Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Posgrado
Maestría en Tecnologías de la Información y Comunicación

# ALGORITMO DE MINADO DE TEXTO DE PUBLICACIONES OBTENIDAS DE GRUPOS ANIMALISTAS EN REDES SOCIALES PARA LA APLICACIÓN CASTRAPP

**Ing. Christopher Abimael Palma Ortíz**
Asesorado por la MSc. Inga. Yuri Asucena Castro Estrada

Guatemala, mayo de 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA


FACULTAD DE INGENIERÍA

# ALGORITMO DE MINADO DE TEXTO DE PUBLICACIONES OBTENIDAS DE GRUPOS ANIMALISTAS EN REDES SOCIALES PARA LA APLICACIÓN CASTRAPP

TRABAJO DE GRADUACIÓN

PRESENTADO A JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

**ING. CHRISTOPHER ABIMAEL PALMA ORTÍZ**
ASESORADO POR LA MSC. INGA. YURI ASUCENA CASTRO ESTRADA

AL CONFERÍRSELE EL TÍTULO DE

**MAESTRO EN TECNOLOGÍAS DE LA INFORMACIÓN Y COMUNICACIÓN**

GUATEMALA, MAYO DE 2022

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA


## NÓMINA DE JUNTA DIRECTIVA

| | |
|---|---|
| DECANA | Inga. Aurelia Anabela Cordova Estrada |
| VOCAL I | Ing. José Francisco Gómez Rivera |
| VOCAL II | Ing. Mario Renato Escobedo Martínez |
| VOCAL III | Ing. José Milton de León Bran |
| VOCAL IV | Br. Kevin Armando Cruz Lorente |
| VOCAL V | Br. Fernando José Paz González |
| SECRETARIO | Ing. Hugo Humberto Rivera Pérez |

## TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

| | |
|---|---|
| DECANA | Inga. Aurelia Anabela Cordova Estrada |
| EXAMINADOR | Mtro. Ing. Edgar Darío Alvarez Cotí |
| EXAMINADOR | Mtro. Ing. Ing. Marlon Antonio Pérez Turk |
| EXAMINADOR | Mtro. Ing. Estudardo Enrique Echeverría Nova |
| SECRETARIO | Ing. Hugo Humberto Rivera Pérez |

# HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

**ALGORITMO DE MINADO DE TEXTO DE PUBLICACIONES OBTENIDAS DE GRUPOS ANIMALISTAS EN REDES SOCIALES PARA LA APLICACIÓN CASTRAPP**

Tema que me fuera asignado por la Dirección de Escuela de Estudios de Postgrado con fecha 02 de agosto de 2019.

**Ing. Christopher Abimael Palma Ortíz**

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.378.2022

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **ALGORITMO DE MINADO DE TEXTO DE PUBLICACIONES OBTENIDAS DE GRUPOS ANIMALISTAS EN REDES SOCIALES PARA LA APLICACIÓN CASTRAPP,** presentado por: **Christopher Abimael Palma Ortíz,** que pertenece al programa de Maestría en artes en Tecnologías de la información y la comunicación después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANA
FACULTAD DE INGENIERÍA

Inga. Aurelia Anabela Cordova Estrada
Decana

Guatemala, mayo de 2022

AACE/gaoc

ESCUELA DE ESTUDIOS DE
POSTGRADO
FACULTAD DE INGENIERÍA

**Guatemala, mayo de 2022**

LNG.EEP.OI.378.2022

En mi calidad de Director de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

**"ALGORITMO DE MINADO DE TEXTO DE PUBLICACIONES OBTENIDAS DE GRUPOS ANIMALISTAS EN REDES SOCIALES PARA LA APLICACIÓN CASTRAPP"**

presentado por **Christopher Abimael Palma Ortíz** correspondiente al programa de **Maestría en artes en Tecnologías de la información y la comunicación** ; apruebo y autorizo el mismo.

Atentamente,

*"Id y Enseñad a Todos"*

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA · FACULTAD DE INGENIERÍA · DIRECCIÓN · ESCUELA DE POSTGRADO ·

**Mtro. Ing. Edgar Darío Álvarez Cotí**
**Director**
**Escuela de Estudios de Postgrado**
**Facultad de Ingeniería**

Guatemala, 10 de julio 2021

**M.A. Edgar Darío Álvarez Cotí**
**Director**
**Escuela de Estudios de Postgrado**
**Presente**

**M.A. Ingeniero Álvarez Cotí:**

Por este medio informo que he revisado y aprobado el **TRABAJO DE GRADUACIÓN** titulado: "ALGORITMO DE MINADO DE TEXTO DE PUBLICACIONES OBTENIDAS DE GRUPOS ANIMALISTAS EN REDES SOCIALES PARA LA APLICACIÓN CASTRAPP" del estudiante **Christopher Abimael Palma Ortíz** quien se identifica con número de carné **201212951** del programa de **Maestría en Tecnologías de la Información y la Comunicación.**

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el *Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014.* Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Atentamente,

MARLON ANTONIO PEREZ TURK
INGENIERO EN CIENCIAS Y SISTEMAS
COLEGIADO No. 4492

**MA. Ing. Marlon Antonio Pérez Türk**
**Coordinador**
**Maestría en Tecnologías de la Información y la Comunicación**
**Escuela de Estudios de Postgrado**

Guatemala, mayo de 2021

Maestro
Edgar Darío Álvarez Cotí
Director
Escuela de Estudios de Postgrado
Presente

Estimado M.A. Álvarez Cotí:

Reciba un cordial y atento saludo, a la vez aprovecho la oportunidad para hacer de su conocimiento que en mi calidad de Asesor del Ingeniero en Sistemas **Christopher Abimael Palma Ortíz** quien se identifica con carnet **201212951**, he revisado el Trabajo de Graduación titulado: **"ALGORITMO DE MINADO DE TEXTO DE PUBLICACIONES OBTENIDAS DE GRUPOS ANIMALISTAS EN REDES SOCIALES PARA LA APLICACIÓN CASTRAPP"** del programa de **Maestría en Tecnologías de la información y comunicación** de esta Escuela de Postgrado, **por lo cual el trabajo revisado cuenta con mi aprobación.**

Agradeciendo de antemano la atención a la presente, me suscribo.

Atentamente,

*"Id Y Enseñad A Todos"*

MSc. Yuri Asucena Castro Estrada
Asesora

YURI ASUCENA CASTRO ESTRADA
Ingeniera en Ciencias y Sistemas
Colegiado No. 14,800

# ACTO QUE DEDICO A:

| | |
|---|---|
| **Dios** | Por toda la sabiduría y entendimiento que me ha brindado durante toda mi vida. |
| **Mi abuela** | Por haber sido mi guía durante mi vida, estoy seguro de que se habría alegrado al verme completar una meta más que vio empezar, que descanse en la gloria de Dios. |
| **Mi madre** | Por haberme apoyado siempre en todas las cosas que emprendo, por aconsejarme y guiarme en cada paso que doy. |
| **Mis amigos** | Mis amigos Lucía Peréz, Freddy Monterroso y Salvador Citalán por haberme apoyado durante la realización del presente documento. |

# AGRADECIMIENTOS A:


| | |
|---|---|
| **Dios** | Por estar siempre presente en cada decisión que tomo en mi vida y por permitirme tener el gran privilegio de estudiar una maestría. |
| **Universidad de San Carlos de Guatemala** | Por brindar una educación de alto nivel y accesible para todos. |
| **Mi asesora** | Por transmitir siempre su conocimiento cada vez que fue necesario, así como el tiempo y apoyo brindado, agradecimientos a la Msc. Inga. Yuri Asucena Castro Estrada. |
| **Mis catedráticos** | Por compartir su conocimiento a través de experiencias y constante participación de la clase, haciendo las clases muy amenas. |

# ÍNDICE GENERAL

# ÍNDICE DE ILUSTRACIONES

## FIGURAS



## TABLAS

# LISTA DE SÍMBOLOS

| Símbolo | Significado |
|---|---|
| **Q** | Quetzales |
| **S** | Segundos |

# GLOSARIO

| | |
|---|---|
| **Algoritmo** | Es un conjunto ordenado y finito de operaciones que tienen como fin un resultado. |
| **Android** | Sistema operativo móvil de código abierto basado en núcleo Linux. |
| ***Azure function*** | Es una solución de Microsoft Azure para ejecutar pequeños fragmentos de código en la nube. |
| **Base de datos NoSQL** | Permiten almacenar datos que no requieren una estructura fija, por lo general en formato JSON. |
| ***Blob storage*** | Es un servicio de Microsoft Azure para almacenar grandes cantidades de datos de objetos no estructurados. |
| **Servicio web** | Tecnología que utiliza un conjunto de protocolos y estándares para enviar o recibir datos entre sistemas o aplicaciones. |

# RESUMEN

En Guatemala existen asociaciones y grupos animalistas en constante publicación de jornadas a bajo costo de servicios veterinarios, por ejemplo, castraciones, vacunaciones y adopciones de mascotas. La aplicación Castrapp recolecta las publicaciones y las agrupa en las 3 categorías mencionadas anteriormente.

La obtención de estas categorías se puede representar por medio de tópicos obtenidos de un conjunto de publicaciones (documentos) aplicando *Machine learning* mediante diferentes tipos de algoritmos, supervisados y no supervisados. Se decidió realizar un minado de texto de las publicaciones obtenidas de Instagram para la aplicación Castrap, seleccionando la rama de los no supervisados e implementando el algoritmo *Latent dirichlet allocation*, conocido por su abreviación como LDA. LDA fue utilizado para la generación de tópicos de publicaciones obtenidas de 6 grupos en la red social Instagram del año 2015 al año 2020, un total de 7,013 publicaciones.

Se definió una cantidad de 3 tópicos para la salida del modelo LDA. De los tópicos generados se eligieron las 5 palabras más representativas, el tópico 0 con vida, hogar, adoptar, familia e información, para el tópico 1 con banco, depósito, amigos, asociación y animales, para el tópico 3 con gato, castración, animal, gatuna y gatuno.

Los resultados del modelo fueron medidos por 2 variables, siendo el puntaje de coherencia con un valor de 0.7507 siendo este muy cercano a 0.7, que es el valor óptimo para el modelo. La segunda variable es la perplejidad con -7.4039

para el entrenamiento y -11.2273 para los datos de prueba, al ser ambas negativas indican que el modelo tiene un buen rendimiento al clasificar nuevas publicaciones.

Los resultados encontrados son similares a los contemplados en la aplicación Castrapp con el algoritmo anterior, realizando un análisis de documentos nuevos, es decir publicaciones nuevas del año 2021, encontrando el porcentaje de tópicos con la probabilidad de pertenencia a cada uno. El nuevo algoritmo tuvo un tiempo de carga de 1.0 segundos en promedio en recuperar las publicaciones de la base de datos mientras que el algoritmo anterior de Castrapp el tiempo de recuperación fue de 2.1 segundos en promedio.

El modelo final LDA fue implementado en una *Azure function* en la nube y realizar reclasificaciones de publicaciones nuevas periódicamente para posteriormente ser almacenadas en una base de datos NoSQL (Azure Cosmos DB) y ser retornadas mediante un *endpoint* a la aplicación Castrapp.

Para evaluar el modelo es necesario contar con un criterio e inferencia humana más a profundidad con más personas, es decir encontrar sentido lógico a los tópicos e inferir un tema.

# PLANTEAMIENTO DEL PROBLEMA Y FORMULACIÓN DE PREGUNTAS ORIENTADORAS

El uso de sitios web de redes sociales se ha convertido en un fenómeno atractivo en nuestro diario vivir (Salloum, Al-Emran, y Shaalan, 2017).

Guatemala es uno de los países de Centroamérica con mayor penetración de internet. En Guatemala con la creciente expansión del Internet y la tecnología móvil se ha incrementado el uso de las redes sociales permitiendo que grupos dedicados a ofrecer jornadas de castración, vacunación y adopción de mascotas lo hagan por medio de publicaciones en redes sociales. La red social más utilizada es Facebook mediante la creación de páginas haciendo uso de Facebook Pages. Dada la cantidad de páginas en Facebook existe una aplicación denominada Castrapp, en la cual se agrupan todas las páginas de Facebook de los grupos y se recolectan publicaciones las cuales son clasificadas mediante un algoritmo de comparación basado en palabras claves de las 3 grandes áreas; castración, vacunación y adopción.

Este algoritmo hace una comparación de palabras clave (previamente configuradas en la aplicación Castrapp) en cada una de las publicaciones obtenidas de las páginas de los grupos en Facebook para luego clasificarla en un área.

Sin embargo, las publicaciones obtenidas no están siendo clasificadas en las áreas correspondientes, por lo que los usuarios no pueden ubicar fácilmente las jornadas dentro de la aplicación. Las publicaciones pueden ser de 3 tipos, ya sea solo texto, texto con imágenes o solo imágenes, muchas veces se incluye un

texto corto y se complementa la información con la imagen. Adicional, cada grupo tiene su propia forma de expresarse al publicar una jornada. También se debe considerar la sintaxis y ambigüedad del lenguaje español, además, algunas veces existen faltas de ortografía. Se debe considerar que no existe una base de conocimiento para este tipo de temas en específico para realizar clasificaciones más acertadas.

El algoritmo de clasificación de la aplicación Castrapp tiene deficiencias al encontrar varias palabras clave dentro de una publicación obteniendo clasificaciones erróneas, además el algoritmo no considera el análisis de imágenes, en donde muchas veces se incluye toda la información de la publicación.

