herramienta utilizada se denomina *Colaboratory* y es propiedad de *Google*. Esta tiene un costo de $9 mensuales, costo incurrido por el desarrollador del proyecto.

1.6 Objetivos

Se utiliza la taxonomía de Bloom de 1956, debido a que utiliza una estructura jerárquica, lo cual facilita la definición de objetivos, desde su nivel más general hasta los niveles más específicos, permitiendo así una visión amplia de lo que se espera lograr.

1.6.1 Objetivo general.

- Evaluar la precisión de distintos modelos de minería de datos para la posible aplicación en el contexto de atención de emergencias del sistema 9-1-1 de Costa Rica.

1.6.2 Objetivos específicos.

- Identificar las fuentes de datos a utilizar como entrada para los modelos de minería de datos.
- Interpretar los datos para constatar que funcionan como datos alternos a datos de emergencias reales.
- Convertir los datos de su forma raíz a datos numéricos que funcionen como entrada para los modelos.
- Aplicar los distintos modelos de minería de datos propuestos.
- Analizar los resultados de los algoritmos de minería de datos.
- Proponer el modelo de clasificación con los mejores resultados.

1.7 Alcances y limitaciones

1.7.1 Alcances

- Documento donde se presenta la comparación y evaluación de distintos modelos de minería de datos para el caso de estudio presentado. Finalmente, se presenta un algoritmo como mejor candidato basado en criterio de precisión y tiempo de ejecución.

1.7.2 Limitaciones

- No se utilizarán datos reales de emergencias.
- Los datos no están en el idioma español.
- No se evalúan todos los algoritmos actualmente en el mercado.
- Por temas de tiempo de procesamiento, no se realiza búsqueda exhaustiva de hiperparámetros.

1.8 Marco de referencia organizacional y socioeconómico

9-1-1 Costa Rica es el sistema oficial de atención de emergencias en Costa Rica. El medio utilizado para la comunicación es la llamada telefónica desde teléfonos fijos o móviles. Un aspecto importante de considerar es que, para realizar una llamada al sistema de emergencias, no es necesario contar con dinero de saldo para llamadas, sea para planes pospago o prepago. Además, tiene cobertura nacional total (Cobertura de Servicio, 2019). Uno de los beneficios del servicio del 9-1-1 es que se puede llamar, inclusive, sin línea celular y puede utilizar cualquier antena, inclusive si no es del operador utilizado. Al realizarse una llamada al sistema

9-1-1, se tiene un formato común para la atención, en general, se deben responder algunas preguntas, las cuales se detallan en la Tabla 2.

Tabla 2: Preguntas a responder durante una emergencia.

Pregunta	**Descripción**
¿Qué sucede?, ¿cuál es la emergencia?	Naturaleza de la emergencia. Se debe brindar información, tal como: la descripción de las personas, los vehículos que intervienen, las armas que hubiera, hace cuánto tiempo ocurrió el incidente, incendio, accidente de tránsito, etc.
¿Dónde sucede la emergencia?	Para atender la emergencia de la forma más oportuna, se debe brindar la dirección de la forma más exacta posible. Preferiblemente utilizando provincia, cantón y distrito. Punto de referencia y descripción del lugar.
Información personal	Indicar nombre completo y número telefónico del que se llama.

Fuente: (Uso del 9-1-1, 2019).

Es importante recalcar que, por seguridad de ambas partes involucradas, todas las llamadas realizadas son grabadas.

1.8.1 Historia.

Tal como es presentado en el sitio del 9-1-1 (2020), la historia del sistema nacional de emergencias 9-1-1 se remota al año 1988. En dicho año, se enfrentó en el país la presencia del huracán Joann, mejor conocido como “Juana”, durante el cual, se evacuaron alrededor de 30 mil personas, la zona más afectada por el huracán fue la zona Sur del país y dejó más de 23 personas fallecidas.

En dado momento, el país no contaba con un sistema telefónico de atención de emergencias. La situación de emergencia llevó al vicepresidente de la República, el ingeniero Jorge Manuel Dengo Obregón, a visitar Estados Unidos, durante este

viaje, conoció el sistema 9-1-1 estadounidense. De regreso de la visita, trajo la idea a Costa Rica de unificar los sistemas de emergencia, lo cual no se logró en ese momento.

Posteriormente, en 1990, el presidente de la Comisión Nacional de Emergencias, el Dr. Humberto Trejos Fonseca, convocó distintas instituciones para traer de vuelta a la luz la idea del Ingeniero Dengo. En una primera propuesta, se pretendían reducir los tiempos de respuesta de las emergencias en la Gran Área Metropolitana, al mismo tiempo que se optimizaban los recursos de las instituciones involucradas.

El Dr. Trejos solicitó un estudio de factibilidad, utilizando como base las necesidades de los distintos entes, requerimientos financieros y recursos humanos. El estudio fue parcialmente financiado por la Canadian Project Preparation Facility.

Una vez aprobado, se diseñó un sistema para satisfacer las necesidades de la población por un periodo estimado de 10 años. En sus inicios, nueve organizaciones formaron parte de programa, las cuales son: Cruz Roja Costarricense, Cuerpo de Bomberos (INS), Guardia Rural, Guardia Civil, Ministerio de Seguridad Pública, hospitales y Centro Nacional de Control de Intoxicaciones, Organismo de Investigación Judicial, Policía de Tránsito, Asociación para el Sordo (inicialmente, fue considerada en el ámbito de consulta).

En 21 de enero de 1994, se inauguró la central telefónica, ubicada en un terreno donado por la Comisión Nacional de Emergencias. Por razones técnicas, se dio inicio utilizando el número 1-2-2, fue hasta el 30 de abril que se empezó a utilizar el 9-1-1. En junio de 1995, el sistema 9-1-1 fue aprobado mediante el Decreto Número 24418. Posteriormente, en diciembre de 1995, el sistema 9-1-1 se adscribe a Instituto Costarricense de Electricidad, mediante la Ley número 7566. Durante los años siguientes, se incorpora al sistema de emergencias el Instituto Nacional de las Mujeres (INAMU), el Patronato Nacional de la Infancia (PANI) y, finalmente, el Instituto WEM (9-1-1 Costa Rica, 2020).

1.8.2 Tipo de negocio y mercado meta.

Tal como se menciona en FILOSOFÍA EMPRESARIAL (MISIÓN, VISIÓN, OBJETIVOS) (2019) el sistema de atención de llamadas e incidentes de emergencia 9-1-1 tiene como objetivo atender de manera oportuna distintas situaciones de emergencia que sucedan en territorio nacional. Para dicha atención, se utiliza una táctica de atención interinstitucional para brindar un número que es capaz de redirigir a su correspondiente ente óptimo para la atención.

1.8.3 Misión, visión y valores.

El sistema de atención de emergencias 9-1-1 tiene como misión garantizar una respuesta interinstitucional en el trámite de las llamadas e incidentes de carácter de emergencia. Principalmente sustentado en infraestructura tecnológica innovadora, gestión de procesos y un equipo de trabajo disciplinado y comprometido (9-1-1 Costa Rica, 2019).

La visión a la que se ajustan es ser reconocidos como el mejor 9-1-1 de América Latina en integración interinstitucional, infraestructura tecnológica, recurso humano y calidad de servicio (9-1-1 Costa Rica, 2019).

1.9 Estado de la cuestión

En la presente sección, se desarrolla una compilación, selección y clasificación de documentos técnicos que son relevantes para el desarrollo de la presente investigación.

1.9.1 Planificación de la revisión.

En esta etapa, se formula una pregunta clara y bien definida sobre el tema de investigación. Se realiza una búsqueda de documentos relacionados con el tema, para conocer qué se ha realizado en el área y, además, obtener una visión clara de las fortalezas y debilidades del tema propuesto. Basado en la revisión, se puede corroborar que el proyecto no es un duplicado de otros estudios y que se brinda un aporte a la comunidad científica.

1.9.1.1 Formulación de la pregunta.

La formulación de la pregunta ayuda a delimitar los alcances de la búsqueda de información y análisis de esta. De esta forma, se logra una mejor precisión sobre los resultados obtenidos.

1.9.1.1.1 Foco de la pregunta.

A pesar de que el proyecto se basa en análisis de texto, es importante, además, incluir en las búsquedas elementos relacionados con análisis de voz para reconocer situaciones de emergencia. Dado que, para hacer análisis de voz, en la mayor parte de las ocasiones, se hace conversión a texto. Por ende, gran parte de esos estudios contienen información muy relacionada con el presente proyecto.

1.9.1.1.2 Amplitud y calidad de la pregunta.

Para el desarrollo de esta sección, se establece la pregunta de investigación de una forma clara y concisa enfocada en el problema que se desea resolver. Para no perder el foco, se hace un listado de palabras clave. Además, para facilitar el desarrollo, se consideran factores clave como lo son la población, aplicación, control y efectos.

1. **Problema**

El medio oficial para el reporte de emergencias es el llamado al 9-1-1. Sin embargo, existen muchas situaciones en las que no es posible entablar una conversación con el operador para explicar claramente la situación, dado que se está en alto riesgo. En otras ocasiones, el problema es a nivel de procesos, principalmente por la saturación de línea en la cual no se puede atender oportunamente a la persona. Es en esa clase de situaciones donde una aplicación que capture texto o sonido ambiente y reconozca la situación podría ser la diferencia entre recibir ayuda oportunamente o no recibirla.

Desde el punto de vista de procesos, el sistema 9-1-1 tiene una gran dependencia sobre el recurso humano necesario para atender las emergencias. Tal como se menciona en la noticia de La Nación (2018), el sistema 9-1-1, al 2018, tenía 175 empleados, de los cuales, 135 se dedican exclusivamente a la atención de llamadas. Por su parte, uno de los principales ingresos del sistema es el impuesto a

las llamadas telefónicas, monto que se ha visto afectado por el uso intensivo de aplicaciones de internet como WhatsApp, que ha llevado a una gran reducción en el uso de llamada común y, con ello, a los ingresos del sistema 9-1-1. Según indica la noticia, al 2018, el sistema 9-1-1 acumulaba una deuda de ¢4000 millones.

En términos generales, la migración a las aplicaciones basadas en la web es masiva y, por ende, la presente investigación se enfoca en estudios que se hayan realizado en relación con técnicas de minería de datos para clasificar texto con respecto a una serie de categorías.

2. Pregunta

Con la anterior definición del problema, se formula la siguiente pregunta de investigación:

¿Qué estudios se han llevado a cabo en el área de minería de datos relacionados con análisis de texto y voz para la clasificación de datos con preferible enfoque en situaciones de emergencia?

3. Palabras clave y sinónimos

Se muestra un listado de palabras claves relacionadas con el tema de investigación para ser utilizadas como entrada a los sitios de búsqueda. Algunas de las palabras están en idioma inglés, debido a que una gran parte de las investigaciones está en este idioma. El listado de palabras se puede observar en la Tabla 3.

Tabla 3: Listado de palabras clave.

Palabra	**Equivalente en inglés**
Minería de datos	*Data mining*
Aprendizaje supervisado	*supervised learning*
Aprendizaje automático	*Machine learning*
Reconocimiento de voz	*Voice recognition*
Clasificación de texto	*Text classification*
Situación de emergencia	*Emergency situation*

Fuente: Elaboración propia

4. Intervención

Observar los resultados de cómo la utilización de técnicas de minería de datos puede ayudar a la clasificación de situaciones de emergencia basado en una entrada textual al algoritmo. Así como extraer y analizar los documentos de mayor relevancia para la investigación.

5. **Control**

No se cuenta con ninguna información inicial. Por ende, se realiza una búsqueda desde cero.

6. **Efectos**

Se espera obtener los documentos y artículos necesarios para comprender los distintos algoritmos de minería de datos utilizados para clasificar situaciones de emergencia basados en entrada de texto y, además, definir cuáles brindan los mejores resultados. En caso de no obtenerse una cantidad considerable de investigaciones relacionadas directamente con situaciones de emergencia, se procede a reducir la búsqueda a clasificación de texto.

7. **Medida de salida**

Con la documentación obtenida, se realiza una revisión de la calidad en sitios especializados para dicho propósito.

8. **Población**

La población objetivo de esta investigación son desarrolladores e investigadores. Además, dado el estado actual, se espera que la investigación abra puertas para que la población en general y distintas instituciones vean el impacto, beneficio y flexibilidad que podría traer el cambio.

9. **Aplicación**

El resultado de esta investigación brinda una puerta para que desarrolladores e investigadores puedan crear aplicaciones de soporte y atención oportuna de emergencias.

10. **Diseño experimental**

Durante el diseño experimental, se realiza un análisis y clasificación de los estudios basado en la calidad y relevancia de la investigación con el tema de

estudio actual. Basado en ello, se obtienen suficientes estudios y, a la vez, se filtran, de forma tal que no se introduzca ruido al estudio.

1.9.1.2 Selección de fuentes.

En esta sección, se especifican las fuentes para la identificación de estudios primarios que se utilizarán para la investigación.

1.9.1.2.1 Definición del criterio de selección de fuentes.

Para la selección de fuentes, se han tomado en cuenta distintos aspectos, entre ellos la disponibilidad de acceso a la información, el respaldo por la comunidad investigadora y el respaldo teórico.

1.9.1.2.2 Lenguaje de estudio.

Para la realización de las búsquedas, se utiliza tanto el idioma español como el inglés. El español para analizar estudios en el idioma de la población objetivo y el inglés, porque la mayoría de los estudios en el área son desarrollados en este idioma.

1.9.1.2.3 Identificación de fuentes.

En esta sección, se presentan las fuentes seleccionadas y se describe cómo se ejecutan las búsquedas.

1. Método de selección de fuentes

El presente estudio se basa principalmente en el respaldo que tiene la fuente en el área de la tecnología, la facilidad de acceso y la relación con el estudio.

2. Cadena de búsqueda

Las cadenas de búsqueda utilizadas tienen una combinación de ‘AND’ y ‘OR’, como se muestra a continuación: *(“data mining” OR “machine learning” OR “supervised learning” OR* “minería de datos” *OR* “aprendizaje automático” *OR* “aprendizaje supervisado”) *AND (“emergency situation” OR “emergency management” OR “emergency call” OR* “situación de emergencia” *OR* “manejo de emeregencias” *OR* “llamado de emeregencia”) *AND (“voice recognition” OR “text clasification” OR “text recognition” OR “speech recognition” OR “asr” OR “automatic speech recognition” OR* “reconocimiento de voz” *OR* “clasificación de texto” *OR*

“reconocimiento de texto” *OR* “reconocimiento automático de habla” *OR* “reconocimiento de voz”).

3. Lista de fuentes

1. *ACM Digital Library.*
2. *IEEE Digital Library.*
3. *Research Gate* (utilizada en caso de no tenerse suficientes registros en las primeros dos fuentes).

1.9.1.2.4 Selección de fuentes después de la evaluación.

En este apartado, se depende de la facilidad para aplicar las cadenas de búsqueda y la calidad de los documentos obtenidos. En adición, se considera de suma importancia la facilidad para acceder al material.

1.9.1.2.5 Comprobación de las fuentes.

Las fuentes seleccionadas, durante muchos años, han tenido impacto principalmente sobre el área de tecnología y son ampliamente utilizadas y reconocidas.

1.9.1.3 Selección de los estudios.

Una vez definidas las fuentes, se definen cuáles trabajos obtenidos en las búsquedas van a ser incluidos en el análisis final.

1.9.1.3.1 Definición del criterio de inclusión y exclusión de estudios.

Se utilizan los criterios presentados en la Tabla 4 como método para incluir o excluir un trabajo. Por su parte, los que cumplan dichos requisitos son candidatos a ser incluidos.

Tabla 4: Criterio de inclusión y exclusión de estudios.

Pregunta de investigación	**Término principal para criterio de inclusión**	**Criterio de exclusión**
¿Qué estudios se han llevado a cabo en el área de minería de datos	“data mining”, “machine learning”, “supervised learning”, “minería de	-Documentos de minería de datos para análisis de texto o texto

relacionados con análisis de texto y voz para la detección de situaciones de emergencia?	datos", "aprendizaje automático", "aprendizaje supervisado", "emergency situation", "emergency management", "situación de emergencia", "manejo de emergencias", "*voice recognition", "text clasification", "text recognition", "speech recognition", "automatic speech recognition",* "reconocimiento de voz", "clasificación de texto", "reconocimiento de texto", "reconocimiento de voz"	proveniente de la voz que no tengan relación con detección de situaciones de emergencias. -Documentos relacionados con detección de situaciones de emergencias que no tengan relación o aplicación de técnicas de minería de datos sobre texto o texto proveniente de voz.

Fuente: Elaboración propia.

1.9.1.3.2 Definición de tipos de estudio

La definición de tipos de estudio se relaciona con la pregunta de investigación, en la Tabla 5, se definen los requisitos para definir los artículos de interés.

Tabla 5: Tipos de estudio.

Pregunta de investigación	**¿Quién?**	**¿Qué?**	**¿Cómo?**	**¿Dónde?**
¿Qué estudios se han llevado a cabo en el área de minería de datos relacionados con análisis de texto y voz para la clasificación de	Persona de cualquier edad y género.	Situaciones de emergencia, minería de datos.	Clasificación, detección, identificación y predicción.	Casas de habitación, sitios públicos, espacios con y sin ruido.

datos con preferible enfoque en situaciones de emergencia?				

Fuente: Elaboración propia.

1.9.1.3.3 Procedimiento para la selección de los estudios.

Para la selección de los estudios, se realizó el siguiente proceso iterativo para cada fuente:

1. Seleccionar la opción de búsqueda avanzada.
2. Utilizar la cadena de búsqueda aplicables para obtener los resultados que satisfacen dichas condiciones.
3. Si la cantidad de estudios es superior a 50 resultados, aplicar el filtro de fechas para obtener publicaciones de los últimos 5 años.
4. Evaluar los resultados obtenidos y aplicar los criterios de exclusión basado en el resumen y palabras clave del artículo.
5. Seleccionar los resultados considerados relevantes para la fuente consultada. Repetir el proceso con las demás fuentes definidas.

1.9.2 Ejecución de la revisión.

1.9.2.1 Ejecución de la selección en la fuente ACM.

1.9.2.1.1 Selección de estudios iniciales.

Tabla 6: Estudios iniciales de la fuente ACM.

#	Título	Autores	Año	URL
1	*CognitiveEMS: a cognitive assistant system for emergency medical services*	Sarah Preum, Sile Shu, Mustafa Hotaki, Ronald Williams, John Stankovic, Homa Alemzadeh	2019	https://dl.acm.org/doi/pdf/10.1145/3357495.3357502
2	*Decision making in assistive environments using multimodal observations*	Yong Lin, Eric Becker, Kyungseo Park, Zhengyi Le, Fillia Makedon	2009	https://dl.acm.org/doi/pdf/10.1145/1579-1-14.1579120
3	*LifeLine: A Device for Detecting Abnormal Patterns*	J. Jenny Li, Peter Krivoshik, Andrey Suvorov, Claudia Fortes, Phillip Kenny	2018	https://dl.acm.org/doi/pdf/10.1145/3243250.3243265

Fuente: Elaboración propia.

1.9.2.1.2 Evaluación de la calidad de los estudios.

Figura 3: Calidad de la fuente ACM.

Fuente: (Journal of the ACM, 2021).

1.9.2.1.3 Revisión de la selección.

Para la selección de los estudios, se realiza una revisión del resumen, inicialmente, si el contenido luce prometedor para la investigación, se procede con el contenido del artículo. En la revisión los artículos, estos fueron ordenados por relevancia.

1.9.2.1.4 Extracción de información.

Para la extracción de la información de los estudios primarios, se consideran los siguientes elementos:

- Detección de situaciones de emergencia utilizando texto o texto proveniente de la voz.
- Uso de aprendizaje supervisado.
- Aprendizaje automático aplicado a detección de situaciones de emergencia.
- Evaluación de algoritmos de minería de datos.
- Evaluación de algoritmos de aprendizaje automático.

Tabla 7: Extracción fuente 1.

Repositorio	**ACM Digital Library**

Título	*CognitiveEMS: a cognitive assistant system for emergency medical services*
Publicación	ACM SIGBED Review. Volumen 16, emisión 2, agosto 2019, páginas 51–60.
Referencia	(Preum et al., 2019)
	Descripción
Área	Asistente cognitivo. Emergencias médicas. Reconocimiento de habla. Procesamiento de lenguaje natural. Sistema de manejo de emergencias.
Resumen	El artículo presenta un modelo de asistente cognitivo para servicios médicos de emergencias donde se utilizan una serie de sensores inteligentes y dispositivos portátiles en conjunto con análisis y colecta de datos en tiempo real para ofrecer sugerencias de cómo proceder en forma dinámica según suceden los acontecimientos. Durante el proceso, se realiza una conversión de voz a texto y basado en dicho texto se extraen los términos biomédicos que, posteriormente, son convertidos a términos médicos, mediante la utilización de herramientas como MetaMap y CLAMP. Sobre los datos obtenidos, el investigador aplica técnicas de procesamiento de lenguaje natural. Al poseer varias fuentes de datos, la idea general de arquitectura es la de un *pipeline,* en el cual se procesan los datos en distintas fases, consiguiendo distinta información para obtener el resultado esperado.
Aspectos a destacar	• Para la conversión voz a texto, se hace uso del API de Google. • Se utilizan herramientas externas para hacer el análisis textual y de lenguaje: METAMAP y CLAMP. No se desarrollan algoritmos de minería de datos.

