



Universidad
Zaragoza

Trabajo Fin de Grado

Análisis de la multimedia en redes sociales

Analysis of the multimedia in social networks

Autor

Jesús Aguas Acín

Directores

Rahul Singh
Javier Fabra Caro

Titulación del autor

Ingeniería Informática

ESCUELA DE INGENIERIA Y ARQUITECTURA
2020

Resumen

Este documento describe el desarrollo de una herramienta capaz de analizar publicaciones en las redes sociales y evaluar sus significados semánticos, es decir, entender su temática para así poder catalogarlos.

Este trabajo fue realizado durante un programa de movilidad UZ-Norteamérica en Estados Unidos, en la universidad San Francisco State University. Gracias a la dirección y asesoramiento del Dr. Rahul Singh y la coordinación desde España por el Dr. Javier Fabra Caro.

Para este estudio se ha utilizado la red social '*Reddit*', debido a la gran cantidad de información que cada publicación contiene, se puede obtener características muy relevantes a partir del título, el texto, los comentarios, e imágenes que los usuarios publican en esta red social. Concretamente se han analizado publicaciones del *subreddit* '*r/opiates*', relacionadas con el consumo de opiáceos y estupefacientes. El objetivo principal de esta herramienta es poder encontrar tópicos interesantes dentro de este ámbito analizando 1000 publicaciones con texto e imágenes.

En el análisis de las publicación se utiliza toda la información que pueda ser útil para categorizarlo, distinguiendo entre información textual y visual, ya que estos tipos distintos de datos serán procesados de manera distinta. Para el análisis del texto, se utiliza la herramienta DUI (Drug Use Insights), y para el análisis de archivos multimedia se utiliza la API: Google Vision API.

Una vez se analizan todas las características relevantes de la publicación, se utiliza un modelo basado en deep learning que es capaz de clasificar esta publicación en una de las categorías especificadas anteriormente y reconociendo así su significado semántico.

Los resultados demuestran la efectividad del modelo, y la gran utilidad que tiene esta herramienta para poder catalogar automáticamente e identificar características comunes en diferentes publicaciones.

Índice

1	Introducción	1
1.1	Motivación	1
1.2	Contexto	2
1.3	Objetivos	3
1.4	Estructura de la memoria	3
2	Recuperación de la información	4
2.1	Búsqueda de la información.....	4
2.1.1	Reddit.....	4
2.2	Extracción de la información	6
2.2.1	Pushshift API	6
2.2.2	Imgur API	6
2.3	Almacenamiento de la información	6
2.3.1	MongoDB	6
3	Análisis multimedia	8
3.1	Cloud Vision API	8
3.2	DUI.....	10
4	Creación del modelo	11
4.1	LDA.....	11
4.2	Metodología	12
5	Resultados	13
5.1	LDA con 5 tópicos	13
5.2	LDA con 10 tópicos	14
5.3	LDA con 15 tópicos	15
6	Conclusiones	16
	Bibliografía	17

Capítulo 1

Introducción

El uso indebido y adicción a las drogas es un peligro significativo para la salud en todo el mundo que afecta tanto a la salud pública y como al bienestar económico.

Estados Unidos se enfrenta particularmente a una grave crisis, en parte debido al uso indebido de opiáceos; solo en 2017, hubo 70.237 muertes por sobredosis de drogas que involucraron abuso de oxicodona, hidromorfona y fentanilo [1] y la tasa de mortalidad diaria durante 2018 se estima en más de 130 personas [2].

El modelado epidemiológico, análisis y diseño de la intervención para el uso indebido de sustancias es especialmente difícil ya que los métodos tradicionales de recolección de datos como encuestas e informes médicos no son en tiempo real ni reflejan el contexto subyacente en un caso individual de adicción. Además, tales métodos dependen de la decisión de los participantes para revelar información que a menudo es profundamente personal.

En este contexto, las publicaciones en las redes sociales que contienen información que los propios usuarios publican sin filtrar pueden ser una fuente de datos novedosa y valiosa para el análisis de adicciones. Actualmente hay pocos, si es que hay alguno, sistemas públicos que permitan el análisis de datos epidemiológicamente relevantes de las redes sociales centrados en el abuso de sustancias.

1.1 Motivación

El abuso de sustancias y la adicción es una importante contemporánea crisis en la salud.

Modelar su epidemiología y el diseño de intervenciones efectivas requiere un análisis de datos en tiempo real junto con los medios para contextualizar los patrones de adicción a escala individuo-comunidad. En este contexto, las redes sociales han comenzado considerarse como una fuente novedosa de información reportada por los usuarios en tiempo real. Sin embargo, la capacidad de los epidemiólogos para usar dicha información se ve significativamente obstaculizada por la falta de algoritmos y software disponibles públicamente para la extracción, análisis y modelado de la información sobre adicciones. Crear una herramienta que sea capaz de modelar esta información puede resultar de gran utilidad para la comunidad científica.