Existen casos similares en donde se ha obtenido información de redes sociales como Twitter y Facebook para su clasificación o agrupación, en el caso de los autores Salloum, Al-Emran y Shaalan (2017) en su estudio *Mining Text in New Channels: A Case Study from Facebook* se obtuvieron publicaciones de 16 páginas de Facebook de canales de noticias internacionales, las cuales fueron preprocesadas, se aplicaron diferentes técnicas de minado de texto así como la determinación de distintos tópicos y técnicas de *clustering* para la agrupación de datos obteniendo como resultado la clasificación de los datos en tres temas principales: Río de Janeiro, USA elections, y UK leaves the European Union, similar al objetivo que busca Castrapp, clasificar las publicaciones en diferentes áreas.

También el autor Ruiz (2015) de acuerdo con su tesis de maestría titulada *Optimización de la búsqueda de intereses personales en Twitter utilizando modelado de tópicos* hace uso de la red social Twitter en donde implementa el modelado de tópicos con el fin de proponer el diseño de un sistema que optimice

el proceso de determinar los intereses personales en Twitter. El autor menciona que Twitter es una buena fuente de información porque los intereses no están sesgados, comparado con Facebook, ya que en esta última las personas muestran una personalidad distinta a la real. Cabe mencionar que Castrapp no necesita tomar en cuenta intereses personales para clasificación de las publicaciones obtenidas de Instagram.

Así mismo los autores Kwok y Yu (2013) en su estudio *Spreading Social Media Messages on Facebook An Analysis of Restaurant Business-to-Consumer Communications*, analizaron los comentarios obtenidos de la páginas de Facebook de 10 cadenas de restaurantes y dos operadores independientes, en donde encontraron patrones interesantes, tales como los tipos de mensajes que reciben más me gusta, clasificación de mensajes en ventas, marketing y convencionales así como la cantidad de comentarios recibidos en un mensaje mediante técnicas de minado de texto.

La categorización correctamente de cada una de las publicaciones permitirá que los usuarios de la aplicación puedan ubicar fácilmente las jornadas y asistir a cada una de ellas, permitiendo que se den más posibilidades de adoptar, mantener saludable y castrar a las mascotas de los usuarios.

Por lo que se propone la siguiente pregunta principal: ¿Cómo optimizar el proceso de clasificación de publicaciones obtenidas de grupos animalistas de Instagram en la aplicación Castrapp?

Adicionalmente, las siguientes preguntas auxiliares se proponen:

- ¿Qué algoritmo de minado de texto se ajusta mejor a las necesidades de la aplicación Castrapp?

- ¿Qué algoritmo permite extraer el texto de las publicaciones que contienen imágenes en la aplicación Castrapp?

- ¿Qué métricas validan el correcto funcionamiento del algoritmo de clasificación de las publicaciones obtenidas de los grupos animalistas de Instagram en la aplicación Castrapp?

# OBJETIVOS

## General

Implementar un método de minado de texto que permita optimizar el proceso de clasificación de publicaciones obtenidas de grupos animalistas de Instagram en la aplicación Castrapp.

## Específicos

- Identificar y aplicar un método de minado de texto que se ajuste a las necesidades de la aplicación Castrapp.

- Identificar e implementar un algoritmo que permita extraer el texto de las publicaciones que contienen imágenes en la aplicación Castrapp.

- Determinar las métricas que validan el correcto funcionamiento del algoritmo de clasificación de las publicaciones obtenidas de los grupos animalistas en la aplicación Castrapp.

# MARCO METODOLÓGICO

- Tipo de investigación

Se realizó una investigación de sistemas para impulsar la inteligencia de negocios de tipo cuantitativa, ya que se aplicó el minado de texto a publicaciones obtenidas de Instagram de los grupos animalistas seleccionados para su clasificación y se implementó en el algoritmo de clasificación de publicaciones en la aplicación Castrapp.

Esta investigación obedece al tipo cuantitativa ya que los resultados obtenidos por medio del método *Latent Dirichlet Allocation* (LDA) son meramente estadísticos y con base a ello es posible agrupar las publicaciones en tópicos, entre los resultados se tienen los siguientes apartados:

  - Clústeres de palabras por tópico
  - Frecuencia de palabras por tópico
  - Distribución de tópicos por documento

- Diseño de investigación

La investigación se define como experimental dado que LDA al ser parte de *Machine lea rning* necesita de un modelo, dicho modelo fue entrenado para mejorar los agrupamientos y resultados.

Para entrenar el modelo fue necesario realizar múltiples corridas y ejecuciones del modelo, hasta lograr tener resultados adecuados. La herramienta *Machine Learning Studio* y las *Function Apps* (Python) de Azure

permitió ejecutar varias corridas con diferentes datos de prueba para entrenar el modelo y mejorar los resultados.

Una vez el modelo fue entrenado lo suficiente como para clasificar las publicaciones esté fue implementado en el algoritmo de clasificación de la aplicación Castrapp.

- Alcance

El alcance de la investigación fue descriptivo ya que se implementó un análisis a los datos por medio de un algoritmo de minado de texto, este algoritmo permitió clasificar publicaciones en distintos temas (tópicos).

El algoritmo de minado de texto se hizo mediante la técnica de *clustering* utilizando el método *Latent Dirichlet Allocation* (LDA) perteneciente *a Machine learning*.

Las áreas que se abarcaron fueron las siguientes:

- Castraciones
- Vacunaciones
- Adopciones

Para ello se tomarán en cuenta las siguientes variables que se representan en la tabla I.

Tabla I. **Variables, subvariables e indicadores**

| Variables | Definiciones | Subvariables | Indicadores |
|---|---|---|---|
| Clasificación de la publicación | Una publicación puede ser clasificada en una categoría (tópico) generada por el modelo LDA. | • Temas (tópicos) por publicación. | • Porcentaje de pertenencia a cada tópico.<br>• Palabras representativas de cada tópico. |
| Contenido de las publicaciones | Formado por los campos de la estructura de la publicación, entre los que se encuentran el texto de la publicación, imágenes, nombre de la página entre otros. | • Publicaciones con texto (numérica).<br>• Publicaciones con imágenes (numérica).<br>• Publicaciones contaminadas, es decir que no cumplen con los requerimientos para su | • Promedio y Media de palabras por todos los documentos.<br>• Promedio y Media de palabras por todos los documentos por tópico.<br>• Palabras de conexión removidas (*stop words*). |

| | | | |
|---|---|---|---|
| | | análisis (numérica). | |

Continuación tabla I.

| | | | |
|---|---|---|---|
| Modelado de tópicos utilizando Latent Dirichlet Allocation | Permite clasificar un texto mediante la extracción de temas del texto. | • Clústeres de palabras por tópico<br>• Frecuencia de palabras por tópico<br>• Distribución de tópicos por documento | 1. Cantidad de publicaciones de prueba para el entrenamiento del modelo.<br>2. Cantidad de temas (tópicos). |

| | | |
|---|---|---|
| | • Puntaje de Coherencia (*Coherence score*)<br>• Perplejidad (*Perplexity*)<br>• Topic Diff<br>• Rho | Número de iteraciones de entrenamiento del modelo. |

Fuente: elaboración propia, empleando Microsoft Excel.

La perplejidad (*perplexity*) fue utilizada como base de comparación, entre menor sea indica un mejor rendimiento del modelo.

El puntaje de coherencia (*coherence score*) mide la distancia relativa entre las palabras en un tópico, midiendo el grado de similitud semántica entre las palabras de puntuación alta probabilidad en el tópico.

*Topic diff* determina cuánto cambiaron los temas después de cada iteración y *rho* es la velocidad de actualización, es decir *rho* controla cuánto afecta el resultado el nuevo set de datos. Cuando ya no existen más cambios significativos entre cada iteración significa que el modelo ha convergido.

- Técnicas de recolección de información

Las técnicas de recolección de datos se dividen en 2 áreas que son, la primera para la investigación de los algoritmos y métodos de minado de texto y la segunda para los datos experimentales o de prueba para entrenar el modelo de *machine learning*.

Para el área de investigación de algoritmos y métodos de minado de texto se utilizaron fuentes tales como libros y ensayos relacionados a *machine learning*, minado de texto y *latent dirichlet allocation*.

Para la obtención de datos experimentales para entrenamiento del modelo de minado de texto se hará uso de *instaloader* para extraer publicaciones de Instagram.

- Procedimiento metodológico

Para llevar a cabo la presente investigación se llevaron a cabo las siguientes fases:

  - Fase I Investigación documental: fue necesario investigar sobre *machine learning* y el minado de texto para la correcta clasificación de las publicaciones en la aplicación Castrapp, para ello se tomaron en cuenta los siguientes puntos:

    - El método que se acopla a los requerimientos *Latent Dirichlet Alocation* (LDA), este pertenece al modelado de tópicos, es más utilizado y ampliamente implementado en más librerías versus otros como *latent simentic indexing* y *non-negative matrix factorization*.

    - Dentro del método *latent dirichlet allocation* se investigó que el método permite procesar texto sin formato, el cual aplica al texto de las publicaciones compartidas en la red social Instagram.

- En esta fase se analizó el proceso que sigue el método *latent dirichlet allocation*, el flujo del proceso consta de los siguientes pasos.

  - ✓ Obtener un conjunto de datos, es decir un *dataset* el cual lleva por nombre una colección de documentos.

  - ✓ *Blackbox*, es decir el algoritmo principal que se encarga de clasificar y encontrar patrones en el texto analizado (LDA).

  - ✓ Generación de resultados, entre los cuales se encuentran los tópicos, la frecuencia de palabras por tópico, y la distribución de temas por documento.

- Se investigaron los lenguajes de programación, siendo los más utilizados para *machine learning* los lenguajes Python y R.

- Se recopilaron librerías existentes y su grado de desarrollo e implementación, entre las que se encontró *vawpal wabbit* para ejecutar LDA, la cual existe para su implementación dentro de *machine learning studio* de Azure y *gensim* para Python en una *Function App* de Azure.

- Fase II Extracción de datos: en esta fase se realizó la solicitud de obtención de publicaciones a los distintos grupos de Instagram mediante *instaloader* (librería de Python).

- Se definieron los grupos adecuados. Los grupos objetivos son aquellos que cuenten con una cuenta registrada en Instagram. Entre los principales se encuentran descritos en la tabla II.

Tabla II. **Grupos animalistas en Instagram**

| **Nombre del grupo** | **Nombre de la cuenta en Instagram** |
|---|---|
| AMA Asociación de Amigos de los Animales | amaguatemala |
| Comunidad Gatuna | comunidadgatuna |
| GuateUnidaVoluntarios | guateunidavoluntarios |
| Animals Hope | animalshope14 |
| AlbergueMascotasMixco | alberguemascotasmixco |
| YoAmoALaCreacionGuate | yoamoalacreacionguate |

Fuente: elaboración propia, empleando Microsoft Excel.

- Los principales campos que se recuperaron de las publicaciones de cada grupo son:

  - ✓ *['node']['__typename']*
  - ✓ *['node']['edge_media_preview_like']['count']*
  - ✓ *['node']['edge_media_to_comment']['count']*
  - ✓ *['node']['owner']['username']*
  - ✓ *['node']['owner']['edge_followed_by']['count']*
  - ✓ *['node']['edge_media_to_caption']['edges'][0]['node']['text']*
  - ✓ *['node']['display_url']*

En dado caso la imagen contenga texto se procedió con la fase III, de lo contrario solo se publicaron los resultados obtenidos en el archivo CSV.

- Cada una de las publicaciones fue almacenada en un archivo con formato CSV en el servidor en donde se realizará el minado de texto con LDA. Las columnas (encabezados) del archivo CSV es similar al siguiente:

  - ✓ *number*
  - ✓ *comments*
  - ✓ *filename*
  - ✓ *followers*
  - ✓ *likes*
  - ✓ *post_id*
  - ✓ *text*
  - ✓ *text_image*
  - ✓ *time*
  - ✓ *username*

- Fase III Extracción de texto incrustado en la imagen.

  - Las publicaciones que contenían imágenes fueron procesadas por la herramienta *Computer Vision* de Microsoft para la extracción del texto para su posterior análisis. Las publicaciones que no contaban con imágenes no fueron procesadas por el método de extracción de texto de imágenes.

- Fase IV Limpieza de datos: en esta fase se efectuó una limpieza de publicaciones de aquellas que cuenten con alguna de las siguientes características, es decir se descartaran.

    - Texto con caracteres no reconocidos por el lenguaje
    - Publicaciones sin texto ni imágenes

Esta fase es muy importante ya que al aplicar una limpieza se logra descontaminar los datos que posteriormente fueron analizados con el método LDA.

- Fase V Ejecución del método *latent dirchlet allocation*: se ejecutó el método *latent dirchlet allocation* mediante la librería *gensim* estableciendo los siguientes parámetros propios del método.

    - Iteraciones
    - Pasadas
- Fase VI Publicación de publicaciones clasificadas y resultados del método LDA: los resultados son almacenados en una base de datos NoSQL con la siguiente estructura.

Figura 1. **Estructura de las publicaciones clasificadas**


```
{
   "filename": "tmp\\2021-03-23_21-37-10_UTC.json",
   "time": "2021-03-23 15:37:10",
   "text": "PULGUITA tiene una …",
   "likes": 101,
   "comments": 3,
```

```
    "username": "amaguatemala",
    "followers": 10435,
    "post_id": "2536003744392743667",
    "text_image": "",
    "category": 0,
    "id": "4da0ee4d-74fd-4481-8758-804f188e263a",
    "display_url": "https://scontent-...",
    "_rid": "jm1EAPY2Y-IkAAAAAAAAAA==",
    "_self": "dbs/jm1EAA==/colls/jm1EAPY2Y-I=/docs/jm1EAPY2Y-
IkAAAAAAAAAA==/",
    "_etag": "\"4a00d754-0000-0100-0000-60680e730000\"",
    "_attachments": "attachments/",
    "_ts": 1617432179
}
```


Fuente: elaboración propia.

