Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: Medicina (TFM-Med)

Análisis de la depresión y la ansiedad causadas por un aborto usando datos de Twitter

---

Autor: Laura Planas Simón

Directores: Davide Cirillo y Laia Subirats Maté

Profesor: Ferran Prados Carrasco

Barcelona, 25 de enero de 2022

Créditos/Copyright

Ficha del trabajo final

Título del trabajo:	Análisis de la depresión y la ansiedad causadas por un aborto usando datos de Twitter
Nombre del autor:	Laura Planas Simón
Nombre del colaborador/a docente:	Davide Cirillo y Laia Subirats Maté
Nombre del PRA:	Ferran Prados Carrasco
Fecha de entrega (mm/aaaa):	01/2022
Titulación o programa:	Máster en Ciencia de Datos
Área del Trabajo Final:	Área Medicina (TFM-Med)
Idioma del trabajo:	Castellano
Palabras clave	Análisis de sentimiento, Twitter, Aborto

Agradecimientos

En primer lugar quiero expresar mi especial gratitud a mis dos tutores, Laia Subirats Maté y Davide Cirillo, por el apoyo brindado y los buenos consejos que he recibido a lo largo del proyecto.

Quiero extender mi agradecimiento a Nataly Buslón y Diego Saby de la unidad Social Link Analytics del Barcelona Supercomputing Center, y en especial a Ángela Leis del Research Programme on Biomedical Informatics (Hospital del Mar y Universitat Pompeu Fabra) por su extenso trabajo previo y en el que se ha basado este trabajo.

También me gustaría agradecer a Women's Brain Project [1] por la iniciativa de realizar una Hackathon con el objetivo de avanzar en la comprensión de las diferencias de género que existen hoy en día en la medicina, donde colaboraron en la Universidad de Zurich y ETH y el Barcelona Supercomputing Center. Gracias a esta iniciativa surgieron las ideas que han hecho hoy posible este y muchos otros proyectos.

Mando mi más sincero apoyo a todas las mujeres que han vivido algo tan duro como un aborto espontáneo, pero que aun así comparten sus experiencias y ayudan a otras mujeres a través de las redes sociales. Este proyecto no habría sido posible sin vosotras.

Por último pero no menos importante, también quiero agradecer a mi pareja y a mi familia por siempre ser mi mayor inspiración y apoyo en todo lo que hago.

Abstract

La falta de apoyo generalizada por parte del personal sanitario las semanas siguientes a que una mujer haya padecido un aborto involuntario ha causado que éstas se vean obligadas a buscar el apoyo y la información necesarias en la comunidad online. Por este motivo se ha observado un aumento de actividad en redes sociales, donde las mujeres han empezado a compartir sus experiencias tras haber sufrido un aborto espontáneo. En este proyecto se han extraído datos de Twitter relacionados con los efectos psicológicos que puede haber causado un aborto involuntario para una mujer, para a continuación aplicar un análisis de sentimiento y una extracción de temas. Con alrededor de 19,000 tweets obtenidos se ha podido determinar si existe una relación entre el hecho de sufrir un aborto espontáneo y desarrollar efectos negativos en la salud mental, tales como depresión o ansiedad, además de comprender en profundidad la conversación alrededor de la pérdida del embarazo. Finalmente, se han representado los resultados obtenidos en una visualización adecuada para su fácil comprensión y divulgación.

Palabras clave: Análisis de sentimiento, Extracción de temas, Twitter, Aborto, Salud mental

The generalized lack of support from healthcare professionals in the posterior weeks after a woman has suffered a miscarriage has caused women to seek the needed support and information in the online community. For this reason, an increase in activity on social networks has been observed, where women have begun to share their experiences after having suffered a miscarriage. The aim of this project has been to extract data from Twitter related to the psychological effects that a miscarriage may have caused a woman, to then apply sentiment analysis and topic extraction. With around 19,000 extracted tweets it has been possible to determine if a relationship between having an abortion and suffering a negative impact on mental health, such as anxiety or depression, exists. In addition, the results have allowed an in-depth comprehension of the conversation around pregnancy loss. Finally, the results have been represented in a suitable visualization for easy comprehension and divulgation.

Keywords: Sentiment Analysis, Topic Extraction, Twitter, Miscarriage, Mental health

Índice

Listado de Figuras

Listado de Tablas

Listado de Códigos

Capítulo 1

Introducción

1.1. Contexto y justificación del trabajo

Un aborto involuntario es la razón más común para perder un bebé durante el embarazo. Las estimaciones varían, pero la organización March of Dimes, basada en la salud materna e infantil, indica que alrededor del 15 % de todos los embarazos acaban en un aborto involuntario [2].

Aun así, la pérdida de un bebé a través de un aborto de estas características, ya sea en el primer trimestre del embarazo o más adelante, sigue siendo un tema tabú alrededor del mundo, ya que está ligado al estigma y la vergüenza, tal y como indica la Organización Mundial de la Salud (WHO) [3].

Existe extensa información clínica sobre el diagnóstico y tratamiento de un aborto espontáneo, pero las implicaciones psicológicas no son tan comúnmente discutidas. Aun así, recientes investigaciones han identificado la presencia de trastornos psicológicos en mujeres que han padecido un aborto, más en concreto ansiedad, depresión [4] y síndrome de estrés postraumático [5].

Aunque diversas asociaciones de obstetricia recomiendan realizar un seguimiento en la salud mental de las mujeres que hayan sufrido un aborto [6], éstas expresan no haberse sentido apoyadas e informadas por parte de los profesionales sanitarios [7], como se indica en esta encuesta.

Por este motivo, las mujeres se ven obligadas a buscar formas de apoyo alternativas, como por ejemplo en la comunidad online. Un estudio cualitativo de dos foros online que tratan sobre el aborto recurrente ha mostrado que las mujeres acuden a las comunidades online y a las redes sociales, tanto para obtener información sobre el tema, como apoyo emocional tras la pérdida del embarazo [8].

Los autores del estudio sugieren que existen redes sociales que pueden ser una fuente de

información sin explotar sobre experiencias de las mujeres que han sufrido un aborto. Ante este contexto nace la idea de este trabajo: extraer datos de Twitter relacionados con los efectos psicológicos causados por un aborto involuntario, aplicar un análisis de sentimiento sobre ellos, extraer los temas de los que hablan y analizar diferentes aspectos demográficos, para finalmente obtener unos resultados que serán representados en las visualizaciones más adecuadas para su comprensión y divulgación.

1.2. Motivación personal

La elección de la temática para este trabajo reside en mi interés personal en la salud mental, especialmente centrada en problemas que padecen las mujeres. Desde mi punto de vista, existe un estigma y falta de visibilidad en problemas típicamente asociados a las mujeres, como por ejemplo el aborto, las complicaciones y secuelas del embarazo, los problemas asociados a la menstruación y muchos más.

Por ese motivo considero extremadamente importante la labor de este trabajo, que busca informar y dar visibilidad a un problema que afecta a muchas mujeres hoy en día: los problemas de salud mental asociados a un aborto involuntario.

Por otro lado, considero que la parte técnica asociada a este trabajo me aportará una experiencia muy valiosa profesionalmente, ya que las redes sociales son una fuente de información en auge, y es importante saber cómo explotar los datos disponibles.

1.3. Objetivos del trabajo

El objetivo principal de este trabajo es el de identificar la aparición de problemas de salud mental, como ansiedad o depresión, en usuarios de Twitter que hablen sobre sus experiencias reales de aborto involuntario, y caracterizar de qué hablan estos usuarios para comprender mejor los problemas reales alrededor de la pérdida del embarazo.

Con tal de conseguir este objetivo principal, también se definen los siguientes objetivos secundarios:

- **Extraer datos** de Twitter mediante código desarrollado en Python.
- **Encontrar las palabras clave** óptimas para conseguir información relacionada con los efectos psicológicos de un aborto.
- **Preparar los datos** que se analizarán posteriormente, con técnicas de limpieza y transformación, para conseguir un conjunto de datos apto para su explotación en un análisis de sentimiento.

- **Realizar un análisis de sentimiento** sobre los datos obtenidos, obteniendo tanto la polaridad como subjetividad de los tweets obtenidos.
- **Extraer los temas** de los que hablan los tweets obtenidos para comprender de qué hablan los usuarios que mencionan el aborto espontáneo.
- **Analizar otros aspectos relevantes** de los datos obtenidos, como qué día de la semana y a qué hora se ha publicado el tweet.
- **Crear visualizaciones** que ayuden a mostrar los resultados obtenidos en el análisis.

1.4. Metodología

A través de los datos extraídos de Twitter se quiere obtener conocimiento sobre los posibles efectos psicológicos en mujeres que hayan sufrido un aborto, empezando por extraer los datos de Twitter, aplicando un análisis de sentimiento y mostrando los resultados en una visualización de datos adecuada.

Por lo tanto, se trata de un proyecto de minería de datos en el que se desarrollará un producto nuevo (los resultados del análisis y su visualización), así que podemos usar la metodología Cross Industry Standard Process for Data Mining (CRISP-DM), que actualmente es la metodología estándar en proyectos de estas características [9].

La metodología CRISP-DM imita la naturaleza cíclica que presentan los proyectos de minería de datos, ya que éstos rara vez acaban una vez se despliega una solución. De esta forma, el ciclo de vida de la metodología CRISP-DM consiste en 6 fases diferentes, que se pueden observar en la Figura 1.1. Como se observa en la figura, la secuencia de fases no es rígida, y moverse entre fases hacia delante y atrás siempre es requerido.

En nuestro proyecto se aplicarán las 6 fases de la metodología de la siguiente manera:

1. **Comprensión del negocio**: la fase inicial se centra en definir el alcance del proyecto, definiendo los objetivos y requisitos, convirtiendo el problema inicial en un proyecto de minería de datos estructurado. El resultado de esta primera fase para este proyecto se encuentra en el capítulo 1 del presente documento.
2. **Comprensión de los datos**: esta fase tiene como objetivo la obtención y comprensión de los datos a usar en el proyecto. Por lo tanto, en este caso se creará un código en Python que permita la extracción de datos de Twitter dadas una serie de palabras clave y a continuación se realizará una exploración preliminar de los datos para comprender sus principales características.

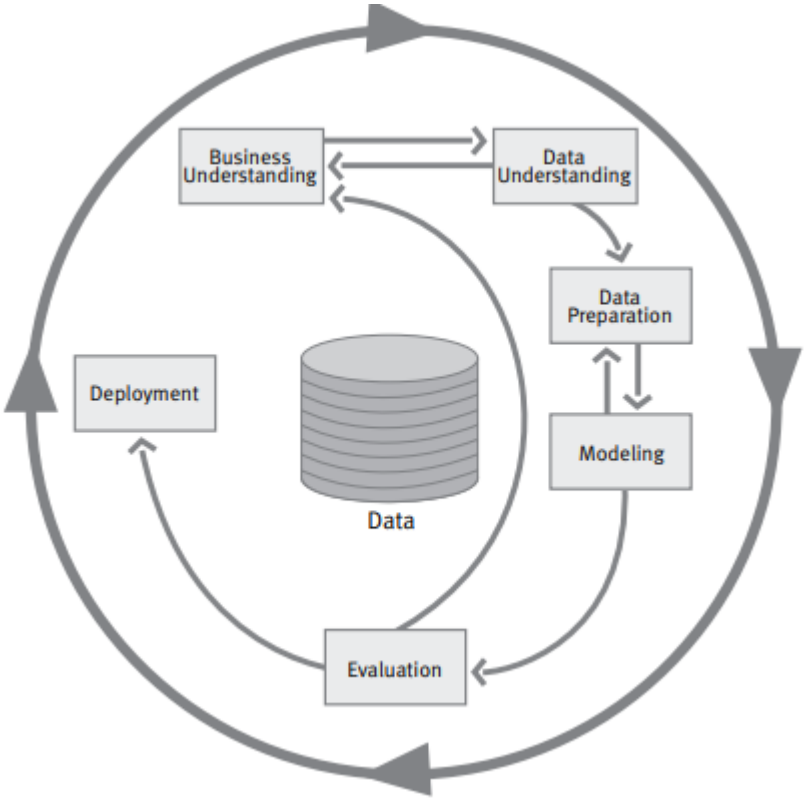


Figura 1.1. Fases de CRISP-DM

3. **Preparación de los datos**: una vez obtenidos los datos, el siguiente paso será su explotación mediante un modelo. Por lo tanto, los datos deberán pasar por un proceso de preparación para ser adecuados para ese proceso. En este caso, ya que los datos que se van a obtener de Twitter serán textos, deberemos aplicar técnicas de procesamiento del lenguaje natural (NLP) [10], además de técnicas comunes de limpieza y preparación de datos.