	• Utiliza llamadas de radio y también reportes de cuidados de los pacientes. • Las herramientas utilizadas en el artículo realizan un análisis léxico, sintáctico y semántico.

Fuente: Elaboración propia.

Tabla 8: Extracción fuente 2.

Repositorio	**ACM Digital Library**
Título	*Decision making in assistive environments using multimodal observations*
Publicación	*PETRA '09: Proceedings of the 2nd International Conference on Pervasive Technologies Related to Assistive Environments.* Junio 2009, artículo No.: 6, páginas 1–8.
Referencia	(Lin, Becker, Park, Le y Makedon, 2009)
	Descripción
Área	Inteligencia artificial. Metodologías computacionales. Sistema de respuesta ante emergencias.
Resumen	El artículo trata la necesidad de movilidad para mejorar la eficiencia de los sensores que comúnmente son estáticos. Por ende, se presenta un modelo robótico, el cual utiliza procesos de decisión parcialmente obsérvales de Markov. El modelo propuesto incorpora la captura de audio y posterior conversión a texto. Sin embargo, el audio y el texto son principalmente utilizados para discernir sonidos como gemidos y no oraciones o palabras como tal.
Aspectos a destacar	• El motor de reconocimiento de habla utilizado es Sphinx, software libre. • Se hace reconocimiento de ruidos más que de oraciones relacionadas con la emergencia.

Fuente: Elaboración propia

Tabla 9: Extracción fuente 3.

Repositorio	**ACM Digital Library**
Título	*LifeLine: A Device for Detecting Abnormal Patterns*
Publicación	*PRAI 2018: Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence*. Agosto 2018, páginas 76–81.
Referencia	(Li, Krivoshik, Suvorov, Fortes y Kenny, 2018)
	Descripción
Área	Inteligencia artificial. Aprendizaje máquina. Detección de movimiento.
Resumen	El artículo presenta una aplicación o dispositivo para detección de situaciones de peligro mediante el uso de sensores y algoritmos de inteligencia artificial y aprendizaje máquina. Una vez reconocida la situación, el dispositivo atiende la emergencia de forma autónoma, enviando alarmas por varios medios de comunicación. El dispositivo se enfoca principalmente en detección de movimiento. A pesar de que integra reconocimiento de habla, este no presenta la forma en que es realizado. Por otra parte, los datos obtenidos son enviados a otro servidor y son utilizados para interpretar futuras situaciones mediante inteligencia artificial.
Aspectos a destacar	No se brinda una explicación del método de reconocimiento de voz utilizado y los modelos de clasificación sobre estos.

Fuente: Elaboración propia.

1.9.2.2 Ejecución de la selección en la fuente IEEE.

1.9.2.2.1 Selección de estudios iniciales.

Tabla 10: Estudios iniciales de la fuente IEEE.

#	Título	Autores	Año	URL

1	*A Behavior Tree Cognitive Assistant System for Emergency Medical Services*	Sile Shu, Sarah Preum, Haydon M. Pitchford, Ronald D. Williams, John Stankovic, Homa Alemzadeh	2019	https://ieeexplore.ieee.org/document/8968233

Fuente: Elaboración propia.

1.9.2.2.2 Evaluación de la calidad de los estudios.

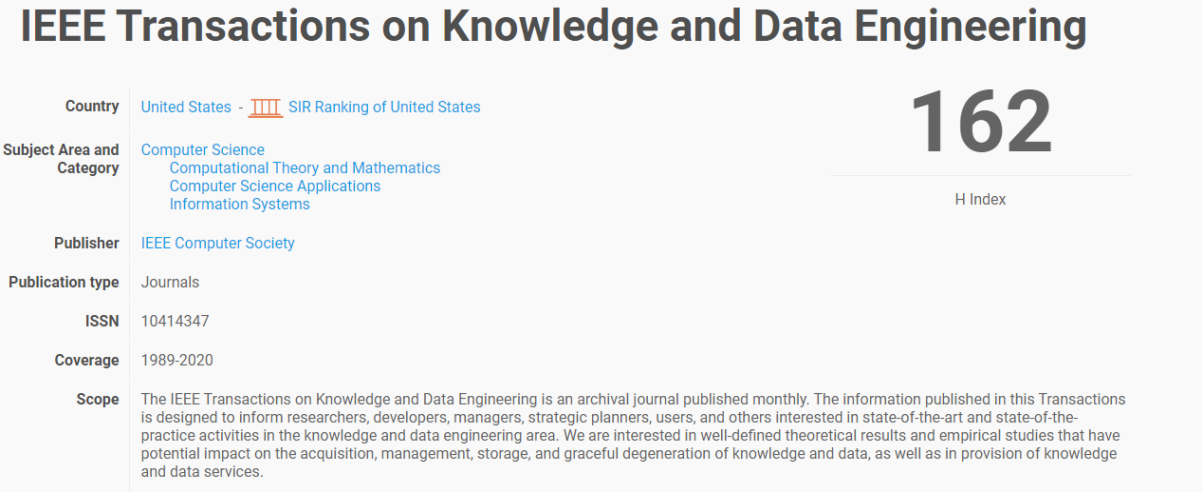


Figura 4: Calidad de la fuente IEEE.

Fuente: (IEEE Transactions on Knowledge and Data Engineering, 2021)

1.9.2.2.3 Revisión de la selección.

Para la selección de los estudios, se realiza una revisión del resumen, inicialmente, si el contenido luce prometedor para la investigación, se procede con el contenido del artículo. En la revisión los artículos, estos fueron ordenados por relevancia.

1.9.2.2.4 Extracción de información

Para la extracción de la información de los estudios primarios se consideran los siguientes elementos:

- Detección de situaciones de emergencia utilizando texto o texto proveniente de la voz.
- Uso de aprendizaje supervisado.

- Aprendizaje automático aplicado a detección de situaciones de emergencia.
- Evaluación de algoritmos de minería de datos.
- Evaluación de algoritmos de aprendizaje automático.

Tabla 11: Extracción fuente 4.

Repositorio	**IEEE Digital Library**
Título	A Behavior Tree Cognitive Assistant System for Emergency Medical Services
Publicación	IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Noviembre 2019, páginas 6188-6195.
Referencia	(Shu, et al., 2020)
	Descripción
Área	Servicios de emergencias médicas. Árbol de comportamiento. Procesamiento de lenguaje natural. Aprendizaje máquina.
Resumen	El artículo presenta un sistema asistente cognitivo para servicios médicos de emergencias (EMS). El marco de trabajo propuesto se basa en arboles de comportamiento (BT) que combina el uso de aprendizaje automático, reconocimiento de habla y procesamiento de lenguaje natural para inferir la situación y tomar acciones al respecto. El motivo expuesto sobre el uso de BT es porque es modulable, rápida respuesta y principalmente la habilidad de adaptarse y aprender mediante aprendizaje reforzado.
Aspectos a Destacar	• Se utilizan los árboles de comportamiento como modelo de análisis. • Se utiliza el API de Google para la conversión de voz a texto.

Fuente: Elaboración propia.

1.9.2.3 Ejecución de la selección en la fuente Research Gate

1.9.2.3.1 Selección de estudios iniciales

Tabla 12: Estudios iniciales de la fuente Research Gate.

#	Título	Autores	Año	URL
1	A Brief Survey of Machine Learning Algorithms for Text Document Classification on Incremental Database	Nihar M. Ranjan, Midhun Chakkaravarthy	2020	https://www.researchgate.net/publication/350451142_A_Brief_Survey_of_Machine_Learning_Algorithms_for_Text_Document_Classification_on_Incremental_Database/stats
2	Text Classification Using Data Mining Techniques: A Review	Oluwakemi Christiana Abikoye, Samuel Oladeji Omokanye , Taye Oladele Aro	2018	https://www.researchgate.net/profile/Idongesit-Eteng/publication/326059800_An_Online_Voting_System_for_Colleges_and_Universities/links/5b35fd44a6fdcc8506dba334/An-Online-Voting-System-for-Colleges-and-Universities.pdf#page=4

Fuente: Elaboración propia.

1.9.2.3.2 Evaluación de la calidad de los estudios

Tal como se menciona, el uso de *Research Gate* se plantea en caso de no tenerse suficientes estudios. Al presentarse la situación se procede a la búsqueda de resultados en dicha fuente. Sin embargo, las fuentes obtenidas son citadas según su publicación. Además, tal como se presenta los resultados no están relacionados con emergencias.

1.9.2.3.3 Revisión de la selección

Para la selección de los estudios, se realiza una revisión del resumen, inicialmente, si el contenido luce prometedor para la investigación, se procede con el contenido del artículo. En la revisión los artículos, estos fueron ordenados por relevancia.

1.9.2.3.4 Extracción de información.

Para la extracción de la información de los estudios primarios, se consideran los siguientes elementos:

- Detección de situaciones de emergencia, utilizando texto o texto proveniente de la voz.
- Uso de aprendizaje supervisado.
- Aprendizaje automático aplicado a detección de situaciones de emergencia.
- Evaluación de algoritmos de minería de datos.
- Evaluación de algoritmos de aprendizaje automático.

Tabla 13: Extracción fuente 1.

Repositorio	**Research Gate**
Título	*A Brief Survey of Machine Learning Algorithms for Text Document Classification on Incremental Database*
Publicación	*Test Engineering and Management*. Volumen 83, junio 2020, páginas 25246 – 25251.
Referencia	(Nihar y Midhun, 2021)
	Descripción
Área	Aprendizaje máquina.

	Aprendizaje supervisado. Clasificación de texto.
Resumen	El artículo recopila una serie de algoritmos de minería de datos para la clasificación de texto. Entre los algoritmos presentados están: • Máquina de soporte vectorial • Redes neuronales. • Bosques aleatorios. • Sistemas difusos. • Teoría probabilista. El artículo no se enfoca en brindar ejemplos, sino, más bien, en brindar una perspectiva en los métodos y autores que han trabajado en ellos. Finalmente, propone un modelo hibrido. Para este modelo híbrido, se presenta el uso de aprendizaje semisupervisado.
Aspectos a destacar	• Excelente resumen de algoritmos de valor para la investigación. • Ofrece un punto de entrada para una posible elección de algoritmos.

Tabla 14: Extracción fuente 2.

Repositorio	**Research Gate**
Título	*Text Classification Using Data Mining Techniques: A Review*
Publicación	*Information Systems Education Journal.* Volumen 22, mayo 2018, páginas 1–8.
Referencia	(Oluwakemi, Samuel y Taye, 2018)
	Descripción
Área	Aprendizaje máquina. Minado de texto. Evaluación de rendimiento. Aprendizaje supervisado. Clasificación de texto.

Resumen	El artículo explica una serie de algoritmos para la clasificación de texto. Entre los algoritmos presentados se tienen: • Redes bayesianas. • K vecinos más cercanos. • Máquina de soporte vectorial. • Redes neurales artificiales. • Árboles de decisión A pesar de que el artículo brinda referencias a investigaciones, no se tienen ejemplos o similar. Es un trabajo teórico que resume una serie de algoritmos. Finalmente, brinda una breve explicación de cómo evaluar con referencia a otros estudios.
Aspectos a destacar	• Brinda una serie de algoritmos que pueden ser de utilidad para la presente investigación. • A pesar de que no es aplicativo, las referencias son de utilidad para obtener fuentes confiables.

Fuente: Elaboración propia.

1.9.4 Resumen de los resultados

Como parte del análisis en la *ACM, IEEE* y *Research Gate*, el total de estudios analizados es de 35 distribuidos entre las 3 entidades. Basado en ellos, se filtró y redujo hasta obtener un total de 6 estudios. En la Tabla 15, se presenta un resumen de los resultados obtenidos.

Tabla 15: Análisis de resultados.

Fuente	Estudios	Relevantes	Primarios
ACM Digital Library	15	7	3
IEEE	8	4	1
Research Gate	12	5	2
Total	35	16	6

Fuente: Elaboración propia.

Capítulo 2. Marco conceptual

De forma metódica, se genera una nube de palabras para obtener una visión general de los términos más frecuentes. Para la generación de la nube de palabras, se utiliza el contenido del estado de la cuestión. Tal como se presenta en Figura 5, la cantidad de palabras es muy grande, pero sobre ella sobresalen muchas con un mayor tamaño, tamaño que indica la frecuencia de apariciones de la palabra.

Figura 5: Nube de palabras.

Fuente: Elaboración propia utilizando el sitio https://www.nubedepalabras.es

Con base en los términos más frecuentes en la nube de palabras, se procede a crear un diagrama de términos y distintas relaciones de importancia, este se

puede observar en la Figura 6.

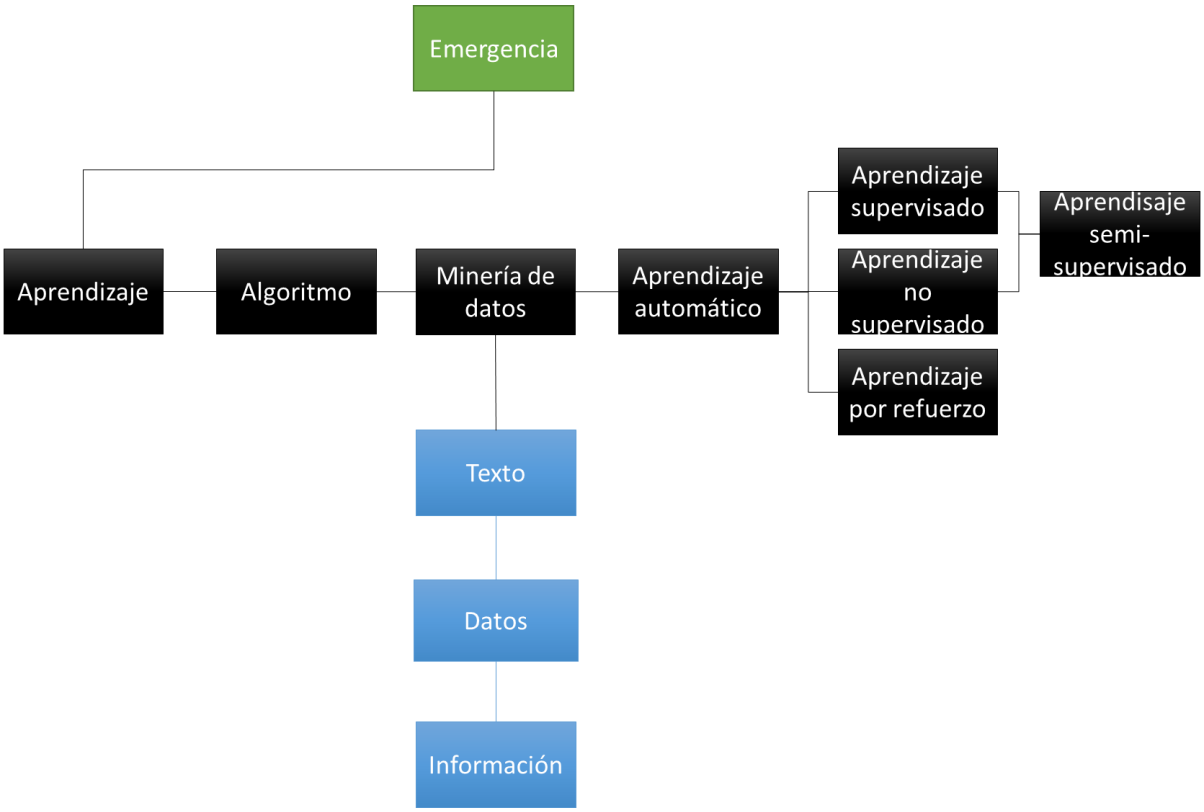


Figura 6: Diagrama de términos.

Fuente: Elaboración propia.

Además, muchos conceptos de importancia para la investigación no están presentes como términos en el estado de la cuestión. Por ende, la siguiente lista presenta los distintos términos a ser explicados en el presente capítulo y el correspondiente orden:

- Emergencia.
- Emergencia vital.
- CRISP-DM.
- Lenguaje de programación.
- Python.
- Jupyter.
- Google Colaboratory.
- Dato.
- Información.
- Conjunto de datos.

- Usabilidad.
- Algoritmo.
- Datos de entrenamiento.
- Datos de prueba.
- Limpieza de datos.
- Lematización.
- Precisión.
- term-frequency-matrix.
- TF-IDF.
- Modelo de clasificación.
- Minería de datos.
- Aprendizaje máquina.
- Hiperparámetro.
- Sobreajuste (uderfitting).
- Subajuste (overfitting).
- Máquina de soporte vectorial.
- Redes bayesianas.
- Regresión logística multinomial.
- Bosques aleatorios.
- K-vecinos más cercanos.

2.1 Conceptos de emergencias

2.1.1 Emergencia

Una emergencia se puede definir como: “Situación de peligro o desastre que requiere una acción inmediata” (Real Academia Española, 2020). En Costa Rica, en caso de presentarse una emergencia, el número de contacto oficial es el 9-1-1, al igual que países como Estados Unidos, México, Canadá y muchos otros más. Sin embargo, el 9-1-1 no es un número de uso global. Por ejemplo, en Europa y parte de Asia el número utilizado es 112.

2.1.2 Emergencia vital

Entiéndase por emergencia vital a las condiciones clínicas que impliquen un posible riesgo de muerto o secuela funcional grave (Superintendencia de Salud de

Chile, s.f.). Cualquier elemento fuera de la anterior definición es considerado una emergencia no vital.

2.2 CRISP-DM

Según Hipp y Wirth's (2000), CRISP-DM (*Cross Industry Standard Process for Data Mining*) se define como un modelo de proceso integral para el desarrollo de proyectos de minería de datos. Además, mencionan que el proceso es independiente del sector y la tecnología también. CRISP-DM se divide en seis fases principales, cada una de ellas dividida en tareas más pequeñas. La seis principales fases son:

1. Entendimiento del negocio.

2. Entendimiento de los datos.

3. Preparación de los datos.

4. Modelado.

5. Evaluación.

6. Despliegue.

A pesar de que se definen una serie de fases, el proceso tiene un enfoque “cíclico”, tal como se presenta en la Figura 7.

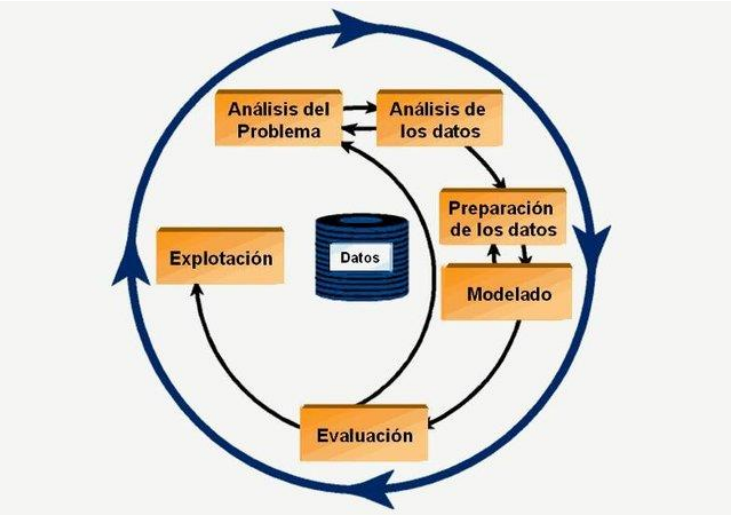


Figura 7: Ciclo de vida de CRISP-DM.

Fuente: (Ordoñes, Horacio y Grass Boada, 2011).

2.3 Ambientes y programación

2.3.1 Lenguaje de programación

De acuerdo con López (2020):

> Un lenguaje de programación, en palabras simples, es el conjunto de instrucciones a través del cual los humanos interactúan con las computadoras. Un lenguaje de programación nos permite comunicarnos con las computadoras a través de algoritmos e instrucciones escritas en una sintaxis que la computadora entiende e interpreta en lenguaje de máquina.

2.3.2 Python

Holden (2018) en el sitio oficial de *Python* lo define como un lenguaje interpretado de alto nivel, dinámico, orientado a objetos y de propósito general que puede ser utilizado en una basta cantidad de aplicaciones. Interpretado hace referencia al uso de un intérprete para traducir el código a lenguaje máquina sin necesidad de un compilado previo.

2.3.3 Jupyter Notebook

Jupyter (2021) indica que *Jupyter Notebook* es un proyecto de tipo código abierto para la interacción con distintos lenguajes de programación. Haciendo énfasis en *Python*, Julia y R. *Jupyter* ofrecen una variedad de funcionalidades para hacer de la interacción con código una tarea más amena. Su objetivo principal es apoyar el desarrollo de proyectos de ciencia de datos. Fue creado en 2014 a partir de *IPython*.

Python puede ser instalado directamente en una máquina y ejecutar el código presente en este proyecto. Sin embargo, *Jupyter* ofrece la funcionalidad de mostrar el código como un documento en el cual el usuario puede ejecutar y agrupar secciones con base en las características. Además, permite agregar comentarios con formato para facilitar su comprensión. En la Figura 8, se presenta un ejemplo simple de la utilización y los beneficios de utilizar una herramienta como *Jupyter.* Inicialmente, se puede observar “**Prueba de comentario con formato negrita**” seguido de una numeración y viñetas. A la derecha, se puede observar cómo luce dicho texto, una vez completada la edición. Posteriormente, se puede observar código *Python*, donde se realiza la suma de 12 + 12 + 24 y se imprime el resultado. Si, por ejemplo, se desea compartir dicho documento con un usuario no experto,

este contiene toda la información, comentarios, código y resultados. Al contrario, con una terminal *Python* pura, para saber los resultados, es necesario volver a ejecutar los comandos.