1.2 Contexto

Este trabajo se realizó durante el último semestre de mi grado en Ingeniería Informática en la Universidad de Zaragoza, mientras participaba en un programa de movilidad en Estados Unidos, en la San Francisco State University. Concretamente entre Febrero y Junio de 2020. Todo comenzó cuando estaba conversando con Rahul Singh, doctor en Computer Science, que me daba clase de Big Data en la universidad.

Estuve conversando con él acerca de mis estudios y le comenté que debía realizar un trabajo de final de grado y fue entonces cuando me habló de un proyecto en el que tenía previa experiencia y que le parecía interesante para realizar un estudio y elaborar una herramienta. Me habló del tema y me dio a conocer el problema existente y cómo podía abordarlo. Me resultó muy interesante, estudié el tema en profundidad y finalmente decidí trabajar en ello.

El primer mes del trabajo fue dedicado al estudio del problema y la evaluación de las distintas redes sociales que nos podían aportar la información que necesitábamos, así como las herramientas que nos permitirían procesar esa información y así poder trabajar con ella. Realizábamos reuniones muy de vez en cuando, le comentaba mis progresos y dudas después de las clases y seguía buscando la mejor forma de abordar este proyecto.

Durante este tiempo descubrí una herramienta muy eficaz para analizar imágenes, Google Vision API, de la cual elaboré varios informes para que mi director evaluase la eficacia de esta herramienta y determinase si era adecuada para este trabajo. También mi director, Rahul Singh, me recomendó utilizar una herramienta elaborada por él en colaboración con uno de sus ex-alumnos, llamada DUI (Drug Use Insights). Además también existían dudas acerca de la fuente de información, tanto Twitter como Reddit eran dos redes sociales candidatas para este trabajo, sin embargo, Reddit acabó siendo un claro vencedor al ofrecer mayor cantidad de información relevante y más accesible que Twitter.

Una vez se establecieron las fuentes de información y las herramientas para procesarlas, el siguiente paso fue empezar y aprender a utilizar estas herramientas para poder enfocarlas a nuestros objetivos y elaborar algoritmos que permitiesen la transformación de la información en resultados interpretables y concluyentes para el estudio.

Por último, y para completar el apartado del contexto, cabe mencionar que durante este tiempo ocurrió la pandemia mundial del COVID-19, la cual dificultó significativamente la comunicación con mi director y la progresión del proyecto.

1.3 Objetivos

El objetivo principal de este proyecto es crear un modelo que sea capaz de identificar los tópicos que comparten las publicaciones en las redes sociales en el ámbito del consumo de drogas.

Dentro de ese principal objetivo, se pueden distinguir distintos objetivos establecidos durante el transcurso del proyecto:

- Investigación sobre las fuentes de información y su obtención
- Investigación sobre las herramientas de procesado de información y su utilización
- Crear una gran base de datos con toda la información
- Creación y entrenamiento del modelo LDA.
- Evaluación de resultados.

1.4 Estructura de la memoria

La memoria del TFG está organizada en capítulos, cada uno detalla una fase del proyecto.

- **Capítulo 1. Introducción:** Sitúa el TFG en su contexto espacial y temporal, elaborando los motivos que llevaron a tomar la elección de trabajar en este tema y se establecen objetivos.
- **Capítulo 2. Recuperación de la información:** Durante esta etapa del proyecto se evalúan las fuentes de información y se describe los métodos utilizados para obtener la información de dichas fuentes.
- **Capítulo 3. Análisis multimedia:** En este capítulo se detalla el proceso de transformación de los datos sin procesar en información relevante y utilizable por el modelo LDA.
- **Capítulo 4. Creación del modelo:** Describe el modelo LDA utilizado para convertir la información procesada en resultados.
- **Capítulo 5. Resultados:** Se muestran los resultados obtenidos por el modelo y se realiza una evaluación crítica destacando los puntos fuertes y débiles del modelo.
- **Capítulo 6. Conclusiones:** Esta sección concluye el trabajo resumiendo el trabajo realizado, analizando los resultados obtenidos y elaborando una conclusión final.

Capítulo 2

Recuperación de la información

A lo largo de este capítulo se detallará el proceso de recuperación de la información, es decir, el conjunto de técnicas y procedimientos utilizados para acceder y extraer la información almacenada en las redes sociales.

Primero se detallará el proceso de evaluación de las distintas fuentes de información, así como la búsqueda de información relevante para poder trabajar con ella. A continuación, se explicará en detalle las técnicas y procedimientos utilizados para extraer esta información y almacenarla en una base de datos local, para su posterior estudio y uso en próximas fases.

2.1 Búsqueda de la información

Esta es la primera fase del proyecto, consiste en la evaluación de las distintas fuentes de información, dentro de las distintas redes sociales, con el objetivo de encontrar una conjunto de datos que se adecúe a las exigencias de este trabajo y que permita obtener unos resultados satisfactorios. Por su propia definición es entiende que es difícil hacer tal estimación nada más empezar con el proyecto. Por lo que la mejor forma de abordar este problema es dedicando muchas horas a estudiar cada una de las redes sociales y buscar fuentes de información que cumplan varios requisitos: que permitan extraer grandes cantidades de información, que esta información sea relevante para el estudio, y que tenga algún tipo de uniformidad que permita trabajar con ella utilizando las mismas técnicas y procedimientos.