4. **Modelado**: se quiere realizar un análisis de sentimiento en los datos extraídos, obteniendo de esta forma tanto la polaridad como subjetividad de los tweets obtenidos, información que nos permitirá lograr nuestro objetivo principal.

5. **Evaluación del modelo**: con los resultados obtenidos del análisis de sentimiento se debe comprobar que se están logrando los objetivos definidos en el alcance del proyecto. Se deben evaluar la calidad del modelo y de los resultados obtenidos antes de proceder a la fase de despliegue final.

6. **Despliegue**: normalmente la obtención de un modelo y del análisis de los datos no es el final del proyecto, ya que el conocimiento obtenido debe ser representado de una forma comprensible para el consumidor. Por lo tanto, con los resultados obtenidos de

nuestro análisis de sentimiento se crearán visualizaciones de datos adecuadas para su fácil comprensión y divulgación.

1.5. Planificación del trabajo

En el siguiente **diagrama de Gantt** (Figura 1.2) se pueden observar las diferentes fases del proyecto, juntamente con sus tareas y las fechas exactas de su duración.

El código de color indica las tareas dedicadas a la definición del proyecto (**azul**), al análisis del estado del arte (**rosa**), a las tareas relacionadas con el desarrollo del proyecto (**verde**), a la creación de visualizaciones (**naranja**) y a la finalización de la memoria y la defensa (**amarillo**).

En el diagrama también se pueden observar las fechas exactas en las que se deben entregar las diferentes PACs para el seguimiento de proyecto, representadas como rombos amarillos.

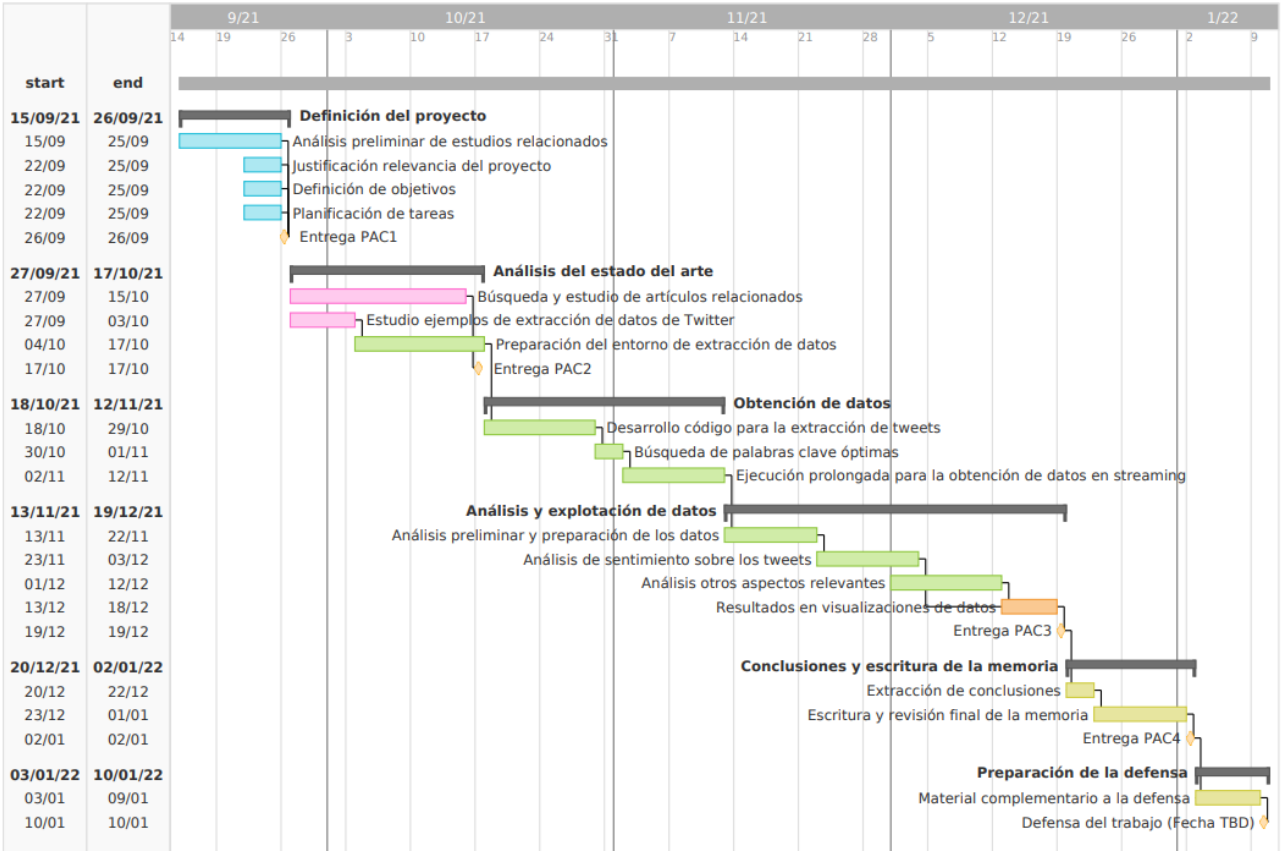


Figura 1.2. Diagrama de Gantt de la planificación del proyecto

Capítulo 2

Estado del arte

Las redes sociales, como por ejemplo Twitter o Instagram, se han convertido en una herramienta de comunicación muy popular entre los usuarios de Internet en los últimos años. Experiencias típicamente consideradas privadas, como son los temas relacionados con la salud mental u otro tipo de enfermedades, son cada vez más discutidas y compartidas en las redes sociales. Estos datos pueden proporcionar información de primera mano sobre las experiencias personales vividas y las implicaciones en la salud mental de las personas que hayan publicado sobre el tema.

Por lo tanto, en las redes sociales se encuentran una gran cantidad de datos textuales sin procesar que contienen las opiniones y experiencias sobre la vida diaria de sus autores. Los textos de las redes sociales son una fuente de información muy valiosa para la minería y análisis de datos, ya que permiten la investigación de una gran variedad de temas, y más en concreto nos permiten centrarnos en el ámbito de la medicina y la salud mental.

Esto implica que son múltiples los trabajos de investigación que se han publicado usando datos de redes sociales para analizar y explotar distintas cuestiones de interés en el ámbito de la medicina, y en esta sección se analizarán en concreto los relacionados con el aborto espontáneo.

2.1. Estudios similares

Encontramos una gran variedad de investigaciones que usan datos extraídos de Twitter u otras redes sociales para analizar las implicaciones psicológicas de algunas enfermedades. Si realizamos una búsqueda en PubMed con los términos “Social Networking” y “Mental Health”, podemos observar un crecimiento en el número de trabajos publicados en los últimos años (Figura 2.1).

Por ejemplo, en este reciente artículo de Ramírez-Cifuentes D. y colaboradores [11], se han usado tweets para caracterizar a los usuarios hispanohablantes de Twitter que muestran signos

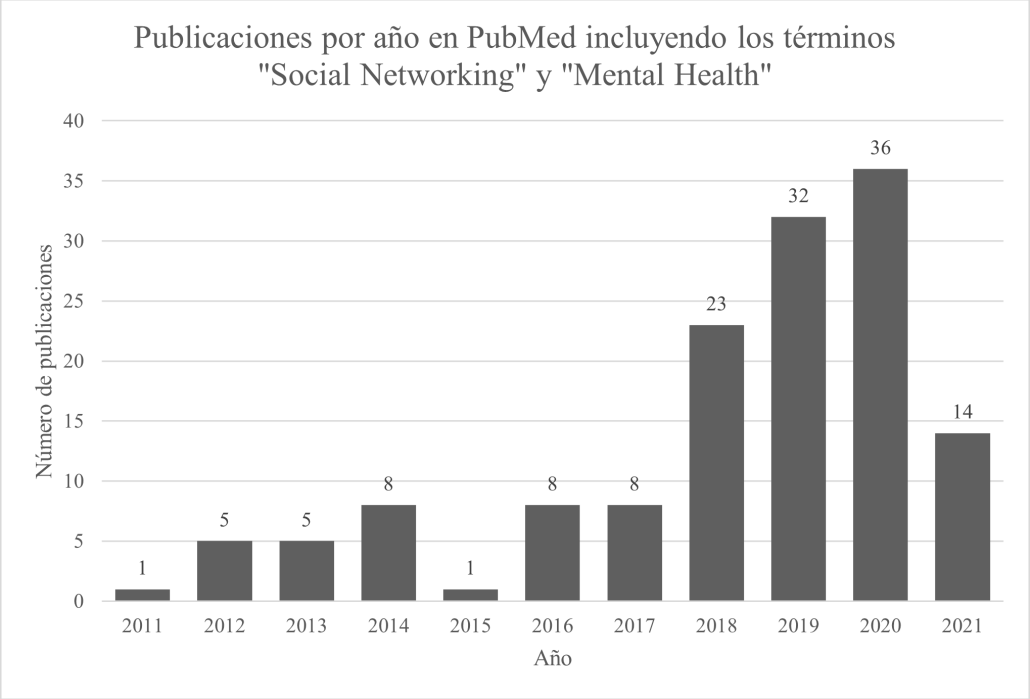


Figura 2.1. Recuento de publicaciones en PubMed incluyendo los términos Social Networking y Mental Health

de anorexia. En esta investigación se han analizado los textos, patrones de publicación, relaciones sociales e imágenes de usuarios que pasaron por diferentes etapas de anorexia nerviosa. En este estudio se compararon dos grupos de tweets, uno de usuarios que estuviesen sufriendo de anorexia nerviosa, y por otro lado un grupo de control de usuarios elegidos aleatoriamente. Con estos grupos de usuarios se usaron técnicas de aprendizaje no supervisado para diferenciar usuarios que se encontrasen al comienzo de la enfermedad, de usuarios ya más adentrados en la enfermedad, y se generaron modelos para la detección de tweets e imágenes relacionadas con la anorexia nerviosa.

Otro ejemplo similar sería el trabajo de Leis A. y colaboradores [12]. En este estudio se extrajeron características lingüísticas y patrones de comportamiento de usuarios de Twitter que pudieran mostrar posibles signos de depresión. Para realizar este estudio se compararon y analizaron 3 grupos de tweets diferentes: un conjunto de datos de usuarios depresivos, un conjunto de tweets escogidos manualmente escritos por los usuarios depresivos, y por último un grupo de control de un grupo de usuarios elegidos aleatoriamente. Los resultados de este estudio mostraron que los usuarios depresivos se expresan de forma diferente, tanto en su forma de escribir tweets como en su interacción con las redes sociales.

Otros ejemplos significativos son el trabajo de Leis A. y colaboradores [13] sobre el análisis de los cambios lingüísticos y en el comportamiento de personas pasando por un tratamiento contra

la depresión usando datos de Twitter, el trabajo de Ramírez-Cifuentes D. y colaboradores [14] sobre la detección de planeamiento del suicidio en las redes sociales, o el trabajo de Ferrara E. y Yang Z. [15] sobre la medición del contagio emocional en las redes sociales.