Los datos de las publicaciones fueron expuestos mediante un *endpoint* para que la aplicación móvil pueda consultarlos.

- Fase VII Pruebas y resultados.

  - Se realizaron varias corridas del flujo total con publicaciones de prueba, para posteriormente verificar los siguientes atributos encontrados con el método LDA, en donde deben existir valores mayores que cero:

    - ✓ Clústeres de palabras por tópico

- ✓ Frecuencia de palabras por tópico
- ✓ Distribución de tópicos por documento

- Se cambiaron las publicaciones por nuevas y se realizaron nuevos entrenamientos, esto permitió verificar resultados y al mismo tiempo entrenar al modelo.

- Fase VIII Publicación del modelo como servicio web: finalmente, luego de entrenar el modelo (fases anteriores) se publicó un servicio web para que pueda ser invocado por la aplicación Castrapp. Las peticiones se realizan por medio de HTTP utilizando el verbo GET.

- Fase IX Desarrollo del algoritmo en la aplicación: el desarrollo del algoritmo consistió en actualizar el algoritmo de clasificación para que la fuente de los datos sea mediante una consulta HTTP a los *endpoints* de los datos localizados en la base de datos NoSQL.
  - Se desarrolló una aplicación híbrida desde cero, tomando en cuenta una interfaz amigable con los usuarios, así como los requerimientos no funcionales y funcionales.

- Instrumentos de recolección de información

Para el desarrollo de la investigación se utilizaron distintos instrumentos que ayudaron en la recolección de la información necesaria para completar la investigación, entre los instrumentos utilizados están:

- Fuentes secundarias

  - Artículos y/o ensayos científicos

- Trabajo de graduación de universidades
- Tutoriales y videos que aporten al desarrollo

- Fuentes primarias

  - Recolección de fotografías de placas de vehículos para realizar pruebas y validar la efectividad de DVRGT.

# INTRODUCCIÓN

El uso y crecimiento de Internet ha permitido que las redes sociales crezcan a un ritmo acelerado y con ello se genere una gran cantidad de datos, en su mayoría no estructurados. Miles de millones de datos no estructurados, los cuales a simple vista carecen de significado, pero si son analizados en conjunto es posible obtener información importante.

Dada la gran cantidad de datos no estructurados y la creciente expansión de Internet, es posible hacer uso del aprendizaje automático mediante *machine learning* y el minado de texto. Existen diferentes tipos de aprendizaje en *machine learning* así como métodos y técnicas, en este caso se aplica un aprendizaje clásico, mediante el aprendizaje no supervisado y el método de *clustering*, el cual es uno de los muchos caminos y ramas que *machine learning* posee. Adicionalmente dentro del método de *clustering* existen varios algoritmos, siendo *latent dirichlet allocation* el que mejor se acopla a las necesidades de la investigación.

En Guatemala existen leyes de protección de animales, así como organizaciones, entidades, personas y grupos dedicados a brindar jornadas a bajo costo con la finalidad de evitar sobrepoblación de animales y enfermedades, al mismo tiempo apoyar a las personas que no pueden costear un servicio de una veterinaria. Las jornadas sobre estos temas son publicadas en Facebook o Instagram por los distintos grupos, estas no poseen un estándar en cuanto a escritura y algunas veces la información está embebida en las imágenes adjuntas.

La aplicación Castrapp recolecta estas publicaciones de los grupos animalistas y mediante un algoritmo básico clasifica dichas publicaciones y las presenta al usuario clasificadas. Este algoritmo presenta deficiencias como la incorrecta clasificación de las publicaciones y no toma en consideración el texto de las imágenes.

Mediante la implementación de un modelo de aprendizaje automático es posible clasificar las publicaciones y dar mejores resultados a los usuarios. Dado que es necesario realizar una clasificación de texto no estructurado se utiliza el método de *clustering* mediante LDA para realizar las clasificaciones y dar resultados, así como preparar un modelo para los futuros análisis.

A continuación, se detalla de forma resumida cada uno de los capítulos de la investigación

En la sección del marco metodológico se detallan aspectos tales como el tipo de estudio seleccionado, el diseño que se aplicará, el alcance, así como variables, las cuales se subdividen en subvariables e indicadores, los cuales permitirán validar los resultados. Adicional se incluyen las técnicas de recolección de información, las fases que llevará el estudio (desde la extracción de publicaciones hasta la publicación de resultados) y las técnicas de análisis de información.

Capítulo I: se presentan casos similares en donde se ha utilizado con anterioridad el minado de texto y LDA para la clasificación de información, en donde los resultados fueron satisfactorios y valiosos. Se expone la importancia de tener un modelo de aprendizaje automático.

Capítulo II: se da a conocer las deficiencias de la aplicación Castrapp actual y de la utilización de algoritmos básicos de comparación. Se da a conocer como el no tomar en cuenta el texto de las imágenes no permite clasificar publicaciones visuales.

Capítulo III: se exponen los alcances de la investigación, los cuales se detallan desde una perspectiva investigativa, técnica y de resultados.

Capítulo IV: se detallan cada uno de los conceptos necesarios para comprender la rama seleccionada de *machine learning*, es decir aprendizaje automático no supervisado con el método de *clustering*. En este capítulo se explica el método *latent dirichlet allocation* y cuáles son sus parámetros y salidas.

Capítulo V: se exponen cada uno de los resultados obtenidos, los cuales son analizados conforme a los resultados del minado de texto mediante las técnicas de análisis de información presentadas en el capítulo cinco.

Capítulo VI: se analizan los resultados de la implementación del método LDA mediante los resultados que este devuelve, así como la comparación con el método anterior. Se analiza si los resultados son correctos y si se acoplan a los objetivos de la investigación.

# 1. ANTECEDENTES

La aplicación Castrapp recopila publicaciones de los grupos animalistas de Facebook y las categoriza en tres grandes áreas, castraciones, vacunaciones y adopciones (Palma, 2018). La clasificación se lleva a cabo con un algoritmo básico de clasificación utilizando palabras claves, por lo que las clasificaciones muchas veces son incorrectas. Adicional no existe ninguna base de conocimientos que pueda soportar estos temas en Guatemala, la cual puede construirse mediante *text mining*.

*Text mining* ha sido aplicado en diversas fuentes de datos, tales como redes sociales, entre ellas Facebook y Twitter. A continuación, se describen algunos estudios.

Los autores Salloum, Al-Emran y Shaalan (2017), realizan una comparación entre las diferencias de *data mining* y *text mining* donde detallan que *text mining* es una subparte de *data mining*, adicional explican que *data mining* trata de descubrir patrones interesantes en almacenes de datos masivas mientras que *text mining* son nombres de procedimientos para extraer datos interesantes y con significado e información de texto sin estructura.

Diferentes técnicas pueden ser aplicadas previamente al análisis y ejecución de *text ming*, por ejemplo, los autores Kwok y Yu (2013) aplicaron una estrategia importante que es la de dividir los mensajes en cuatro tipos: solo texto, conteniendo una URL, con video embebido, y mostrando una fotografía. Adicional también dividieron los mensajes en ventas y márketing y mensajes

convencionales. La clasificación de publicaciones es una técnica que puede ser aplicada a cualquier publicación obtenida de redes sociales.

*Text mining* se ha transformado en un campo con bastante tendencia y que ha sido incorporado en distintos campos de estudio tal como lingüística computacional, restablecimiento de datos de minería e información (Salloum, Al-Emran, Monem y Shaalan, 2017).

Los autores Salloum, Al-Emran, Monem y Shaalan (2017), explican los diferentes estudios e investigaciones que han sido aplicados mediante *text mining*, entre ellos el caso de la página de SAMSUNG Mobile en Facebook, en donde durante el período de 3 meses se recolectaron 128,371 comentarios considerados como *data* para posteriormente aplicar análisis conceptual que fue utilizado por el análisis de contenido y en última instancia, el análisis estadístico de conglomerados que se realizó mediante un análisis relacional, en donde los autores concluyen que los datos obtenidos de redes sociales se integran mediante el análisis estadístico de conglomerados y se realizan en función del resultado del análisis conceptual.

Otra área de aplicación en donde se ha utilizado *text mining* es la industria de la pizza, los autores He, Zha y Li (2013) analizaron el texto sin estructura de las páginas en Facebook y perfiles en Twitter de tres grandes cadenas de pizza: Pizza Hut, Domino's Pizza y Papa John's Pizza. Primero recolectaron todos los datos cuantitativos manualmente de las páginas/perfiles tales como número de fans/seguidores, número de publicaciones, comentarios, compartidos y *likes* así como la frecuencia de postear publicaciones. Segundo, aplicaron *text mining* para descubrir nuevos conocimientos y patrones, y para adquirir una comprensión más a detalle de cómo las tres cadenas de pizza están utilizando las redes sociales en la práctica.

Así mismo el autor Ruiz (2015) realiza una investigación experimental en donde utilizó el algoritmo LDA (*Latent dirichlet allocation*) para crear un modelo y así encontrar las probabilidades de que una palabra en un documento esté relacionada con un tema cualquiera, así mismo también utilizó TF IDF (*Term frequency inverse document frequency*) para obtener las frecuencias de palabras en un texto así, como los pesos en relación con el resto de los textos.

Por lo tanto, una vez conocidas las diversas implementaciones de *text mining* es posible aplicarlo a los temas de castraciones, vacunaciones y adopciones de mascotas en las publicaciones obtenidas de los grupos animalistas de Instagram en Guatemala.

# 2. JUSTIFICACIÓN

Este trabajo de graduación corresponde a la línea de investigación de sistemas para impulsar la inteligencia de negocios dado que mediante el minado de texto (*text mining*) se encuentran patrones en texto no estructurado para la clasificación de contenido de publicaciones de grupos animalistas de Instagram en Guatemala.

Las mascotas y animales juegan un rol importante en nuestra sociedad, en donde muchos viven en las calles y otros pocos en casas. Muchos grupos animalistas mediante páginas de Facebook e Instagram crean jornadas de adopciones, vacunaciones y castraciones de animales. Con ello se pretende reducir la cantidad de animales en la calle mediante la castración (prevención) y permitir dar un respiro a los refugios sobrepoblados mediante la adopción de mascotas y por medio de las vacunaciones a mantener la buena salud de animales que viven en una familia.

Dada la gran cantidad de publicaciones que los grupos animalistas publican en las redes sociales; se ha incrementado un desorden debido a las diversas publicaciones de diversos grupos animalistas. La aplicación Castrapp recopila dichas publicaciones, pero no se clasifican correctamente. Es por ello por lo que se propone un algoritmo mediante el minado de texto para extraer patrones importantes de las publicaciones para crear una base de conocimientos y así clasificar con mejores resultados las publicaciones.

Mediante el análisis de patrones se creará una base de conocimientos para obtener resultados más acertados en la clasificación de publicaciones en las áreas de castración, vacunación y adopción de mascotas. Dicha base de conocimientos es posible utilizarla en aplicaciones móviles o sitios web tal como en la aplicación Castrapp, esto permitirá a los usuarios localizar futuras publicaciones más rápidamente al estar clasificadas en el área adecuada.

Más adelante se espera que otro tipo de información pueda ser extraída de las publicaciones, como por ejemplo enfermedades, maltrato o abandono de animales, así como análisis en tiempo real de los comentarios que los usuarios hagan dentro de las páginas.

# 3. ALCANCES

## 3.1. Resultados

Se implementó un método de clasificación de jornadas animalistas para la aplicación Castrapp por medio del algoritmo de minado de publicaciones, permitiendo que la información sobre las jornadas sea localizada fácilmente por los usuarios, así como a la correcta difusión de estas por parte de los grupos.

Se realizó la extracción de texto de imágenes que permitió el proceso de minado de texto en aquellas publicaciones que poseen datos importantes incrustados dentro de la imagen.

Se creó una base de datos de conocimientos que permitió y permitirá identificar situaciones como maltrato, abandono entre otros.

## 3.2. Técnicos

Se desarrolló en el lenguaje Python un algoritmo que permitió minar publicaciones de forma óptima mediante LDA.

Se diseñó y se implementó una arquitectura en la nube que permitió minar publicaciones obtenidas de Instagram de los grupos animalistas mediante *APIs* de redes sociales o de terceros.

Se creó un servicio web REST que permitió ser consultado desde Internet y devolvió las publicaciones ya clasificadas según las categorías provistas por el modelo LDA.

Se implementó un algoritmo de extracción de texto de imágenes.

Se analizó el historial de los resultados obtenidos en el proceso de minado de texto en publicaciones.

### 3.3. Investigativos

Se investigó el comportamiento en el minado de publicaciones mediante *latent dirichlet allocation*.

Se investigaron algoritmos que permitan la extracción de texto en imágenes.

Se clasificaron las métricas en función del algoritmo de minado de publicaciones.

# 4. MARCO TEÓRICO

## 4.1. Redes sociales

Una red social representa una estructura social, está formada por actores y conexiones, es decir una conexión punto a punto entre dos individuos, estas identifican las relaciones sociales existentes entre los actores en un contexto social, por contexto social se entiende como un pueblo, una escuela, una ciudad, una clase entre otros (Arnaboldi, Passarella, Conti y Dunbar, 2015).

Las redes sociales poseen características tales como vértices (*vertex*) y bordes (*edges*), en donde cada individuo es un vértice y las relaciones entre ellos son los bordes (Ugander, Karrer, Backstrom y Marlow, 2011). Las redes sociales son parte del diario vivir.

### 4.1.1. Facebook

Facebook es una red social con presencia en varios países del mundo, se encuentra compitiendo con Google para convertirse en el sitio web más utilizado en el mundo, contando con 1.4 billones de usuarios mensualmente (Golbeck, 2015).