2.1.1 Reddit

La red social elegida fue Reddit, concretamente el repositorio *r/opiates*.

Primero aclarar que se escogió una fuente de información en inglés por varios motivos, el primero es que mi director era de habla inglesa, por lo que esa era la mejor forma de que pudiese evaluar las fuentes y seguir el proceso entendiendo los resultados. Otro motivo muy relevante es que las publicaciones en este idioma eran mucho más numerosas que en cualquier otro idioma, debido a que el inglés es el idioma con más aceptación universal. Además de que este estudio estaba enfocado a estudiar la drogadicción en Estados Unidos (habla inglesa) donde este problema es muy grave respecto a otros países.

El motivo por el que se escogió Reddit entre otras redes sociales es debido a que destaca por ser el foro mas popular para hablar sobre temáticas concretas y permite distintas formas de expresar las ideas, como texto, imágenes y videos, además de que cada publicación contiene mucha más información que ya que el limite de caracteres de cada publicación es de hasta 40.000 caracteres, comparado con Twitter, cuyo límite es de 280 caracteres.

Una vez seleccionada Reddit como red social, hay que seleccionar un subreddit, es decir, la temática del foro del que se extraerán las publicaciones. Debido al caso de estudio de este proyecto, se consideró el subreddit *r/opiates* como el más apropiado debido a que las publicaciones tratan de todo relacionado con opiáceos, desde fármacos, adicción, hasta rehabilitación y es el más popular entre la comunidad en este ámbito.

Las publicaciones pueden contener solo texto, solo imágenes o pueden contener texto e imágenes. En este proyecto se han utilizado publicaciones que contienen ambos tipos de multimedia con la ambición de poder combinar la semántica de ambos.

A continuación se muestra un ejemplo de una publicación de Reddit con texto e imágenes:

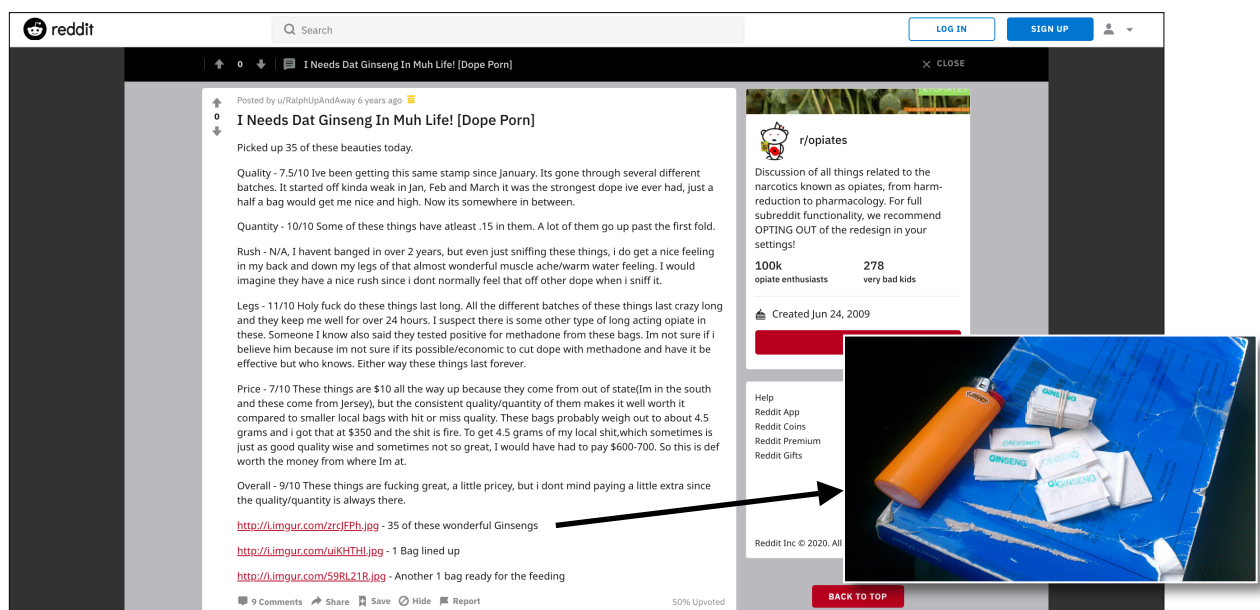


Figure 2.1: Ejemplo de una publicación en Reddit r/opiates, contiene un título, texto, comentarios y una o más imágenes (link a Imgur, donde se almacenan las imágenes) [3]

2.2 Extracción de la información

Una vez establecida la fuente de información, el siguiente paso es extraer la información relevante de cada publicación. Para ello se utilizaron dos APIs distintas, Pushshift API para obtener información acerca de la publicaciones en Reddit, y Imgur API para obtener las imágenes de la página web Imgur, donde estaban almacenadas las imágenes mencionadas en las publicaciones.