Si nos centramos en el tema del aborto, la literatura disponible disminuye considerablemente. Es importante diferenciar los trabajos que tratan sobre la legislación del aborto de los que tratan sobre las implicaciones psicológicas de un aborto espontáneo en las mujeres. Sobre el primer caso encontramos el trabajo de Graells-Garrido E. y colaboradores [16], que tiene como objetivo el análisis de la representación del debate sobre el aborto en Twitter, en los países Chile y Argentina. Este trabajo se centra en comprender cómo los grupos demográficos se ven representados en las discusiones en las plataformas de micro-blogging, usando como tema de discusión la legislación del aborto en estos dos países, pero no realiza un análisis del impacto psicológico del aborto en las mujeres.

Por otro lado, si nos centramos en analizar las implicaciones psicológicas de un aborto espontáneo, encontramos el trabajo de Mercier R.J. y colaboradores [17], que trata sobre un estudio cualitativo sobre las experiencias compartidas sobre el aborto espontáneo en publicaciones de la red social Instagram. Para la realización del estudio se extrajeron 200 publicaciones aleatorias que incluyeran descripciones de experiencias personales sobre el aborto espontáneo y que usaran el hashtag #ihadamiscarriage. Estas publicaciones se analizaron usando un análisis de contenido dirigido que fue llevado a cabo por un equipo multidisciplinar, que identificó temas comunes entre las publicaciones, codificando así todas las publicaciones y creando un documento con el significado de las diferentes codificaciones. En los resultados de este estudio se observaron emociones complejas y a menudo conflictivas alrededor del aborto, que en muchas usuarias duraron durante meses e incluso años. Como conclusión se obtuvo que las mujeres publican sobre sus abortos por una variedad de razones, pero especialmente para encontrar apoyo en la comunidad en línea, para romper el tabú alrededor del tema del aborto y como mecanismo para lidiar con la pérdida del embarazo.

Este estudio se basa en un estudio previo [8], que analizó dos foros online sobre la pérdida del embarazo recurrente y que descubrió que las mujeres normalmente usan las comunidades en línea tanto para obtener información como para obtener apoyo emocional en la situación del aborto espontáneo. Los autores de esta investigación señalaron que otras plataformas de redes sociales convencionales pueden ser una fuente de datos sin explotar sobre la experiencia del paciente con un aborto espontáneo.

Por lo tanto, ya que el estudio existe en la red social Instagram, el siguiente paso es analizar los estudios similares en la red social Twitter, en la que se basa el presente proyecto. Para este caso encontramos dos artículos significativos, muy recientes y centrados en la comunidad angloparlante.

El trabajo de Cesare N. y colaboradores [18], publicado en 2019, tiene el objetivo de caracterizar a los usuarios que hablan sobre el tema del aborto espontáneo y los partos prematuros en Twitter, encontrando las tendencias y los factores impulsores de estos sucesos, y analizando el impacto emocional descrito por las mujeres que han publicado sus experiencias. Este estudio se realizó extrayendo alrededor de 300.000 tweets sobre el aborto espontáneo y usando el algoritmo Latent Dirichlet Allocation (LDA) para identificar los temas principales discutidos. A parte del uso de este algoritmo, también se realizó trabajo manual para categorizar una parte de los tweets en una de las siguientes categorías autoexcluyentes, indicando el sentimiento detrás de cada publicación: dolor/tristeza/depresión, ira, alivio, aislamiento, preocupación y neutral. Los resultados de esta investigación encontraron 8 temas principales en los tweets analizados, y se detectó que las características más comunes obtenidas del análisis de sentimiento fueron el dolor y la preocupación, estando presentes en alrededor de la mitad de las publicaciones. Por lo tanto, se pudo comprobar que existe un posible impacto físico, psicológico y emocional sobre las mujeres que han sufrido un aborto espontáneo.

Por otro lado, también encontramos el trabajo de Klein A.Z. y colaboradores [19], publicado en 2020, que tuvo dos objetivos principales. En primer lugar, se quiso evaluar si las mujeres publican sobre el aborto espontáneo, pérdida del embarazo y parto prematuro, entre otros, en Twitter, y por otro lado, se quiso desarrollar un método de procesamiento del lenguaje natural que pudiera detectar automáticamente usuarias con este perfil para realizar estudios observacionales a gran escala. En esta investigación se usó una base de datos de 400 millones de tweets públicos sobre usuarios que han anunciado su embarazo en Twitter. Se crearon expresiones regulares concretas para poder extraer de este conjunto de tweets los que estuvieran hablando de complicaciones en el resultado del embarazo, y adicionalmente dos anotadores clasificaron manualmente una pequeña parte de los tweets entre los que anunciaban el resultado del embarazo y los que no. Los resultados de este estudio muestran que las mujeres sí hablan de sus experiencias adversas en el embarazo en Twitter, y por otro lado, que las técnicas de NLP creadas en el estudio son capaces de detectar automáticamente usuarios para usar en estudios observacionales a gran escala.

Estos trabajos nos muestran el potencial que tienen los datos extraídos de las redes sociales, y en concreto de Twitter, para el análisis de aspectos relevantes de la medicina, en los que es muy interesante analizar la percepción y las experiencias de los pacientes para comprender mejor las implicaciones psicológicas de algunos sucesos o enfermedades.

2.2. Tecnologías para el análisis de datos de Twitter

El presente proyecto se basa en la aplicación de técnicas de análisis de sentimiento sobre datos textuales obtenidos de Twitter. La tecnología escogida para el desarrollo del proyecto será Python, así que se pueden analizar las técnicas más actuales para la extracción de datos de Twitter, para el procesamiento de lenguaje natural (NLP) y para el análisis de sentimiento de un texto.

En primer lugar, los datos de Twitter se pueden extraer a través de la API de Twitter. A través de un software, que en nuestro caso será código de Python, nos podemos conectar a Twitter a través de un punto de conexión (endpoint) de la API. Los datos de Twitter son diferentes respecto a las otras redes sociales, ya que reflejan información que los usuarios han decidido compartir públicamente. Por lo tanto, con la API de Twitter podemos acceder solamente a los datos públicos de Twitter, como los tweets o la información de un usuario público.

Con tal de usar la API de Twitter, en primer lugar se debe registrar una aplicación en el portal del desarrollador de Twitter, que nos generará las credenciales necesarias para obtener datos de Twitter. Existen diversas herramientas para obtener datos de Twitter a través de su API [20], pero en nuestro caso analizaremos las principales librerías de Python que nos permiten realizar esta tarea. Las 4 librerías soportadas por Twitter para Python son TwitterAPI, python-twitter, Tweepy y Twython. Estas 4 librerías permiten el acceso a datos de Twitter a través de OAuth (Open Authentication) y permiten obtener datos de forma normal o por streaming a través de la API de Twitter. Aun así, la librería más popular (basado en su valoración en Github) es Tweepy, que destaca por su facilidad de uso y su buena documentación, además de ser una de las librerías más actualizadas.

Para la tarea de procesamiento del lenguaje natural, incluyendo las técnicas de procesado del texto y de análisis de sentimiento, las principales librerías que se usan en Python son NLTK (Natural Language Tool Kit), TextBlob y SpaCy, mientras que para la extracción de temas existen modelas en las librerías scikit-learn y Gensim.

La librería NLTK, que es la más conocida para problemas de procesamiento del lenguaje natural y análisis de sentimiento, normalmente se recomienda para su uso en la investigación y el aprendizaje, pero no está optimizada para entornos en producción. Por otro lado, TextBlob es una librería construida encima de NLTK, y es más accesible y fácil de usar a cambio de no tener un rendimiento altamente optimizado. Esta librería es normalmente la más recomendada para realizar estudios de subjetividad y polaridad en un análisis de sentimiento de un texto. Por último, la librería SpaCy está diseñada para ser muy rápida y usada en entornos de producción, pero no está tan recomendada para problemas de análisis o investigación.

Capítulo 3

Desarrollo

Como se menciona en el apartado 1.4 del presente documento, la metodología usada para el desarrollo de este proyecto ha sido CRISP-DM [9]. Por lo tanto, los apartados que se presentan a continuación para describir el desarrollo del proyecto siguen las fases definidas por esta metodología.

En concreto, en este apartado se describen los pasos para la obtención y comprensión de los datos, su preparación y transformación, el modelado de un análisis de sentimiento y de una extracción de temas, y finalmente la evaluación de los modelos mencionados.

3.1. Extracción de los datos de Twitter

3.1.1. API de Twitter, limitaciones y proceso de streaming

Los datos usados para este estudio han sido extraídos de la red social Twitter, que pone a disposición de los desarrolladores una API que permite obtener tweets en tiempo real con algunas limitaciones definidas [21].

La conexión a la API de Twitter se ha realizado mediante el código de Python desarrollado para este proyecto, donde se ha realizado una conexión a Twitter a través de un punto de conexión (*endpoint* en inglés) de la API. Como se ha comentado en el presente estado del arte, existen diferentes librerías de Python que facilitan la conexión y extracción de datos mediante la API de Twitter. En este proyecto se ha usado la librería Tweepy, que destaca por su facilidad de uso y buena documentación, además de ser la librería más recientemente actualizada [22].

Para poder extraer los datos de Twitter se requiere la creación de una aplicación en el portal de desarrollador de Twitter, que generará las credenciales necesarias para conectarse a Twitter y obtener datos [23].

Existen dos formas de extraer datos de la API de Twitter, usando lo que se conoce como

Search API, que permite la extracción de tweets que ya han ocurrido en una sola búsqueda, o usando la **Streaming API**, que permite abrir un canal de comunicación constante con Twitter y extraer tweets que están ocurriendo en tiempo real.

Los dos métodos se encuentran limitados al usar la versión gratuita de la API de Twitter. En el caso de usar la Search API, la limitación solamente permite obtener los 5,000 últimos tweets dada una palabra de búsqueda, además de tener una limitación de 180 peticiones cada 15 minutos. Por otro lado, la Streaming API, aunque extraiga datos a tiempo real también presenta limitaciones, ya que Twitter no tiene la capacidad de devolver al momento todos los tweets que se generan con una palabra de búsqueda concreta. La cantidad de tweets recibida respecto a la original se ve afectada por el tráfico que esté experimentando Twitter, por lo que en un proceso de streaming se espera obtener entre un 1 % y un 40 % de los tweets generados en tiempo real [24].

Aun así, la extracción de datos con streaming permite acercarse más al comportamiento real de los usuarios que están generando contenido en tiempo real, así que es la opción que se ha usado en el desarrollo de este proyecto.

3.1.2. Formato de los datos de Twitter

Los datos de Twitter son diferentes respecto a los de otras redes sociales, ya que reflejan información que los usuarios han decidido compartir públicamente. Por lo tanto, con la API de Twitter solamente se puede acceder a los datos públicos de Twitter, como los tweets o la información de un usuario público.

En el proyecto actual, solamente se han extraído datos de tweets, que es el bloque atómico básico de todo lo que forma Twitter. El objeto Tweet tiene una larga lista de atributos en su raíz, como por ejemplo la fecha o el identificador del tweet, pero también contiene objetos más complejos en su interior, como un objeto usuario (User) o entidades (Entities). En concreto, se han extraído tweets originales y respuestas, pero se han excluido los retweets durante el proceso de *streaming*.

En la plataforma de desarrollador de Twitter se puede encontrar la lista completa de elementos dentro de un Tweet [25]. Por lo tanto, ya que un tweet contiene elementos complejos en su interior, su representación se da en un formato de dato semiestructurado. En concreto, toda la información obtenida al recibir información de la API de Twitter viene en formato JSON [26].