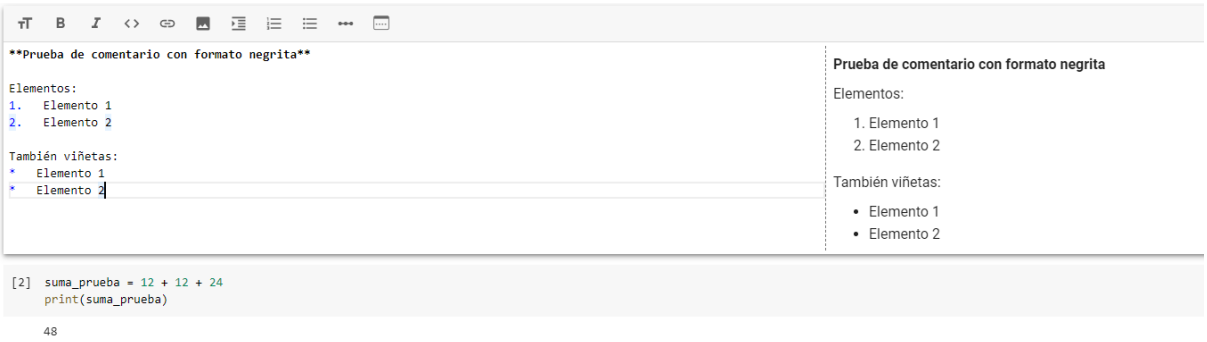


Figura 8: Ejemplo de utilización del lenguaje Python mediante Jupyter Notebooks.

Fuente: Elaboración propia

2.3.4 Google Colaboratory

Python es un lenguaje de programación, *Jupyter* es una herramienta para mejorar la usabilidad y entendimiento de los procesos realizados en *Python*. Ambas necesitan una máquina para correr. Puede ser una computadora de uso personal, en la nube u otras. *Colaboratory* es un servicio en la nube de Google en la cual se tiene acceso a ambos ambientes, tanto *Python* como *Jupyter*.

Tal como se menciona en What is Colaboratory (s.f.), Algunos de los beneficios que se pueden mencionar del uso de *Colaboratory* es que ofrece una versión gratuita que para propósitos simples es suficiente. Entre las limitaciones de la versión gratis está que un ambiente no puede mantenerse corriendo por más de 12 horas. Además, el acceso a GPU (Unidad de procesamiento gráfico, del inglés *Graphical Process Unit*) y TPU (Unidad de procesamiento tensorial, del inglés *Tensor Process Unit*) es limitado. El uso de GPU y TPU es importante cuando se desea realizar procesamiento paralelo, de esta forma los tiempos de ejecución se reducen en gran medida. Con respecto a la versión PRO, esta ofrece tiempos de ejecución extendidos, inclusive mayores a 24 horas. Acceso a más memoria RAM, GPU y TPU.

2.4 Fuentes

2.4.1 Dato

De acuerdo con el Diccionario de la Real Academia Española (2020), “información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho”. Los datos, en términos de minería de datos, no ofrecen valor como tal, requieren de procesamiento y análisis para convertirse en información.

2.4.2 Información

Según el Diccionario de la Real Academia Española, “comunicación o adquisición de conocimientos que permiten ampliar o precisar los que se poseen sobre una materia determinada”. En términos generales, la relación entre dato e información se puede interpretar, tal como lo presenta la Figura 9.

Figura 9: Relación entre dato e información.

Fuente: Elaboración propia.

2.4.3 Conjunto de datos

Cambridge Dictionary (s.f.) define un conjunto de datos como una colección de conjuntos separados de información que una computadora trata como una única unidad. Por lo general, los conjuntos de datos corresponden a elementos (columnas) separadas por algún delimitador, donde cada variable representa una columna y cada celda un valor para dicha columna. Una fila corresponde a un registro de relación entre las distintas columnas.

2.4.4 Usabilidad

El término usabilidad hace referencia a la facilidad de uso. Por ejemplo, el sitio web Kaggle.com, el cual contiene cientos de conjuntos de datos ofrece una métrica que indica la usabilidad del conjunto. Para ello, tal como se menciona en Product Feedback (s.f), utilizan una serie de parámetros para categorizar un conjunto de datos y brindar una categoría que es un número entre 1 y 10. Entre

dichos parámetros utilizan la calidad y completitud de la documentación, referencias, disponibilidad, metadatos asociados, entre otros.

2.5 Conceptos de minería de datos

2.5.1 Algoritmo

Un algoritmo se puede definir como un “conjunto ordenado y finito de operaciones que permite hallar la solución de un problema” (Real Academia Española, 2020). En términos generales, un algoritmo necesita una entrada, un proceso intermedio de operaciones y, finalmente, brinda una salida.

2.5.2 Datos de entrenamiento y prueba

Cuando se aplican algoritmos de minería de datos, es muy común encontrarse con los términos datos de entrenamiento y datos de prueba. Lo anterior por una simple razón, al iniciar un proyecto de minería de datos, con un conjunto de datos específico, se desea obtener un resultado. Sin embargo, para lograr dicho objetivo, es necesario entrenar el modelo con los datos que se tienen. Si se utiliza el 100 % de los datos, es imposible saber la precisión de este, ya que no se tiene forma de probar, a menos que se lleve a producción. Es por dicha razón que el proceso común es realizar una separación del conjunto de datos en dos partes: datos de entrenamiento y datos de prueba. Los datos de entrenamiento son utilizados para, como su nombre lo indica, entrenar el modelo. Por su parte, los datos de prueba se utilizan para validar qué tan capaz es el modelo ya entrenado de funcionar con datos que no conoce.

2.5.3 Limpieza de datos

Tal como lo indica Tableau (s.f.), la limpieza de datos hace referencia al proceso de arreglo y remoción de datos incorrectos, corruptos, formateados incorrectamente, duplicados o incompletos. Una correcta limpieza de los datos puede marcar la diferencia entre un modelo con excelentes resultados y un modelo con resultados mediocres. Es por ello que es uno de los pasos más importantes en los proyectos de minería de datos.

2.5.4 Lematización y derivación

The Stanford Natural Language Processing Group (s.f.) hace referencia a que ambas, lematización y derivación, buscan reducir las formas flexivas de las palabras y, de alguna forma, reducirlas a la forma base. Sin embargo, con la derivación, por

lo general, solamente se recortan las palabras al final, llevando así a posibles problemas gramaticales. Por su parte, la lematización busca realizar un análisis de vocabulario y morfología, con el objetivo de realizar una correcta reducción de la palabra.

2.5.5 Precisión

La precisión se define como “el grado en el cual un resultado de una medición, cálculo y especificación llega a un valor correcto o estándar” (Oxford Léxico, s.f.). Para efectos de la presente investigación, la precisión es utilizada como medida de calidad bajo distintos cálculos.

2.5.6 TF-IDF (Term frequency – Inverse document frequency)

TF-IDF significa en español frecuencia de término – frecuencia inversa de documento. Tal como lo menciona Zhang, Yoshida y Tang (2011), la idea fundamental detrás de TF-IDF es la asignación de peso a los términos basado en las apariciones en documentos. Por ende, un término que aparece en muchos documentos debería tener un peso mayor a un término que aparece en menos documentos. Para el cálculo del peso, se utiliza la fórmula presente en la Figura 10, donde: W_{ij} hace referencia al peso del término i en el documento j; N corresponde al número de documentos en la colección; $th_{i,j}$ es la frecuencia del término i en el documento j; finalmente, df_i es la frecuencia del término en los documentos.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

Figura 10: Fórmula para calcular el peso de un término.

Fuente: (Zhang, Yoshida y Tang, 2011)

2.5.7 Modelo de clasificación

Novakovi et al. (2017) tratan en su investigación la evolución de los modelos de clasificación. Como parte de ella, explican lo que es un modelo de clasificación, iniciando con que es una de las tareas más comunes en aprendizaje máquina y es un problema de clasificar instancias no conocidas en una de las categorías-clases predefinidas. Las funciones objetivo son discretas. La clasificación de un objeto está basada en encontrar similitudes con objetos predeterminados que son miembros de

distintas clases. Dichas similitudes son determinadas mediante sus características. En los modelos de clasificación, el número de clases es conocido *a priori*.

Según Aggarwal y Zhai (2012), un problema de clasificación puede ser definido de la siguiente manera. Si tenemos un conjunto de datos de entrenamiento D = {X1,...,XN }, de forma tal que cada registro está marcado con un valor de clase dibujado de un grupo de K valores discretos distintos indexados por {1…K}. Los datos de entrenamiento son utilizados para construir un modelo de clasificación que relaciona las características con su correspondiente registro de la clase marcada.

2.5.8 Minería de datos

La minería de datos es utilizada para descubrir patrones y relaciones entre los datos, con un énfasis en grandes bases de datos. La minería de datos está en las fronteras de distintos campos, tales como administración de base de datos, inteligencia artificial, aprendizaje máquina, reconocimiento de patrones y visualización de datos (Friedman, 1997).

2.5.9 Aprendizaje automático

El aprendizaje automático es una rama de los algoritmos computacionales que están diseñados para emular la inteligencia humana, aprendiendo del ambiente que lo rodea. El aprendizaje automático ha sido de forma exitosa en distintos campos, tales como el reconocimiento de patrones, visión por computadora, ingeniera espacial, finanzas, entretenimiento, biología computacional, aplicaciones médicas, entre otras (Naqa y Murphy, 2015). Tal como lo menciona Bonaccorso (2017), el aprendizaje máquina se divide en tres grandes áreas, las cuales son el aprendizaje supervisado, no supervisado y por refuerzo. Existe una nueva aplicación denominada aprendizaje semisupervisado y es una mezcla entre supervisado y no supervisado.

2.5.10 Tipos de aprendizaje

2.5.10.1 Aprendizaje supervisado.

Bonaccorso (2017) menciona que el aprendizaje supervisado se refiere al uso de algoritmos, en el cual se provee una entrada y la salida esperada. Basado en esta entrada, el algoritmo puede optimizar sus parámetros y trata de reducir la pérdida, con el objetivo de obtener una diferencia entre valor actual y esperado lo

más cercano a cero. Además, el mismo estudio menciona algunos ejemplos de usos de aprendizaje supervisado, entre los cuales se tienen:

- Análisis predictivo.
- Detección de *spam*.
- Detección de patrones.
- Procesamiento de lenguaje natural.
- Análisis de sentimiento.
- Clasificación de imágenes.

2.5.10.2 Aprendizaje no supervisado.

Por su parte, Bonaccorso (2017) indica que el aprendizaje no supervisado se basa en el hecho de que no tiene supervisor y, por ende, no tiene medida de error. Es principalmente utilizado para agrupar un conjunto de elementos basados en su similitud o distancia. Dentro de las principales aplicaciones que se tienen se mencionan:

- Segmentación de objetos.
- Detección de similitudes.
- Etiquetado automático.

2.5.11 Hiperparámetro

Tal como lo indican Tong y Hong (2020), los hiperparámetros hacen referencia a parámetros que no pueden ser modificados, durante la fase de entrenamiento de los modelos. Pueden ser incluidos en la construcción de la estructura del modelo. Por ejemplo, en redes neuronales, el número de capas ocultas, la función de activación, entre muchos otros.

Cada modelo tiene una serie de hiperparámetros para configurar el modelo antes de ejecutar el entrenamiento. Hoy en día, con los avances en poder computacional, se tiende a optimizar hiperparámetros, mediante el uso de estrategias de prueba continua con parámetros predefinidos. Por ejemplo, se desea optimizar el número de árboles en un modelo de bosque aleatorio. Por ende, se desea entrenar el modelo con 100, 1000, 10 000 árboles para saber cuál ofrece los mejores resultados. Conforme crece el número de hiperparámetros, la cantidad de

combinaciones de posibles valores también crece; aumentando así la complejidad de la optimización de estos.

2.5.12 Sobreajuste (*over-fitting*) y subajuste (*under-fitting*)

Es importante indicar que la idea fundamental tras un algoritmo de aprendizaje supervisado es ajustar una función a los datos de entrenamiento. Tal como lo explican Khalaf y Zaman (2014), el sobreajuste es un problema clave en los modelos de aprendizaje supervisado. Este se detecta cuando los datos calzan de forma tan perfecta, de manera que el ruido y las peculiaridades son memorizadas por el modelo. El problema surge al presentar nuevos datos al modelo, ya que, el rendimiento cae. Es más común caer en sobreajuste al utilizar conjuntos de datos pequeños. En la Figura 11, se puede observar cómo la función trata de ajustarse de manera perfecta a los datos.

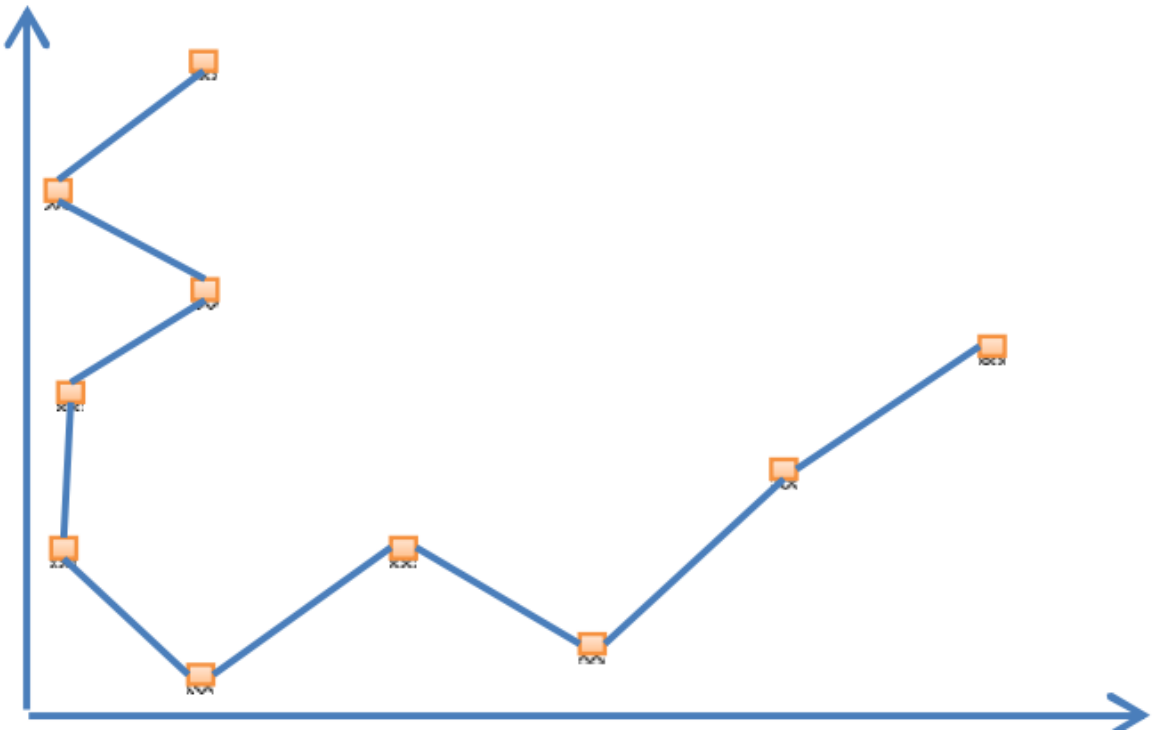


Figura 11: Ejemplo de sobreajuste.

Fuente: (Khalaf & Zaman, 2014)

Por su parte, el subajuste, hace referencia a lo opuesto del sobreajuste. En este caso, el modelo no es capaz de detectar las peculiaridades de los datos. En la Figura 12, se puede observar que los datos están repartidos por el espacio y la función no se ajusta a estos.

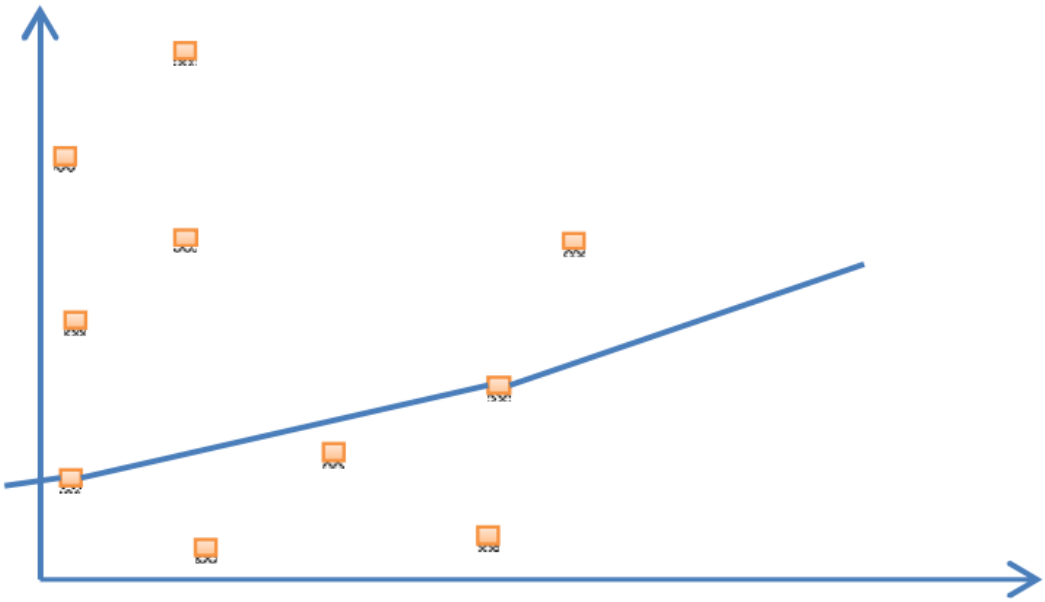


Figura 12: Ejemplo de subajuste.

Fuente: (Khalaf & Zaman, 2014)

2.5.13 Técnicas de minería de datos.

2.5.13.1 Bosques aleatorios.

Breiman (2001), en su artículo, define los bosques aleatorios como una combinación de árboles predictores, los cuales dependen de una muestra independiente de valores de un vector aleatorio con la misma distribución para todos los árboles.

2.5.13.2 Máquinas de soporte vectorial.

Las máquinas de soporte vectorial son matemáticamente complejas y computacionalmente intensivas. El propósito de este trabajo no es adentrarse en la complejidad del algoritmo. Sin embargo, es importante explicar su funcionamiento de alto nivel. Tal como lo menciona Betancourt (2005), una máquina de soporte vectorial mapea los puntos de entrada a un espacio de características mayor. Posteriormente, encuentra un hiperplano mayor que los separe y maximice el margen entre las clases en este espacio.

2.5.13.3 Regresión logística multinomial.

La regresión logística multinomial, también conocida por los nombres: regresión logística multiclase, Logit multinomial, entre otras, es una generalización de la

regresión logística para problemas multiclase. Pando y San Martín (2004) mencionan que la regresión logística multinomial es utilizada en modelos con variables dependientes de tipo nominal con más de dos categorías. Las variables dependientes pueden ser tanto continuas como categóricas.

2.5.5.13.4 Redes bayesianas multinomiales.

A modo de definición base, se puede utilizar el trabajo realizado por López, García et al. (2007), en el cual se define una red bayesiana como un grafo dirigido acíclico que codifica relaciones probabilísticas de dependencia e independencia condicional y que codifica el modelo, con base en las evidencias muestrales, mediante la regla de Bayes.

Por su parte, como lo indican Jiang et al. (2016), con las redes bayesianas multinomiales se asume que todos los atributos son independientes de cualquier otro dado el contexto de la clase. Además, este ignora cualquier dependencia entre los atributos.

2.5.13.5 K vecinos más cercanos

El algoritmo de los K vecinos más cercanos se basa en la ubicación de los puntos sobre un espacio n-dimensional. Mitchell (1997) indica que este algoritmo asume que todos los puntos están en un espacio n-dimensional y los vecinos más cercanos son definidos en términos de la distancia euclidiana estándar.

Capítulo 3. Marco metodológico

3.1 Tipo de investigación

Dado que la presente investigación se enfoca en la evaluación de distintos modelos de minería de datos para la posible aplicación en el contexto de atención de emergencias, se considera que la investigación es de tipo evaluativa.

3.2 Alcance investigativo

De acuerdo con Namakforoosh (2005), en la investigación exploratoria, “el investigador conoce poco el área de estudio, donde todavía no se ve la necesidad de generar hipótesis”. Además, se menciona que “el objetivo de una investigación exploratorio es la generalización de ideas y perspectivas”.

La presente investigación se define como investigación exploratoria, esto porque, a pesar de que existen distintas investigaciones en relación con el tema, no se encontraron estudios con una relación total. Por otra parte, inclusive con la existencia de estudios similares a nivel global, a nivel nacional, no se tienen investigaciones de este tipo.

3.3 Enfoque

Para el desarrollo de esta investigación, se propone un enfoque alternativo. En el enfoque alternativo, basado en las ideas propuestas por Naranjo (2020), se enfatiza en que lo cuantitativo y lo cualitativo nunca han estado separados. Para lograr enfocar el estudio de forma alternativa, se hace explícito el encuadre ontológico, epistemológico y axiológico.

Iniciando con la ontología, perteneciente a la filosofía y encargada de estudiar/demostrar la existencia, es el “área de la filosofía que estudia lo que significa existir” (Cambridge Dictionary, s.f.). Basados en las áreas de interés de la investigación, principalmente la atención de emergencias, la minería de datos y el reconocimiento de voz, en la Figura 13, se puede observar la ontología correspondiente a cada una de ellas.