Facebook tiene tantas funcionalidades que es difícil de describirlo con una sola definición, pero en general Facebook es un sitio web en donde una persona puede registrarse en línea y compartir con otros amigos en línea. Dentro de estas funcionalidades se encuentran el poder crear usuario, grupos, páginas, desarrollo de aplicaciones, manejo de publicidad, juegos entre otros.

#### 4.1.1.1. Páginas en Facebook

Organizaciones o negocios pueden contar con páginas en Facebook, estas permiten entregar mensajes a cualquiera que las siga (Crookes, 2017).

Cualquier persona mediante un dispositivo inteligente y acceso a Internet puede crear una página en Facebook y publicar en ella, las personas que sigan dicha página pueden acceder a las publicaciones, en dado caso la página sea pública esta no requiere seguir la página, mientras que si es privada está requerirá previamente seguir la página para visualizar las publicaciones o contenido de esta.

Según Abram y Karasavas (2018), las páginas cuentan con una categoría, a continuación, son listadas:

- Negocio local o lugar
- Compañía, organización o institución
- Marca o producto
- Artista, banda o figura pública
- Entretenimiento
- Causa

#### 4.1.1.2. Publicaciones de Facebook

Las publicaciones en Facebook son usualmente textos que contienen links, fotos, videos e información de ubicación (Golbeck, 2015).

Usuarios y así como páginas pueden crear publicaciones y compartirlas públicamente o con grupo de usuarios específicos, en el caso de las páginas depende de la configuración propia, es decir, si es pública o privada.

### 4.1.2. Instagram

Instagram es una red social donde subir y compartir fotografías radica como su principal función, no de palabras (Zimmerman y Ng, 2019).

Es decir que en esta red social incluso podremos encontrar información incrustada en texto dentro de las fotografías cargadas, dado que esta red social está orientada a ser compatible con dispositivos móviles. Desde antes de ser adquirida por Facebook su crecimiento no se ha detenido, esta fue desarrollada el 6 de octubre del 2010 y adquirida por Facebook en abril del 2012 (Miles, 2019).

#### 4.1.2.1. Perfiles en Instagram

Los perfiles en Instagram son similares a los perfiles de Facebook, para ello es necesario abrir una cuenta en Instagram para iniciar con la creación de uno (Zimmerman y Ng, 2019).

Dentro del perfil es posible compartir fotografías y seguir otros perfiles, así como otros perfiles pueden seguir nuestra cuenta, los perfiles pueden ser otras personas o marcas (Zimmerman y Ng, 2019).

#### 4.1.2.2. Publicaciones en Instagram

Similar a las publicaciones de Facebook estas pueden contener *hashtags*, *likes* y comentarios, así como una o más imágenes asociadas diferenciándose por no tener textos pesados o de gran longitud dentro de la descripción de la publicación (Zimmerman y Ng, 2019).

### 4.1.3. *Social graph*

Es un gráfico que representa las relaciones entre individuos, este puede representar una red social (Arnaboldi, Passarella, Conti y Dunbar, 2015).

Por medio de está gráfica es posible visualizar todas las relaciones (*edges*) existentes entre los vértices (conocidos como individuos). Los grafos permiten representar un conjunto de entidades y sus relaciones mediante abstracciones matemáticas (Sakr y Pardede, 2011).

Los grafos están presentes en muchos conjuntos de datos, desde redes sociales en donde los grafos permiten representar las relaciones entre amigos, las rutas para viajes, así como los destinos, entre otros.

#### 4.1.3.1. Bases de datos orientadas a grafos

Existen bases de datos orientadas a grafos tales como Neo4J o HyperGraphDB. Estas bases de datos nacieron dado que el modelo relacional carece de soporte para estructuras de tipo árbol o grafo.

XML ha emergido como una solución que permite representar estructuras de tipo árbol o grafo, similar al formato JSON.

## 4.2. *Natural language Processing* (NLP)

El procesamiento del lenguaje natural (*natural language processing*) aprovecha herramientas, técnicas y algoritmos para procesar y comprender los datos basados en el lenguaje natural, estos generalmente no están estructurados (Sarkar, 2019).

Las publicaciones obtenidas de redes sociales cumplen con tener texto no estructurado, generalmente la mayor parte del contenido obtenido de Internet no se encuentra estructurado. El lenguaje natural es aquel que los humanos han desarrollado y evolucionado mientras lo practican durante el tiempo (Sarkar, 2019).

Los datos no estructurados, tales como texto, imágenes y videos contienen riqueza de información, la cual puede ser aprovechada para encontrar patrones y tomar decisiones correctas con base a los datos (Sarkar, 2019).

## 4.3. *Machine learning*

*Machine learning* permite tomar decisiones de una gran cantidad. Etaati (2019) define que en la actualidad ya no solo las industrias y empresas grandes tienen acceso a tecnologías para implementar *machine learning*, ahora es posible utilizar herramientas provistas por empresas tales como Microsoft.

Machine learning es solo una parte de lo que es inteligencia artificial (*Artificial intelligence*, en inglés). En la siguiente gráfica provista por Etaati (2019) se puede observar en donde se encuentra *machine learning* con respecto a otros subconjuntos de AI.

## 4.4. Clasificación utilizando el análisis *single-label*

*Single label* se da cuando los entrenamientos se representan con una simple etiqueta denominada λ de un conjunto disjunto de *labels* L (Tsoumakas, Katakis y Vlahavas, 2009).

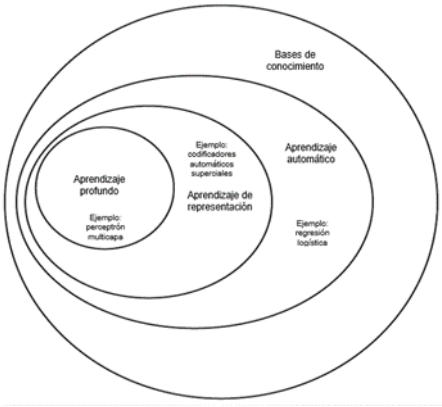
Este tipo de clasificación se da cuando se analiza un documento y se etiqueta a un solo tema, por ejemplo, las reacciones a una película religiosa se pueden etiquetar como religión.

## 4.5. Clasificación utilizando el análisis *multi-label*

El análisis *multi-label* es el más utilizado, este se define cuando se asocia a un documento múltiples etiquetas (Tsoumakas, Katakis y Vlahavas, 2009).

Continuando con el ejemplo anterior, con el análisis *multi-label* la película religiosa se puede etiquetar como religión y también como película.

Figura 2**. Subconjuntos de inteligencia artificial**




Fuente: Etaati (2019). *Machine Learning with Microsoft Technologies.*

Una definición clara provista por Etaati (2019) quien hace referencia a Raschka (2014), menciona que *machine learning* se define como un conjunto de herramientas con el fin de aplicarlos en datos y así encontrar un sentido mediante el uso de algoritmos.

Un algoritmo es una secuencia de instrucciones ordenadas que se ejecutan para transformar una entrada en una salida, es decir que un algoritmo comprende una serie de pasos que se ejecutan para obtener un resultado.

Con *machine learning* es posible encontrar sentido a los datos para obtener un mejor panorama, permitiendo entender lo que pasó, por qué pasó, qué pasará en un tiempo futuro, y cómo hacer que pase (Etaati, 2019).

*Machine learning* cuenta con dos diferentes enfoques, aprendizaje supervisado (*supervised learning*) y aprendizaje no supervisado (*unsupervised learning*).

### 4.5.1. *Supervised learning*

El aprendizaje supervisado busca predecir un valor o grupo por medio de datos de fechas anteriores (Etaati, 2019).

Es decir que es posible predecir, por ejemplo, la cantidad de alumnos que ingresarán a la universidad este año (valor) o cuando se analizan grupos, es posible predecir, por ejemplo, cuando un cliente de un banco se convertirá en una nueva categoría, por ejemplo, cliente tipo A.

### 4.5.2. *Unsupervised learning*

Con el Aprendizaje no supervisado ya se tiene una idea de la respuesta antes y al momento de estar creando y entrenando el modelo, el objetivo final es encontrar un patrón natural en los datos y así extraer información valiosa (Etaati, 2019).

Por medio de la clasificación o agrupamiento sin supervisión es posible resumir información para descubrir estructuras en los datos, permitiendo dividir un conjunto de objetos sin etiquetar en grupos o clasificaciones en donde todos los objetos que pertenecen a un grupo deben ser coherentes y homogéneos (Yang, 2016).

La clasificación de contenido en páginas web del perfil del usuario es un ejemplo de una clasificación no supervisada.

## 4.6. Modelado de tópicos

El modelado de tópicos extrae varios conceptos distintos o tópicos de un corpus grande que tiene varios tipos de documentos y cada documento tiene uno o más conceptos. Por conceptos se entiende como opiniones, hechos, correos y otros similares (Sarkar, 2019).

Es decir que entre estos conceptos se puede abarcar contenido de redes sociales, tales como publicaciones y comentarios, mediante técnicas matemáticas y estadísticas es posible encontrar estructuras semánticas latentes en un corpus (Sarkar, 2019).

En el modelado de tópicos se generan *clusters* o grupos de términos que son distintos uno de otro, estos *clusters* generan temas o conceptos (Sarkar, 2019).

Estos conceptos o temas formados previamente pueden utilizarse para conocer el tema principal de un corpus y al mismo tiempo encontrar conexiones semánticas entre las palabras frecuentes en varios documentos.

Existen 3 métodos de mayor uso que son:

- *Latent semantic indexing*
- *Latent dirichlet allocation*
- *Non-negative matrix factorization*

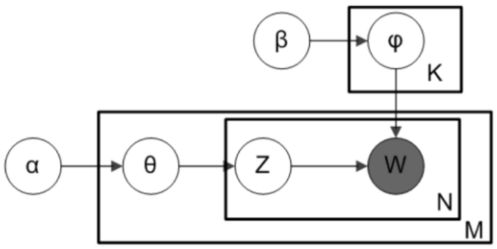
Se elige *latent dirchlet allocation* por el alto grado de utilización en el mercado, existiendo variedad de librerías para su uso en lenguajes como Python o R. *Non-negative matrix factorization* es el más reciente y efectivo, pero al ser reciente existen menos recursos con los cuales trabajar adecuadamente (Sarkar, 2019).

### 4.6.1. *Latent Dirichlet Allocation* (LDA)

Es un modelo generativo probabilístico (Sarkar, 2019), este inicialmente desarrollado por los genetistas en Reino Unido, esto debido a que necesitaban inferir la estructura de la población por medio de la secuencia de genes, luego fue popularizado por Stanford Researchers en 2003 (Lane, Howard y Hapke, 2019).

En LDA se asume que cada documento tiene una combinación de temas. A continuación, se observa una figura sobre la notación en LDA.

Figura 3. **Notación utilizada en LDA**




- K es el número de tópicos
- N es el número de palabras en el documento
- M es el número de documentos a analizar
- α es el parámetro de concentración Dirichlet anterior de la distribución del tópico por documento.
- β es el mismo parámetro de la distribución de palabras por documento
- ϕ(k) es la distribución de palabras para el tópico K
- Θ(i) es la distribución tópicos para el documento i
- z(i,j) es la asignación del tópico para w(i,j)
- w(i,j) es la palabra j en el documento i.
- ϕ y Θ son distribuciones Dirichlet, z y w son multinomios.

Fuente: Sarkar (2019). *Text Analytics with Python - A Practitioner's Guide to Natural Language Processing.*

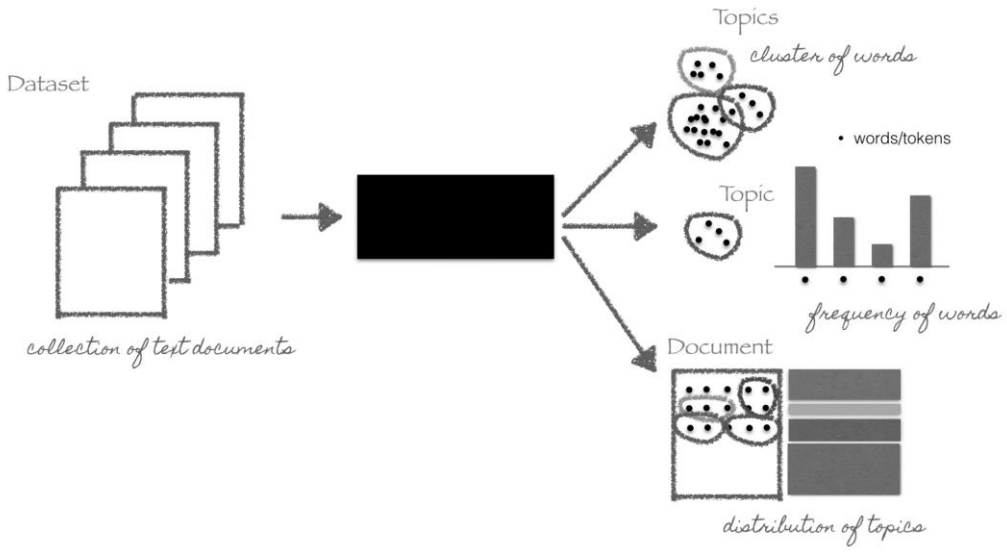
En la siguiente figura se observan los parámetros que necesita LDA, así como los documentos y palabras a analizar, el cuadro negro es el procedimiento como tal, es decir donde se realiza todo el proceso de LDA, la salida son clústeres de palabras, frecuencia de palabras y distribución de tópicos, así como una mezcla de tópicos para cada documento.