A continuación podemos ver un ejemplo de un Tweet real (Figura 3.1) y su renderización en formato JSON (Código 3.1), viendo algunos de los atributos raíz (solamente los campos más fundamentales del tweet) y los objetos hijos en su interior (denotados con la notación {}):

Figura 3.1. Ejemplo de Tweet

Código 3.1. Renderización del tweet de la Figura 3.1 en formato JSON

```
{
"created_at": "Fri Oct 15 18:49:13 +0000 2021",
"id": 1449084965187112967,
"id_str": "1449084965187112967",
"text": "We remember them today and everyday #WaveOfLight #
   BabyLossAwarenessWeek https://t.co/ypUY8UZXYx",
"user": {},
"entities": {}
}
```

En el siguiente apartado se especifican en más detalle los campos que se han seleccionado para el análisis del proyecto actual.

3.1.3. Fechas en que se realizó el streaming

Se han seleccionado dos períodos diferentes para la extracción de datos de Twitter en tiempo real. En primer lugar, el 15 de octubre de cada año se celebra el "Pregnancy And Infant Loss

Rememberance Day", que pretende dar visibilidad y concienciar sobre la realidad de muchas mujeres que sufren de problemas en el embarazo, tales como el aborto espontáneo o el embarazo ectópico entre otros [27].

La celebración anual de este día genera diferentes iniciativas, como el "Pregnancy and Infant Loss Awareness Month" en Estados Unidos, o la "Baby Loss Awareness Week" en el Reino Unido. Durante este período también se celebran encendidas de velas, conocidas como "Wave of Light" [28] juntamente con otras iniciativas orientadas a crear conciencia sobre la prevalencia de la pérdida del embarazo y la muerte infantil.

Dado que es una fecha en la que se genera mucha conversación alrededor de la pérdida del embarazo, el primer streaming se inició el día 15 de octubre de 2021, que era un viernes, y estuvo obteniendo tweets hasta el final del fin de semana (domingo 17 de octubre de 2021).

La palabra clave usada para la extracción de tweets en este período fue un hashtag sobre la semana de concienciación de la pérdida de un bebé: **#BabyLossAwarenessWeek**. Este fue el hashtag principal usado para la celebración de esta semana de concienciación, así que a través de él se pueden obtener un gran número de tweets relacionados con el aborto espotáneo.

En segundo lugar, también se quiso analizar una semana aleatoria sin ningún evento relacionado con la pérdida del embarazo, para observar la conversación del día a día sobre el tema y comparar las diferencias entre los tweets publicados en los dos periodos de tiempo. Por lo tanto, el segundo período de streaming se llevó a cabo del 27 de octubre al 4 de noviembre de 2021.

Las palabras clave que se usaron para este segundo streaming fueron: **miscarriage**, **baby loss** y **pregnancy loss** (aborto espontáneo, pérdida del bebé y pérdida del embarazo). En este caso, la selección de palabras clave se basó en los artículos mencionados en el presente estado del arte, en los que se han extraído datos de otras redes sociales con el fin de caracterizar a los usuarios que hablan del aborto espontáneo con estas palabras clave principalmente [18][19].

Dado que los dos conjuntos de datos van a ser mencionados de forma continuada en el presente documento, el primer conjunto de tweets será mencionado como datos de "Awareness" y el segundo conjunto de tweets será mencionado como datos de "Streaming".

Finalmente, al acabar el proceso de recolección de tweets de los dos conjuntos de datos se han obtenido **7,433** tweets el conjunto de Awareness y **11,747 tweets** para el conjunto de Streaming.

3.2. Transformación de los datos

3.2.1. Paso de datos semiestructurados a estructurados

Los datos extraídos directamente de Twitter se encuentran en un formato semiestructurado, más en concreto en formato JSON. Con tal de poder analizar y usar estos datos en un modelo se deben convertir a datos estructurados, es decir, a formato tabular.

De los objetos complejos que se encuentran dentro del objeto Tweet, como el usuario, las entidades y la localización solo son necesarios algunos campos en concreto, así que se ha realizado una transformación para aplanar los datos actuales (transformación normalmente conocida como *flatten*).

De esta forma se consigue transformar unos datos con diferentes niveles de profundidad a unos datos con un solo nivel en la raíz. Por ejemplo, todos los campos dentro del objeto complejo usuario saldrán un nivel hacia afuera y pasarán a tener el prefijo "user" delante del nombre del campo, como se puede observar en el ejemplo de la Figura 3.2.

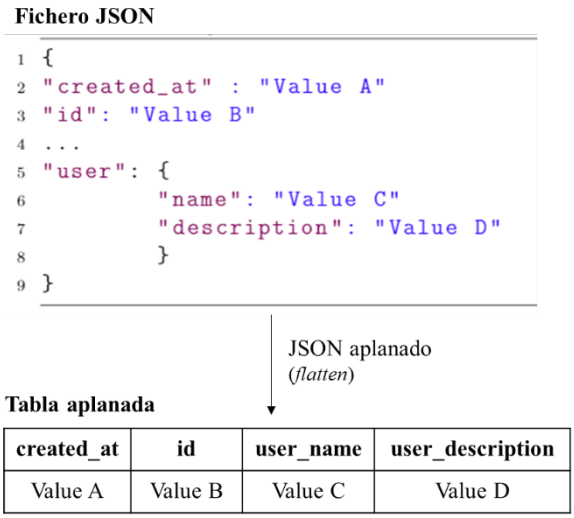


Figura 3.2. Ejemplo de aplanado de datos en formato JSON

3.2.2. Selección de columnas

Una vez aplanados los datos, el número de columnas ha crecido y se pueden observar un gran número de campos que no son necesarios para el análisis y modelado a realizar. Por lo tanto, en la Tabla 3.1 se listan los campos que se han seleccionado para este proyecto, juntamente con una breve explicación de lo que contienen. El ejemplo mostrado en la tabla proviene del tweet de la Figura 3.1.

Ya que los datos han sido aplanados, muchos de estos campos contienen el prefijo "user", "extended_tweet" (una de las entidades) o "place".

Tabla 3.1. Lista de columnas seleccionadas

Nombre del campo	**Tipo**	**Contenido**	**Ejemplo**
created_at	Fecha y hora	Fecha y hora de creación del tweet	Fri Oct 15 18:49:13 +0000 2021
id	Numérico	ID del tweet	1449084965187112967
id_str	Texto	ID del tweet	1449084965187112967
text	Texto	Texto del tweet	We remember them today and everyday #WaveOfLight #BabyLossAwarenessWeek
truncated	Booleano	Indica si el tweet está truncado o no, ya que ahora existen tweets de más de 40 caracteres que son truncados	False
source	Texto	Fuente de publicación (móvil, ordenador)	Twitter for Android
lang	Texto	Idioma del tweet (detectado automáticamente)	en
in_reply_ to_status_id	Numérico	Si es una respuesta, el ID del tweet al que responde, si no es Null	NaN (no es una respuesta)
user_id	Numérico	ID del usuario	212646190
user_id_str	Texto	ID del usuario	212646190
user_name	Texto	Nombre completo del usuario	Vivian Maeda
user_screen_ name	Texto	Nombre de usuario (el que aparece con la @ normalmente)	blippingbird
user_location	Texto	Localización del perfil del usuario	Scotland
user_url	Texto	URL del perfil del usuario	https://t.co/upueVKCalP
user_description	Texto	Descripción del perfil del usuario	Proud mum of Luca and Matti-8 x marathons
user_verified	Booleano	Indica si el usuario está verificado o no	False
user_followers_ count	Numérico	Recuento de seguidores del usuario	1672
user_friends_ count	Numérico	Recuento de usuarios a los que sigue el usuario	2242

user_favourites_count	Numérico	Recuento de tweets favoritos	20013
user_statuses_count	Numérico	Recuento de tweets escritos y retweeteados por el usuario	15850
user_created_at	Fecha y hora	Fecha y hora de creación del usuario	Sat Nov 06 17:09:57 +0000 2010
user_geo_enabled	Booleano	Indica si la geolocalización está activada para este usuario	True
user_lang	Texto	Lenguaje que ha especificado el usuario	en
place_id	Texto	ID representativo de un lugar	NaN
place_url	Texto	URL representando el lugar o metadatos del mismo	NaN
place_place_type	Texto	Tipo de localización	Country
place_name	Texto	Representación corta del nombre del sitio	NaN
place_full_name	Texto	Representación larga del nombre del sitio	NaN
place_country_code	Texto	Código corto del país en el que se encuentra el lugar	NaN
place_country	Texto	Nombre del país en el que se encuentra el lugar	NaN
place_bounding_box_coordinates	Texto	Coordenadas que forman una caja que contiene el lugar	NaN
extended_tweet_full_text	Texto	Si el tweet está truncado, el texto completo del tweet	NaN
extended_tweet_entities_hashtags	Lista de textos	Lista de hashtags obtenidos del tweet	[‘#WaveOfLight’, ‘#BabyLossAwarenessWeek’]
extended_tweet_entities_urls	Lista de textos	Lista de URLs que aparecen en el tweet	[ ]
extended_tweet_entities_media	Lista de objetos JSON	Lista de objetos multimedia que aparecen en el tweet (fotos, videos o gifs)	[{photo object}] (resumido, muy largo)

3.2.3. Transformaciones comunes

Se han aplicado las siguientes transformaciones sobre las columnas que no contienen el texto del tweet:

- **Filtrado de tweets que no estén en inglés**. Aunque en uno de los parámetros del proceso de *streaming* ya se ha especificado la búsqueda de tweets en inglés, ya que la

detección del idioma es normalmente automática, se ha realizado un nuevo filtrado para eliminar posibles tweets en otros idiomas.

- **Transformación de campos de fecha**. Como se ha podido observar en el análisis de los campos del apartado anterior, los campos de fecha vienen en un formato que no es adecuado para ser usado directamente, así que en este proceso se van a crear los campos 'day', 'month', 'day_of_week', 'hour' y 'date' para poder realizar un análisis temporal de los conjuntos de datos.

- **Obtención de columna indicando si el tweet es una respuesta**. Existe un campo que indica el identificador del tweet al que responde el tweet observado, que se encontrará vacío si el tweet actual no es una respuesta. A partir de esta columna se ha creado una nueva columna booleanda llamada 'is_reply', que indica si el tweet actual es una respuesta a otro tweet o no.

- **Extracción del tipo de contenido multimedia que contiene el tweet**. La lista de objetos que contiene toda la información sobre el contenido multimedia del tweet sigue siendo muy compleja, así que solamente se ha extraído el tipo de contenido multimedia, que puede ser foto, vídeo o gif, y se ha guardado esta información en las nuevas columnas 'has_photos', 'has_videos' y 'has_gifs'.

3.2.4. Transformaciones en el texto

Cuando se trabaja con texto es de vital importancia realizar un proceso de limpieza adecuado del mismo para poder explotar al máximo la información disponible. Por lo tanto, se han realizado un seguido de transformaciones sobre el campo que contiene el texto del tweet, que será el que usaremos para el modelado.

El objetivo final de esta transformación es pasar de tener un texto crudo, con signos de puntuación, nombres de usuario, hashtags y muchos otros elementos no aprovechables, a tener una lista de palabras útiles para el análisis de sentimiento y la extracción de temas.

Dado que dependiendo de si un tweet está truncado o no el texto se encuentra en un campo diferente, la primera transformación ha sido agrupar todos los textos en un mismo campo. Una vez obtenido el texto de cada tweet en un mismo campo se han realizado las siguientes transformaciones, algunas comunes al procesar cualquier tipo de texto, y otras específicas al procesar texto de un tweet:

1) Conversión de todo el texto a minúscula

Ya que se quieren identificar como iguales las mismas palabras, tanto si están en minúscula como en mayúscula, el primer paso es la conversión de todo el texto a minúscula [29, Capítulo 2.2.3].