2.2.1 Pushshift API

Pushshift API que permite obtener todo tipo de información de la publicación solicitada en Reddit. La documentación de esta herramienta fue esencial para aprender a manejar y realizar las peticiones concretas a la API. [4]

Para este proyecto se extrajeron 2000 publicaciones con texto e imágenes, el título de la publicación, el texto y los comentarios se pudieron extraer con esta herramienta sin ningún problema con esta herramienta.

2.2.2 Imgur API

Sin embargo, para obtener las imágenes de cada publicación se tuvo que utilizar una API distinta, ya que éstas estaban almacenadas en otra página web distinta de Reddit, la cual solo contenía enlaces a Imgur, una página web exclusivamente dedicada a almacenar las imágenes de los usuarios.

Para obtener las imágenes se utilizó Imgur API, que al igual que Pushshift API, dispone de una documentación que facilita su uso. [5]

2.3 Almacenamiento de la información

Con toda la información relevante de cada publicación obtenida, se requiere de una forma de almacenamiento que permita guardar todos los datos que se utilizaran durante el estudio.

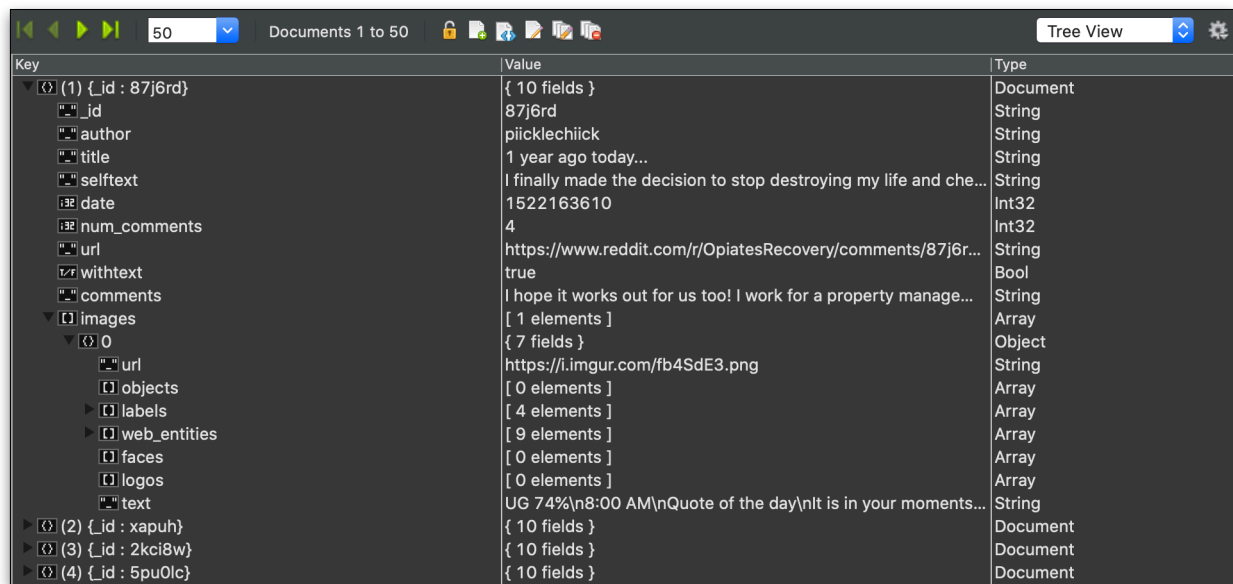
Para este tipo de problema el tipo de base de datos más adecuado es una base de datos no relacional, ya que únicamente se almacenaran publicaciones, las cuales son un árbol de atributos dentro de atributos, por lo que una base de datos relacional significaría numerosas relaciones innecesarias y que con una base de datos documental se pueden ahorrar. Por lo que la base de datos documental escogida fue MongoDB.

2.3.1 MongoDB

Quizás la base de datos documental más popular, MongoDB se caracteriza por su facilidad de uso y el potencial para almacenar miles de documentos asegurando la consistencia de los datos y una gran flexibilidad para realizar cambios en el modelo.

Este tipo de bases de datos están formadas por ‘documentos’, los cuales no es necesario que compartan la misma estructura interna, es decir, que cada documento puede tener distintos atributos a otro. Esta cualidad es esencial para este proyecto ya que permite crear múltiples documentos sin que el número de imágenes de cada publicación sea un problema.

Dentro de cada documento se almacena toda la información relevante de cada publicación, además de otros atributos que informan del estado de la publicación y se han añadido con el objetivo de facilitar su manipulación.