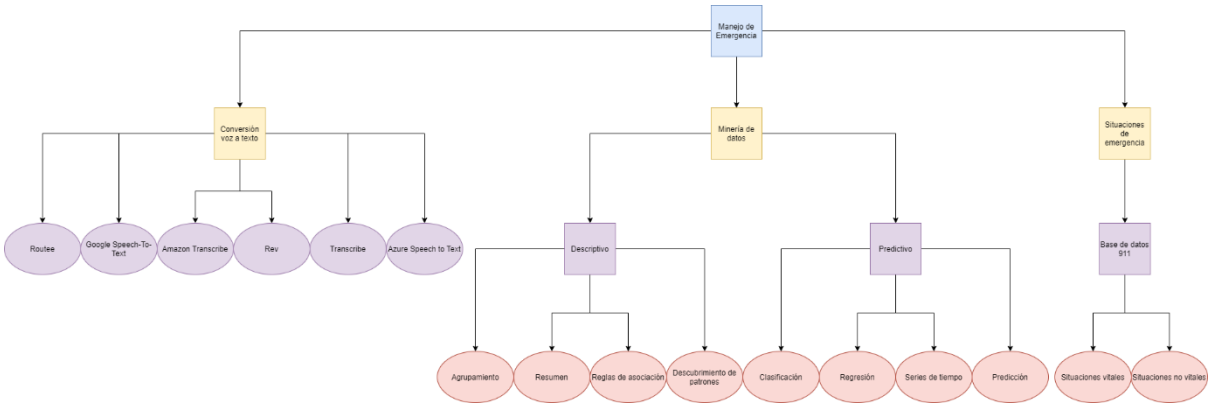


Figura 13: Esquema conceptual.

Fuente: Elaboración propia. Utilizando: Draw.io.

La epistemología también perteneciente a la filosofía y se encarga de estudiar cómo sabemos, es “la parte de la filosofía que trata sobre el estudio de cómo sabemos las cosas” (Cambridge Dictionary, s.f.). El papel del investigador en esta investigación es involucrado. Esto porque, para el desarrollo de este trabajo, es

necesario tener contacto con distintos elementos, tales como: creación de escenarios, grabación, encuestas, *scripts*, entre otros.

Finalmente, desde la dimensión axiológica, se pretende realizar un abordaje probabilístico para la evaluación de la precisión.

Tabla 16: Evaluación de precisión de los modelos.

Evaluación del modelo	**Criterio – Referencia valor precisión (P)**
Sobresaliente	P > 0.85
Aceptable	0.75 < P <= 0.85
Malo	0 <= P <= 0.75

Fuente: Elaboración propia.

3.4 Diseño

El diseño de la investigación es de tipo experimental, para lograr los resultados, se utiliza la metodología CRISP-DM de la cual se explica en secciones posteriores. La siguiente lista corresponde a un resumen de los pasos necesarios para lograr dicho resultado:

- Comprensión de negocio.
- Obtención de datos.
- Se inicia el proceso de limpieza de datos, el cual abarca los siguientes aspectos:
    - Eliminar signos de puntuación, comillas, apostrofes, palabras vacías, etiquetas HTML (lenguaje de marcas de hipertexto, del inglés *HyperText Markup Language*), escape de caracteres, saltos de línea, espacios extra.
    - Convertir todo el texto a minúscula.
- Una vez limpios los datos, se procede a realizar la lematización para obtener la palabra raíz de todas las palabras.
- Se crea un vector TF-IDF.
- Se ejecutan los distintos algoritmos de minería de datos para clasificar los textos, según la situación.
- Se interpretan los resultados y el nivel de confianza para cada uno de los algoritmos.

- Se realiza un análisis general de los resultados y se elaboran las conclusiones.

3.5 Población y muestreo

Al no tenerse datos reales o ficticios sobre situaciones de emergencia, se procedió a utilizar un conjunto de datos alternativos, el cual cuenta con características similar a lo que se supone podría ser un conjunto de datos real. Con características similares se hace referencia a cantidad de clases y longitud del contenido.

3.6 Técnicas de análisis de información

La estructura general de la investigación se basa en el estándar CRISP-DM (del inglés *Cross Industry Standard for Data Mining*). Se elige este estándar por el hecho de que brinda un amplio soporte y estructura en el desarrollo de proyectos de minería de datos. En la Figura 14, se presenta el estándar CRISP-DM y cada una de sus etapas.

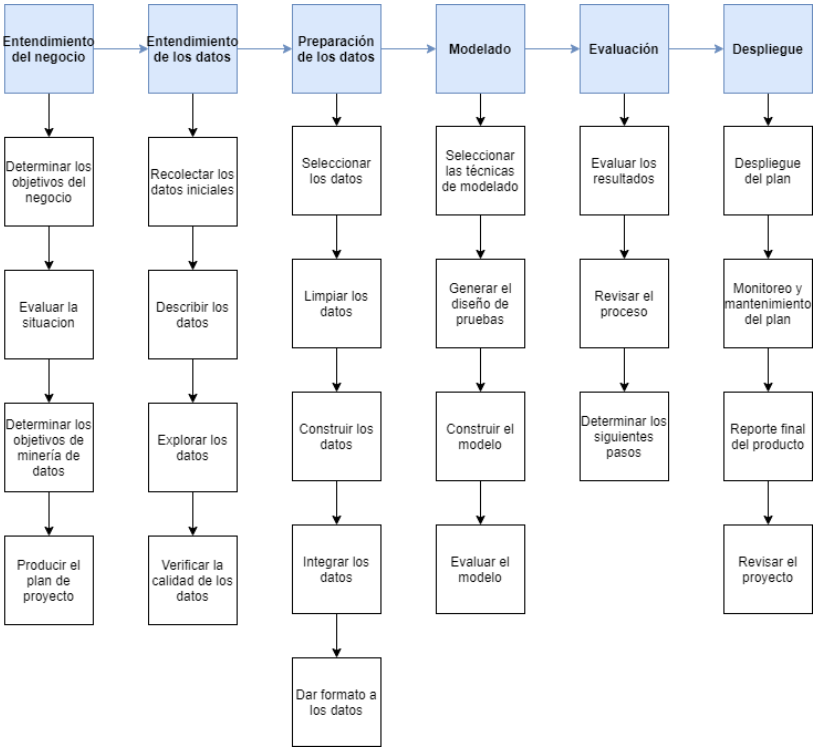


Figura 14: Estándar CRISP-DM.

Fuente: Elaboración propia basado en Hipp y Wirth's (2000).

Capítulo 4. Análisis del diagnóstico

Con el objetivo de llevar a cabo la investigación de forma organizada y ordenada, tal como se menciona en la sección 3.7, se hace uso de estándar CRISP-DM. Por ende, el presente capítulo presenta las distintas etapas desde el entendimiento del negocio hasta la evaluación del modelo, dejando así el despliegue como punto extra en la sección de conclusiones y recomendaciones. Esto llevando control de los resultados y de las distintas alteraciones del proceso.

A pesar de que no se cuenta con datos reales de emergencias, se describe y explica el motivo de la elección de datos alternativos como fuente de datos para la investigación. Además, a pesar de que se utilizan datos alternativos, se realizan explicaciones, sugerencias y comparaciones para constatar la posible funcionalidad de los datos en la clasificación de emergencias, ya que el objetivo real del negocio es la atención de emergencias y no cualquier aspecto alterno que pueda traer consigo el uso de datos alternativos.

4.1 Entendimiento del negocio

4.1.1 Objetivos del negocio.

El objetivo principal para el negocio es la evaluación de distintos modelos de minería de datos para la posible aplicación en el contexto de atención de emergencia del sistema 9-1-1 en Costa Rica.

4.1.2 Evaluación de la situación.

Los avances en distintas áreas como sensores, redes Wi-Fi, móviles, posicionamiento, calidad de audio y sonido han sido gigantes. En adición, los sistemas computacionales son cada vez más poderosos y accesibles para la población común, dando acceso barato a la minería de datos, macrodatos, inteligencia artificial, aprendizaje máquina, entre otros. Sin embargo, a pesar de todo el avance, los trabajos en temas de atención de emergencias basados en minería de datos son muy pocos y se reducen aún más en el ámbito nacional.

El sistema nacional de atención de emergencias no brinda datos sobre las emergencias que atienden, puesto que estos son de carácter confidencial y no están disponibles siquiera para uso académico. Por dicha razón, se hace uso de datos alternativos con características similares.

4.1.3 Objetivos de minería de datos

Los objetivos, en términos de minería, son:

- Ejecutar, al menos, cinco algoritmos de minería de datos, los cuales utilizan la misma entrada de datos.
- Evaluar la precisión de los algoritmos de minería de datos propuestos.
- Elegir un algoritmo como candidato a mejor algoritmo basado en el tiempo de ejecución y la precisión. Para dicho algoritmo, explicar el motivo de la elección con criterio sólidos.

4.1.4 Plan del proyecto

El proyecto, al estar basado en un estándar como lo es CRISP-DM, está definido por una serie de etapas, seis para ser exactos. En la cual, se lleva a cabo un proceso incremental y cíclico para lograr el objetivo, iniciando con el entendimiento del negocio y finalizando con el despliegue del plan.

Para lograr los objetivos de minería de datos, se requieren distintos aspectos en sincronía, para que todo suceda de forma organizada. Inicialmente, es indispensable la utilización de una herramienta de programación para desarrollar cada uno de los algoritmos. Para evitar algún impacto sobre los resultados, se utiliza la misma herramienta. En términos de lenguaje de programación, se utiliza *Python* versión 3.7. Dicho lenguaje es ejecutado mediante *Jupyter Notebook*, el cual está alojado en la herramienta de *Google* denominada *Colaboratory*.

4.2 Entendimiento de los datos

4.2.1 Recolección inicial de datos.

Al no tenerse datos reales, uno de los principales desafíos es obtener un conjunto de datos para poder trabajar sin sufrir grandes alteraciones. Inicialmente, se hizo una búsqueda en distintos sitios, pues no es solamente un tema de tipo de datos, algunos de las características que se desea son:

- Los párrafos son de longitud moderada a baja.
- El texto no debe estar sobrecargado de símbolos extraños.
- La cantidad de registros debe ser grande, preferiblemente del rango de los cientos de miles.
- Los datos no deben mezclar idiomas.
- Los datos deben ser datos clasificados/categóricos.

Basado en los anteriores criterios, se eligieron dos conjuntos de datos, el primero denominado *Yahoo! Answers*, el segundo denominado *AG News*. Con estos dos conjuntos, se inicia la investigación y uno de ambos debe ser elegido como el conjunto de datos a utilizar.

4.2.2 Características de los datos.

Para los dos conjuntos de datos seleccionados, las características son:

4.2.2.1 Yahoo! Answers.

El conjunto de datos *Yahoo! Answers Comprehensive Questions and Answers* (Yahoo! Answers preguntas y respuestas comprensivas) tiene sus raíces en el *Yahoo! Webscope program*. Dicho sitio contiene distintos conjuntos de datos de utilidad para la comunidad. Estos solamente están disponibles para tareas de carácter no comercial. El conjunto de datos padre contiene 4 483 032 preguntas y respuestas. Además de lo mencionado, el conjunto de datos contiene metadatos (no para todos los registros) con información, tal como la mejor respuesta, categoría y subcategoría.

Posteriormente, basado en dicho conjunto de datos es que *Yahoo! Answers* fue creado y, además, utilizado en Zhang, Zhao y LeCun (2015). El nuevo conjunto de datos extrajo las 10 principales categorías del conjunto raíz. Para cada clase, se tienen 140 000 registros para entrenamiento y 6000 registros para pruebas. En la Tabla 17, se pueden observar las categorías en las que se clasifican las preguntas y respuestas.

Tabla 17: Categorías utilizadas en el conjunto de datos Yahoo! Answers.

Número de clase	Nombre de clase
1	Cultura y sociedad
2	Ciencia y matemática
3	Salud
4	Educación y Referencia
5	Computación e informática
6	Deporte
7	Negocios y finanzas

8	Música y entretenimiento
9	Relaciones y familia
10	Gobierno y política

Fuente: Elaboración propia basado en Zhang, Zhao y LeCun (2015)

4.2.2.2 AG News conjunto de datos de clasificación.

Con respecto al conjunto de datos *AG News,* los datos raíz fueron obtenidos mediante el motor de búsqueda denominado *ComeToMyHead*, el cual obtuvo más de un millón de resultados posterior a que se inició su ejecución en 2004 en un periodo de un año. Este obtuvo los artículos utilizando más de 2000 fuentes de datos. Dichos datos raíz están estructurados en una tabla de base de datos con la estructura mostrada en la *Tabla 18*.

Tabla 18: Estructura de tabla raíz del conjunto de datos AG NEWS.

Campo	Tipo	Null	Llave	Defecto
Fuente	VARCHAR(32)		PRI	
URL	VARCHAR(255)		PRI	
Título	TEXT	YES	MUL	NULL
Imagen	VARCHAR(255)	YES		NULL
Categoría	VARCHAR(32)		PRI	
Descripción	TEXT	YES		NULL
Ranquin	INT(11)	YES		NULL
Fecha publicación	TIMESTAMP	YES		TIMESTAMP ACTUAL
Vídeo	VARCHAR(255)	YES		NULL

Fuente: Elaboración propia basado en Zhang, Zhao y LeCun (2015)

Posteriormente, basado en dicho conjunto de datos es que *AG News* conjunto de datos de clasificación fue creado y, además, utilizado en Zhang, Zhao y LeCun (2015). Este nuevo conjunto de datos es una versión simplificada de conjunto raíz en el cual solamente se utilizan las cuatro principales clases/categorías del conjunto de datos padre. En total, son 120 000 registro para entrenamiento con 30 000 registros por clase, y un total de 7600 registros para pruebas del modelo. Además, cuenta con 3 columnas una para la categoría, título y descripción.

4.2.3 Exploración de los datos.

Para la correcta utilización y selección de los datos, es necesario pasar por un proceso de validación y limpieza. El objetivo de esta investigación es evaluar modelos para la detección de situaciones de emergencia. Por ende, tal como ya se mencionó, la idea fundamental es que los datos tengan una estructura similar a lo que podría ser una conversación sobre una emergencia.

Durante una emergencia, dado que es una conversación entre humanos y no datos obtenidos mediante la web, se espera que no contenga símbolos extraños, como caracteres HTML, mezcla de muchos idiomas, entre otros. Por dicha razón, la exploración de los datos es un punto muy importante para la investigación.

4.2.3.1 Categorías y cantidad de registros.

Un aspecto importante a considerar es la cantidad de categorías en cada conjunto de datos. Por un lado, *Yahoo! Answers* está basado en 10 categorías. Por su parte, *AG News* contiene 4 categorías de clasificación. En la Figura 15, se puede observar la cantidad de registros de entrenamiento para ambos conjuntos de datos, un aspecto importante de observar es que tan solo una categoría de *Yahoo! Answers* contiene más registros que todas las categorías de *AG News* juntas.

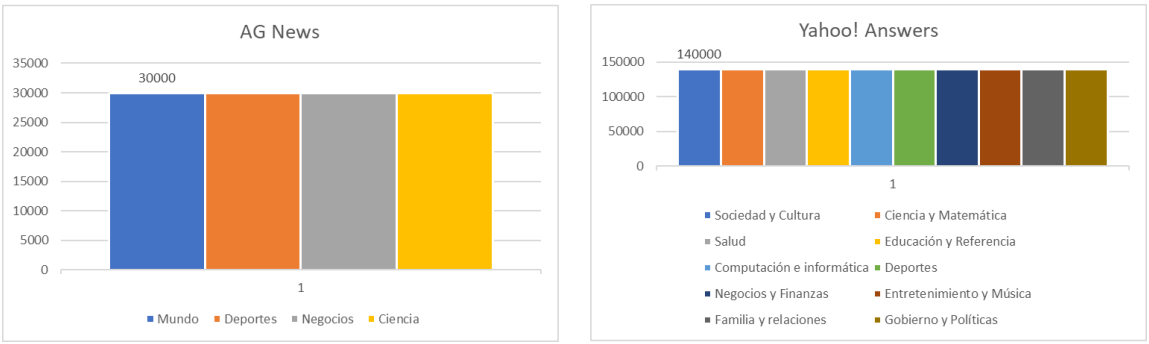


Figura 15: AG News y Yahoo! Answers. Comparativa de clases y cantidad de registros de entrenamiento.

Fuente: Elaboración propia.

4.2.3.2 Estructura de los datos.

Tal como se puede observar en la Figura 16, el conjunto de datos contiene un total de 4 columnas, 3 de ellas corresponde a datos necesarios, como el título, pregunta y mejor respuesta. Por ende, para su utilización, es necesario unir las columnas en una sola.

	class	title	question	best_answer
0	6	What is the best off-road motorcycle trail ?	long-distance trail throughout CA	i hear that the mojave road is amazing! \...
1	3	What is Trans Fat? How to reduce that?	I heard that tras fat is bad for the body. Wh...	Trans fats occur in manufactured foods during ...
2	7	How many planes Fedex has?	I heard that it is the largest airline in the ...	according to the www.fedex.com web site:\nAir ...
3	7	In the san francisco bay area, does it make se...	the prices of rent and the price of buying doe...	renting vs buying depends on your goals. <br /...
4	5	What's the best way to clean a keyboard?	I have very small stuff stuck under my keyboar...	There are commercial kits available, but a can...
...	...	...	...	...
1399994	3	do all these ads on tv of yoko etc regarding h...	NaN	I increased my height 2 feet afterwards so yes...
1399995	7	Ways to sell your video games?	Like if you want to sell your video games how ...	ebay, electronic boutique, babbages or flea ma...
1399996	3	is it normal to have nots in your breast or bo...	NaN	It can be normal as long as they are not cance...
1399997	1	Who can speak Hindi??	If you can write it here!!	Main hindi bol sakti hoon.kahiye.
1399998	5	where can i find websites were i can have a v...	NaN	★★ HELP WITH SEARCHING ★★\n\n I used to have ...

Figura 16: Conjunto de datos Yahoo! Answers.

Fuente: Elaboración propia.

Con respecto al conjunto de datos *AG News*, en la Figura 17, se puede observar que contiene 3 columnas. De ellas, dos contienen datos que deben ser unidos en una sola columna. Específicamente el título y la descripción.

	class	title	description
0	3	Wall St. Bears Claw Back Into the Black (Reuters)	Reuters - Short-sellers, Wall Street's dwindli...
1	3	Carlyle Looks Toward Commercial Aerospace (Reu...	Reuters - Private investment firm Carlyle Grou...
2	3	Oil and Economy Cloud Stocks' Outlook (Reuters)	Reuters - Soaring crude prices plus worries\ab...
3	3	Iraq Halts Oil Exports from Main Southern Pipe...	Reuters - Authorities have halted oil export\f...
4	3	Oil prices soar to all-time record, posing new...	AFP - Tearaway world oil prices, toppling reco...
...	...	...	...
119995	1	Pakistan's Musharraf Says Won't Quit as Army C...	KARACHI (Reuters) - Pakistani President Perve...
119996	2	Renteria signing a top-shelf deal	Red Sox general manager Theo Epstein acknowled...
119997	2	Saban not going to Dolphins yet	The Miami Dolphins will put their courtship of...
119998	2	Today's NFL games	PITTSBURGH at NY GIANTS Time: 1:30 p.m. Line: ...
119999	2	Nets get Carter from Raptors	INDIANAPOLIS -- All-Star Vince Carter was trad...

Figura 17: Conjunto de datos AG News.

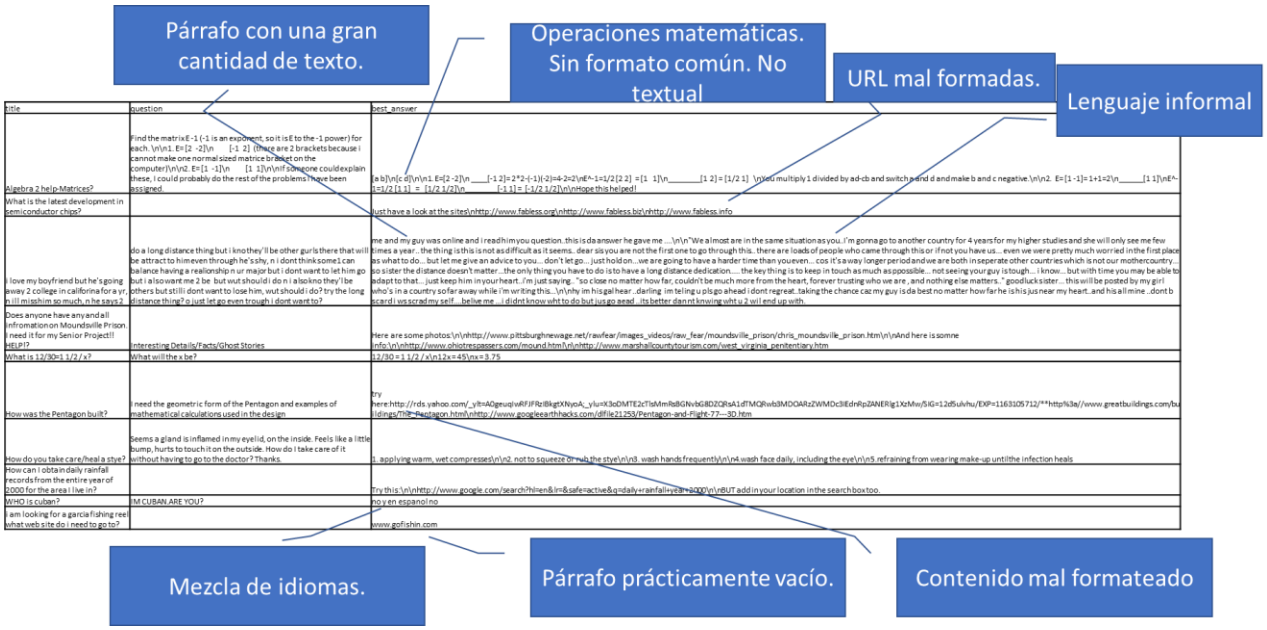
Fuente: Elaboración propia.