Sarkar (2019) expone los siguientes pasos generales que el procedimiento de LDA debe aplicar:

- Inicializar los parámetros (alfa y beta).
- Para cada documento, aleatoriamente inicializar cada palabra (W) para cada uno de los tópicos (K).
- Iniciar un proceso iterativo que incluya los siguientes incisos para cada documento (D), para cada palabra (W) del documento y para cada tópico (T):

    - Calcular P(T|D), que es la proporción de las palabras en un documento (D) asignadas a un tópico (T).
    - Calcular P(W|T), que es la proporción de asignaciones al tópico (T) sobre todos los documentos que tienen la palabra (W).
    - Reasignar la palabra (W) con el tópico (T) con probabilidad P(T|D) * P(W|T), considerando todas las otras palabras y sus asignaciones en los tópicos.

Figura 4. **Framework de LDA**




Fuente: Sarkar (2019). *Text Analytics with Python - A Practitioner's Guide to Natural Language Processing.*

En la figura 4 se observa todo el proceso de LDA, desde el ingreso de las colecciones de documentos textuales hasta la salida de los clústeres de palabras por tópico, los tópicos y la distribución de tópicos por documento.

## 4.7. *Serverless computing*

También conocido como funciones como servicio (FaaS) que permiten al proveedor de la nube administrar el contenedor donde la función se encuentre alojada, esto permite a los usuarios solo preocuparse por escribir el código, subirlo y ejecutarlo sin preocupaciones sobre la arquitectura (Stigler y Stigler, 2018).

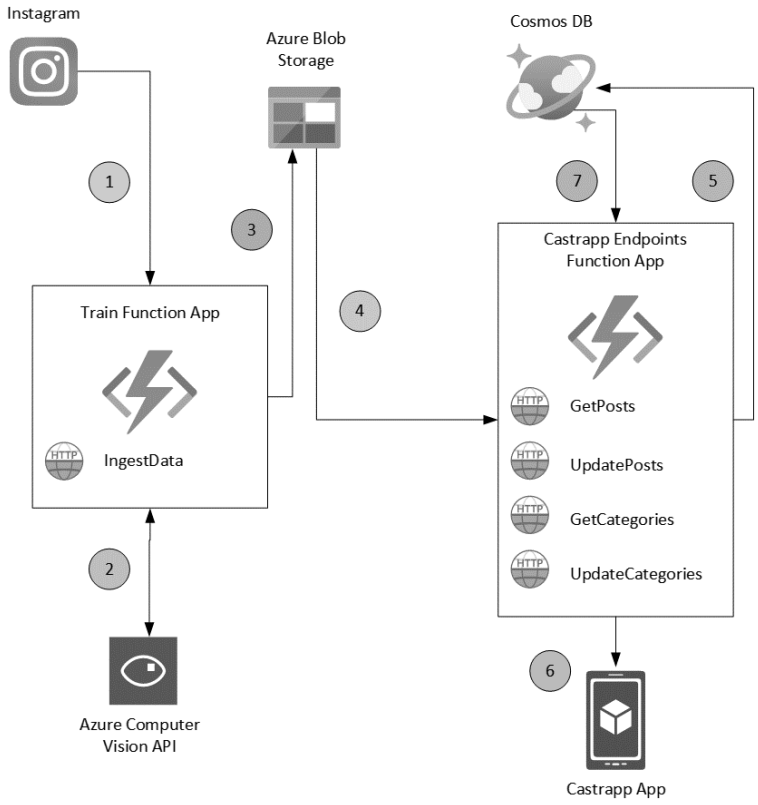
Por medio de *serverless computing* es posible soportar una gran carga de peticiones en dado caso la función sea solicitada por varios usuarios, es decir que esta reciba muchas peticiones a través de Internet.

# 5. PRESENTACIÓN DE RESULTADOS

## 5.1. Diseño y desarrollo de la arquitectura

Para soportar todo el análisis y entrenamientos para la generación del modelo LDA utilizando la librería *gensim* se diseñó e implementó la siguiente arquitectura detallada por medio de la figura 5.

Figura 5. **Diagrama de arquitectura**




Fuente: elaboración propia, empleando Function Castrapp Endpoints.

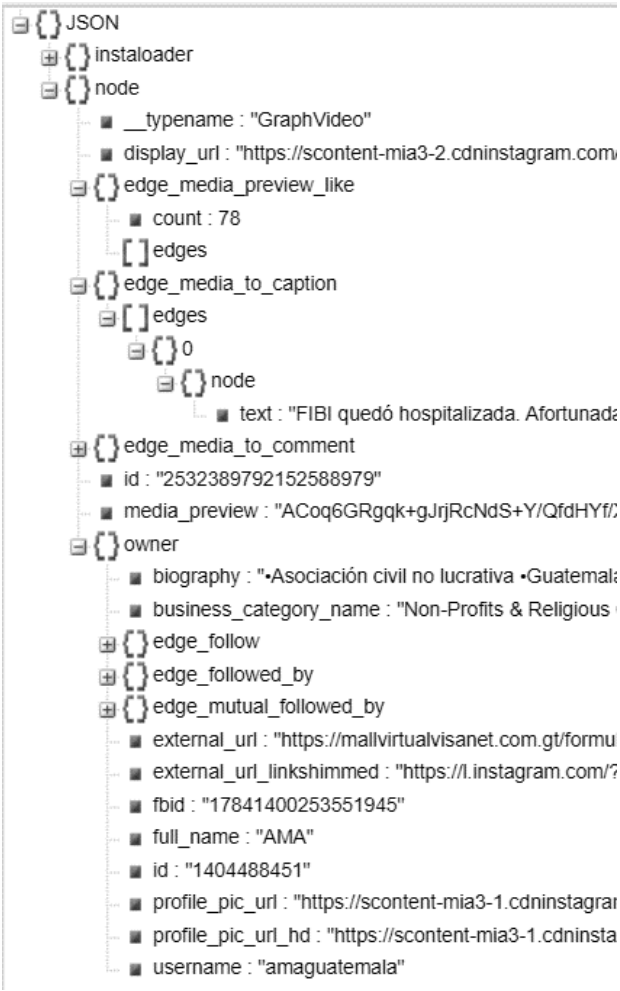
En la figura 5 se ejemplifica el proceso completo desde la obtención de los datos hasta la entrega de estos ya clasificados a la aplicación móvil. En el paso 1 los datos son obtenidos de Instagram mediante la librería *instaloader* mediante una Azure *Function* con el lenguaje Python. Dentro de la función nombrada *Train Function App* se realiza la transformación de las publicaciones, realizando limpieza de datos dentro de los cuales se realiza una eliminación de palabras de conexión conocidas como *stop words* así como la extracción del texto de las imágenes mediante Azure *Computer Vision* API en el paso 2.

Una vez el texto es transformado se realizan los entrenamientos del modelo LDA utilizando la librería *gensim* y el modelo final es almacenado en *Blob Storage* de Azure en el paso 3. La *Azure Function Castrapp Endpoints* en Python es la encargada de leer el modelo LDA desde el *Blob Storage* y realizar las reclasificaciones de publicaciones periódicamente y almacenarlas en la base de datos *Cosmos DB*, así como también las categorías, esta aplicación también se encarga de servir las categorías y publicaciones a la aplicación móvil Castrapp por medio de sus *endpoints*.

## 5.2. Estructura de los datos obtenidos de Instagram

Los datos obtenidos de Instagram mediante la librería *instaloader* para Python tienen la siguiente estructura.

Figura 6. **Estructura de datos obtenida de Instagram**




Fuente: elaboración propia, empleando Jsonviewer.

Algunos de los campos de la estructura de datos fueron omitidos para fines representativos. El campo utilizado para el entrenamiento del modelo está ubicado en *text* dentro de node en la posición 0 del arreglo *edges* del objeto *edge_media_top_caption*.

## 5.3. Entrenamientos del modelo LDA

Para el entrenamiento del modelo se realizaron diferentes corridas con diferente cantidad de datos y con parámetros variables. A continuación, se describen los ejemplos de algunos entrenamientos y se explica cada uno de los resultados obtenidos.

### 5.3.1. Palabras de conexión

Para realizar el correcto conocimiento de tópicos es necesario eliminar palabras de conexión o *stop words* en inglés, en el caso de las publicaciones animalistas se detectaron adicional a las palabras de conexión de la librería *nltk.corpus* las siguientes:

Tabla III. **Palabras de conexión adicionales utilizados**

| Stop words adicionales |
|---|
| 'de', 'es', 'estan', 'estamos', 'con', 'en', 'un', 'de', 'que', 'para', 'los', 'las', 'no', 'por', 'para', 'si', 'mas', 'hoy', 'gracias', 'deben', 'llegar', 'favor', 'com', 'seguir', 'nombre', 'ama', 'ser', 'gmail', 'puedes', 'poder', 'haber', 'hacer', 'deber', 'tener', 'contar', 'cuenta', 'dia', 'hora', 'dar', 'zona', 'domingo' |

Fuente: elaboración propia.

### 5.3.2. Extracción de texto de la imagen

Para extraer texto de una imagen se utiliza el servicio de Microsoft *Vision* en Azure.

La figura 7 es enviada para análisis y el texto recibido del API es: *Triple felina rabia leucemia vacunación felina 4* Domingo Abril Agenda al correo comunidadgatunagt@gmail.com.

Figura 7. **Imagen para análisis con Microsoft Vision Service**


Fuente: Gatuna (2021). *Vacunación gatuna*.

### 5.3.3. Perplejidad, puntaje de coherencia, rho y *topic diff*

Puntaje de coherencia (*coherence score*) mide la distancia relativa entre las palabras en un tópico, midiendo el grado de similitud semántica entre las palabras de puntuación alta (probabilidad) en el tópico, y perplejidad (*perplexity*) es una medida de que tan bien un modelo de probabilidad predice una muestra.

*Topic diff* determina cuánto cambiaron los temas después de la iteración actual y rho es la velocidad de actualización, es decir rho controla cuánto afecta el resultado el nuevo set de datos. Cuando ya no existen más cambios significantes entre cada iteración significa que el modelo ha convergido.

Los valores para el puntaje de coherencia (*coherence score*) son de 0.0 a 1.0, la tabla IV detalla el significado de cada valor.

Tabla IV. **Rango para puntajes de coherencia**

| **Puntaje de coherencia** | **Significado** |
|---|---|
| 0.3 o menor | Malo |
| 0.4 | Bajo |
| 0.55 | Bien, pero puede mejorar |
| 0.65 | Muy bueno |
| 0.7 | Modelo bastante acertado |
| 0.8 | Poco probable |
| 0.9 | Posiblemente incorrecto |

Fuente: elaboración propia, empleando Microsoft Excel.

Perplejidad por palabra o *perplexity per word* puede tomar los valores por lo general negativos y es posible analizarla con datos de entrenamiento y de prueba. Dentro de la librería *gensim* se específica que entre más pequeño sea este valor mejor será el modelo.

### 5.3.4. Entrenamiento con 3507 publicaciones

Para tener el 50 % de las publicaciones se abarca de la fecha 12/05/2019 hasta 30/11/2020 con un total de 3507 publicaciones. Se utilizan 1500 iteraciones y 50 pasadas, parámetros propios de la librería *gensim*.

Tabla V. **Resultados de entrenamiento con 3507 publicaciones**

| **Número** | **Tópico** | **Palabra** | **P** |
|---|---|---|---|
| 0 | 0 | Banco | 0.026627 |
| 1 | 0 | Deposito | 0.019028 |
| 2 | 0 | Industrial | 0.015912 |
| 3 | 0 | Animales | 0.014686 |
| 4 | 0 | Numero | 0.014541 |
| 5 | 1 | Hogar | 0.008318 |
| 6 | 1 | Adoptar | 0.008097 |
| 7 | 1 | Ayudar | 0.007988 |
| 8 | 1 | Gatito | 0.007945 |
| 9 | 1 | Ayuda | 0.007669 |
| 10 | 2 | Gato | 0.020616 |
| 11 | 2 | Castración | 0.01967 |
| 12 | 2 | Gatito | 0.018797 |
| 13 | 2 | Seguro | 0.011163 |
| 14 | 2 | Salud | 0.01097 |

Fuente: elaboración propia, empleando Microsoft Excel.

Se definieron 3 tópicos como resultado del modelo, dentro de esos tópicos en la tabla VI se eligieron las 5 palabras más representativas, es decir con porcentaje más alto, que representan a cada tópico.

En la tabla VI se observa cómo *topic diff* y *rho* van variando durante el entrenamiento del modelo LDA con 3507 publicaciones.

Tabla VI. **Extracto de *topic diff* y rho con 3507 publicaciones**

| |
|---|
| topic diff=0.061682, rho=0.147839 |
| topic diff=0.061292, rho=0.146249 |
| topic diff=0.060934, rho=0.146249 |
| topic diff=0.060559, rho=0.144710 |
| topic diff=0.060216, rho=0.144710 |
| topic diff=0.059855, rho=0.143218 |
| topic diff=0.059531, rho=0.143218 |
| topic diff=0.059175, rho=0.141771 |
| topic diff=0.058907, rho=0.141771 |
| topic diff=0.058517, rho=0.140368 |
| topic diff=0.058235, rho=0.140368 |
| topic diff=0.057882, rho=0.139005 |
| topic diff=0.057593, rho=0.139005 |

Fuente: elaboración propia, empleando Microsoft Excel.

Tabla VII. ***Perplexity* y *coherence score* para modelo con 3507 publicaciones**

| **Variable** | **Resultado** |
|---|---|
| *Perplexity* por palabra | -7.104350832527324 |
| *Coherence score* | 0.7481032509190125 |

Fuente: elaboración propia, empleando Microsoft Excel.