Key	Value	Type
(1) {_id : 87j6rd}	{ 10 fields }	Document
_id	87j6rd	String
author	piicklechiick	String
title	1 year ago today...	String
selftext	I finally made the decision to stop destroying my life and che...	String
date	1522163610	Int32
num_comments	4	Int32
url	https://www.reddit.com/r/OpiatesRecovery/comments/87j6r...	String
withtext	true	Bool
comments	I hope it works out for us too! I work for a property manage...	String
images	[1 elements]	Array
0	{ 7 fields }	Object
url	https://i.imgur.com/fb4SdE3.png	String
objects	[0 elements]	Array
labels	[4 elements]	Array
web_entities	[9 elements]	Array
faces	[0 elements]	Array
logos	[0 elements]	Array
text	UG 74%\n8:00 AM\nQuote of the day\nIt is in your moments...	String
(2) {_id : xapuh}	{ 10 fields }	Document
(3) {_id : 2kci8w}	{ 10 fields }	Document
(4) {_id : 5pu0lc}	{ 10 fields }	Document

Figura 2.2: Ejemplo de un documento (representa una publicación en Reddit) en MongoDB. Se pueden observar almacenados todos los atributos relevantes mencionados, como el título, texto, comentarios, y un array de imágenes, en las que por cada imagen se almacenan los resultados obtenidos por la herramienta Google Vision API [ver Sección 3.1]

Capítulo 3

Análisis multimedia

Durante este capítulo se explicará y mostrará en detalle los dos tipos de análisis realizados de cada publicación según el tipo de datos.

Como se ha explicado anteriormente, todas las publicaciones están formadas por dos tipos distintos de multimedia: texto e imágenes. Ambos no pueden ser analizados con la misma herramienta ya que su naturaleza es distinta, por lo que es necesario tener dos herramientas distintas que permitan obtener características relevantes para cada uno de los tipos de información.

Para el análisis de imágenes se utiliza la herramienta Cloud Vision API.

Para el análisis de texto se utiliza la herramienta DUI (Drug Use Insights).

3.1 Cloud Vision API

Esta herramienta creada por Google permite a los desarrolladores integrar fácilmente funciones de detección visual de características dentro de aplicaciones. Estas características incluyen etiquetado de objetos, detección de rostros y de sus expresiones, reconocimiento óptico de caracteres (OCR), identificación de entidades web y etiquetado de contenido explícito.

Toda esta información se puede extraer de cada imagen mediante la API, la cual requiere la posesión de una cuenta en Google Cloud Platform y un pago cuando se excede el uso básico permitido.

En este enlace se puede encontrar todo el código para realizar el análisis de una imagen mediante Google Vision API a partir de la url de una imagen como parámetro, está muy bien comentado por lo que se puede entender a la perfección como funciona. [6]

A continuación se muestra un ejemplo de la eficacia de esta herramienta mediante una página web dedicada exclusivamente a la demostración de esta API. [7]

Pongamos por ejemplo la siguiente publicación [ver Figura 3.1] en Reddit acerca de un llavero que es otorgado por la confraternidad internacional de Narcóticos Anónimos como premio por estar un largo periodo de tiempo en abstinencia de drogas.