4.2.3.4 Tipos de datos – variabilidad.

Por el hecho de que se tiene una cantidad muy grande de datos, para dar un vistazo general, es prácticamente imposible analizar cada registro. Por ende, se filtraron 30 registros aleatorios sobre el total de registros de entrenamiento. Posteriormente, para no mostrar una lista extensa de filas, se hace una reducción a alrededor de 10 registros, en los cuales se tengan más peculiaridades.

En la Figura 18, se pueden observar 10 registros del conjunto de datos *AG News*, entre los aspectos más importantes de recalcar se tienen (por orden de líneas):

- Operaciones matemáticas complejas y con una gran cantidad de símbolos de difícil organización.
- Se tienen direcciones URL (Localizador de Recursos Uniforme, del inglés *Uniform Resource Locator*) mal formadas y no se tiene el contenido del sitio apuntado por dicha dirección.
- Párrafos de gran extensión comparados con la media. Además, se tiene lenguaje informal o mal escrito.
- Contenido mal formateado en el cual no se tienen espacios. Por otra parte, no brinda información más que la dirección a visitar.
- Se hace mezcla de idiomas, lo cual incremente el nivel de complejidad para tareas como lematización.
- Párrafos prácticamente vacíos y sin pregunta.



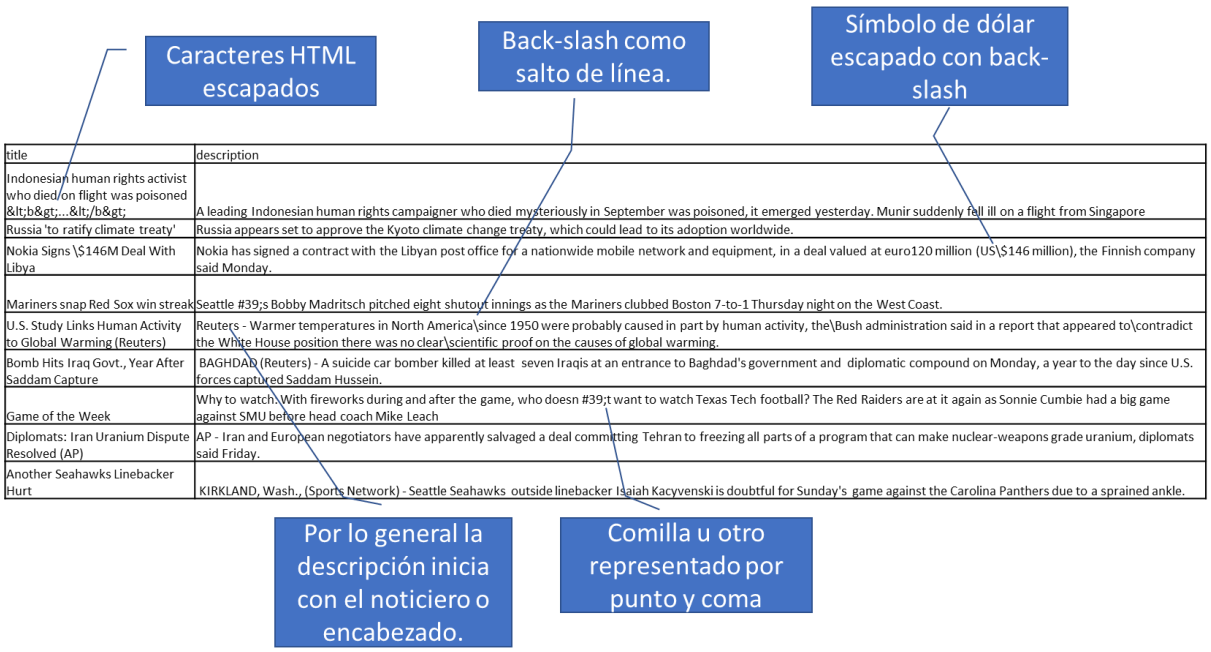
title	question	best_answer
Algebra 2 help-Matrices?	Find the matrix E -1 (-1 is an exponent, so it is E to the -1 power) for each .\n\n1. E=[2 -2]\n [-1 2] (I have are 2 brackets because I cannot make one normal sized matrice bracket on the computer)\n\n2. E=[1 -1]\n [1 1]\n\nif someone could explain these, I could probably do the rest of the problems? have been assigned.	a b]\n\nc d]\n\n1. E=[2 -2]\n ____[-1 2]=2*2-(-1)(-2)=4-2=2\nE^-1=1/2[2 2] =[1 1]\n_____[1 2] = [1/2 1] \nYou multiply 1 divided by ad-cb and switch a and d and make b and c negative.\n\n2. E=[1 -1]=1+1=2\n_____[1 1]\nE^-1=1/2 [1 1] = [1/2 1/2]\n [-1 1] = [-1/2 1/2]\n\nHope this helped!
What is the latest development in semiconductor chips?		Just have a look at the sites\nhttp://www.fabless.org\nhttp://www.fabless.biz\nhttp://www.fabless.info
I love my boyfriend but he's going away 2 college in california for a yr, n i'll miss him so much, n he says 2	do a long distance thing but i know they'll be other guris there that will be attract to him even through he's shy, n i dont think some1 can balance having a realionship n ur major but i dont want to let him go but i also want me 2 be but wut should i do n i also kno they'll be others but still i dont want to lose him, wut should i do? try the long distance thing? o just let go even trough i dont want to?	me and my guy was online and i read him you question. this is da answer he gave me ..\n\n"We almost are in the same situation as you..I'm gonna go to another country for 4 years for my higher studies and she will only see me few times a year.. the thing is this is not as difficult as it seems.. dear sis you are not the first one to go through this.. there are loads of people who came through this or if not you have us... even we were pretty much worried in the first place as what to do... but let me give an advice to you... don't let go... just hold on...we are going to have a harder time than you even...cos it's a way longer period and we are both in seperate other countries which is not our mothercountry... so sister the distance doesn't matter...the only thing you have to do is to have a long distance dedication..... the key thing is to keep in touch as much as possible... not seeing your guy is tough.... i know... but with time you may be able to adapt to that... just keep him in your heart..I'm just saying.. "so close no matter how far, couldn't be much more from the heart, forever trusting who we are , and nothing else matters.." good luck sister... this will be posted by my girl who's in a country so far away while I'm writing this..\n\nhy im his gal hear..darling im telling u pls go ahead i dont regret..taking the chance coz my guy is da best no matter how far he is his jus near my heart..and his all mine..dont b scard i ws scrad my self....bellue me....i didnt know wht to do but jus go aead...its better dan nt knwing wht u 2 wil end up with..
Does anyone have any and all infromation on Moundsville Prison. I need it for my Senior Project!! HELP!?	Interesting Details/Facts/Ghost Stories	Here are some photos:\n\nhttp://www.pittsburghnewage.net/~rawfear/images_videos/raw_fear/moundsville_prison/chris_moundsville_prison.htm\n\nAnd here is somne info:\n\nhttp://www.ghosttrappers.com/mound.htm\n\nhttp://www.marshallcountytourism.com/west_virginia_penitentiary.htm
What is 12/30=1 1/2 / x?	What will the x be?	12/30 = 1 1/2 / x\n\n12x=45\nx=3.75
How was the Pentagon built?	I need the geometric form of the Pentagon and examples of mathematical calculations used in the design	try here:http://rds.yahoo.com/_ylt=A0geuqlwRFJFRzIBkgtXNyoA;_ylu=X3oDMTE2cTlkMmRqBGNvbGlDZQRsA1dfMQRwb3MDOARzZWMDc3lEdnRpZANERlg1XzMw/SIG=12d5ukvhu/EXP=1163105712/**http%3a//www.greatbuildings.com/buildings/The_Pentagon.html\nhttp://www.googlearthhacks.com/diffile21253/Pentagon-and-Flight-77---3D.htm
How do you take care/heal a stye?	Seems a gland is inflamed in my eyelid, on the inside. Feels like a little bump, hurts to touch it on the outside. How do I take care of it without having to go to the doctor? Thanks.	1. applying warm, wet compresses\n\n2. not to squeeze or rub the stye\n\n3. wash hands frequently\n\n4.wash face daily, including the eye\n\n5.refraining from wearing make-up until the infection heals
How can I obtain daily rainfall records from the entire year of 2000 for the area I live in?		Try this:\n\nhttp://www.google.com/search?hl=en&lr=&safe=active&q=daily+rainfall+year+2000\n\nBUT add in your location in the search box too.
WHO is cuban?	IM CUBAN ARE YOU?	no y en espanol no
I am looking for a garcia fishing reel what web site do I need to go to?		www.gofishin.com

Figura 18: Yahoo! Answers, 10 registros obtenidos de 30 aleatorios, características de los datos.

Fuente: Elaboración propia.

Por su parte, el conjunto de datos *AG News* contiene algunas características peculiares, basados en los 9 registros que se filtraron, tal como se puede observar en la Figura 19. Entre ellas se tienen:

- Caracteres HTML escapados en lugar del valor normal.
- Algunos símbolos como el $ (dólar) tienden a tener una barra invertida al frente. Puede ser escapado o salto de línea es ese lugar.
- Barra invertida en lugares aleatorios. Apunta a ser salto de línea.
- Tendencia a estar el noticiero o titular al inicio de la descripción seguido por un guion.



title	description
Indonesian human rights activist who died on flight was poisoned ...	A leading Indonesian human rights campaigner who died mysteriously in September was poisoned, it emerged yesterday. Munir suddenly fell ill on a flight from Singapore
Russia 'to ratify climate treaty'	Russia appears set to approve the Kyoto climate change treaty, which could lead to its adoption worldwide.
Nokia Signs \$146M Deal With Libya	Nokia has signed a contract with the Libyan post office for a nationwide mobile network and equipment, in a deal valued at euro120 million (US\$146 million), the Finnish company said Monday.
Mariners snap Red Sox win streak	Seattle #39;s Bobby Madritsch pitched eight shutout innings as the Mariners clubbed Boston 7-to-1 Thursday night on the West Coast.
U.S. Study Links Human Activity to Global Warming (Reuters)	Reuters - Warmer temperatures in North America\since 1950 were probably caused in part by human activity, the\Bush administration said in a report that appeared to\contradict the White House position there was no clear\scientific proof on the causes of global warming.
Bomb Hits Iraq Govt., Year After Saddam Capture	BAGHDAD (Reuters) - A suicide car bomber killed at least seven Iraqis at an entrance to Baghdad's government and diplomatic compound on Monday, a year to the day since U.S. forces captured Saddam Hussein.
Game of the Week	Why to watch: With fireworks during and after the game, who doesn #39;t want to watch Texas Tech football? The Red Raiders are at it again as Sonnie Cumbie had a big game against SMU before head coach Mike Leach
Diplomats: Iran Uranium Dispute Resolved (AP)	AP - Iran and European negotiators have apparently salvaged a deal committing Tehran to freezing all parts of a program that can make nuclear-weapons grade uranium, diplomats said Friday.
Another Seahawks Linebacker Hurt	KIRKLAND, Wash., (Sports Network) - Seattle Seahawks outside linebacker Isaiah Kacyvenski is doubtful for Sunday's game against the Carolina Panthers due to a sprained ankle.

Figura 19: AG News, 9 registros obtenidos de 30 aleatorios, características de los datos.

Fuente: Elaboración propia.

4.2.3 Verificación de calidad de los datos.

4.2.4.1 Usabilidad.

Con respecto al término usabilidad, este puede ser muy subjetivo. Por ende, para efectos de los conjuntos de datos, se hace referencia a usabilidad basado en el puntaje asignado a cada conjunto de datos por el sitio Kaggle.com. En la Tabla 19, se puede observar la comparativa para este aspecto en ambos conjuntos de datos.

Tabla 19: Usabilidad de los conjuntos de datos.

Conjunto de datos	Usabilidad
Yahoo! Answers	2.4
AG News	7.1

Fuente: Kaggle.com

4.2.4.2 Calidad basada en análisis exploratorio.

Tomando como referencia los resultados e información obtenida en la sección 4.2.3 de Exploración de datos, en conjunto con las características esperadas se procede a brindar el criterio sobre la calidad de los datos. Inicialmente, la usabilidad nos da un indicio de que el conjunto de datos *Yahoo! Answers* tiene problemas o complicaciones para lograr un conjunto de datos útil y aplicable en distintos algoritmos. En la Tabla 20, se presenta una comparativa entre ambos conjuntos de datos.

Tabla 20: Comparativa entre Yahoo! Answers y AG News*.*

Característica	Yahoo! Answers	AG News	Comentario
Usabilidad	2.4	7.1	La diferencia es de alrededor de un 67 % a favor de AG News. Nos brinda una idea global de los problemas que podemos enfrentar con Yahoo! Answers.
Idioma	Variado	Inglés	En el conjunto de datos Yahoo! Answers se encuentran respuestas y preguntas en distintos idiomas, esto principalmente por ser abierto para su uso. A diferencia de AG News que es un ambiente más formal y controlado.

Longitud de los registros	Muy variado.	Poco variado. En general similares.	Con Yahoo! Answers se encuentran registros prácticamente vacíos y otros con párrafos de gran extensión.
Símbolos	Gran cantidad y no siguen un patrón simple.	Pocos, por lo general etiquetas HTML.	
Tamaño del conjunto de datos	Grande	Mediano	AG News puede inclusive considerarse pequeño en comparación con Yahoo! Answers. Esto porque tan solo una categoría de Yahoo! Answers contiene la totalidad de registros de AG News.

Fuente: Elaboración propia.

Una vez realizada la comparación, podemos resumir que la calidad de los datos de AG News es superior en prácticamente todos los aspectos. Además, es importante recalcar que se desea un conjunto de datos que pueda tener alguna relación con lo que sería un conjunto de datos de llamadas al 9-1-1. Con respecto a *AG News*, en general, no contiene gran cantidad de símbolos extraños, el idioma tiende a ser constante, la longitud podría ser perfectamente la longitud de una conversación corta.

Un aspecto importante del conjunto de datos *Yahoo! Answers* es la cantidad de registros. Al ser este trabajo relacionado con la utilización de técnicas de minería de datos para clasificar situaciones, la cantidad de datos es un punto a considerar, esto por el hecho de que se trabaja con modelos de clasificación supervisado, por ende, tener suficientes datos de entrenamiento es necesario para poder obtener un modelo con un amplio lenguaje de conocimiento, sin caer en sobreajuste o subajuste. A pesar de lo anterior, con los recursos que se cuenta para desarrollar la investigación, el gran tamaño del conjunto de datos *Yahoo! Answers* y la limitación de tiempo, este se vuelve un trabajo inmanejable para el tiempo estipulado, debido al tiempo de procesamiento que puede llevar el entrenamiento de los modelos; y no

solo el entrenamiento, tan solo buscar los parámetros a utilizar puede ser una tarea de días.

4.3 Preparación de los datos

4.3.1 Selección de los datos.

Una vez realizado el análisis de los datos y tomando en consideración aspectos como el tiempo para desarrollar la investigación, calidad de los datos, tamaño del conjunto de datos, entre otros. Se toma la decisión de utilizar el conjunto de datos *AG News* como conjunto de datos oficial para el desarrollo e implementación de los modelos de minería de datos. Esto por el hecho de que el conjunto de datos *AG News* tiene un balance entre calidad de los datos y cantidad de registros para realizar entrenamiento y prueba.

4.3.2 Limpieza de datos.

El proceso de limpieza de datos es un aspecto fundamental, esto porque de ello depende gran parte de los resultados en la aplicación de los modelos de minería de datos. Además, al utilizarse el vector TF-IDF, si no se realiza una correcta limpieza de los datos, es muy sencillo que muchos de los términos sean elementos extraños del conjunto de datos que, en realidad, no deben ser considerados. La presenta investigación se basa en aprendizaje supervisado para clasificación de texto. Al trabajar con texto, es importante que las palabras que significan lo mismo vayan juntas. Por ejemplo, en la Tabla 21, se puede observar una serie de palabras y su correspondiente palabra raíz. Al unir todas las palabras en una sola, esto conlleva a incrementar el peso de la palabra y no un peso individual.

Tabla 21: Palabras y su palabra raíz.

Palabra	Raíz
Hi, Hello, HELLO, hi, hello	*Hello*
Peter, Peter´s, peter, peTER	*Peter*
Increase, ncreased, increasing	*increase*

Fuente: Elaboración propia.

En los siguientes puntos, se presenta el proceso llevado a cabo, para lograr la limpieza de los datos.

4.3.2.1 Carga de datos.

Como punto inicial se realiza la carga de los datos, la herramienta utilizada es *Google Colaboratory* para ejecutar *script* de *Python* en un *Notebook*. Para obtener los datos, se hace uso de una conexión con *Google Drive*, lugar donde se almacenan los resultados, datos, documentos y más.

Los datos provienen de un archivo CSV (valores separados por coma, del inglés *Comma-Separated Values*), en total son dos archivos, uno para pruebas y otro para entrenamiento. En la Figura 20, se presenta la estructura general de los datos.

	Class Index	Title	Description
0	3	Wall St. Bears Claw Back Into the Black (Reuters)	Reuters - Short-sellers, Wall Street's dwindli...
1	3	Carlyle Looks Toward Commercial Aerospace (Reu...	Reuters - Private investment firm Carlyle Grou...
2	3	Oil and Economy Cloud Stocks' Outlook (Reuters)	Reuters - Soaring crude prices plus worries\ab...
3	3	Iraq Halts Oil Exports from Main Southern Pipe...	Reuters - Authorities have halted oil export\f...
4	3	Oil prices soar to all-time record, posing new...	AFP - Tearaway world oil prices, toppling reco...

Figura 20: Estructura inicial de los datos.

Fuente: Elaboración propia.

Para los distintos modelos de minería de datos, es necesario utilizar todo el texto que se tenga como una única entidad por fila. Por ende, las columnas Título (*Title*) y Descripción (*Description*) se unen bajo una sola columna. Posteriormente, estas dos columnas iniciales se eliminan. Con ello, la estructura de la tabla luce como la presente en la Figura 21.

	class	content
0	3	Wall St. Bears Claw Back Into the Black (Reute...
1	3	Carlyle Looks Toward Commercial Aerospace (Reu...
2	3	Oil and Economy Cloud Stocks' Outlook (Reuters...
3	3	Iraq Halts Oil Exports from Main Southern Pipe...
4	3	Oil prices soar to all-time record, posing new...

Figura 21: Estructura de los datos posterior a unión de columnas.

Fuente: Elaboración propia.

4.3.2.2 Proceso de limpieza.

Basados en el análisis exploratorio y revisión de datos, se puede observar algunos de los puntos a ejecutar, en cuanto a limpieza de datos. Sin embargo, además de ello, hay más aspectos que se deben mejorar y ejecutar. El proceso de limpieza inicia con el reemplazo de todos los saltos de línea por un espacio vacío. Esto porque el salto de línea es especificado mediante "\\". En la Tabla 22, se presenta un registro ejemplo de cómo luce el antes y el después de eliminar los caracteres que especifican el salto de línea.

Tabla 22: Antes y después de eliminar los saltos de línea.

Antes	Después
What's New With Google News\\Google News has added a whole bunch of features while we weren't lookin'. First off there's a new pull-down menu at the top of the page which easily allows you access to the top stories across all the Google News properties. If you look at that ...	*What's New With Google News Google News has added a whole bunch of features while we weren't lookin'. First off there's a new pull-down menu at the top of the page which easily allows you access to the top stories across all the Google News properties. If you look at that ...*

Fuente: Elaboración propia.

Como siguiente paso, en la limpieza de datos, se tiene la eliminación de las comillas. En este caso, la comilla doble que, en términos generales, no aportan valor al análisis textual. En la Tabla 23, se presenta un registro ejemplo de cómo luce antes y después de eliminar las comillas.

Tabla 23: Antes y después de eliminar las comillas dobles.

Antes	Después
Kerry Challenges Bush Record on Issues DETROIT - Sen. John Kerry accused President Bush on Wednesday of presiding over an "excuse presidency," challenging	*Kerry Challenges Bush Record on Issues DETROIT - Sen. John Kerry accused President Bush on Wednesday of presiding over an excuse presidency, challenging Bush's credibility on jobs,*

Bush's credibility on jobs, the record national deficit and the war in Iraq...	*the record national deficit and the war in Iraq...*

Fuente: Elaboración propia.

Una vez se eliminan las comillas, se procede a convertir todos los registros a minúsculas, ya que esto puede llegar a afectar al tener palabras como *Hello, hello, heLLo,* la cual es la misma y, por ende, debería ser identificada como tal. En la Tabla 24, se puede observar un registro ejemplo del antes y después

Tabla 24: Antes y después de convertir el texto a minúsculas.

Antes	Después
Kerry Challenges Bush Record on Issues DETROIT - Sen. John Kerry accused President Bush on Wednesday of presiding over an "excuse presidency," challenging Bush's credibility on jobs, the record national deficit and the war in Iraq...	*kerry challenges bush record on issues detroit - sen. john kerry accused president bush on wednesday of presiding over an excuse presidency, challenging bush's credibility on jobs, the record national deficit and the war in iraq...*

Fuente: Elaboración propia.