### 5.3.5. Entrenamientos con 7013 publicaciones

Para tener el 100 % de los datos se abarca del 01/01/2015 al 30/11/2020, aproximadamente 5 años de datos obtenidos de los grupos mencionados anteriormente en Instagram. Para fines de explicación cada modelo será identificado por una versión. Para cada una de las siguientes versiones se utilizan 1500 iteraciones y 50 pasadas, parámetros propios de la librería *gensim*.

#### 5.3.5.1. Versión 1 del modelo

Se inicia retirnando las siguientes palabras de conexión iniciales:

- De
- Es
- Están
- Estamos
- Con
- En
- Un
- De
- Que
- Para
- Los
- Las
- No
- Por

- Para
- Si
- Mas
- Hoy
- Gracias
- Deben
- Llegar
- Favor
- Com
- Seguir
- Nombre
- Ama
- Ser
- Gmail
- Puedes
- Poder
- Haber
- Hacer
- Deber
- Tener
- Contar
- Cuenta
- Día
- Hora
- Dar
- Zona
- Domingo

Estas surgen al realizar un muestreo de las palabras que podrían llegar a afectar en las publicaciones. La versión 1 se muestra a continuación en la tabla VIII.

Tabla VIII. **Tópicos de entrenamiento del modelo versión 1**

| **Número** | **Tópico** | **Palabra** | **P** |
|---|---|---|---|
| 0 | 0 | Banco | 0.025769815 |
| 1 | 0 | Depósito | 0.01907579 |
| 2 | 0 | Amigos | 0.017300244 |
| 3 | 0 | Asociación | 0.016385593 |
| 4 | 0 | Animales | 0.01606939 |
| 5 | 1 | Gatito | 0.022093423 |
| 6 | 1 | Gato | 0.019296741 |
| 7 | 1 | Castración | 0.015024771 |
| 8 | 1 | Gatuna | 0.012023885 |
| 9 | 1 | Salud | 0.009766747 |
| 10 | 2 | Vida | 0.011559698 |
| 11 | 2 | Ayudar | 0.009502799 |
| 12 | 2 | Hogar | 0.008660078 |
| 13 | 2 | Adoptar | 0.007688133 |
| 14 | 2 | Perrito | 0.006554659 |

Fuente: elaboración propia, empleando Microsoft Excel.

Se definieron de igual manera 3 tópicos y se seleccionaron las 5 palabras más representativas.

A continuación, en la tabla IX se muestran los resultados de *topic diff* y *rho* para esta iteración con el 100 % de los datos, con un total de 7013 publicaciones.

Tabla IX. **Extracto de *topic diff* y rho del modelo versión 1**

| |
|---|
| topic diff=0.097403, rho=0.140710 |
| topic diff=0.134898, rho=0.139338 |
| topic diff=0.097041, rho=0.139338 |
| topic diff=0.109521, rho=0.139338 |
| topic diff=0.096402, rho=0.139338 |
| topic diff=0.133463, rho=0.138005 |
| topic diff=0.096034, rho=0.138005 |
| topic diff=0.108374, rho=0.138005 |
| topic diff=0.095427, rho=0.138005 |
| topic diff=0.132078, rho=0.136709 |
| topic diff=0.095057, rho=0.136709 |
| topic diff=0.107265, rho=0.136709 |
| topic diff=0.094483, rho=0.136709 |

Fuente: elaboración propia, empleando Microsoft Excel.

Tabla X. **Perplejidad y puntaje de coherencia del modelo versión 1**

| **Variable** | **Resultado** |
|---|---|
| Perplexity por palabra | -7.3583343608152765 |
| Coherence Score | 0.7572291872920296 |

Fuente: elaboración propia, empleando Microsoft Excel.

#### 5.3.5.2. Versión 2 del modelo

Se retiraron 3 palabras más del listado de palabras de conexión: gatito, ayuda y ayudar y con ello se generó la versión 2 del modelo.

Tabla XI**.** ***Perplexity* y *coherence score* del modelo versión 2**

| **Variable** | **Resultado** |
|---|---|
| Perplexity por palabra | -7.392534138227789 |
| Coherence Score | 0.8055956667395899 |

Fuente: elaboración propia, empleando Microsoft Excel.

Los tópicos encontrados son los siguientes:

Tabla XII. **Tópicos de entrenamiento con 7013 publicaciones versión 2**

| **Número** | **Tópico** | **Palabra** | **P** |
|---|---|---|---|
| 0 | 0 | Vida | 0.01165 |
| 1 | 0 | Hogar | 0.008623 |
| 2 | 0 | Adoptar | 0.007609 |
| 3 | 0 | Perrito | 0.006107 |
| 4 | 0 | Ir | 0.005764 |
| 5 | 1 | Gato | 0.020137 |
| 6 | 1 | Castración | 0.016063 |
| 7 | 1 | Gatuna | 0.010536 |
| 8 | 1 | Cirugía | 0.009544 |
| 9 | 1 | Salud | 0.009372 |
| 10 | 2 | Banco | 0.026248 |
| 11 | 2 | Depósito | 0.01943 |
| 12 | 2 | Asociación | 0.01669 |
| 13 | 2 | Amigos | 0.016667 |
| 14 | 2 | Animales | 0.016368 |

Fuente: elaboración propia, empleando Microsoft Excel.

#### 5.3.5.3. Versión 3 del modelo

Se retiran las palabras necesitar, ir, ver y solo. Los valores para *perplexity* y *coherence score* son mostrados en la tabla XIII.

Tabla XIII. ***Perplexity* y *coherence score* del modelo versión 3**

| **Variable** | **Resultado** |
|---|---|
| *Perplexity* por palabra | -7.403904165272241 |
| *Coherence score* | 0.7507037250120788 |

Fuente: elaboración propia, empleando Microsoft Excel.

Los siguientes tópicos son el resultado del entrenamiento:

Tabla XIV. **Tópicos de entrenamiento del modelo versión 3**

| **Número** | **Tópico** | **Palabra** | **P** |
|---|---|---|---|
| 0 | 0 | Banco | 0.025951 |
| 1 | 0 | Depósito | 0.01921 |
| 2 | 0 | Amigos | 0.01744 |
| 3 | 0 | Asociación | 0.016501 |
| 4 | 0 | Animales | 0.016182 |
| 5 | 1 | Vida | 0.012195 |
| 6 | 1 | Hogar | 0.009065 |
| 7 | 1 | Adoptar | 0.008025 |
| 8 | 1 | Perrito | 0.00641 |

Continuación tabla XIV.

| | | | |
|---|---|---|---|
| 9 | 1 | Familia | 0.005397 |
| 10 | 2 | Gato | 0.019737 |
| 11 | 2 | Castración | 0.015417 |
| 12 | 2 | Gatuna | 0.011458 |
| 13 | 2 | Salud | 0.009952 |
| 14 | 2 | Cirugia | 0.00933 |

Fuente: elaboración propia, empleando Microsoft Excel.

#### 5.3.5.4. Versión 4 del modelo

Se agregan perrito, Guatemala, querer y esperar a la lista de palabras de conexión no deseadas.

Tabla XV. **Perplexity y Coherence score del modelo versión 4**

| **Variable** | **Resultado** |
|---|---|
| Perplexity por palabra | -7.422213377919046 |
| Coherence Score | 0.7525846808293911 |

Fuente: elaboración propia, empleando Microsoft Excel.

En la tabla XVI se describen los tópicos para el modelo versión 4.

Tabla XVI. **Tópicos de entrenamiento del modelo versión 4**

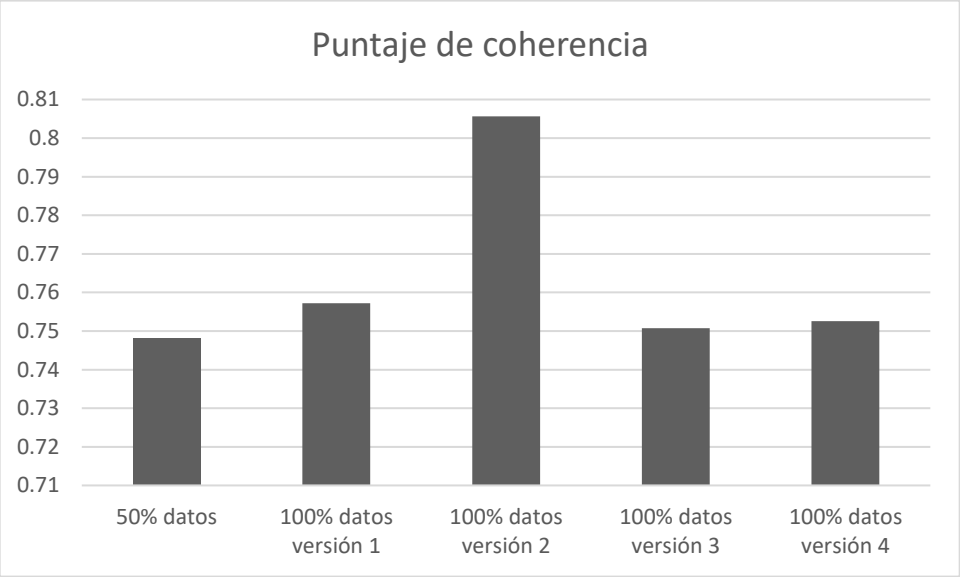
| Número | Tópico | Palabra | P |
|---|---|---|---|
| 0 | 0 | Vida | 0.013655341 |
| 1 | 0 | Hogar | 0.010575749 |
| 2 | 0 | Adoptar | 0.009109301 |
| 3 | 0 | Familia | 0.005524444 |
| 4 | 0 | Información | 0.005477937 |
| 5 | 1 | Banco | 0.018374905 |
| 6 | 1 | Depósito | 0.013601415 |
| 7 | 1 | Amigos | 0.011694854 |
| 8 | 1 | Asociación | 0.011682382 |
| 9 | 1 | Animales | 0.011457458 |
| 10 | 2 | Gato | 0.010470577 |
| 11 | 2 | Castración | 0.007447002 |
| 12 | 2 | Animal | 0.007098621 |
| 13 | 2 | Gatuna | 0.006412162 |
| 14 | 2 | Gatuno | 0.006066528 |

Fuente: elaboración propia, empleando Microsoft Excel.

### 5.3.6. Comparación entre las diferentes versiones del modelo

En la figura 8 en el eje Y se encuentra el rango de valores para el puntaje de coherencia y en el eje X los distintos entrenamientos realizados, iniciando por el entrenamiento con 3507 publicaciones (50 % de los datos) y continuando con las distintas versiones utilizando 7013 publicaciones (100 % de los datos). Se observa que el modelo con versión 3 tiene el puntaje de coherencia más cercano a 0.7.

Figura 8. **Puntaje de coherencia de los modelos entrenados**

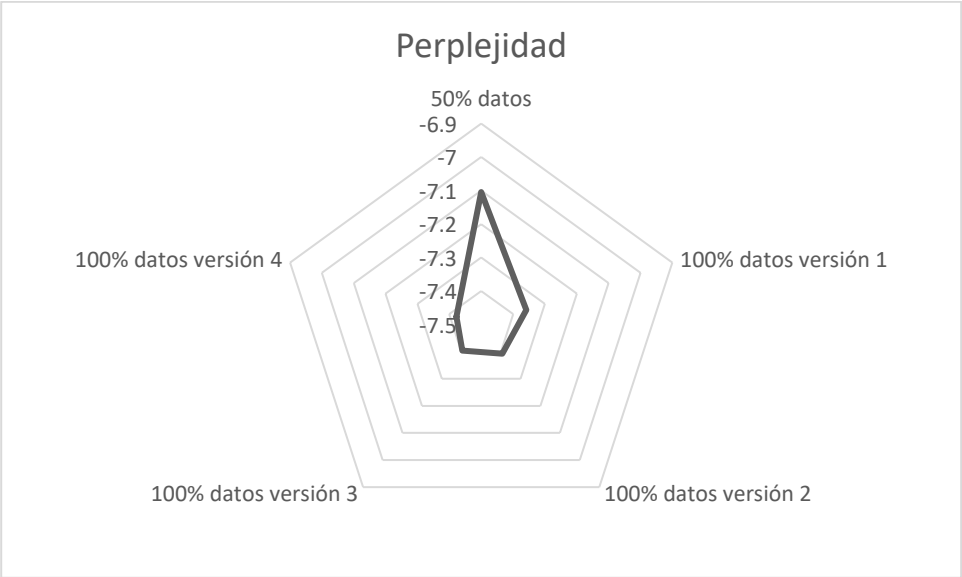



Fuente: elaboración propia, empleando Microsoft Excel.

De la misma manera se grafican los valores obtenidos para la perplejidad (*perplexity*) en la figura 9.

En la figura 9 se muestra una gráfica de radar, el rango muestra los valores que tomó la perplejidad durante los entrenamientos para cada uno de los entrenamientos y modelos generados, iniciando con el cálculo con 3507 publicaciones (50 % de los datos) y luego con las diferentes versiones con 7013 publicaciones (100 % de los datos).

Figura 9. **Gráfica de radar para perplejidad**




Fuente: elaboración propia, empleando Microsoft Excel.

También se analiza la perplejidad para datos del año 2021, específicamente los meses enero, febrero y marzo con un total de 250 publicaciones, estos se detallan en la tabla XVII.

Tabla XVII. **Perplejidad para datos de prueba de los meses enero, febrero y marzo del 2021**

| **Variable** | **Resultado** |
|---|---|
| Perplexity por palabra | -11.22728172256678 |

Fuente: elaboración propia, empleando Microsoft Excel.