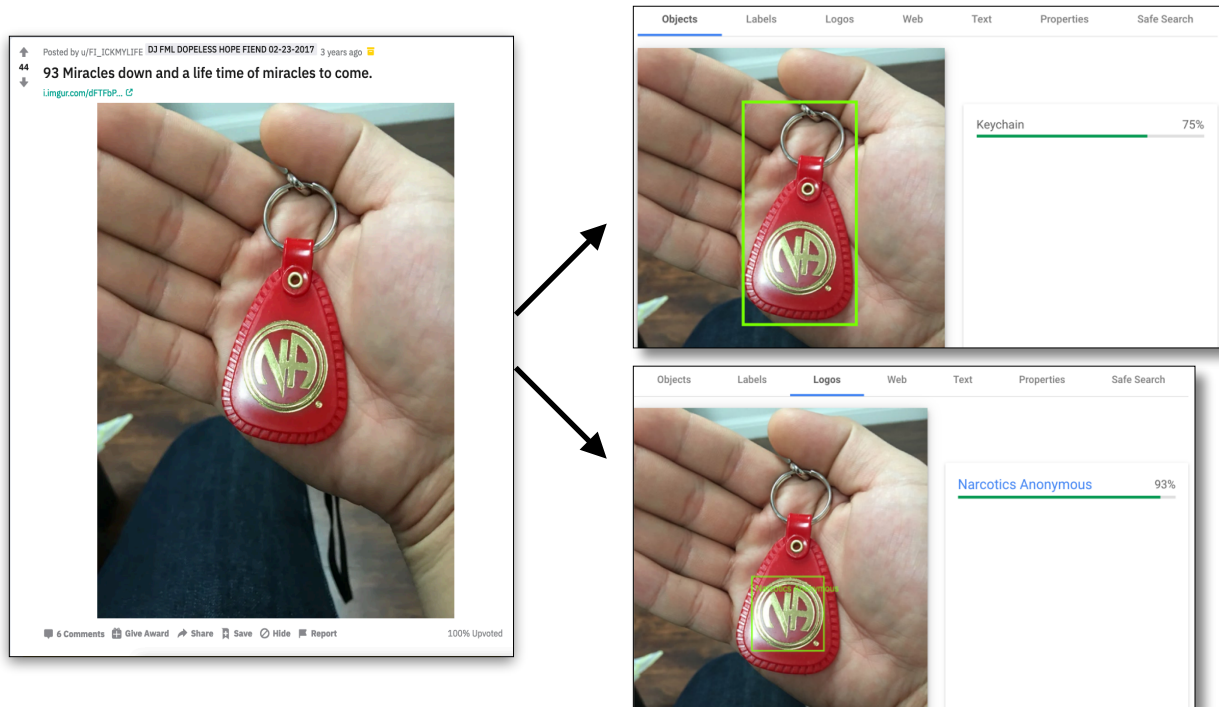


Figura 3.1: A la izquierda se muestra una publicación en Reddit [8] la cual contiene una imagen que muestra un llavero con el logo de Narcóticos Anónimos, a la derecha se muestran dos capturas de pantalla del análisis realizado por Google Vision API, la de arriba muestra el resultado de la identificación de objetos y la de abajo el reconocimiento de logos en la imagen.

Al analizar esta imagen mediante Google Vision API obtenemos muchas características acerca de la imagen, agrupadas por categorías, en este caso se muestran las categorías Objetos y Logos ya que muestran los resultados más relevantes para entender el contenido de la imagen. [Figura 3.1]

De entre todas las funcionalidades que permite realizar esta herramienta hay algunas que no son necesarias para este proyecto, como por ejemplo, el etiquetado de contenido explícito. Por lo que las funcionalidades que se consideraron esenciales para el estudio fueron las siguientes:

- Identificación de objetos
- Etiquetado de imagen
- Reconocimiento de logos
- Identificación de entidades web
- Reconocimiento facial
- Reconocimiento de texto

3.2 DUI

Esta herramienta fue creada por Zachary Prince & Deeptanshu Jha (con la dirección de Rahul Singh), ex-alumnos de la San Francisco State University (mi universidad durante el programa de movilidad). Se trata de una herramienta en fases avanzadas de desarrollo diseñada para ayudar a epidemiólogos en el análisis de texto para descifrar el significado dentro del dominio de la drogas

DUI proporciona información acerca de los términos usados para referirse a distintos tipos de drogas, efectos en su uso, emociones y sensaciones, eventos y lugares y en general, cualquier término relacionado con este ámbito.

Esta herramienta fue utilizada, con la autorización de sus creadores, para analizar el título, texto y comentarios de cada publicación para así poder obtener los términos clave relacionados con el consumo de drogas.

Estos términos clave fueron combinados junto con los extraídos por la herramienta citada anteriormente para analizar imágenes, creando así un conjunto de características y palabras clave de cada publicación que posteriormente utilizamos para crear el modelo de aprendizaje profundo supervisado de clasificación. [ver Sección 4]

Capítulo 4

Creación del modelo

En este apartado se detallará el último paso del proyecto, la elaboración del modelo.

Este consistirá en realizar un LDA (Linear Discriminant Analysis) de los datos extraídos en apartados anteriores para poder identificar conjuntos de términos, que podríamos entender como temáticas o categorías, dentro del lenguaje utilizado en el dominio de las drogas.

4.1 LDA

LDA es un método utilizado en machine learning para encontrar una combinación lineal de rasgos que caracterizan o separan dos o más clases de objetos o eventos. Utilizaremos esta combinación resultante como un clasificador lineal, de modo que podremos identificar distintas categorías según los términos que la componen.

LDA hace predicciones al estimar la probabilidad de que un nuevo conjunto de entradas pertenezca a cada clase. La clase que obtiene la mayor probabilidad es la clase salida y se realiza una predicción. El modelo utiliza el teorema de Bayes para estimar las probabilidades.

Una característica de este modelo es que se debe especificar el número de grupos (categorías) resultantes del modelo. Como no conocemos el número óptimo de tópicos, la solución más razonable es inspeccionar los resultados [Sección 5] y quedarse con el modelo que devuelva los resultados más satisfactorios.

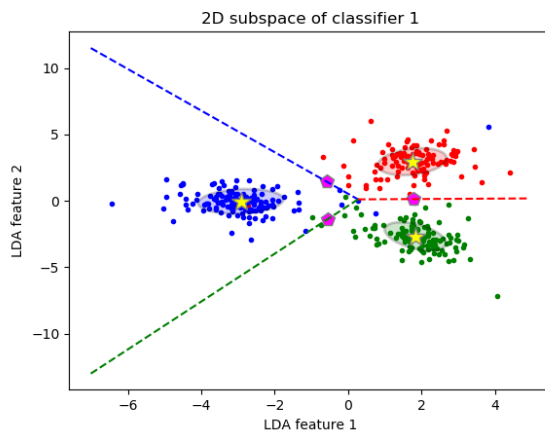


Figura 4.1: Se muestra un ejemplo de un análisis LDA en un problema de clasificación multiclase en el que a partir de dos características de los datos de entrada (LDA feature 1 & LDA feature 2) se logran separar los datos en 3 clases o tópicos

4.2 Metodología

Para implementar el análisis LDA se ha utilizado Python ya que es el lenguaje utilizado por excelencia en proyectos relacionados con machine learning. Hay muchos motivos, entre ellos, la sencillez con la que permite ejecutar acciones complejas con sentencias muy simples, hace que la comunidad se haya puesto de acuerdo en elaborar la mayoría de librerías y funciones de machine learning en este lenguaje de programación interpretado. De este modo, Python ofrece múltiples formas de realizar esta tarea, en este caso se ha escogido utilizar la librería de Scikit-Learn, ya que dispone de sencillas funciones que realizan de forma transparente todo el cálculo que hay detrás de este complejo método.