En inglés es común el uso de "s" para indicar pronombres posesivos. Por ejemplo: *bush's credibility* (credibilidad de Bush), este apóstrofe no brinda valor al conjunto de datos y el análisis en sí. Por dicha razón, es eliminado y reemplazado por vacío, de esta forma, se obtiene *Bush*, en lugar de ***Bush's*** y los distintos modelos los puede considerar la misma palabra. En la Tabla 25, se muestra un ejemplo del antes y después de eliminar los pronombres posesivos. Puede que la oración luzca gramáticamente incorrecta, sin embargo, lo importante es la reducción de las palabras a su forma simple. Los distintos modelos se basan a nivel de palabras, por ende, la gramática y semántica no tienen valor en el resultado final, tal como si lo tiene la correcta reducción de las palabras.

Tabla 25: Antes y después de eliminar los pronombres posesivos.

Antes	Después
Kerry Challenges Bush Record on Issues DETROIT - Sen. John Kerry	*kerry challenges bush record on issues detroit - sen. john kerry accused*

accused President Bush on Wednesday of presiding over an excuse presidency, challenging Bush's credibility on jobs, the record national deficit and the war in Iraq...	*president bush on wednesday of presiding over an excuse presidency, challenging bush credibility on jobs, the record national deficit and the war in iraq...*

Fuente: Elaboración propia.

Por el hecho de realizar muchos reemplazos, a lo largo del documento, dichos reemplazos generan, por lo general, espacios en blanco, además de los espacios que pueda existir previamente. Los espacios en blanco no aportan ningún valor al estudio. Por ende, se proceden a eliminar todos los espacios y reemplazarlos por un espacio único. Para lograr el reemplazo, se utiliza la expresión regular "/s+", tal como se puede observar en la Tabla 26.

Tabla 26: Eliminación de espacios extra.

Antes	Después
John Kerry accused President Bush on Wednesday of presiding over an excuse presidency, challenging Bush credibility on jobs, the record national deficit and the war in Iraq...	*john kerry accused president bush on wednesday of presiding over an excuse presidency, challenging bush credibility on jobs, the record national deficit and the war in iraq...*

Fuente: Elaboración propia.

El siguiente paso en la limpieza de datos se relaciona con las etiquetas HTML, que se encuentran en muchos de los documentos. La mayor parte de las etiquetas se encuentran en formato escapado, lo que quiere decir que, en lugar de observase "<", se tiene "*lt;*".

Como primer paso para eliminar las etiquetas HTML, se deben convertir los caracteres escapados a su carácter real, tal como se puede observar en la Tabla 27.

Tabla 27: Antes y después de convertir caracteres escapados a versión normal.

Antes	Después
Staples Profit Up, to Enter China Market NEW YORK (Reuters) - Staples	*Staples Profit Up, to Enter China Market NEW YORK (Reuters) -*

***Inc.** ylt**;A HREF=“http://www.investor.reuters.com/FullQuote.aspx?ticker=SPLS.O target=/stocks/quickinfo/fullquote”ygt**; SPLS.**Oylt;/Aygt**;*	*Staples Inc. <A HREF=“http://www.investor.reuters.com/FullQuote.aspx?ticker=SPLS.O target=/stocks/quickinfo/fullquote”>SPLS.O</A>*

Fuente: Elaboración propia.

Una vez convertidos los caracteres, se puede hacer uso de la biblioteca *Beautiful Soup* para eliminar las etiquetas HTML y conservar solamente el contenido.

Tabla 28: Antes y después de eliminar las etiquetas HTML.

Antes	Después
Staples Profit Up, to Enter China Market NEW YORK (Reuters) - Staples Inc. <A HREF=“http://www.investor.reuters.com/FullQuote.aspx?ticker=SPLS.O target=/stocks/quickinfo/fullquote”>SPLS.O</A>	*Staples Profit Up, to Enter China Market NEW YORK (Reuters) - Staples Inc. SPLS.O*

Fuente: Elaboración propia.

En términos generales, los atributos de las etiquetas HTML no brindan valor al estudio. Usualmente, se tienen atributos como clases, id, valor, dirección URL, entre otras. Para todos ellos, a nivel general, no brindar aporte al estudio y es por dicha razón que son eliminados. Por el contrario, al dejarlos, la aparición de las etiquetas puede conllevar a ser palabras clave con gran peso, si aparecen en muchas ocasiones. Dicha situación es no deseada.

El siguiente paso en la limpieza de datos es la eliminación de los signos de puntación. Como se ha mencionado, cualquier símbolo o carácter puede afectar los resultados. Por ende, eliminar todos los signos de puntación es de suma importancia. Más aún, considerando que los signos de puntuación aparecen de forma muy repetitiva en los documentos, lo anterior conlleva a que los signos tengan gran peso, espacio que debe ser utilizado por palabras con valor y no signos que no

aportan a la investigación. En la Tabla 29, se presenta el antes y después de eliminar los signos de puntación.

Tabla 29: Antes y después de eliminar los signos de puntación.

Antes	Después
the top U.S. office products retailer, on Tuesday reported a 39 percent jump in quarterly profit, raised its full-year forecast and said it plans to enter the fast-growing Chinese market, sending its shares higher.	*the top US office products retailer on Tuesday reported a 39 percent jump in quarterly profit raised its full-year forecast and said it plans to enter the fast-growing Chinese market sending its shares higher.*

Fuente: Elaboración propia.

Como parte del proceso de limpieza de datos, se procede a ejecutar la lematización. En la Tabla 30, se presenta el antes y después de realizar el proceso de lematización, se resaltan en amarillo los puntos que fueron modificados para el registro en cuestión.

Tabla 30: Antes y después de aplicar lematización.

Antes	Después
the race is on second private team sets launch date for human spaceflight (space com) space com - toronto canada -- a second team of rocketeers competing for the #36 10 million ansari x prize a contest for privately funded suborbital space flight has officially announced the first launch date for its manned rocket	*the race be on second private team set launch date for human spaceflight (space com) space com - toronto canada -- a second team of rocketeers compete for the #36 10 million ansari x prize a contest for privately fund suborbital space flight have officially announce the first launch date for its man rocket*

Fuente: Elaboración propia.

Una vez completa la lematización, se puede establecer que los datos están en su forma raíz. El proceso de limpieza está prácticamente completo. Sin embargo, en ambiente de análisis textual, hay un aspecto que tiene gran peso sobre los datos, pero no tiene valor para los resultados, en inglés se denominan *stop words*, la

traducción utilizada es palabras vacías. En inglés, se pueden mencionar palabras como: *the*, *is*, *and*, entre otras; en español, se tienen palabras, tales como la, le, es, o, entre otros. Como se menciona, dichas palabras no tienen valor para el modelo, pero, en caso de mantenerse, pueden afectar los resultados, ya que aparecen mucho en los textos. Por dicha razón, el presente paso corresponde a eliminar las palabras vacías. En la

Tabla 31, se presenta el antes y después de eliminar las palabras vacías. Tal como se puede observar, dichas palabras son muy regulares en los textos y, por ende, pueden afectar mucho los resultados.

Tabla 31: Antes y después de eliminar las palabras vacías.

Antes	Después
the race be on second private team set launch date for human spaceflight (space com) space com - toronto canada -- a second team of rocketeers compete for the #36 10 million ansari x prize a contest for privately fund suborbital space flight have officially announce the first launch date for its man rocket	*race second private team set launch date human spaceflight (space com) space com - toronto canada -- second team rocketeers compete #36 10 million ansari x prize contest privately fund suborbital space flight officially announce first launch date man rocket*

Fuente: Elaboración propia.

Como parte de todo el proceso de limpieza de datos, durante el desarrollo y cambios, muchos espacios en blanco fueron introducidos. Por dicha razón, el último paso ejecutado en la limpieza de datos es la eliminación de todos los espacios en blanco excedentes.

4.3.3 Construir los datos.

Por el hecho de que se tiene el conjunto de datos completo para trabajar, el último paso previo al modelado es la construcción de los datos. No es necesario realizar la integración, ya que, por defecto, se tienen todos. Por construcción, a pesar de que se tiene el conjunto de datos, se hace referencia a la creación de los vectores que utilizan los distintos modelos.

El paso más importante de todo el proceso es la correcta limpieza de datos, esto porque, al tener los datos limpios, se elimina gran parte del ruido. Inclusive, si se tienen datos limpios, utilizar todas y cada una de las palabras para la entrada de

los modelos conlleva a un tiempo de entrenamiento sumamente alto. Además, el aumento en la cantidad de palabras no tiene una relación directa con la precisión. Debido a que, por ejemplo, muchas de las palabras puede que no tengan peso sobre la categoría. Es por dicha razón que, para los modelos de clasificación de texto, mediante aprendizaje supervisado, se tiende a utilizar técnicas como la de conteo de palabras o vectorización. En la presente investigación, se hace uso de la técnica de TF-IDF.

Capítulo 5. Propuesta de solución

5.1 Modelado

5.1.1 Selección de técnicas de modelado.

Para el desarrollo de esta investigación, se seleccionaron un total de 5 modelos de minería de datos. El problema a resolverse es un problema de clasificación y, para ello, se tienen las clases de cada registro. Por ende, el problema de minería de datos es un problema de aprendizaje supervisado. Los algoritmos de minería de datos utilizados son:

- Bosque aleatorio.
- Regresión logística multinomial.
- Redes bayesianas multinomiales.
- K Vecinos más cercanos.
- Máquina de soporte vectorial.

5.1.2 Diseño de pruebas.

El diseño de pruebas para la presente investigación y para todos los modelos en cuestión se basa en la precisión obtenida sobre los datos de pruebas para el modelo ya entrenado. Como punto inicial, es importante mencionar que los datos utilizados en los distintos modelos ya estaban previamente separados en datos de prueba y datos de entrenamiento. Estos no fueron alterados en su distribución porcentual para, en caso de ser necesario, poder comparar esta investigación con otras investigaciones que utilizan el mismo conjunto de datos. Anteriormente, en la sección 4.2 Entendimiento de los datos, se mencionan las características del conjunto de datos. Entre ellas que cuenta con 120 000 registros de entrenamiento y 7600 registros de prueba. El conjunto cuenta con 4 clases, por ende, se tienen 30

000 mil registros de entrenamiento y 1900 de prueba. Porcentualmente, un 93 % de los datos corresponden a entrenamiento y solamente el 7 % a datos de prueba.

En cuanto al diseño de pruebas, este se lleva a cabo de una manera incremental. Con incremental se hace referencia a un incremento controlado del número de características a utilizar como entrada del modelo. Como parte de la preparación de los datos, el resultado final es un vector de frecuencia de términos; en dicha sección, no se indica la cantidad total de entradas en dicho vector. Sin embargo, para probar los algoritmos, se utilizan vectores con la cantidad de términos descrita en Tabla 32. Las pruebas se realizan de forma incremental, iniciando con 300 términos y finalizando con 2000 términos. Para cada vector, se ejecutan los 5 algoritmos analizados, lo cual conlleva a un total de 20 pruebas.

Tabla 32: Cantidad de términos para cada prueba.

	Prueba 1	*Prueba 2*	*Prueba 3*	*Prueba 4*
Número de registros	300	800	1200	2000

Fuente: Elaboración propia.

Al probar un algoritmo de minería de datos, algunos son computacionalmente complejos, como, por ejemplo, las máquinas de soporte vectorial. Otros son menos intensivos como la regresión logística multinomial. La complejidad no es sinónimo de calidad. Sin embargo, para efectos de investigación, un aspecto de suma importancia es el tiempo de ejecución. La razón detrás de elegir pruebas entre 300 y 2000 es tanto por una estrategia incremental, como por temas de tiempo de ejecución, al tener algoritmos con tiempos de días.

No solamente la cantidad de términos es un aspecto a considerar. Además de ello, cada algoritmo tiene una serie de parámetros que se pueden ajustar, algunos de estos parámetros son valores continuos, lo cual quiere decir que se pueden tener N posibilidades. Para cada uno de los algoritmos acá estudiados, se puede hacer uso de bibliotecas para ajustar dichos parámetros con el método de “fuerza bruta”. Se definen todas las posibilidades a combinar y, posteriormente, la biblioteca se encarga de obtener la combinación con mejores resultados.

En términos generales, la estrategia suena muy convincente por el hecho de que ajustar parámetros es una tarea compleja. Sin embargo, si se toma en consideración que se tienen 4 pruebas distintas, con un máximo de 2000 términos,

para 5 algoritmos distintos, en cuestión de varios días, se pueden obtener los resultados para todos. No obstante, si a lo anterior se le agrega el uso de “fuerza bruta” para encontrar los mejores hiperparámetros, los tiempos de ejecución pueden fácilmente ascender a semanas o meses. Es por dicha razón que no se hace uso de bibliotecas para obtener hiperparámetros. Por lo contrario, se hace uso de un ajuste manual único y pruebas para validar precisión.

Tal como se indica, el método utilizado para comparar los resultados de los algoritmos es mediante la precisión. Específicamente, se hace uso de la función *accuracy_score* parte de la biblioteca Métricas de *scikit-learn.*

La función *accuracy_score* calcula la precisión, ya sea la fracción (predeterminada) o el recuento (normalizar = Falso) de las predicciones correctas (3.3. Metrics and scoring: quantifying the quality of predictions, s.f.). Para calcular la precisión, se utiliza: $precisón(x, y) = \frac{1}{n}\sum_{i=0}^{n-1} 1(x = y)$, en la cual “x” es la predicción y “y” es valor real. Dentro del paréntesis, se tiene (x = y), lo cual, al ser iguales, se tiene como resultado 1 caso contrario 0; “n” indica el número total de registros.

5.1.3 Construcción de los modelos.

Para cada uno de los cinco algoritmos en cuestión, se procede a construir el modelo, cada uno de ellos con sus correspondientes hiperparámetros y ajustes correspondientes.

5.1.3.1 Bosques aleatorios.

Para la creación del bosque aleatorio, se tienen una serie de hiperparámetros a configurar. En la Tabla 33, se presentan los parámetros a utilizar como entrada al modelo.

Tabla 33: Hiperparámetros para la creación del bosque aleatorio.

Parámetro	Valor	Descripción
Bootstrap	True	Utilizado para indicar al modelo que utilice un subconjunto de los datos para generar los distintos árboles. Durante la generación, distintas secciones del conjunto de datos son utilizada para cada árbol. Por otra parte, si este parámetro es definido como falso,

		todo el conjunto de datos es utilizado para generar cada árbol.
Random_state	8	Al definirse Bootstrap como verdadero, el parámetro *random_state* es utilizado para configurar la variabilidad del particionado de datos para la generación de los árboles. Al mismo tiempo, es utilizado para el muestreo de las características durante la división en cada nodo.
Max_depth	100	Parámetro utilizado para indicar la profundidad máxima de los árboles. Es útil para definir un límite de particionado y de esta forma controlar la profundidad de los árboles.
Max_features	Sqrt	Al utilizarse sqrt (raíz cuadrada) como parámetro de *max_features*, lo que se indica es que el número de características a considerar es igual a la raíz cuadrada del número de características. Por ejemplo: si se utilizan 120 000 características la cantidad máxima a utilizar es 346.
Min_Sample_Leaf	1	Para este parámetro, se utiliza el valor por defecto. 1 indica que debe existir al menos un registro que se pueda clasificar con este nodo para que este sea considerado hoja.
Min_sample_split	5	Este parámetro es utilizado para indicar al modelo la cantidad mínima de registros que debe tener un nodo interno para que pueda ser dividido.
N_estimators	1000	Los bosques aleatorios utilizan un conjunto de N árboles para promediar el resultado final del modelo. El valor por defecto para la cantidad de estimadores es 100. Sin embargo, luego de realizar varias pruebas incrementales, se notó que el modelo mejoraba inclusive hasta llegar a los 1000 estimadores. Por dicha razón, se configura la cantidad de estimadores con valor 1000.

Fuente: Elaboración propia basado en Scikit Learn (s.f.)

El modelo se ejecuta mediante la creación de un objeto *RandomForestClassifier* y asignando los parámetros anteriormente mencionados. En la Figura 22, se puede observar la creación del modelo con los parámetros descritos. En adición a los hiperparámetros, se tienen, además, dos parámetros extra, *n_jobs* = -1 para indicar que utilice tantos procesos en paralelo como sea posible. El parámetro *verbose* es para indicar al proceso que se desea ver el estado de la ejecución mientras está sucediendo.

```
rfc = RandomForestClassifier(random_state=8,
    bootstrap= True,
    max_depth= 100,
    max_features= 'sqrt',
    min_samples_leaf= 1,
    min_samples_split= 5,
    n_estimators= 1000,
    n_jobs=-1,
    verbose=True)
```

Figura 22: Creación de modelo Bosque Aleatorio. Fuente: Elaboración propia.

Finalmente, una vez creado el modelo el proceso de entrenamiento es muy sencillo de ejecutar. Simplemente, se debe utilizar el método *fit*, al cual se le debe pasar el vector de características y las etiquetas, ambas del conjunto de entrenamiento. Tal como se observa en la Figura 22, el modelo se asignó a una variable de nombre “rfc”, por ende, para ejecutar el entrenamiento, se utiliza *rfc.fit(features_train, labels_train).*

Posterior al entrenamiento, se tiene la fase de predicción, en la cual se utiliza el modelo entrenado con datos de prueba, para con ellos validar la precisión de este. Ejecutar la predicción es muy sencillo, solamente se debe ejecutar *rfc.predict(features_test),* de esta forma, se obtienen los resultados para el conjunto de prueba. Dado que se tienen las predicciones y también los valores reales, con ello, se puede validar la precisión de este. En la Figura 23, se presenta la salida obtenida al crear el modelo, realizar el entrenamiento y predecir los datos de prueba.

```
[Parallel(n_jobs=-1)]: Using backend ThreadingBackend with 4 concurrent workers.
[Parallel(n_jobs=-1)]: Done  42 tasks      | elapsed:   42.5s
[Parallel(n_jobs=-1)]: Done 192 tasks      | elapsed:  2.8min
[Parallel(n_jobs=-1)]: Done 442 tasks      | elapsed:  6.2min
[Parallel(n_jobs=-1)]: Done 792 tasks      | elapsed: 11.0min
[Parallel(n_jobs=-1)]: Done 1000 out of 1000 | elapsed: 13.9min finished
[Parallel(n_jobs=4)]: Using backend ThreadingBackend with 4 concurrent workers.
[Parallel(n_jobs=4)]: Done  42 tasks      | elapsed:    0.1s
[Parallel(n_jobs=4)]: Done 192 tasks      | elapsed:    0.3s
[Parallel(n_jobs=4)]: Done 442 tasks      | elapsed:    0.7s
[Parallel(n_jobs=4)]: Done 792 tasks      | elapsed:    1.2s
[Parallel(n_jobs=4)]: Done 1000 out of 1000 | elapsed:    1.5s finished
[Parallel(n_jobs=4)]: Using backend ThreadingBackend with 4 concurrent workers.
The training accuracy is:
[Parallel(n_jobs=4)]: Done  42 tasks      | elapsed:    0.9s
[Parallel(n_jobs=4)]: Done 192 tasks      | elapsed:    5.0s
[Parallel(n_jobs=4)]: Done 442 tasks      | elapsed:   10.9s
[Parallel(n_jobs=4)]: Done 792 tasks      | elapsed:   18.2s
```

Figura 23: Salida del proceso de creación, entrenamiento y predicción para el bosque aleatorio.

Fuente: Elaboración propia.

5.1.3.2 Máquinas de soporte vectorial.

Las máquinas de soporte vectorial son un algoritmo de procesamiento complejo, el cual puede consumir mucho tiempo, dependiendo la cantidad de datos. La biblioteca utilizada creada por *Scikit Learn*, si se hace referencia a la documentación en ella se menciona que el tiempo de entrenamiento incrementa, al menos de forma cuadrática, con respecto al número de registros. Además, para conjuntos de datos, en la escala de decenas de miles, resulta impráctico (Scikit Learn, s.f.). En la Tabla 25, se presenta los hiperparámetros utilizados para entrenar el modelo.

Tabla 34: Hiperparámetros utilizados para la máquina de soporte vectorial.

Parámetro	Valor	Descripción
Random_state	8	Utilizado para controlar la variabilidad de la generación de valores para la mezcla de datos de estimaciones.
Cache_size (MB)	10000	Con la plataforma utilizada mediante el uso de la versión paga se logra incrementar la cantidad posible de cache para utilizar. Por defecto el valor

		es 200MB. Para la ejecución de este proceso se utilizaron 10.000MB, alrededor de 10GB de cache.
Probability	*True*	Parámetro utilizado para habilitar estimaciones de probabilidad.
Kernel	*rbf*	RBF (función de base radial, del inglés *Radial Basis Function). Kernel* define el tipo de *kernel* que se desea utilizar en el modelo.

Fuente: Elaboración propia basado en Scikit Learn (s.f.)

Una vez definidos los hiperparámetros a utilizar para la creación del modelo, se procede a la creación del objeto mediante el uso de SVC (Clasificación de vector de soporte, del inglés *Support Vector Classification*).

```
svm =svm.SVC(random_state=8,
             verbose=True,
             cache_size=10000,
             probability=True,
             kernel='rbf')
```

Figura 24: Creación del objeto de la máquina de soporte vectorial.

Fuente: Elaboración propia.