2) Eliminación de nombres de usuario (@username)

Los nombres de usuario que se encuentran dentro del texto del tweet contienen información que no se necesita para el tipo de análisis que se quiere realizar a posterior. Por lo tanto, en esta transformación se eliminan todos los nombres de usuario que puedan aparecer en el texto, que se encuentran precedidos siempre por un símbolo '@'.

3) Eliminación de hashtags (#hashtag)

Igual que en el caso anterior, los hashtags son tan diversos que no es posible automatizar su separación en diferentes palabras y no se puede aprovechar la información que contienen para el análisis posterior. Por lo tanto, también se han eliminado los hashtags del texto del tweet, que vienen siempre precedidos por el símbolo '#'.

4) Eliminación de hipervínculos (https://...)

Dentro del texto del tweet también aparecer hipervínculos, que siempre van precedidos por el conjunto de letras 'http'. Por lo tanto, los hipervínculos también se han eliminado del texto del tweet.

5) Eliminación de dígitos

En el tipo de análisis que se quiere realizar en este proyecto, los dígitos no aportan información para comprender el sentimiento o el tema del que trata un tweet. Por lo tanto, se han eliminado los dígitos del texto.

6) Eliminación de signos de puntuación y espacios extra

Los signos de puntuación, que se encuentran pegados a las palabras en un texto, pueden entorpecer la agrupación de palabras iguales en el proceso de modelado. Por ese motivo, en esta transformación se eliminan todos los signos de puntuación y espacios extra que se encuentren en el texto [29, Capítulo 2.2.3].

7) Eliminación de palabras de 1 carácter

Las palabras de un solo carácter normalmente representan conectores, que son palabras que no aportan información para comprender el sentimiento o el tema de un texto. Por lo tanto, se han eliminado estas palabras para no entorpecer el modelado posterior.

8) Tokenización

Dada una secuencia de palabras y un elemento de separación claro, el proceso de tokenización es la tarea de dividir la secuencia en lo que se conoce como tokens. En el caso de la tokenización de un tweet, se estará dividiendo el texto del tweet en palabras usando los espacios como elemento de separación [29, Capítulo 2.2.1].

Dado que se han eliminado los signos de puntuación en un paso previo a este, el proceso de tokenización será más sencillo, ya que no se deben tener en cuenta los apóstrofes en las contracciones que normalmente se usan en inglés.

A partir de este paso se pasa de tener un texto conteniendo todo el tweet, a tener una lista de palabras del tweet.

9) Eliminación de 'stopwords'

Encontramos que algunas palabras extremadamente comunes que aparecen en el texto no aportan valor a la hora del modelado. Estas palabras son comúnmente conocidas como 'stopwords' [29, Capítulo 2.2.2].

En las librerías usadas para el procesamiento del lenguaje natural, normalmente se incluye una lista de 'stopwords' del idioma inglés, por lo que se pueden eliminar estas palabras de la lista de palabras que se obtiene de un tweet.

10) Lematización

Por razones gramaticales, en los documentos aparecen diversas formas verbales de una palabra, o familias de palabras con palabras de significados parecidos. Es muy útil en el proceso de preparación de los datos extraer la raíz de este tipo de palabras para poder agruparlas como iguales en el modelado posterior [29, Capítulo 2.2.4].

Existen normalmente dos técnicas para la extracción de estas raíces: el truncado de palabras (*stemming*) y la lematización (*lemmatization*). En el proceso de *stemming* se refiere normalmente a un proceso heurístico crudo que corta los extremos de las palabras con el objetivo de intentar obtener la raíz de estas. En muchas ocasiones este proceso no puede conseguir la raíz de una palabra solamente truncando la misma, como por ejemplo en el caso de los verbos irregulares.

Por otro lado, el proceso de *lemmatization* se refiere a buscar correctamente la raíz de las palabras gracias al uso de vocabulario y análisis morfológico de las palabras, para eliminar de forma más inteligente la parte flexiva de las palabras y obtener una raíz más estable. Por ejemplo, si se quisiera obtener la raíz de la palabra 'saw' (verbo vió en inglés), el proceso de lemmatization extraería el verbo 'see' (verbo ver en inglés) como raíz, mientras que el proceso de stemming extraería una 's' como resultado.

Por lo tanto, ya que en el texto de los tweet se encuentra una variedad de lenguaje muy diversa, se ha aplicado un proceso de lematización sobre las palabras para obtener una lista con solamente la raíz de las mismas.

Al finalizar todas estas transformaciones, el resultado es una lista de las raíces de las palabras útiles extraídas del texto de los tweets, que es la que permitirá el modelado posterior.

3.3. Modelado y evaluación

En primer lugar se ha aplicado un análisis de sentimiento, para obtener tanto la polaridad como la subjetividad y comprender el sentimiento que deprenden los tweets que hablan sobre la pérdida del embarazo. A continuación, se ha aplicado una extracción de temas sobre los tweets para comprender mejor de qué están hablando los mismos y comprender mejor la conversación alrededor del aborto espontáneo.

3.3.1. Análisis de sentimiento

El análisis de sentimiento es una técnica de procesamiento del lenguaje natural (NLP) que se usa para determinar la polaridad y subjetividad de datos textuales. Este tipo de análisis se suele usar para detectar sentimiento en datos de redes sociales, entender la reputación de una marca o entender mejor a los usuarios, entre otras aplicaciones.

El análisis de sentimiento puede ser implementado con una de las dos opciones disponibles usando machine learning: con aprendizaje supervisado o aprendizaje no supervisado.

El análisis de sentimiento usando aprendizaje no supervisado implica usar un enfoque basado en reglas léxicas para analizar un texto, mientras que el enfoque supervisado usa un modelo de clasificación a partir de métodos tradicionales de machine learning. El uso de aprendizaje supervisado para crear un modelo de clasificación tiene la limitación de la gran cantidad de datos necesarios para el entrenamiento, ya que para crear un clasificador suficientemente potente para identificar el sentimiento de nuevos textos se necesitaría un gran conjunto de datos textuales.

Por este motivo, normalmente en un análisis de sentimiento con textos no enfocados en un tema muy concreto, se usan técnicas basadas en reglas de aprendizaje no supervisado. Como se ha comentado en el estado del arte, para Python existen principalmente dos librerías de código abierto para la aplicación de un modelo no supervisado de análisis de sentimiento sobre el texto: NLTK (VADER) y TextBlob.

La librería NLTK usa VADER (Valence Aware Dictionary and Sentiment Reasoner) [30], que es una librería para el análisis de sentimiento preconstruida, de código abierto y protegida bajo la licencia MIT [31]. Esta librería se basa en una lista de características léxicas (normalmente palabras) que se etiquetan como positivas o negativas según su significado semántico para calcular el sentimiento de un texto. Esta librería solamente permite realizar análisis de la polaridad, pero ya que está optimizada para datos de redes sociales está altamente indicada al usar datos de Twitter.

Por otro lado, la librería TextBlob [32] es una librería de procesamiento de datos simplificada que permite realizar un análisis tanto de polaridad como de subjetividad, pero su potencia no es tan alta al usar datos de redes sociales, ya que ignorará las palabras con las que no esté

familiarizada y los resultados del análisis de sentimiento no serán tan precisos.

3.3.1.1. Polaridad

La polaridad de un texto indica si este expresa un sentimiento positivo, negativo o neutral a partir de una puntuación flotante dentro del rango [-1, 1]. Las reglas para determinar el sentimiento de un texto a partir de su puntuación de polaridad son [33]:

- Sentimiento positivo: $polaridad >= 0{,}05$
- Sentimiento neutral: $(polaridad > -0{,}05)\ y\ (polaridad < 0{,}05)$
- Sentimiento negativo: $polaridad <= -0{,}05$

Se ha usado la librería NLTK para obtener la polaridad de cada tweet y se han obtenido 4 parámetros diferentes [33]:

- **Puntuación compuesta (*compound*)**: este valor es calculado a partir de la suma de las puntuaciones de polaridad para cada palabra del texto y normalizando el resultado final para estar entre -1 (extremo más negativo) y +1 (extremo más positivo). Esta métrica resulta muy útil en los casos en que se quiera obtener una representación unidimensional de sentimiento para una oración determinada.
- **Proporciones Positiva, Negativa y Neutral (*pos, neg y neu*)**: son las proporciones del texto que pertenece a cada categoría de sentimiento, así que la suma de estas tres puntuaciones debe sumar 1. Esta representación de la polaridad es muy útil para la analizar la proporción en la que un texto presenta un sentimiento semánticamente.

En el Código 3.2 se pueden observar varios ejemplos de los parámetros obtenidos al calcular la polaridad de una frase.

Código 3.2. Ejemplos de parámetros obtenidos del análisis de polaridad con NLTK

```
1 VADER is smart, handsome, and funny.
2 {'pos': 0.746, 'neu': 0.254, 'neg': 0.0, 'compound': 0.8316}
3
4 VADER is not smart, handsome, nor funny.
5 {'pos': 0.0, 'neu': 0.354, 'neg': 0.646, 'compound': -0.7424}
```

3.3.1.2. Subjetividad

La información textual existente se puede clasificar a grandes rasgos en dos tipos: hechos y opiniones. Los hechos son expresiones objetivas sobre entidades y eventos, mientras que las opiniones son normalmente expresiones subjetivas que hablan de sentimientos o valoraciones sobre las entidades y eventos.

Por lo tanto, el análisis de subjetividad de un texto indica el nivel de subjetividad que este presenta, usando una puntuación flotante dentro del rango [0,1], donde 0 indica un texto muy objetivo y 1 indica un texto muy subjetivo. En este caso, se ha usado la librería TextBlob para obtener las puntuaciones de subjetividad de cada tweet [32].

En el Código 3.3 se pueden observar varios ejemplos de la subjetividad obtenida al analizar un texto.

Código 3.3. Ejemplos del parámetro obtenido del análisis de subjetividad con TextBlob

```
1 I think this movie is great!
2 {'subjectivity': 0.75}
3
4 Yesterday we saw a movie.
5 {'subjectivity': 0.0}
```

3.3.2. Extracción de temas (LDA)

La extracción de temas es una técnica de aprendizaje automático que organiza y entiende grandes conjuntos de datos textuales a través de asignar categorías a partir del tema presente en el texto.

Un tema se define como un patrón de repeticiones de términos en un conjunto de palabras. Por ejemplo, el conjunto de palabras salud, médico, paciente y hospital pertenecen al tema "atención médica", mientras que las palabras granja, cultivos y trigo pertenecen al tema "agricultura".

La extracción de temas usa técnicas de procesamiento del lenguaje natural para descomponer el lenguaje humano de modo que se puedan encontrar patrones y estructuras semánticas dentro de los textos para extraer conocimientos.

Los dos enfoques más comunes para el análisis de temas con aprendizaje automático son el modelado de temas (*topic modeling*) y la clasificación de temas (*topic classification*). Igual que en el caso del análisis de sentimiento, la primera estrategia usa aprendizaje no supervisado y la segunda aprendizaje supervisado. Para poder aplicar un modelo de clasificación para detectar los temas de un texto se deben conocer los temas presentes en el texto de antemano, hecho que no se cumple con el análisis que se trata de realizar en este proyecto.

Por lo tanto, se usará el modelado de temas para extraer los diferentes temas que se encuentran en el texto de los tweets usando un modelo de aprendizaje no supervisado. Estas técnicas pueden inferir patrones y agrupar expresiones similares sin necesidad de definir etiquetas o de entrenar datos de antemano.