Todo el código utilizado para este último paso se puede encontrar en el archivo LDA.py dentro del repositorio del proyecto [10]. En este archivo primero se recogen todos los datos extraídos en etapas anteriores y se agrupan en documentos, los cuáles después se pasan al algoritmo que crea el modelo LDA según el número de tópicos especificados. Finalmente se muestran los resultados por pantalla, formateados con una función que se encarga de mostrar los tópicos junto con un top 10 de los términos con mayor frecuencia en ese tópico, esto quiere decir, los 10 términos que mejor representan a la categoría a la que pertenecen.

Todas las funciones y clases utilizadas pertenecientes a la librería de Scikit-learn fueron previamente encontradas y estudiadas gracias a la documentación sobre la librería. [9]

Capítulo 5

Resultados

A partir del proceso anterior se obtienen los resultados del modelo LDA, los cuáles muestran un número determinado de categorías o tópicos junto con un conjunto de palabras que representan los términos más frecuentes de cada grupo.

Debido a la naturaleza cualitativa de los resultados es complicado cuantificar la eficacia del modelo, por lo que la única forma de evaluarlos es realizando un análisis de coherencia mediante observación. Por lo que para cada número distinto de tópicos se debe evaluar manualmente la eficacia del modelo para así poder determinar cuál es el más adecuado en este ámbito.

A continuación se mostrarán los resultados del modelo para 5, 10 y 15 tópicos y se evaluará la validez y eficacia de cada uno.

5.1 LDA con 5 tópicos

A continuación se muestra la salida del modelo LDA para 5 tópicos:

```
LDA:
Topic 1:
opioids dope goods packaged drug_use_general food weed administration acquisition water
Topic 2:
person hair opioids meme internet weed shoe clothing hand drug_use_general
Topic 3:
art coffee illustration shirt sleeve arts water visual photography design
Topic 4:
leg skin close arm joint injection hand histamine human flesh
Topic 5:
goods packaged font black line meter logo text opioids white
```

Figura 5.1: Salida por pantalla del modelo LDA para 5 tópicos

En este caso, se puede observar que los tópicos no parecen tener mucho sentido ya que están compuestos de términos que no parecen guardar ningún tipo de relación entre ellos a simple vista. Esto podría deberse a que el número de tópicos establecidos es demasiado bajo por lo que múltiples tópicos se combinan en uno para ajustarse al requisito de 5 tópicos. De modo que se acaban mezclando términos de tópicos que deberían ser distintos.

5.2 LDA con 10 tópicos

A continuación se muestra la salida del modelo LDA para 10 tópicos:

```
LDA:
Topic 1:
meme food internet image opioids coin cuisine rice person humour
Topic 2:
shoe spoon tableware cutlery utensil kitchen sneakers shoes metal fork
Topic 3:
shirt sleeve betty earrings paraphernalia clothing museum games legal outerwear
Topic 4:
histamine leg reaction filters micron knee injection human swelling ankle
Topic 5:
goods packaged font black meter opioids line product logo text
Topic 6:
water blue administration vehicle font wheel opioids text device packaged
Topic 7:
plant poppy flower water opium tree leaf botany opioids stem
Topic 8:
dog hair nail opioids grooming dope goods packaged drug_use_general person
Topic 9:
opioids dope drug_use_general weed administration person acquisition recovery_general hair finance
Topic 10:
wood floor flooring art design stain car hardwood sport jean
```

Figura 5.2: Salida por pantalla del modelo LDA para 10 tópicos

Este número de tópicos parece mejorar el resultado anterior como se predecía, por ejemplo, se puede reconocer que el primer tópico está relacionado con memes y humor en internet.

Sin embargo, se siguen encontrando términos que no guardan mucha coherencia con el resto de la clase, probablemente porque aún sigue ocurriendo el mismo problema que el primer resultado [Sección 5.1] y se debe aumentar más el número de tópicos.