## 5.4. Clasificación utilizando el modelo LDA final generado

Una publicación referente a jornada de castración es analizada con el modelo LDA versión 3 generado en el entrenamiento con el 100 % de los datos, dado que este tiene el mejor resultado para el puntaje de coherencia. El texto de la publicación detalla que se realizará una castración el 11 de abril del 2021 dirigida a gatitos con confirmación de asistencia previa a un precio de Q. 175.00, también especifican los cuidados que se deben tener previo a la jornada.

Los resultados de la clasificación son los mostrados en la tabla XVIII, en donde en la primera columna aparece el número de tópico y en la segunda el porcentaje de pertenencia a dicho tópico.

Tabla XVIII. **Resultados de clasificación utilizando LDA**

| **Tópico** | **Porcentaje (P)** |
|---|---|
| 0 (vida, hogar, adoptar, familia, información) | 0.23622571 |
| 1 (banco, depósito, amigos, asociación, animales) | 0.32264173 |
| 2 (gato, castración, animal, gatuna, gatuno) | 0.44113258 |

Fuente: elaboración propia, empleando Microsoft Excel.

Los resultados de la clasificación utilizando el algoritmo anterior de la aplicación Castrapp se detalla en la tabla XIX, cabe mencionar que en este algoritmo los tópicos o temas ya están definidos en tres: castración, vacunación y adopciones.

Tabla XIX. **Resultados de clasificación utilizando algoritmo anterior**

| **Tópico** | **Porcentaje (P)** |
|---|---|
| Castración | 1.00 |
| Vacunación | 0.00 |
| Adopción | 0.00 |

Fuente: elaboración propia, empleando Microsoft Excel.

Una publicación relacionada a vacunación felina publicada por el grupo Comunidad Gatuna en Instagram es analizada con el modelo LDA generado. El texto de la publicación especifica que se realizará una jornada de vacunación el 7 de marzo del 2021 en zona 2 de la ciudad de Guatemala, se especifican también los paquetes y las vacunas que contienen, así como los precios y el correo para confirmar asistencia.

Los resultados de la clasificación son mostrados a continuación en la tabla XX utilizando el modelo LDA, y en la tabla XXI el algoritmo anterior de la aplicación Castrapp.

Tabla XX. **Resultados de clasificación para segunda publicación**

| **Tópico** | **Porcentaje (P)** |
|---|---|
| 0 (vida, hogar, adoptar, familia, información) | 0.3412825 |
| 1 (banco, depósito, amigos, asociación, animales) | 0.60074496 |
| 2 (gato, castración, animal, gatuna, gatuno) | 0.05797259 |

Fuente: elaboración propia, empleando Microsoft Excel.

Tabla XXI. **Resultados de clasificación utilizando algoritmo anterior de Castrapp**

| **Tópico** | **Porcentaje (P)** |
|---|---|
| Castración | 0.00 |
| Vacunación | 1.00 |
| Adopción | 0.00 |

Fuente: elaboración propia, empleando Microsoft Excel.

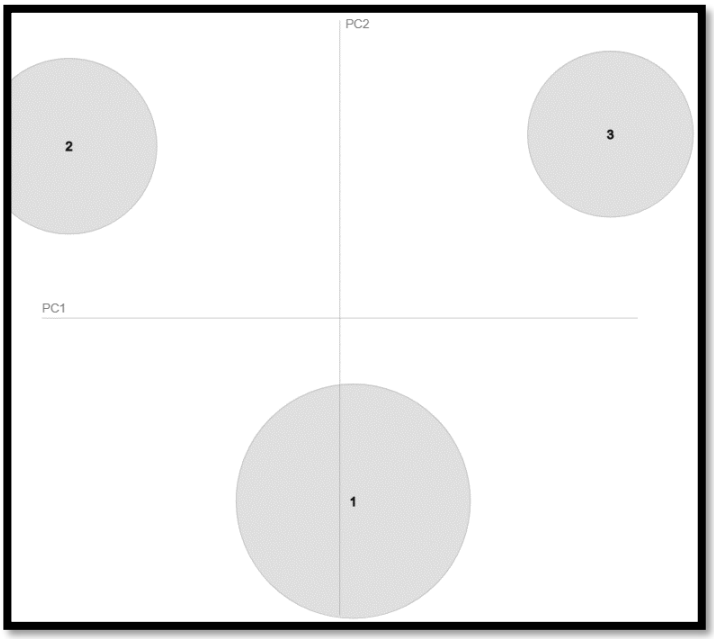
El nuevo algoritmo tuvo un tiempo de carga de 1.0 segundos en promedio en recuperar las publicaciones de la base de datos.

El algoritmo anterior de Castrapp realiza una clasificación interna en el dispositivo al recibir las publicaciones, el tiempo de recuperación fue de 2.1 segundos en promedio.

### 5.4.1. Mapa de distancia inter tópicos del modelo final

A continuación, en la figura 10 se grafica el *Intertopic Distance Map*. Esta gráfica nos permite identificar cuán alejados están los tópicos o si ellos tienen palabras en común que hagan que exista un traslape entre tópicos y los resultados puedan ser inciertos debido a esto. En la figura 10 se observa que los tópicos no tienen ningún traslape, es decir no comparten palabras en común que puedan afectar los resultados.

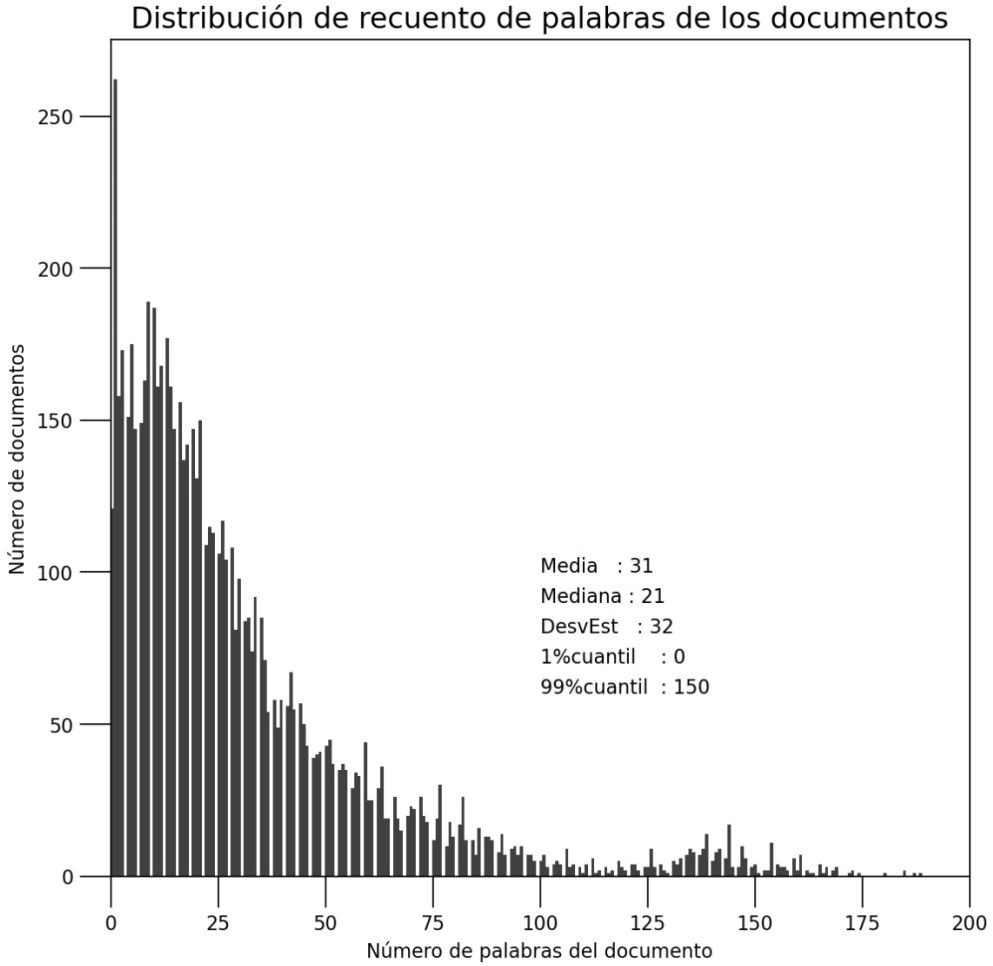
Figura 10**.** **Mapa de distancia inter tópicos (usando escalado multidimensional)**




Fuente: elaboración propia, empleando Python.

En la figura 11 se muestra la distribución de recuento de palabras del documento, es decir de las publicaciones.

Figura 11. **Distribución de recuento de palabras de las publicaciones**



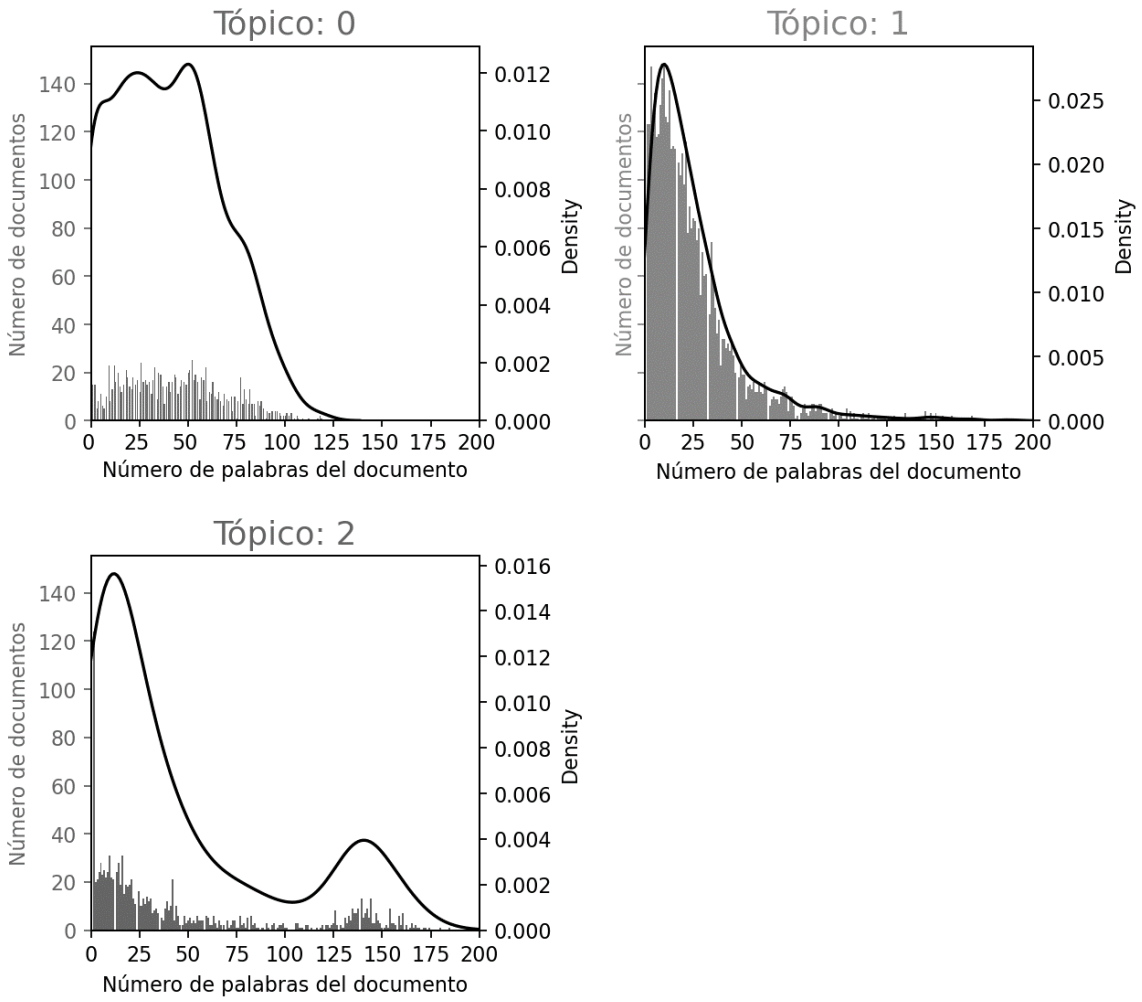

Fuente: elaboración propia, empleando Python.

Con base a la figura 11 se observa que 31 palabras es la media que existe en las publicaciones analizadas, 21 es la mediana. La desviación estándar es de 32 palabras entre las publicaciones.

En la figura 12 se observa una gráfica similar pero separada por tópico, en ella se puede observar que el tópico 1 es quien muestra una tendencia similar a

la figura 12, el tópico 0 y tópico 1 muestran por su parte una menor cantidad de palabras en los documentos.

Figura 12. **Distribución de recuento de palabras de los documentos por tópico dominante**




Fuente: elaboración propia, empleando Python.

## 5.5. Diferencias entre las versiones de cada modelo generado

En la tabla XXII se detallan los valores encontrados para cada uno de los modelos LDA generados.

Tabla XXII. **Versiones de los modelos LDA**

| | Versión 1 | Versión 2 | Versión 3 | Versión 4 |
|---|---|---|---|---|
| Perplexity por palabra | -7.3583 | -7.3925 | -7.4039 | -7.4222 |
| Coherence score | 0.7572 | 0.8056 | 0.7507 | 0.7526 |

Fuente: elaboración propia, empleando Microsoft Excel.

Cada una de las versiones varían por la cantidad de palabras de conexión removidas, estas fueron seleccionadas con base a la mayor probabilidad de aparición en cada entrenamiento.

## 5.6. Criterios de selección del modelo

Dos variables fueron utilizadas para seleccionar y evaluar el mejor modelo, adicional de las pruebas de clasificación realizadas con las publicaciones y el criterio humano del investigador.