Existen diferentes enfoques para obtener los temas de un texto, como por ejemplo "Term Frequency and Inverse Document Frequency" (TD-IDF) [34] o técnicas de factorización matricial no negativa (NMF) [35], aunque el modelo más popular hoy en día es el conocido como Latent Dirichlet Allocation (LDA) [36], que es el que se usará en este proyecto.

El modelo LDA funciona asumiendo que un documento se ha generado a partir de seleccionar un conjunto de temas, y a continuación un conjunto de palabras para cada tema. Por lo tanto, para encontrar los temas dentro de un texto aplica un proceso de ingeniería inversa basado en las distribuciones de probabilidad de las palabras del documento.

3.3.2.1. Preparación del texto para el modelado

Antes de poder aplicar el modelo LDA sobre los tweets disponibles se deben aplicar ciertas transformaciones sobre el texto.

En primer lugar, se debe identificar qué elemento simboliza un documento en el contexto de los datos del problema, para comprender qué es lo que forma el corpus. En el problema actual

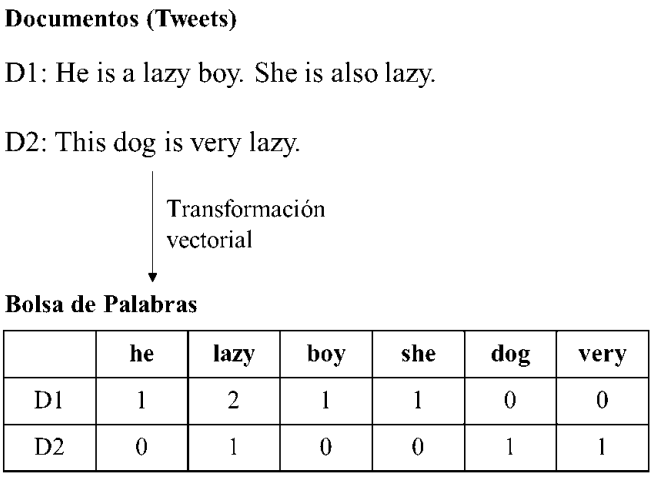
cada uno de los tweets es lo que representa un documento, por lo que el corpus será la unión de todo el texto de todos los tweets de un conjunto de datos.

Una vez definido el corpus se debe crear lo que se conoce como Bolsa de Palabras (Bag of Words) [37]. Una bolsa de palabras es un modelo que se usa para representar los datos textuales antes de modelar texto con técnicas de machine learning. Es necesario aplicar esta técnica porque los textos son datos no organizados, y técnicas como el machine learning esperan entradas y salidas de tamaños definidos. Por lo tanto, los algoritmos de machine learning no pueden trabajar directamente con texto crudo, así que este se debe convertir a una representación numérica o vectorial.

Una bolsa de palabras, por lo tanto, es una representación de un texto que describe la ocurrencia de las palabras dentro de un documento. Este modelo necesita dos cosas:

- **Un vocabulario de palabras conocidas**, que a grandes rasgos es un diccionario de todas las palabras que aparecen en el corpus.
- **Una medida de presencia de las palabras conocidas**. La medida de presencia es normalmente el recuento de veces que aparece una palabra o la frecuencia de aparición de una palabra respecto al total de palabras. En la mayoría de las aplicaciones es favorable usar frecuencias por encima del recuento para evitar que las palabras más frecuentes dominen en el entrenamiento del modelo, pero el funcionamiento de LDA requiere el uso del recuento de palabras como medida de presencia de las palabras.

La base detrás de esta representación es que dos documentos serán similares si tienen contenido similar (es decir, palabras similares). Con esta información un modelo de machine learning ya puede aprender sobre el significado de un documento.



	he	lazy	boy	she	dog	very
D1	1	2	1	1	0	0
D2	0	1	0	0	1	1

Figura 3.3. Ejemplo de transformación vectorial de varios textos a una Bolsa de Palabras

En la Figura 3.3 se puede observar un ejemplo de la transformación que sufre un texto al ser convertido en una bolsa de palabras.

3.3.2.2. Aplicación y evaluación del modelo

Una vez transformados los textos de los tweets en una representación de bolsa de palabras se puede aplicar el modelo LDA para encontrar un número de temas determinado.

Aunque se están aplicando técnicas de aprendizaje no supervisado, existen algunas medidas que ayudan a evaluar el correcto funcionamiento del modelo.

La primera medida usada normalmente para la evaluación de modelos de texto es la perplejidad (*perplexity*). Esta medida captura cómo de sorprendido se encuentra un modelo al observar nuevos datos que no ha visto nunca. Sin embargo, estudios recientes han demostrado que la perplejidad y el juicio humano para identificar un tema a menudo no están correlacionados [38].

Esta limitación en la medida de perplejidad ha causado la investigación y aparición de medidas más cercanas al juicio humano, como la medida de coherencia (*coherence*) de un tema [39].

La medida de coherencia de un tema puntúa el grado de similitud semántica entre las palabras con puntuación más alta dentro del tema. Estas medidas ayudan a distinguir los temas que son semánticamente interpretables de temas que son puramente un producto de la inferencia estadística [40]. Existen diferentes medidas para calcular la coherencia de un tema, pero en el proyecto actual se ha usado la medida **C_v** [41], al ser la más comúnmente usada y la que aporta por defecto la librería Gensim.

3.3.2.3. Hiperparámetros

El modelo LDA usa dos hiperparámetros en su aplicación, normalmente llamados Alfa y Beta [36].

El parámetro Alfa controla la similitud de los documentos, por lo que un valor bajo representa documentos como una mezcla de pocos temas, mientras que un valor alto resulta en representaciones con más temas para cada documento.

Por otro lado, el parámetro Beta representa la similitud entre temas, por lo que un valor bajo representa temas que son más diferentes, ya que se obtienen menor temas, y un valor alto representa obtener más temas, pero más similares entre ellos.

Al usar el modelo de LDA de la libería Gensim, existe la opción de dejar que el modelo busque los mejores valores para Alfa y Beta permitiéndole aprender vectores asimétricos [42], por lo que estos no se han modificado.

Por otro lado, un modelo LDA también necesita que se especifique el número de temas que el modelo va a tener, ya que el este no puede decidir por sí mismo el número de temas que debe

buscar.

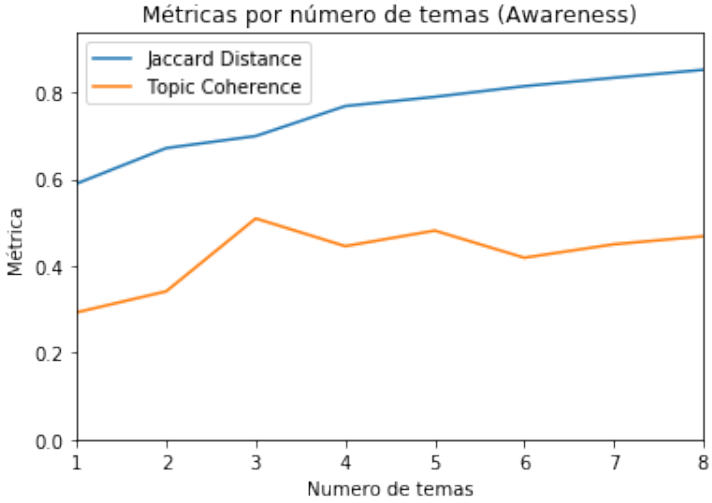


Figura 3.4. Métrica de coherencia y distancia de Jaccard para diferente número de temas en el modelo LDA (Awareness)

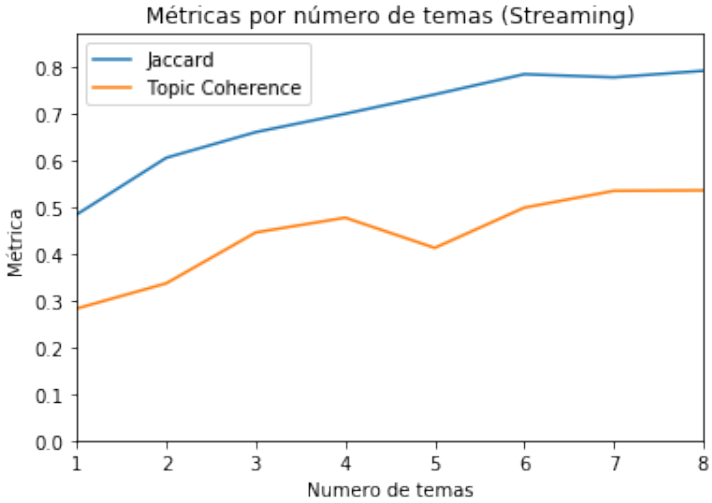


Figura 3.5. Métrica de coherencia y distancia de Jaccard para diferente número de temas en el modelo LDA (Streaming)

Para decidir el mejor número de temas se quiere maximizar la medida de coherencia de un tema y minimizar la distancia semántica entre temas, es decir, que se repitan el menor número de palabras entre conjuntos. Para calcular la distancia entre temas se ha usado la distancia de Jaccard [43].

Para obtener el número idóneo de temas para el modelo se ha realizado una prueba con diferentes números de temas (de 1 a 8 temas) y se han obtenido los resultados de las Figuras 3.4 y 3.5 para los dos conjuntos de datos.

Observando las figuras se ha determinado que el número de temas idóneo para el conjunto de datos Awareness es 3 y para el conjunto de datos Streaming es 4. Se puede observar que la distancia entre temas siempre crece al aumentar el número de temas, por lo que la mejor estrategia a seguir es usar el mínimo número de temas con una coherencia alta.

Por último, al aplicar los modelos con los hiperparámetros especificados se han obtenido los pesos para las palabras dentro de cada uno de los temas que se pueden observar en el Código 3.4.

Código 3.4. Pesos para las palabras con más apariciones en los temas identificados

```
1  // Temas Awareness
2  [(0, '0.042*"loss" + 0.038*"baby" + 0.023*"pregnancy" + 0.020*"day" + 0.017*
       "infant" + 0.016*"awareness" + 0.015*"light" + 0.014*"lose" + 0.012*"week
       " + 0.011*"today"'),
3  (1, '0.045*"baby" + 0.035*"singh" + 0.025*"loss" + 0.024*"support" + 0.024*"
       need" + 0.023*"suffer" + 0.022*"life" + 0.022*"hear" + 0.021*"raise" + 0.
       020*"daughter"'),
4  (2, '0.069*"miscarriage" + 0.024*"woman" + 0.010*"say" + 0.009*"go" + 0.008*
       "know" + 0.008*"people" + 0.008*"like" + 0.008*"get" + 0.008*"year" + 0.0
       07*"fuck"')]
5
6  // Temas Streaming
7  [(0, '0.070*"baby" + 0.070*"loss" + 0.029*"sorry" + 0.011*"love" + 0.010*"
       lose" + 0.007*"go" + 0.007*"family" + 0.006*"day" + 0.006*"know" + 0.006*
       "take"'),
8  (1, '0.029*"miscarriage" + 0.028*"baby" + 0.019*"loss" + 0.018*"pregnancy" +
        0.017*"woman" + 0.015*"help" + 0.013*"get" + 0.012*"please" + 0.012*"
       dedicate" + 0.011*"mean"'),
9  (2, '0.070*"miscarriage" + 0.012*"get" + 0.010*"say" + 0.010*"know" + 0.010*
       "abortion" + 0.009*"go" + 0.009*"woman" + 0.009*"pregnancy" + 0.008*"
       people" + 0.008*"like"'),
10 (3, '0.062*"miscarriage" + 0.022*"woman" + 0.011*"vaccine" + 0.011*"death" +
        0.009*"result" + 0.009*"year" + 0.009*"week" + 0.008*"first" + 0.008*"
       covid" + 0.008*"study"')]
```

3.4. Sumario de productos obtenidos

A partir de los dos procesos de recolección de datos descritos a lo largo de este capítulo se han obtenido dos conjuntos de tweets enfocados en la conversación sobre la pérdida del embarazo y el aborto espontáneo.