5.3 LDA con 15 tópicos

A continuación se muestra la salida del modelo LDA para 15 tópicos:

```
LDA:
Topic 1:
toilet coin comedy graphics clinic constipation zoom opioids art clip
Topic 2:
spoon shoe cutlery tableware utensil kitchen sneakers shoes men footwear
Topic 3:
macro photography abdomen navel paraphernalia beige black wallpaper product legal
Topic 4:
morphine drug injection filters oxycodone reduction pharmaceutical harm wheel parallel
Topic 5:
black font line logo text photography brand white organism meter
Topic 6:
text device zoom legal friend night writing paraphernalia wine multimedia
Topic 7:
coffee emblem water appliance tea sea horizon extract filter drink
Topic 8:
hair person dope human goods packaged imgur black drug_use_general forehead
Topic 9:
person hair meme internet clothing opioids black dope drug_use_general photography
Topic 10:
wood floor flooring design stain car art hardwood sport jean
Topic 11:
dog breed shih companion tzu snout puppy canidae carnivore mammal
Topic 12:
spoon packaged goods blue flooring reaction person cop tableware hair
Topic 13:
plant poppy flower opium stem botany leaf screenshot flowering wildflower
Topic 14:
goods packaged opioids dope drug_use_general weed product food administration water
Topic 15:
leg arm skin joint close human flesh histamine reaction hand
```

Figura 5.3: Salida por pantalla del modelo LDA para 15 tópicos

Este resultado parece interesante ya que por fin se han obtenido tópicos directamente relacionados con drogas, como es el caso del tópico 4, sin embargo la mayoría de tópicos, aunque tengan coherencia, no están directamente relacionados con este tema, por ejemplo, el tópico 13 está relacionado con plantas.

También se observan algunos casos de mezcla de tópicos, como es el caso del tópico 2, que mezcla términos relacionados con utensilios de cocina y calzado.

Es lógico que cuanto mayor sea el número de tópicos más sentido y coherencia tengan los términos entre ellos, sin embargo el objetivo es encontrar un modelo aceptable que contenga el menor número de tópicos posibles.

Al realizar pruebas con mayor número de tópicos, los resultados no muestran una mejora tan notable como la que ha habido hasta llegar a 15 tópicos por lo que se puede concluir que este resultado es el óptimo para el estudio.

Capítulo 6

Conclusiones

Este último capítulo resume el trabajo realizado, analizando los resultados obtenidos y elaborando una conclusión final.

En primer lugar quiero agradecer a mi director Rahul Singh por guiarme a lo largo del proyecto y a Javier Fabra Caro por su coordinación con la administración desde la Universidad de Zaragoza. Este proyecto se realizó con las expectativas de encontrar características y tópicos importantes dentro del mundo de las drogas, un problema que a día de hoy sigue afectando a millones personas.

Respecto a la evaluación de los resultados obtenidos, se puede observar que el mejor resultado se da al utilizar 15 tópicos [ver Sección 5.3], a partir de allí la diferencia en mejora al aumentar el número de tópicos no es significativa. Una conclusión importante que saco de este proyecto y que no consideraba al principio del mismo es encontrarme con que hay más tópicos no relacionados con el mundo de las drogas que al contrario. Esto se puede deber a varios motivos, la herramienta DUI está en desarrollo por lo que aún se podría mejorar ciertos aspectos. Además también es probable que la mayoría de imágenes que publicaban los usuarios no estaban directamente relacionadas con el mundo de las drogas, y eran elementos que se encontraban en segundo plano pero que la herramienta Google Vision API los ha reconocido y se les ha otorgado la misma importancia que elementos relacionados con la drogadicción y que son relevantes para el estudio.

Al completar este trabajo también he cumplido el objetivo personal de realizar un estudio completo utilizando los conocimientos aprendidos a lo largo de mi carrera universitaria aplicado a tratar de comprender un problema humanitario grave para la salud. A lo largo del proyecto he mejorado mi capacidad para resolver problemas más allá de los estudiados en clase y considero estar mejor preparado para realizar proyectos similares que puedan tener aplicaciones igual de interesantes.

Todo el directorio de trabajo con todo el código utilizado se puede encontrar en el repositorio “TFG” en mi cuenta de Github [10].

Bibliografía

- [1] Scholl, L. et al. (2019) Drug and opioid-involved overdose deaths—United States, 2013–2017. *MMWR Morb. Mort. Wkly Rep.*, 67, 1419.
- [2] CDC. (2018) CDC/NCHS, National Vital Statistics System, Mortality. CDC WONDER, Atlanta, GA: US Department of Health and Human Services.
- [3] u/RalphUpAndAway, “I Needs Dat Ginseng In Muh Life!”, <https://www.reddit.com/2563z9>
- [4] Baumgartner, Jason M. “*Pushshift Documentation*” *Release 4.0*, <https://readthedocs.org/projects/reddit-api/downloads/pdf/latest/>
- [5] Imgur, “*Imgur API Documentation*” <https://apidocs.imgur.com/>
- [6] Google, “*Cloud Vision API Documentation*”, <https://cloud.google.com/vision/docs>
- [7] Google, “*Drag and Drop*”, <https://cloud.google.com/vision/docs/drag-and-drop>
- [8] u/FI_ICKMYLIFE, “93 Miracles down and a life time of miracles to come”, <https://www.reddit.com/6dos7r>
- [9] Scikit-Learn, “*Scikit-learn documentation*”, https://devdocs.io/scikit_learn/
- [10] Jesús Aguas, “TFG”, Github, <https://github.com/jesusaguas/TFG>