Posterior a la creación del modelo, se inicia el entrenamiento de este. Basados en el nombre de la variable anteriormente asignada “svm”, el entrenamiento del modelo se realiza mediante el llamada de la función *fit*, *svc_0.fit(features_train, labels_train)*, al cual se le pasa la matriz de valores y el vector de etiquitas. No se utiliza la verbosidad para la salida durante la ejecución porque eso podría afectar el tiempo de ejecución para un algoritmo de alta complejidad computacional.

5.1.3.3 Regresión logística multinomial.

El tercer algoritmo a implementar es la regresión logística multinomial, la cual es una variante de la regresión logística binaria. La biblioteca utilizada para ejecutar este algoritmo es, al igual que para los anteriores modelos, *Scikit Learn*, la cual forma parte de los modelos lineales de la paquetería. En la Tabla 35, se presenta los hiperparámetros utilizados en la configuración del modelo.

Tabla 35: Hiperparámetros de configuración de modelo de regresión logística multinomial.

Parámetro	Valor	Descripción
Max_iter	1000	Parámetro que indica la cantidad máxima de iteraciones que puede llegar a hacer el algoritmo. Pueden ser menos. Si al alcanzar max_iter el algoritmo no ha convergido entonces el programa levanta una excepción (advertencia) indicando que no hubo convergencia.
Multi_class	Multinomial	Con respecto a *multi_class*, por defecto el valor asignado es auto. Y en efecto para el problema acá resuelto auto utiliza Multinomial por el tipo de datos utilizado. La otra opción admitida para el parámetro es OvR (uno contra todos, del One vs Rest). En caso de utilizarse OvR, el algoritmo aplica una regresión binaria.
Solver	*Newton-cg*	Indica el algoritmo a utilizar en el problema de optimización.
Penalty	L2	Para el solucionador, *Newton-cg* solamente soporta L2. Por ende, se específica que es la penalidad a utilizar.

Fuente: Elaboración propia basado en Scikit Learn (s.f.)

Una vez definidos los hiperparámetros a utilizar, se procede a la creación del objeto del tipo del modelo a realizar. El objeto se denomina *LogisticRegression* y toma como entrada los hiperparámetros anteriormente explicados. En la Figura 25, se muestra la creación del objeto con sus correspondientes hiperparámetros de configuración.

```
LR = LogisticRegression(max_iter=1000,
                        multi_class= 'multinomial',
                        penalty= 'l2',
                        solver= 'newton-cg')
```

Figura 25: Creación del objeto LogisticRegression, base del modelo.

Fuente: Elaboración propia.

Posterior a la creación del modelo la ejecución del entrenamiento, al igual que con los anteriores, solamente se debe llamar la función *fit* del objeto. Dado que el objeto se denomina “LR”, para entrenar el modelo se ejecuta *LR.fit(features_train, labels_train),* pasando como parámetros la matriz de características y sus correspondientes etiquetas.

5.1.3.4 Redes bayesianas multinomiales.

El modelo de red bayesiana multinomial utilizado para esta investigación es mediante *Scikit learn* y su biblioteca *MultinomialNB*. El modelo de redes bayesianas es aplicable para modelos con características discretas, pero también funciona con conteos fraccionales tales como en TF-IDF (Scikit Learn, s.f.).

El modelo de red bayesiana proporcionado por *Scikit Learn* permite configurar 3 hiperparámetros. Tal como se indica en Scikit Learn (s.f.), el parámetro *alpha* indica un parámetro aditivo de suavidad, con valor defecto de 1; *fit_prior* indica cuando aprender la clase antes de las probabilidades, por defecto, con valor verdadero, y, finalmente, *class_prior* con valor por defecto de no, si se especifica las prioridades no son ajustadas acorde a los datos. Para la ejecución de los modelos acá descritos dichos parámetros no fueron modificados. Por ende, se procede a la ejecución del modelo de forma directa. En la Figura 26, se presenta la creación del objeto y el inicio de la ejecución del entrenamiento.

```
multinomial_nb = MultinomialNB()
multinomial_nb = mnbc.fit(features_train, labels_train)
```

Figura 26: Creación del objeto de red bayesiana y entrenamiento de este.

Fuente: Elaboración propia.

5.1.3.5 K vecinos más cercanos

Para el modelo de los K vecinos más cercanos, se hace uso de la biblioteca *KNeighborsClassifier* parte del grupo *neighbors* (vecinos) de *Scikit Learn*. Como parte de los hiperparámetros, el modelo permite alrededor de 8 parámetros. Sin

embargo, para el presente caso, se hace uso de uno de ellos. El parámetro utilizado corresponde a *n_neighbors* del español, número de vecinos. Dicho parámetro indica la cantidad de vecinos a utilizar para las consultas. Por defecto, el valor es 5, para este caso de estudio, se utilizan 400 vecinos, ya que la cantidad de datos es elevada.

Para crear el objeto, se debe utilizar el objeto *KNeighborsClassifier* y pasar los parámetros deseados, tal como se observa en la Figura 28, en la cual se presenta la creación del modelo y su ejecución/entrenamiento.

```
knnm = KNeighborsClassifier(n_neighbors=400,
                            n_jobs=-1)
knnm.fit(features_train, labels_train)
```

Figura 27: Creación y ejecución del modelo de K vecinos más cercanos.

Fuente: Elaboración propia.

En la Figura 28, se presentan todos los parámetros configurados en el modelo. El parámetro *n_jobs* = -1 se utiliza para hacer uso de tantos hilos de procesamiento como sea posible, de esta forma, reducir los tiempos de ejecución.

```
{'algorithm': 'auto',
 'leaf_size': 30,
 'metric': 'minkowski',
 'metric_params': None,
 'n_jobs': -1,
 'n_neighbors': 400,
 'p': 2,
 'weights': 'uniform'}
```

Figura 28: Parámetros del modelo de K vecinos más cercanos.

Fuente: Elaboración propia.

5.1.4 Evaluación de los modelos

Con respecto a los modelos, desde el punto de vista técnico, solamente uno de ellos que, para futuras investigaciones, puede ser considerable a reemplazar. Dicho algoritmo es la máquina de soporte vectorial. Tal como se ha explicado en distintas secciones del documento, las máquinas de soporte vectorial son matemática y computacionalmente complejas, lo cual conlleva a tiempos de

ejecución más elevados. Inclusive, en caso de lograr buenos resultados, tiempos de ejecución muy elevados pueden hacer que el algoritmo sea impráctico.

5.2 Evaluación

5.2.1 Evaluación de los resultados.

Para cada uno de los cinco modelos y sus correspondientes cuatro pruebas, se recolectaron tanto los tiempos de ejecución del entrenamiento, como la precisión. Para la precisión, se hace uso de *accurracy_score* como método general. En la Tabla 36, se presentan los resultados obtenidos para cada modelo en términos de precisión.

Tabla 36: Resultados de los modelos. Primera columna representa el modelo, primera fila contiene la cantidad de características en vector.

	300	800	1200	2000
Bosque aleatorio	80 %	85 %	86 %	87 %
Regresión logística multinomial	79 %	86 %	87 %	89 %
Redes bayesianas multinomiales	78 %	84 %	85 %	87 %
K vecinos más cercanos	71 %	83 %	85 %	87 %
Máquina de soporte vectorial	82 %	88 %	89 %	90 %

Fuente: Elaboración propia.

Es importante recalcar que la máquina de soporte vectorial obtuvo los mejores resultados. Seguido de ella, se encuentra la regresión logística multinomial, separada solamente por un punto porcentual. Dichos resultados solamente hacen referencia a la precisión, en la Tabla 37, se presentan los resultados obtenidos con respecto a los tiempos de ejecución, a pesar de que la máquina de soporte vectorial logró buenos resultados en términos de precisión. Con respecto a los tiempos de ejecución, la diferencia es abismal.

Tabla 37: Tiempos de ejecución de los algoritmos. Tiempo expresado en minutos.

	300	800	1200	2000
Bosque aleatorio	22	31	40	50
Regresión logística multinomial	1	3	5	8
Redes bayesianas multinomiales	0.03	0.03	0.03	0.03
K vecinos más cercanos	9	19	27	47

Máquina de soporte vectorial	540	780	1020	1260

Fuente: Elaboración propia.

Con el objetivo de obtener una visión más clara de los resultados, específicamente para los dos algoritmos con mejores resultados, regresión logística y máquina de soporte, en la Figura 29, se muestra en modo de gráfico los tiempos. El objetivo primordial de la gráfica es observar si existe algún patrón con respecto a la tendencia de crecimiento. Sin embargo, a pesar de que los tiempos de ejecución de la máquina de soporte vectorial son muy altos, la tendencia de ambos es prácticamente lineal, con respecto a la cantidad de registros.

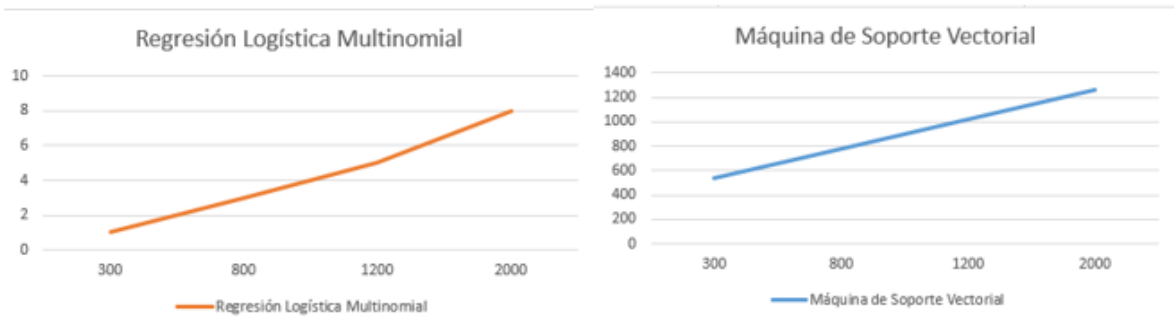


Figura 29: Tiempo de ejecución de: Regresión Logística Multinomial y Máquina de Soporte Vectorial.

Fuente: Elaboración propia.

Con respecto a una visión general de los resultados, en términos de precisión, en la Figura 30, se puede observar el resultado para todos. En dicha figura, un aspecto a notar es que, de forma común, el mayor crecimiento lo obtuvieron los algoritmos al pasar de 300 características a 800 características. El algoritmo con mayor crecimiento es el de K vecinos más cercanos, con alrededor de un 15 % de crecimiento en su primera fase. Posterior a ello, el crecimiento no fue tan significativo.

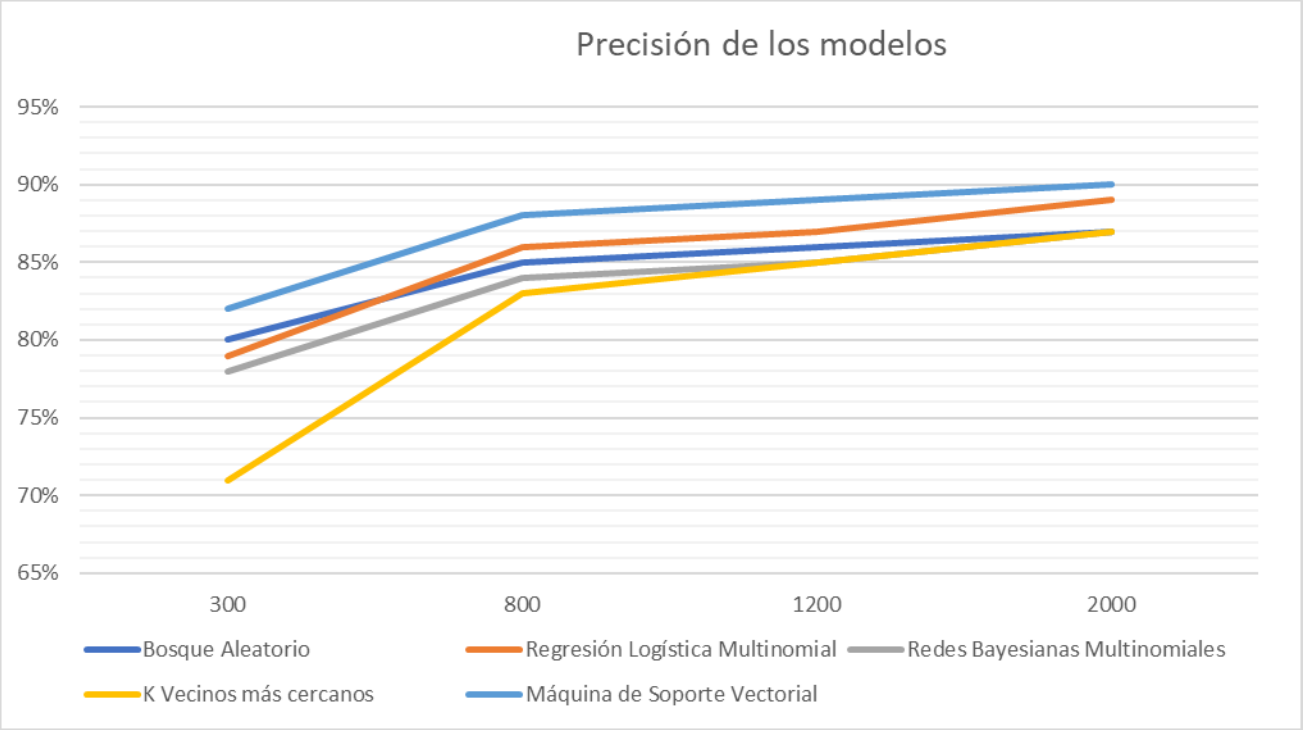


Figura 30: Precisión de los modelos de minería de datos ejecutados.

Fuente: Elaboración propia.

Si se enfoca nuevamente en los dos algoritmos con mejor rendimiento, la diferencia en tiempo de ejecución es de 1252 minutos, considerando el objetivo primordial del proyecto, el cual es la atención de emergencias, y que, en general, las emergencias demandan alta prioridad y rápido tiempo de respuesta. Es por dicha razón que se incluye una última prueba para tratar de igual forma el 90 % de precisión obtenido por la máquina de soporte vectorial. Esto con la idea de obtener un algoritmo con igual o mejor rendimiento, pero con menor tiempo de ejecución. Inclusive si para ello es necesario más características en el vector. El caso a realizar es el algoritmo de regresión logística multinomial con 4000 características, en lugar de 2000.

El resultado obtenido para las 4000 características con el modelo de regresión se puede observar en la Tabla 38. Tal como se presenta, la precisión es

igual al 90 % obtenido por la máquina de soporte vectorial. Sin embargo, el tiempo de ejecución es alrededor de 1500 minutos menor.

Tabla 38: Regresión logística multinomial utilizando 4000 características.

Modelo	Precisión	Tiempo de ejecución (minutos)
Regresión logística multinomial	90 %	14

Fuente: Elaboración propia.

5.2.2 Elección del mejor algoritmo (basado en criterios).

Basado en los criterios de rendimiento y tiempo de ejecución, se procede a elegir un algoritmo como mejor postor, en caso de implementación de una plataforma de reconocimiento de situaciones de emergencia. A pesar de que la máquina de soporte vectorial logra la mejor precisión basado en 2000 características, su seguidor es la regresión logística multinomial con solamente un punto porcentual de diferencia. Sin embargo, al incrementar las características a 4000 el algoritmo, logra un 90 % de precisión. Igualando así la precisión, pero con tiempo de ejecución 99 % menor. Por ende, el algoritmo a elegir como candidato es la regresión logística multinomial. En la Figura 31, se muestra el resultado final para la regresión logística multinomial. En ella, se puede observar el tiempo de ejecución de solamente 14 minutos y la precisión del 90 %.

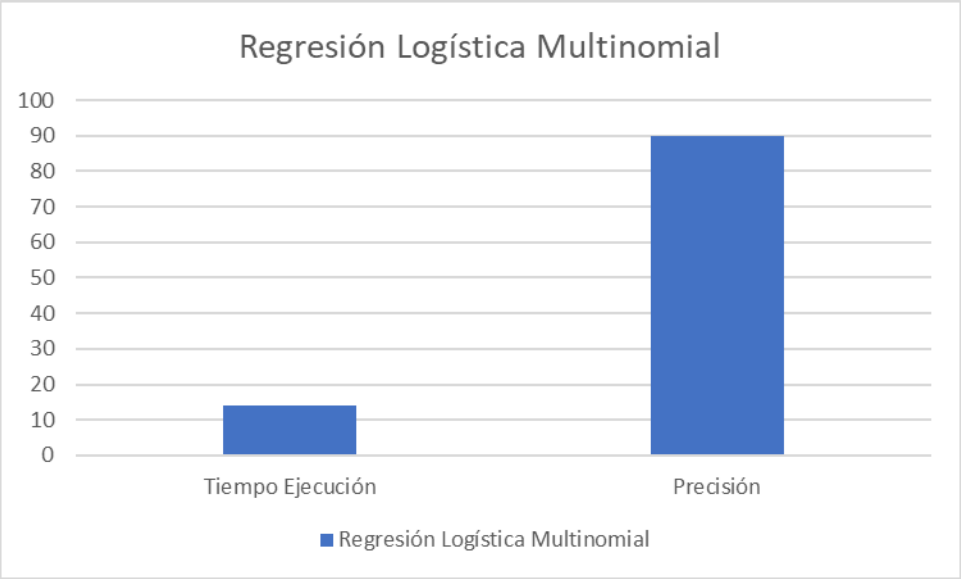


Figura 31: Resultados finales, regresión logística multinomial.

Fuente: Elaboración propia.

Capítulo 6. Conclusiones y recomendaciones

En el presente capítulo, se presenta una serie de conclusiones basada en los objetivos definidos para la investigación. Además, se indican recomendaciones con respecto a las experiencias obtenidas durante la investigación.

6.1 Conclusiones

Conclusiones del objetivo 1: Identificar las fuentes de datos a utilizar como entrada para los modelos de minería de datos. Este objetivo fue alcanzado y se concluye que:

- La identificación de las fuentes de datos fue, sin duda alguna, el paso más complejo con respecto a tiempo consumido y búsquedas realizadas.
- El plan inicial para la investigación siempre estuvo en utilizar datos reales. Sin embargo, posterior a contactar con el 9-1-1 de Costa Rica, se obtuvo respuesta negativa, con respecto a las transcripciones necesarias. Lo anterior por un tema de confidencialidad. Por alrededor de dos meses, se buscó un conjunto de datos de emergencias en inglés o español. El tiempo empleado se volvió inmanejable y, finalmente, fue necesario optar por datos alternativos.
- Una vez que se decidió que la búsqueda de datos alternativos era inviable, se inició un proceso de investigación con respecto a qué considerar para un conjunto de datos alternativo. Inicialmente, se debe considerar que debe ser un conjunto de datos categórico. La cantidad de clases no debe ser muy grande, preferiblemente alrededor de 5. El tamaño de los registros no debe ser muy grande con respecto al texto, esto porque, en general, una llamada corresponde a una serie de preguntas estructuradas con un formato específico. Además, al ser una emergencia, no se tienden a mantener largas conversaciones.
- A pesar de todos los avances en temas de datos públicos, en el área de atención de emergencias, lo anterior está reducido a solamente datos estadísticos, impidiendo así progreso y aporte de la comunidad científica.

Conclusiones del objetivo 2: Interpretar los datos para constatar que funcionan como datos alternos a datos de emergencias reales. Este objetivo fue alcanzado y se concluye que:

- Al no tenerse un punto de referencia de cómo luce realmente un conjunto de datos de emergencias real, es complicado poder interpretar para constatar que funciona como alternativo.
- Las llamadas al 9-1-1 tienen una serie de preguntas estructuradas y reducida. Por ende, un registro no debe ser muy extenso.
- La cantidad de categorías o resolución al hablarse de emergencias se reduce a policial, ambulancia, tránsito, bomberos y falsa alarma. Por ende, es preferible un conjunto de datos con alrededor de 5 categorías.

Conclusiones del objetivo 3: Convertir los datos de su forma raíz a datos numéricos que funcionen como entrada para los modelos. Este objetivo fue alcanzado y se concluye que:

- El procesamiento de texto es muy complejo y costoso, generalmente se convierte el texto a vectores. En esta investigación, se utiliza el TF-IDF como vector de términos.
- Uno de los pasos más importantes previo a la ejecución de los modelos es la creación de la matriz TF-IDF. Dicho proceso no solo se refiere a crearlo de forma aleatoria, sino que conlleva tiempo y pruebas. Si se realiza un conteo de cada una de las palabras del conjunto de datos, la cantidad es muy grande y sería demasiado complejo para procesar.
- Para poder saber el número perfecto de palabras a elegir en el vector, sería necesario utilizar todas las posibilidades y ejecutar el algoritmo para cada una de dichas posibilidades. Lo anterior es improbable, en términos de tiempo.
- Sería sencillo pensar que utilizar todas las palabras es una buena opción. Sin embargo, dentro de un conjunto de datos, se tienen muchas palabras que no aportan a un modelo. Por ejemplo: palabras vacías y palabras comunes que no diferencian un elemento de otro. Por ejemplo, la palabra asesinato puede ser muy determinante en un conjunto de datos de noticias. Pero la palabra “hola”, en general, no va a tener peso. Por ende, elegir un mínimo de

apariciones en el conjunto de datos es muy importante. Esto en conjunto con una correcta limpieza de datos.