Para el desarrollo de este proyecto se ha usado un Jupyter Notebook de Python, donde se describe de forma estructurada todo el código necesario para la obtención de los datos, su procesamiento y el modelado posterior, además de la obtención de los resultados que se presentan en este documento.

Este notebook se llama **miscarriage-analysis.ipynb** y se encuentra alojado en el repositorio público GitHub, juntamente con los datos extraídos en este proyecto (consulta del código en Apéndice A).

Estos conjuntos de datos pueden resultar de gran interés para investigaciones futuras sobre el tema, por lo que se han generado diferentes ficheros con los datos a lo largo de las fases del proyecto.

1. **Tweets en crudo**: se han generado dos ficheros en formato JSON con los datos directamente extraídos de la API de Twitter sin modificar. Se encuentran en la carpeta **json_files**.

2. **Datos aplanados y transformados en formato tabla**: se han aplanado los niveles complejos del formato JSON, se han seleccionado las columnas necesarias y se han aplicado diversas transformaciones para obtener dos ficheros en formato CSV. Se encuentran en el directorio **csv_files/transformation**.

3. **Datos con resultados finales**: se han añadido columnas con el resultado del análisis de sentimiento, la clasificación del tema identificado en cada Tweet y la aparición de palabras de depresión en el texto para generar este fichero final, que contiene todos los resultados del presente proyecto. Se encuentran en el directorio **csv_files/results**.

Capítulo 4

Experimentos y resultados

4.1. Análisis general del sentimiento

Una vez se han calculado tanto la polaridad como la subjetividad de cada uno de los tweets, se pueden analizar los conjuntos de datos por diferentes parámetros.

Tabla 4.1. Media de polaridad y subjetividad por conjunto de datos

Conjunto de datos	Polaridad media	Subjetividad media
Awareness	-0.029	0.369
Streaming	-0.068	0.462

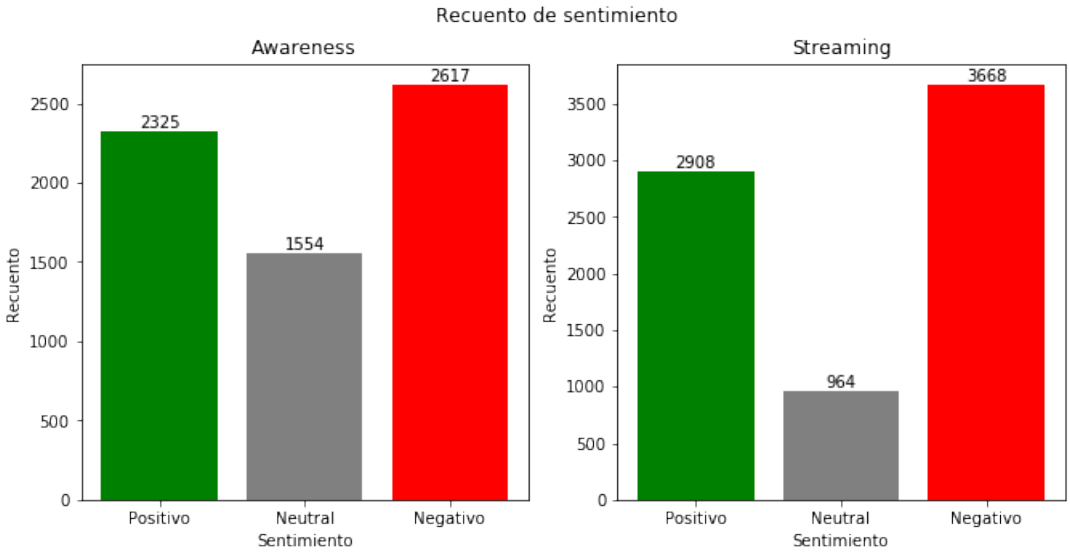


Figura 4.1. Recuento de sentimiento

En la Tabla 4.1 se puede observar la media de polaridad y subjetividad de los dos conjuntos de datos y en la Figura 4.1 se encuentra el recuento de tweets por sentimiento.

Se puede observar que en los dos conjuntos de datos los tweets que predominan son negativos, aunque también se observa un grupo considerable de tweets positivos. No se pueden observar un gran número de tweets neutrales, por lo que los tweets hablando sobre el aborto espontáneo son generalmente positivos o negativos, pero no neutrales.

4.1.1. Análisis temporal

Se han podido observar algunos patrones al analizar la polaridad y la subjetividad media por hora del día.

En relación con la polaridad media por hora, se puede observar en el caso de los datos de Awareness un ligero decrecimiento de la polaridad durante las horas de la tarde, y un decrecimiento más pronunciado en las horas de la noche. Esto indica que los tweets publicados durante estos períodos de tiempo son en general más negativos que los publicados a otras horas del día (Figura 4.2).

Si se analiza lo mismo en el conjunto de datos de Streaming no se puede observar ninguna conclusión tan marcada a causa de las fluctuaciones de la figura. Se puede observar solamente que alrededor de las 16h y 17h hay una subida generalizada de la polaridad de los tweets, por lo que durante esta hora estos son más positivos (Figura 4.2).

Por lo que respecta a la subjetividad, no se puede observar ningún patrón muy marcado, aunque generalizadamente se puede observar una bajada de la subjetividad durante las horas de la mañana y la tarde, por lo que los tweets son más subjetivos por la noche y más objetivos durante el día (Figura 4.3).

También se puede observar que los tweets del conjunto de datos de Streaming son más subjetivos que los del conjunto de Awareness.

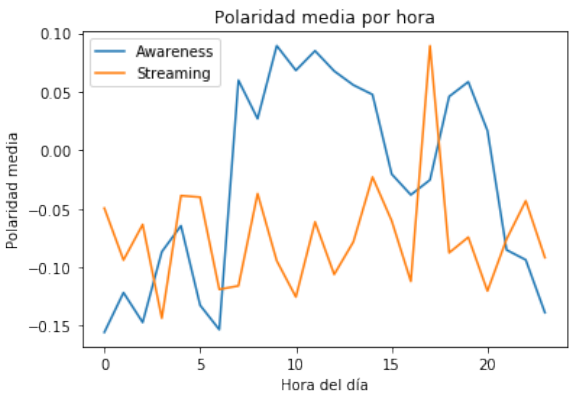


Figura 4.2. Polaridad media por hora del día

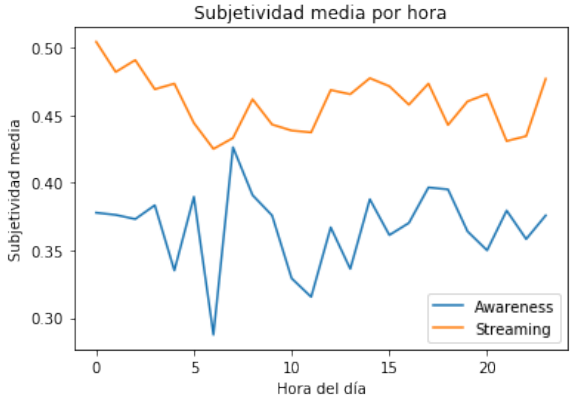


Figura 4.3. Subjetividad media por hora del día

4.1.2. Análisis de correlación

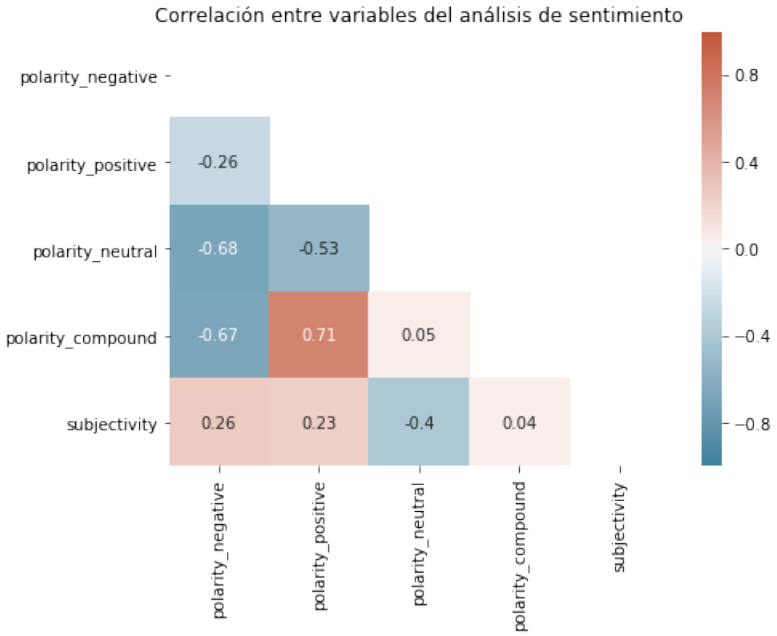


Figura 4.4. Matriz de correlación entre campos de polaridad y subjetividad

Observando el comportamiento temporal durante el día de la polaridad y la subjetividad se puede apreciar lo que parece una ligera relación entre la polaridad y la subjetividad, así que se ha calculado el coeficiente de correlación de Pearson entre los diferentes campos obtenidos de la polaridad y la subjetividad para la combinación de los dos conjuntos de datos (Figura 4.4).

Como era esperable, existe una correlación fuerte entre los propios campos obtenidos del cálculo de la polaridad, ya que la combinación del campo positivo, negativo y neutral es lo que genera el campo combinado (*compound* en inglés).

Se observa también una ligera correlación entre los campos de polaridad con el de subjetividad. Se puede observar que como mayor es la polaridad positiva o negativa, la subjetividad crece proporcionalmente, mientras que como más neutral es un tweet la subjetividad crece de forma inversamente proporcional. Esto indica que los tweets más positivos y los más negativos son también los más subjetivos, mientras que los tweets con un sentimiento neutral son normalmente más objetivos.

4.2. Caracterización de temas identificados

Una vez obtenida la clasificación de tema para cada uno de los tweets de los dos conjuntos de datos, se puede observar gráficamente la representación espacial de los temas identificados.

En los gráficos interactivos presentes en el código desarrollado para este proyecto (consulta del código en Apéndice A) se pueden analizar en detalle las palabras que más aparecen en cada tema identificado por el modelo. En la Figura 4.5 se puede observar como ejemplo el gráfico obtenido para el conjunto de datos de Awareness, y las palabras con más apariciones en el tema 1.