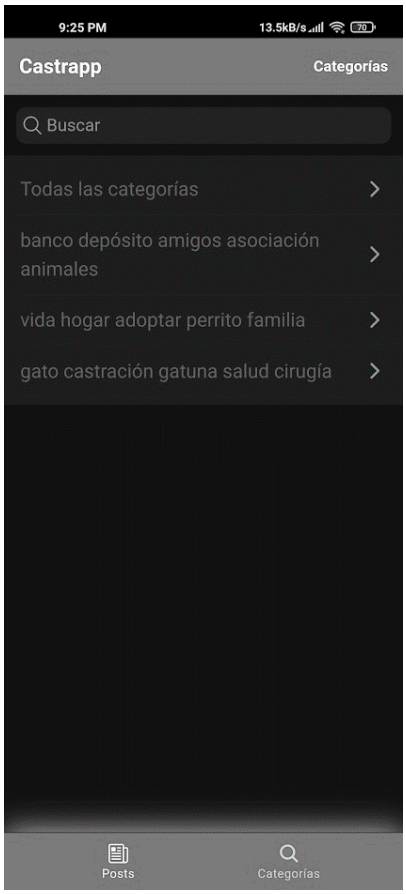
La primera variable fue *perplexity* o perplejidad, esta se calculó primero con los datos de entrenamiento y luego con los datos de prueba, la segunda fue el puntaje de coherencia.

## 5.7. Integración con la aplicación Castrapp

Se desarrolló una aplicación móvil para Castrapp, en donde se incluyeron los *endpoints* desarrollados. Algunas de las capturas son mostradas a continuación.

La sección de categorías fue realizada con base a los tópicos generados, estos se muestran en la Figura 13.

Figura 13. **Categorías en la aplicación Castrapp**




Fuente: elaboración propia, empleando Castrapp.

La sección de publicaciones es mostrada de acuerdo con la clasificación obtenida al aplicar el modelo LDA final. En la se muestra una publicación en la aplicación Castrapp.

Figura 14. **Publicación en la aplicación Castrapp**


Fuente: elaboración propia, empleando Castrapp.

# 6. DISCUSIÓN DE RESULTADOS

## 6.1. Análisis de los tópicos del modelo LDA

En la tabla XXIII se muestran los tópicos junto a las palabras más representativas con el porcentaje de probabilidad, además se agrega una columna más de inferencia humana realizada por el investigador, es decir asumir a cuál tema pertenece el conjunto de tópicos.

Para el tópico 0 se asume que trata de adopciones para hogares o familias dando una mejor vida a las mascotas. Este tema se relaciona con las adopciones de mascotas.

Para el tópico 1 se asume que se trata de asociaciones sobre animales, estás crean publicaciones sobre depósitos bancarios muchas veces para pedir ayuda o en la publicación de jornadas de castraciones o vacunaciones describen el precio y la forma de pago de la jornada.

Por último, el tópico 2 se relaciona a castraciones de gatos, estas implican cirugías. Este tema es el equivalente a castraciones de animales en el algoritmo anterior de la aplicación Castrapp.

Tabla XXIII. **Análisis de los tópicos**

| Tópico | Palabra | Probabilidad | Tema relacionado |
|---|---|---|---|
| 0 | Banco | 0.025951 | Depósito de bancos en las asociaciones |
| 0 | Depósito | 0.01921 | |
| 0 | Amigos | 0.01744 | |
| 0 | Asociación | 0.016501 | |
| 0 | Animales | 0.016182 | |
| 1 | Vida | 0.012195 | Adopciones para un hogar y vida |
| 1 | Hogar | 0.009065 | |
| 1 | Adoptar | 0.008025 | |
| 1 | Perrito | 0.00641 | |
| 1 | Familia | 0.005397 | |
| 2 | Gato | 0.019737 | Castración que implican cirugías para gatos |
| 2 | Castración | 0.015417 | |
| 2 | Gatuna | 0.011458 | |
| 2 | Salud | 0.009952 | |
| 2 | Cirugía | 0.00933 | |

Fuente: elaboración propia, empleando Microsoft Excel.

### 6.1.1. Análisis de la publicación sobre castraciones de mascotas

Una publicación relacionada a castración de gatos publicada por el grupo Comunidad Gatuna, al analizarla con el algoritmo se obtuvo que el tema dominante fue el relacionado a castración de gatos (tópico 3) con un 44.113258 %. Aquí se puede observar que el modelo LDA si está clasificando correctamente las publicaciones para castraciones. Por su parte los tópicos 0 y 1 cuentan con

23.622571 % y 32.264173 % respectivamente. Esto debido a que en las publicaciones de castraciones por lo general se habla de la forma de pago (tópico 0) y las ventajas en la vida de los animales (tópico 1).

### 6.1.2. Análisis de la publicación de vacunación de mascotas

Una publicación relacionada a una jornada de vacunación fue clasificada con el modelo LDA, los resultados marcan al tópico 0 como dominante referida a depósito de banco en las asociaciones con un 60.074496 % de probabilidad, el resultado se da porque en las descripciones de las jornadas de vacunaciones se describen varios tipos de paquetes con diferentes precios.

## 6.2. Interpretación de los tópicos y sus palabras asociadas

La interpretación y validación de estos tópicos fue realizada por el investigador, pero requeriría una investigación en donde la interpretación humana de varias personas se viera relacionada. Los autores Chang, Boyd-Graber, Wang, Gerrish y Blei (2009) realizaron una investigación en donde evaluaron los tópicos de dos formas:

- Palabra intrusa
- Tópico intruso

Por medio de estas dos técnicas preguntaron a varias personas a detectar la palabra intrusa en un tópico y la segunda si un documento hacia sentido con un tópico en específico, con ello lograron validar sus modelos LDA.

## 6.3. Rendimiento del nuevo algoritmo

El nuevo algoritmo utilizando el modelo LDA presenta una mejora en cuanto a tiempo de respuesta en la obtención de publicaciones, disminuyendo en 1.1 segundos el tiempo de respuesta, dado que las publicaciones se encuentran preclasificadas en la base de datos, como contraparte, dicha reclasificación no se realiza en tiempo real por lo que las publicaciones más nuevas estarán disponibles en la siguiente reclasificación.

## 6.4. Efectividad del nuevo algoritmo de clasificación

Los valores óptimos del modelo LDA según la tabla III para el puntaje de coherencia y comparando los resultados obtenidos del entrenamiento, muestra que el modelo LDA versión 3 tiene una efectividad correcta al momento de clasificar una publicación de tipo animalista, para esta versión del modelo LDA se tiene un puntaje de Coherencia de 0.7507, siendo la versión 3 la mejor entre todas las versiones del modelo por ser más cercana a 0.7.

De igual manera para la versión 3 del modelo LDA el valor de perplejidad con los datos de prueba es negativo y menor que la perplejidad calculada con los datos de entrenamiento, demostrando que el modelo versión 3 tiene un correcto rendimiento al clasificar documentos.

## 6.5. Trabajos futuros

En una investigación futura se deben incluir más grupos animalistas al listado de grupos, también explorar otro tipo de algoritmos *de machine learning* de la rama de supervisados, y así puede guiar a los algoritmos a los resultados requeridos o esperados.

# CONCLUSIONES

1. Se identificó e implementó el algoritmo de machine learning, latent dirichlet allocation (LDA) para el minado de texto y la obtención de tópicos de las publicaciones animalistas obtenidas de Instagram para su posterior clasificación acorde a los tópicos encontrados en la aplicación Castrapp por medio de servicios en la nube.

2. Para la extracción de texto de las imágenes se identificó e implementó el servicio en la nube Microsoft Vision provisto como parte de las soluciones de Microsoft Azure, el cual forma parte del algoritmo de entrenamiento para la aplicación Castrapp.

3. Se determinaron las métricas para la validación del algoritmo en la aplicación Castrapp dividiéndose en dos; las provistas por la librería *gensim* siendo estas el puntaje de coherencia con 0.7507 que indica un modelo acertado y la perplejidad con -7.4039 del entrenamiento y -11.2273 en las pruebas cuyos valores muestran un buen rendimiento en la clasificación de tópicos; la segunda métrica fue de carácter intuitivo e inferencia humana, es decir encontrar sentido lógico al conjunto de palabras de cada tópico e inferir un tema con base a esas palabras por el investigador.

4. Por medio de *latent dirchlet allocation* fue posible implementar un método de minado de texto al algoritmo de clasificación de la aplicación Castrapp permitiendo optimizar el proceso de obtención, clasificación, almacenamiento y presentación de las publicaciones animalistas

obtenidas de grupos en Instagram, con ello se redujo el tiempo de respuesta en la obtención de publicaciones en un 47.62 % (1.1 segundos).

# RECOMENDACIONES

1. Las palabras diminutivas o sinónimos debieran ser removidas a nivel de todas las publicaciones previo al entrenamiento, para evitar contaminación de palabras similares o con el mismo significado en los tópicos.

2. Para los entrenamientos del modelo LDA es necesario adquirir un equipo dedicado a ello, por lo que se recomienda migrar las funciones de entrenamiento a algún servicio en la nube, por ejemplo, Microsoft *machine learning Studio* o en su defecto un servidor de alto rendimiento en procesamiento.

3. La sección de categorías de la aplicación Castrapp debería incluir un apartado para comentarios sobre las clasificaciones, y así recibir retroalimentación de los usuarios y mejorar el modelo LDA.

# BIBLIOGRAFÍA


1. Abram, C., y Karasavas, A. (2018). *Facebook for Dummies.* New York, Estados Unidos: John Wiley & Sons, Inc.

2. Alpaydin, E. (2016). *Machine learning: the new AI.* Londres, Inglaterra: MIT press.

3. Arnaboldi, V., Passarella, A., Conti, M. y Dunbar, R. I. (2015). *Online social networks: human cognitive constraints in Facebook and Twitter personal graphs.* Amsterdam, Paises Bajos: Elsevier.

4. Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S. y Blei, D. (marzo, 2009). Reading tea leaves: How humans interpret topic models. *Neural information processing systems, 22*, 288-296.

5. Crookes, D. (2017). *Facebook for Beginner in Easy Steps.* Reino Unido: Easy Steps Limited.

6. Etaati, L. (2019). *Machine Learning with Microsoft Technologies.* Nueva Zelanda, Oceania: Apress.

7. Gatuna, C. (07 de Marzo de 2021). V*acunación gatuna.* Guatemala: Autor.

8. Golbeck, J. (2015). *Introduction to social media investigation: a hands-on approach.* Ámsterdam, Paises Bajos: Syngress - Elseiver.

9. He, W., Zha, S. y Li, L. (marzo, 2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International journal of information management, 33*(3), 464-472.

10. Kwok, L. y Yu, B. (junio, 2013). Spreading social media messages on Facebook: An analysis of restaurant business-to-consumer communications. *Cornell Hospitality Quarterly, 54*(1), 84-94.

11. Lane, H., Howard, C. y Hapke, H. (2019*). Natural Language Processing in Action Video Edition*. New York, Estados Unidos: Manning Publications Co.

12. Miles, J. (2019). *Instagram Power: Build Your Brand and Reach More Customers with Visual Influence: Build Your Brand and Reach More Customers with Visual Influence*. Estados Unidos: McGraw Hill Professional.

13. Palma, C. (2018). *Aplicación móvil para jornadas de castración, vacunación y adopción de mascotas mediante el acceso a los grupos animalistas utilizando la API Graph de Facebook*. Guatemala: Autor.

14. Raschka, S. (13 de abril, 2014). *Implementing a Principal Component Analysis (PCA)*. [Mensaje en un blog]. Recuperado de https://sebastianraschka.com/Articles/2014_pca_step_by_step.html.

15. Ruiz, P. (2015). *Optimización de la búsqueda de intereses personales en twitter utilizando modelado de tópicos* (Tesis de licenciatura). Universidad de San Carlos de Guatemala, Guatemala.

16. Sakr, S. y Pardede, E. (2011). *Graph data management: techniques and applications.* Australia: IGI Publishing.

17. Salloum, S. A., AI-Emran, M., Monem, A. y Shaalan, K. (marzo, 2017). A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J, 2*(1), 127-133.

18. Salloum, S. A., Al-Emran, M. y Shaalan, K. (julio, 2017). Mining social media text: extracting knowledge from Facebook. *International Journal of Computing and Digital Systems, 6*(02), 73-81.

19. Salloum, S. A., Al-Emran, M. y Shaalan, K. (mayo, 2017). Mining text in news channels: a case study from Facebook. *International Journal of Information Technology and Language Studies, 1*(1), 1-9.

20. Salloum, S. A., Mhamdi, C., AI-Emran, M. y Shaalan, K. (abril, 2017). Analysis and classification of Arabic Newspapers' Facebook pages using text mining techniques. *International Journal of Information Technology and Language Studies, 1*(2), 8-17.

21. Sarkar, D. (2019). *Text Analytics with Python - A Practitioner's Guide to Natural Language Processing*. Karnataka, India: Apress.

22. Stigler, M. y Stigler, M. (2018). *Beginning Serverless Computing.* Virginia, Estados Unidos: Apress.

23. Tsoumakas, G., Katakis, L. y Vlahavas, L. (febrero, 2009). Mining multi-label data. *Data mining and knowledge discovery handbook*, 667-685.

24. Ugander, J., Karrer, B., Backstrom, L. y Marlow, C. (2011). *The anatomy of the facebook social graph*. Estados Unidos: arXiv.

25. Yang, Y. (2016). Temporal Data Mining via Unsupervised Ensemble Learning. Amsterdam, Paises Bajos: Elsevier.

26. Zimmerman, J. y Ng, D. (2019). *Social media marketing all-in-one for dummies.* New Jersey, Estados Unidps: John Wiley & Sons.