Conclusiones del objetivo 4: Aplicar los distintos modelos de minería de datos propuestos. Este objetivo fue alcanzado y se concluye que:

- Cambios en el tamaño de la matriz de términos puede cambiar los resultados de un algoritmo en gran medida. Esto, por ejemplo, puede causar que un algoritmo, al aumentar de 1000 a 15 000 términos incremente en 10 % su rendimiento. Sin embargo, puede estancarse en 1500 y no cambiar prácticamente nada, inclusive si se utilizan 5000 términos.
- Es de suma importancia, para obtener resultados aceptables, la limpieza correcta de los datos, a mayor grado de limpieza de los datos menor será el ruido introducido a los algoritmos.
- El algoritmo de máquina de soporte vectorial logra buenos resultados para el conjunto de datos utilizado. Sin embargo, el tiempo de ejecución es extenso pero esperable ya que dicho algoritmo es matemáticamente muy complejo. A pesar de ser un tiempo esperable, es demasiado en comparación con los algoritmos en comparación.

Conclusiones del objetivo 5: Analizar los resultados de los algoritmos de minería de datos. Este objetivo fue alcanzado y se concluye que:

- Los dos algoritmos que lograron los mejores resultados fueron la máquina de soporte vectorial y la regresión logística multinomial.
- Los tiempos de ejecución de la regresión logística multinomial son muy bajos con respecto a la máquina de soporte vectorial.
- La regresión logística multinomial tiene un crecimiento prácticamente lineal, con respecto al incremento en la cantidad de términos.
- Es preferible aumentar la cantidad de términos para la regresión logística multinomial que obtener una precisión similar con la máquina de soporte vectorial, pero con menos términos. Esto por la diferencia en tiempos de ejecución en la cual inclusive incrementando la cantidad de términos en la regresión logística el tiempo de ejecución es mucho menor.

Conclusiones del objetivo 6: Proponer el modelo de clasificación con los mejores resultados. Este objetivo fue alcanzado y se concluye que:

- Basado en las necesidades de la investigación, los criterios para definir un algoritmo como mejor candidato que otro puede variar. En esta investigación, el rendimiento en tiempo de recursos es importante y, por dicha razón, fue utilizado como criterio.
- La complejidad de los algoritmos puede causar que los tiempos de ejecución se incrementen hasta cientos de veces el tiempo de ejecución de otros. Por ende, a menos que la diferencia en precisión sea muy alta entre un algoritmo y otro, el tiempo de ejecución de la fase de entrenamiento debe considerarse como un aspecto primordial para la elección.
- La regresión logística multinomial y la máquina de soporte vectorial logran una precisión muy similar.
- El salto de precisión entre los 300 términos y los 800 términos fue, en general, mayor que el salto entre las restantes opciones de términos. En este caso, entre 800-1200 y 1200-2000. A pesar de que no se continuó incrementando la cantidad de términos, por el anterior patrón, es muy posible que la precisión no va a incrementar en gran medida con el aumento de términos. Por ejemplo, para la regresión logística multinomial, que logró una precisión de 89 % con 2000 términos, al incrementar la cantidad de términos a doble solamente aumentó la precisión un 1%.

Conclusiones del objetivo general: Evaluar la precisión de distintos modelos de minería de datos para la posible aplicación en el contexto de atención de emergencias del sistema 9-1-1 de Costa Rica. Este objetivo fue alcanzado y se concluye que:

Inicialmente, el objetivo de la investigación más allá de evaluar la precisión para una posible aplicación en atención de emergencias era el de realmente evaluar datos de emergencias. Por ende, a pesar de que los datos tienen algunas características que se consideran similares a datos de emergencias, la diferencia en resultado puede ser muy grande por la variabilidad que puede tener la transcripción de voz a texto. Sin embargo, la clasificación de texto es una tarea compleja y los resultados obtenidos fueron muy buenos. Como algoritmo sobresaliente, se definieron todos aquellos que logran más de un 85 % de precisión. Para la presente

investigación, todos lograron más del 85 %, lo cual indica que, en términos generales, todos logran clasificar más de 85 de cada 100 textos. Además de ello, dos de los algoritmos, la máquina de soporte vectorial y la regresión logística, lograron un 90 % de precisión. Lo anterior concluye que 9 de cada 10 textos son clasificados correctamente. Si esto se mapea a emergencias, se puede decir que de cada 10 llamadas de emergencia 9 de ellas son correctamente clasificadas.

En términos generales, es difícil hacer una relación directa entre emergencias y texto de noticias. La transcripción de voz es un tema que puede afectar el resultado, a diferencia de las noticias que están redactadas por un profesional y además no depende de la voz. Por contraparte, el ámbito de la noticia puede tener gran variabilidad, al realizarse una llamada al 9-1-1, esta tiene una estructura predefinida basada en las preguntas estándar. Lo anterior en adición a que las emergencias tienen un vocabulario más reducido y eso podría indicar que se tienen palabras con mayor peso para cada situación o clase. El punto que puede afectar más los resultados obtenidos en esta investigación, en comparación con emergencias reales, es el impacto de la transcripción. Durante una emergencia, una persona puede estar alterada, llorando o con acento difícil de comprender. Dichos aspectos no afectan en la clasificación de noticias textuales, ya que se parte de un texto definido.

6.2 Recomendaciones

Tomando como base la experiencia obtenida con la presente investigación, la siguiente lista contiene una serie de recomendaciones y sugerencias para investigadores interesados en adentrarse en el área.

- Antes de considerar cualquier aspecto, tener claro las posibles complicaciones para obtener datos reales de emergencias, en caso de que sea una necesidad absoluta. Lo anterior porque los datos de emergencias tienden a ser de carácter confidencial. Para la presente investigación, la búsqueda de datos (conjunto de datos) tomó alrededor de 3 meses de revisión, consulta, correos y comunicación. Finalmente, fue imposible obtener un conjunto de datos válido.
- Por un tema de seguridad para el trabajo y, además, para tener la posibilidad de incrementar los recursos computacionales, se recomienda el uso de herramientas web, como, por ejemplo, *Google Drive* para almacenar los

datos y *Google Colaboratory* para realizar procesamiento y mantener el código disponible en cualquier hora y lugar. Se mencionan las anteriores herramientas, ya que fueron las utilizadas en la presente investigación, y, por ende, es con las que se tiene experiencia y confirmación del funcionamiento.

- La cantidad de bibliotecas y herramientas han crecido de forma acelerada, durante los últimos años. Lo anterior se indica para tratar de reutilizar las utilidades existentes tanto como sea posible, de esta forma, reducir los tiempos de investigación y enfocarse en los objetivos del trabajo. Además, la correcta elección del lenguaje de programación puede afectar en gran medida los tiempos de investigación. Lenguajes como *Python* y R han tomado mucha fuerza en los últimos años para temas de minería de datos. Por ende, la cantidad de facilidades es gigante en comparación con otros lenguajes de programación.
- Tal como se indica, la búsqueda de datos puede tomar bastante tiempo. Además de ello, otro aspecto a considerar es el tiempo de ejecución de los algoritmos y los algoritmos elegidos. Inicialmente, con respecto a los algoritmos, en la presente investigación, se ejecutaron 5. Sin embargo, en el mercado existe una basta lista de ellos. Por ende, un buen aporte sería probar los datos con nuevos algoritmos y constatar los resultados. Existen algoritmos con tiempos de ejecución cortos y también, por otra parte, existen algoritmos como la máquina de soporte vectorial que son intensos en el consumo de recursos.
- Dentro del entrenamiento de modelos para un mismo modelo, se tienen distintas configuraciones, denominado hiperparámetros. Para investigaciones de corta duración, el uso de fuerza bruta (probar muchas posibilidades o todas) para encontrar los hiperparámetros óptimos trae consigo un consumo muy alto de recursos computacionales. Con lo anterior, se da un gran incremento en los tiempos de ejecución, dependiendo del algoritmo, dicha búsqueda puede estar al nivel de días o semanas. Sin considerar que, en caso de no obtener resultados esperados, puede ser necesario realizar más limpieza de datos y volver a ejecutar los modelos. Para investigaciones de corta o mediana duración, con muchos algoritmos en prueba, los tiempos de ejecución pueden llegar a ser insostenibles.

- Con respecto a la limpieza de datos, se recomienda considerar como un punto clave, ya que gran parte de los resultados dependen de la correcta limpieza de los datos. Además, la nula o incompleta limpieza de datos trae consigo un incremento considerable en el tiempo requerido para completar la investigación.

Capítulo 7. Reflexiones finales

El presente trabajo surgió basado en la experiencia del autor con la utilización del sistema 9-1-1. A pesar de que 9-1-1 hace todo lo posible por ofrecer el mejor servicio posible, la ausencia de las tecnologías actuales causa dificultades de coordinación. El acceso a Internet cubre cada vez más regiones del país, las herramientas presentes en los dispositivos móviles crecen año con año. Sin embargo, se siguen tramitando las emergencias en prácticamente la misma manera.

Un aspecto que queda en evidencia con el presente trabajo es la falta de datos públicos relacionados con emergencias. Lo anterior traería consigo un gran aporte a la comunidad científica para poder desarrollar investigaciones al respecto y, con ello, mejorar y brindar una atención más oportuna. Aportes en el área se traducen en cantidad de vidas salvadas.

En relación con investigación en el área, basado en las búsquedas realizadas, la variedad de investigaciones es muy reducida. Posiblemente, la falta de investigación se deba al poco acceso a datos que se tiene. Además, dicho factor aparenta ser el causante del estancamiento en el área. En Costa Rica, no se encontró ninguna investigación, con respecto a la aplicación de 9-1-1 mejorado. Un excelente punto inicial sería el acceso a datos, de esta forma, investigadores nacionales e internacionales pueden realizar aportes y no es necesario una inversión por parte del sistema de atención de emergencia.

A modo de reflexión, la tecnología ha impactado prácticamente todas las áreas de la vida diaria. Sin embargo, a pesar de representar la diferencia entre vida o muerte, la atención de emergencias no ha tomado ventaja de todos los beneficios de las tecnologías actuales.

Capítulo 8. Trabajos a futuro

A pesar de que la presente investigación no se desarrolla con los datos reales de emergencia, los resultados son muy prometedores. Por ende, la lista de trabajos que se pueden realizar o iniciar basado en lo acá propuesto puede ser muy amplia, entre ellos se pueden mencionar:

- Si se logra obtener acceso en algún momento a corto plazo, un gran aporte sería la aplicación de la presente investigación a un conjunto de datos real de emergencias. Posiblemente, sea necesario realizar algunos ajustes en secciones como la limpieza de datos.
- Si al aplicar la presente investigación a datos reales se logran buenos resultados, el abanico de posibilidades es muy amplio. Se puede desarrollar una aplicación con directorios internos capaz de realizar preguntas a la persona, de forma escrita o verbal, con ello, tomar la decisión del ente a llamar. Quizá, un paso intermedio puede ser la creación de un agente para brindar soporte al personal del 9-1-1 encargado de atender las emergencias. Dicho agente puede ser utilizado para casos, donde se ha alcanzado la capacidad del 9-1-1 o como forma de confirmación para el personal encargado. Conforme se aumente el desarrollo de sistemas e investigación, la inclusión de sensores y herramientas es inminente, por ejemplo: GPS, cámara, sensores de presión sanguínea, entre otros.

Referencias

3.3. Metrics and scoring: quantifying the quality of predictions. (s.f.). Recuperado de Scikit Learn: https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score

911 and E911 Services. (n.d.). Recuperado de Federal Communications Commission: https://www.fcc.gov/general/9-1-1-and-e9-1-1-services

9-1-1 Costa Rica. (2019, Noviembre 13). *Datos Estadísticos*. Recuperado de 9-1-1 Costa Rica: http://www.911.go.cr/datos-estadisticos/

9-1-1 Costa Rica. (2019, Noviembre 22). *Filosofía empresarial: misión, visión, objetivos*. Recuperado de 9-1-1 Costa Rica: http://www.911.go.cr/filosofia-empresarial/

9-1-1 Costa Rica. (2020, Abril 07). *Reseña Histórica*. Recuperado de 9-1-1 Costa Rica: http://www.911.go.cr/resena-historica/

Aggarwal, C., y Zhai, C. (2012, Enero 07). A Survey of Text Classification Algorithms. *Mining Text Data.* doi:https://doi.org/10.1007/978-1-4614-3223-4_6

Betancourt, G. (2005). LAS MÁQUINAS DE SOPORTE VECTORIAL (SVMs). *Scientia Et Technica, 1*(27). doi:https://doi.org/10.22517/23447214.6895

Bonaccorso, G. (2017). *Machine Learning Algorithms.* Packt Publishing Ltd.

Breiman, L. (2001). Random Forest. *Machine Learning, 45*, 5-32. doi: https://doi.org/10.1023/A:1010933404324

Cambridge Dictionary. (n.d.). *Epistemología*. Recuperado de Cambridge Dictionary: https://dictionary.cambridge.org/es/diccionario/ingles/epistemology

Cambridge Dictionary. (n.d.). *Ontología*. Recuperado de Cambridge Dictionary: https://dictionary.cambridge.org/es/diccionario/ingles/ontology

Cambridge Dictionary. (s.f.). *Cambridge Dictionary*. Recuperado de Dataset: https://dictionary.cambridge.org/es/diccionario/ingles/dataset

Chavarría-González, M. C. (2011). Actualidades en Psicología. Recuperado de La dicotomía cuantitativa/cualitativo, falsos dilemas en la investigación social: https://revistas.ucr.ac.cr/index.php/actualidades/article/view/70/pdf_56

Cobertura de Servicio. (2019, Noviembre 22). Recuperado de 9-1-1 Costa Rica: http://www.911.go.cr/cobertura-de-servicio/

Data Scientist Salaries. (2021, Junio 06). Recuperado de Glassdor, inc: https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm

Dato. (2020). Recuperado de Real Academia Española: https://dle.rae.es/dato#Bskzsq5

Duran, J. (2020, 12 11). *Fake News Detection on Social Media.* Recuperado de ACM Digital Library: https://users.wpi.edu/~kmus/ECE579M_files/ReadingMaterials/fake_news[1828].pdf

Friedman, J. (1997). Data Mining and Statistics: What's the Connection? *Proceedings of the 29th Symposium on the Interface Between Computer*. Recuperado de https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.7990yrep=rep1ytype=pdf

Hipp, J., y Wirth's, R. (2000). CRISP-DM: Towards a Standard Process Model for Data. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*.

Holden, S. (2018, Setiembre 16). *The Python Wiki*. Recuperado de Python: https://wiki.python.org/moin/FrontPage

IEEE Transactions on Knowledge and Data Engineering. (2021, Abril). Recuperado de Scimago Journal y Country Rank: https://www.scimagojr.com/journalsearch.php?q=17362ytip=sidyclean=0

Jiang, L., Wang, S., Li, C., y Zhang, L. (2016). Structure extended multinomial naive Bayes. *Information Sciences, 329*, 346-356. doi:https://doi.org/10.1016/j.ins.2015.09.037

Journal of the ACM. (2021, Abril). Recuperado de Scimago Journal y Country Rank: https://www.scimagojr.com/journalsearch.php?q=23127ytip=sid

Jupyter. (2021, Julio 13). Recuperado de Jupyter: https://jupyter.org/

Khalaf, H., y Zaman, R. (2014). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science*.

La Nación. (2018, Enero 09). *Urge financiar el 911*. Recuperado de La Nación: https://www.nacion.com/opinion/editorial/editorial-urge-financiar-el-911/EJUWQSQDLVDDZN4FUGRSNUFR4E/story/

Li, J., Krivoshik, P., Suvorov, A., Fortes, C., y Kenny, P. (2018, Agosto 15). LifeLine: A Device for Detecting Abnormal Patterns. *Proceedings of the International*

Conference on Pattern Recognition and Artificial Intelligence, 76-81. doi:https://doi.org/10.1145/3243250.3243265

Lin, Y., Becker, E., Park, K., Le, Z., y Makedon, F. (2009, Junio 09). Decision making in assistive environments using multimodal observations. *Proceedings of the 2nd International Conference on Pervasive Technologies Related to Assistive Environments*, 1-8. doi:https://doi.org/10.1145/1579114.1579120

López, J., García, J., De la Fuente, L., y De la Fuente, E. (2007). LAS REDES BAYESIANAS COMO HERRAMIENTAS DE MODELADO EN PSICOLOGÍA. *Anales de Psicología / Annals of Psychology, 23*(2), 307-316. doi:https://doi.org/10.6018/analesps

López, M. (2020, 07 16). *Qué es un lenguaje de programación*. Recuperado de OpenWebinars: https://openwebinars.net/blog/que-es-un-lenguaje-de-programacion/

Mitchell, T. (1997). *Machine Learning.* Nueva York: McGraw-Hill.

Namakforoosh, M. N. (2005). *Metodología de Investigación.* México: Limusa.

Naqa, I., y Murphy, M. (2015). What Is Machine Learning? *Machine Learning in Radiation Oncology*. doi:https://doi.org/10.1007/978-3-319-18305-3_1

Naranjo, L. (2020). Investigación en Informática: el enfoque alternativo. *Technology Inside by CPIC*, *5*, 1–15. Recuperado de https://cpic-sistemas.or.cr/revista/index.php/technology-inside/article/view/35

Nihar, R., y Midhun, C. (2021). A Brief Survey of Machine Learning Algorithms for Text Document Classification on Incremental Database. *Test Engineering and Management, 83*, 25246 – 25251. Recuperado de https://www.researchgate.net/publication/350451142_A_Brief_Survey_of_Machine_Learning_Algorithms_for_Text_Document_Classification_on_Incremental_Database

Novakovi, J., Veljovic, A., Ilic, S., Papic, Z., y Tomovic, M. (2017). Evaluation of Classification Models in Machine Learning. *Theory and Applications of Mathematics y Computer Science*, 39-46. Recuperado de https://uav.ro/applications/se/journal/index.php/TAMCS/article/view/158/126

Oluwakemi, A., Samuel, O., y Taye, A. (2018). Text Classification Using Data Mining Techniques: A Review. *Information Systems Education Journal, 22*, 1-8. Recuperado de https://www.researchgate.net/publication/325930133_Text_Classification_Using_Data_Mining_Techniques_A_Review

Ordoñes, Y., Horacio, D., y Grass Boada, D. (2011). HERMINWEB: Herramienta de Minería de Uso de la Web Aplicado a los Registros del Proxy. *Conference: 9th Latin American and Caribbean Conference for Engineering and Technology*.

Oxford Léxico. (s.f.). *Accuracy*. Recuperado de Oxford Léxico: https://www.lexico.com/definition/accuracy

Pando, V., y San Martín, R. (2004). Regresión logística multinomial. *Cuadernos de la Sociedad Española de Ciencias Forestales, 18*. doi:https://doi.org/10.31167/csef.v0i18.9478

Preum, S., Shu, S., Hotaki, M., Williams, R., Stankovic, J., y Alemzadeh, H. (2019, Julio 16). CognitiveEMS: a cognitive assistant system for emergency medical services. *SIGBED Rev., 16*(2), 51-6. doi:https://dl.acm.org/doi/pdf/10.1145/3357495.3357502

Product Feedback. (s.f.). Recuperado de Kaggle: https://www.kaggle.com/product-feedback/93922

Real Academia Española. (2020). *Algoritmo*. Recuperado de Diccionario de la Real Academia Española: https://dle.rae.es/algoritmo?m=form

Real Academia Española. (2020). *Dato*. Recuperado de Diccionario de Real Academia Española: https://dle.rae.es/dato#Bskzsq5

Real Academia Española. (2020). *Emergencia*. Recuperado de Diccionario de la Real Academia Española: https://dle.rae.es/emergencia

Real Academia Española. (2020). *Información*. Recuperado de Diccionario de la Real Academia Española: https://dle.rae.es/informaci%C3%B3n

Scikit Learn. (s.f.). *sklearn.ensemble.RandomForestClassifier*. Recuperado de Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scikit Learn. (s.f.). *sklearn.linear_model.LogisticRegression*. Recuperado de Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Scikit Learn. (s.f.). *sklearn.naive_bayes.MultinomialNB*. Recuperado de Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

Scikit Learn. (s.f.). *sklearn.svm.SVC*. Recuperado de Scikit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Shu, S., Preum, S., Pitchford, H., Williams, R., Stankovic, J., y Alemzadeh, H. (2020, Enero 27). A Behavior Tree Cognitive Assistant System for Emergency Medical Services. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 6188-6195. doi:https://doi.org/10.1109/IROS40897.2019.8968233

Superintendencia de Salud de Chile. (s.f.). *Urgencia o emergencia vital*. Recuperado de Superintendencia de Salud: http://www.supersalud.gob.cl/difusion/665/w3-propertyvalue-2129.html#:~:text=Es%20toda%20condici%C3%B3n%20cl%C3%ADnica%20que,a%20un%20establecimiento%20de%20salud.

Tableau. (s.f.). *Tableau*. Recuperado de Guide To Data Cleaning: Definition, Benefits, Components, And How To Clean Your Data: https://www.tableau.com/learn/articles/what-is-data-cleaning

The Stanford Natural Language Processing Group. (s.f.). *Stemming and lemmatization*. Recuperado de The Stanford Natural Language Processing Group: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Tong, Y., y Hong, Z. (2020). Hyper-Parameter Optimization: A Review of Algorithms. *Machine Learning*.

Uso del 9-1-1. (2019, Noviembre 22). Recuperado de 9-1-1 Costa Rica: http://www.911.go.cr/uso-del-9-1-1/

What is Colaboratory? (s.f.). Recuperado de Colaboratory: https://colab.research.google.com/notebooks/intro.ipynb

Zhang, W., Yoshida, T., y Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Syst. Appl.*, 2758-2765. doi:10.1016/j.eswa.2010.08.066

Zhang, X., Zhao, J., y LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *CoRR*. Recuperado de Kaggle.

Zhang, X., Zhao, J., y LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. *NeurIPS*, 649-657.