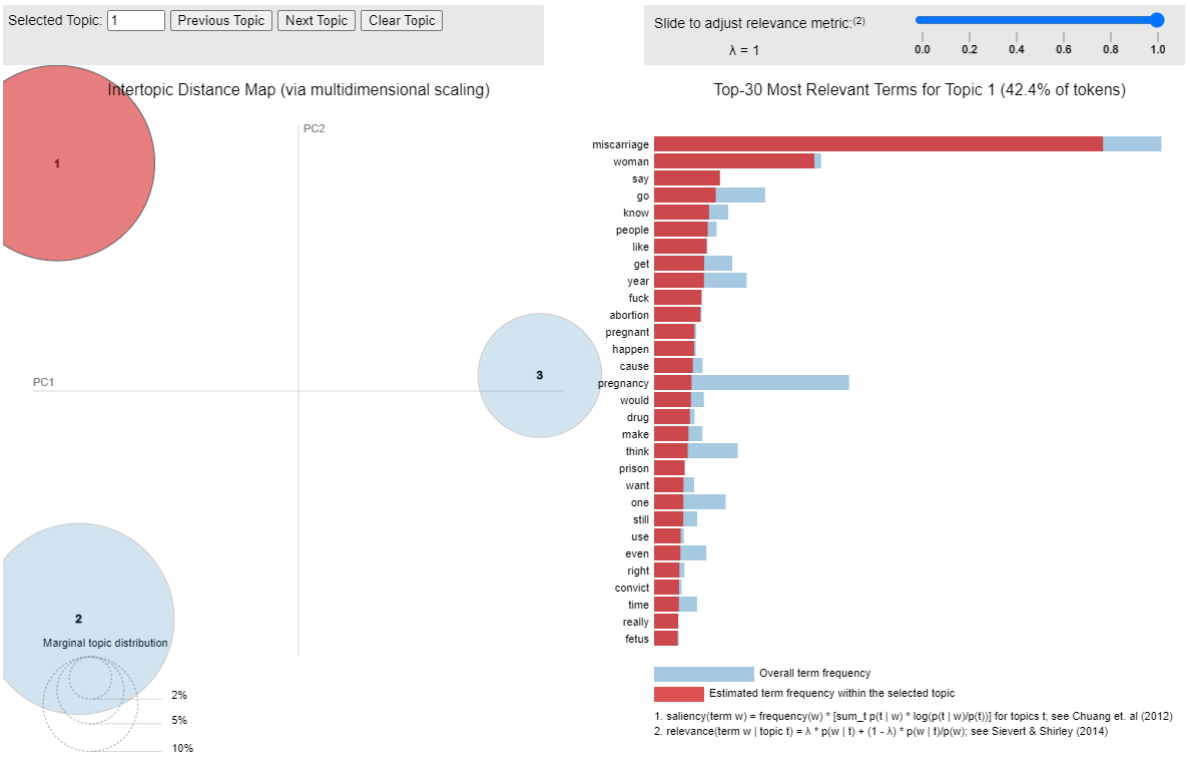


Figura 4.5. Visualización de los temas identificados por LDA (Awareness)

En estas visualizaciones cada uno de los círculos representa uno de los temas identificados. El tamaño del círculo representa cómo de grande es el tema dentro del conjunto de datos, y la distancia entre los círculos representa cómo de diferentes son las palabras que contienen los diferentes temas.

A partir de las palabras con más apariciones en cada tema, en las Figuras 4.6 y 4.10, se ha podido definir un nombre para cada uno de los temas. Para poder comprender mejor el contenido de cada uno de los temas, se han extraído a continuación algunos tweets de cada tema en los dos conjuntos de datos.

Figura 4.6. Palabras más frecuentes por tema (Awareness)

4.2.1. Temas identificados en el conjunto Awareness

Tema 1) Oklahoma case (Caso de Oklahoma)

A principios del mes de octubre salió a la luz una sentencia sobre un jurado que declaró culpable de asesinato a una mujer en Oklahoma después de sufrir un aborto espontáneo causado por el uso de drogas [44].

La sentencia generó una gran controversia y activó la conversación alrededor de las leyes que prohíben el aborto en Estados Unidos y sobre disparidades sociales [45] [46], por lo que durante el mes de octubre se pueden encontrar muchos tweets comentando y opinando sobre el tema referente a este encarcelamiento.

Por lo tanto, no es de extrañar que este sea un tema muy comentado en las redes sociales también durante la semana de concienciación sobre el aborto espontáneo, período en el que se realizó la obtención de datos para este conjunto.

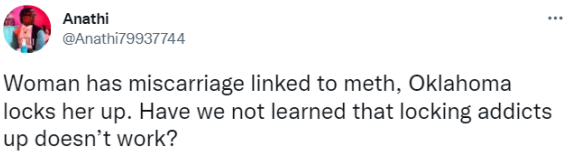


Figura 4.7. Ejemplo de tweet clasificado en el Tema 1: Oklahoma case

Tema 2) Awareness (Concienciación)

Este segundo tema identificado en los datos de este período de tiempo está formado por mensajes de apoyo a mujeres que hayan vivido experiencias de pérdida del embarazo, con tal de normalizar y dar visibilidad sobre el tema.

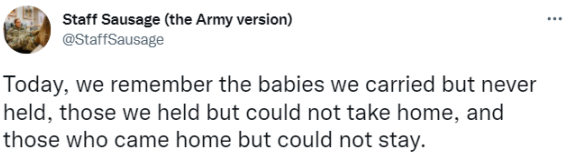


Figura 4.8. Ejemplo de tweet clasificado en el Tema 2: Awareness

Tema 3) Support (Apoyo)

Por último, en este último tema se encuentran mensajes hablando del apoyo necesario para superar una situación de pérdida del embarazo teniendo en cuenta el estigma que aun existe alrededor del tema.

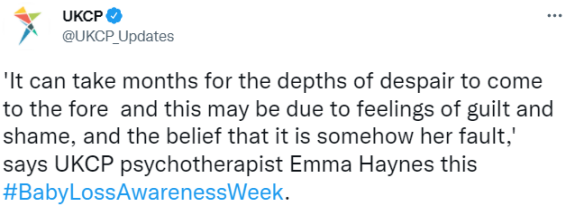


Figura 4.9. Ejemplo de tweet clasificado en el Tema 3: Support

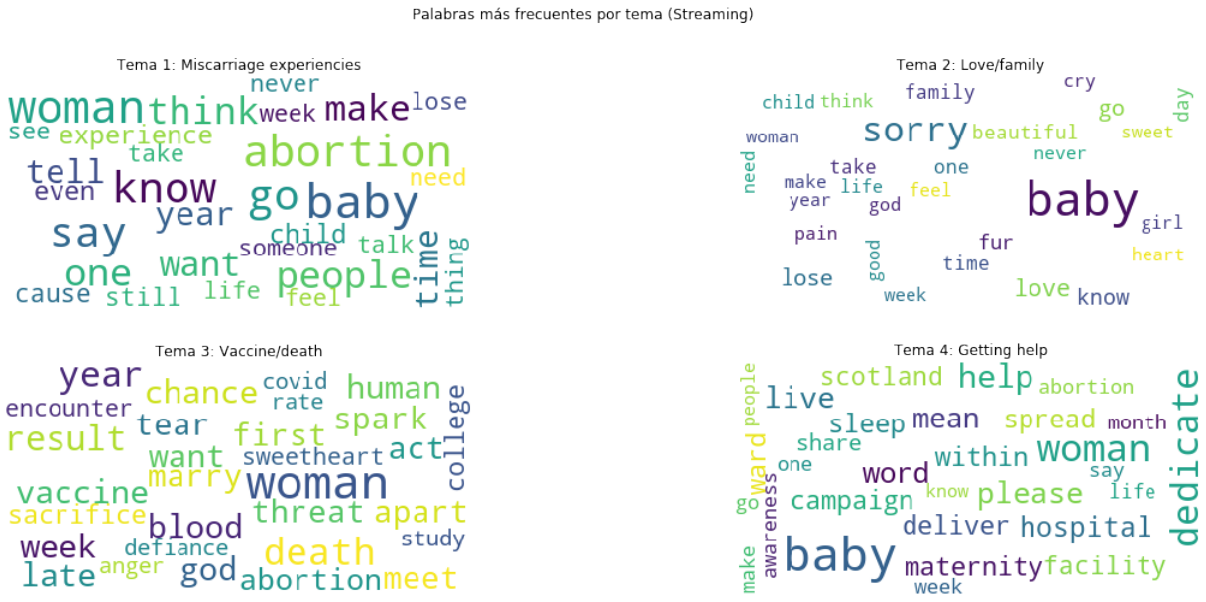


Figura 4.10. Palabras más frecuentes por tema (Streaming)

4.2.2. Temas identificados en el conjunto Streaming

Tema 1) Miscarriage experiencies (Experiencias de aborto espontáneo)

En este tema se han podido observar opiniones y mensajes vitales para comprender las experiencias reales de las mujeres que han sufrido la pérdida del embarazo. Dentro de estos mensajes se han identificado quejas y problemas claros que las mujeres han vivido durante sus abortos espontáneos.

Se ha realizado pequeña recolección de ejemplos de tweets de este tema, donde se pueden identificar problemas claros que afectan negativamente a la salud, tanto física como mental, de

las mujeres que han sufrido un aborto espontáneo. En concreto, se han identificado problemas económicos, problemas de estigma y problemas en el sistema sanitario. El conjunto de tweets recolectados se pueden consultar en el Apéndice B.

Dado que los datos de este conjunto se obtuvieron haciendo streaming de Twitter solamente durante una semana, y en el proceso de streaming no se puede obtener el 100 % de los tweets disponibles en tiempo real, no se puede concluir que esta observación sea significativa respecto a la población, pero sí que permite identificar una nueva línea de investigación para extraer directamente de las experiencias reales de mujeres los problemas existentes alrededor de la pérdida del embarazo.

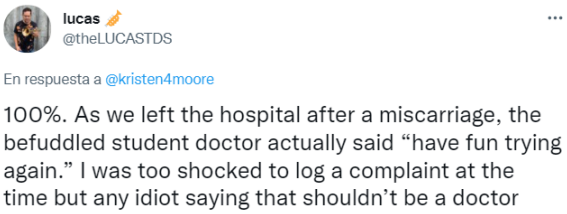


Figura 4.11. Ejemplo de tweet clasificado en el Tema 1: Miscarriage experiencies

Tema 2) Love/family (Amor/familia)

En este tema se encuentran los tweets que mandan mensajes de ánimo a las mujeres que han sufrido la pérdida del embarazo y a sus familias. Se encuentran muchos tweets mandando mensajes sobre el amor y plegarias.

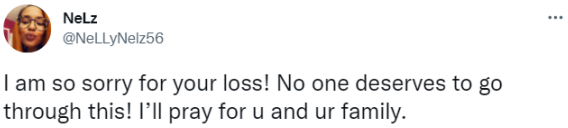


Figura 4.12. Ejemplo de tweet clasificado en el Tema 2: Love/family

Tema 3) Vaccine/death (Vacuna/muerte)

Teniendo en cuenta la situación actual de la pandemia de COVID-19, no es de extrañar que exista un grupo de tweets hablando sobre como la vacunación y la propia enfermedad afecta a las mujeres embarazadas y a las que han sufrido un aborto espontáneo.

En este conjunto de tweets se encuentran en general mensajes informativos sobre la afectación de la vacunación a las mujeres embarazadas, y algunos mensajes de preocupación por casos de aborto espontáneo relacionados con la enfermedad o la vacunación.

Tema 4) Getting help (Conseguir ayuda)

En este conjunto de tweets se encuentran mayoritariamente mensajes informativos sobre como conseguir ayuda en caso de la pérdida del embarazo.

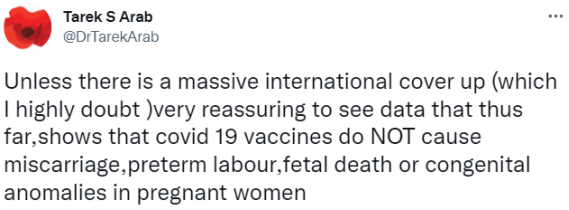


Figura 4.13. Ejemplo de tweet clasificado en el Tema 3: Vaccine/death

Figura 4.14. Ejemplo de tweet clasificado en el Tema 4: Getting help

4.3. Aparición de síntomas de depresión en los tweets

Uno de los objetivos principales de este proyecto es el análisis de la aparición de síntomas de depresión, ansiedad u otros problemas relacionados con la salud mental a través del texto de los tweets extraídos.

Se han seleccionado de los conjuntos de tweets los que presentan alguna de las 20 palabras más usadas por pacientes que sufren de depresión en entornos clínicos. Estas palabras fueron identificadas y seleccionadas conjuntamente por un psicólogo y un médico de familia con experiencia clínica y se basaron en la definición y las características generales de la depresión de acuerdo con el 'Diagnostic and Statistical Manual of Mental Disorders' [47]. La lista de palabras en inglés se puede encontrar en la Figura 4.15, juntamente con la traducción al castellano.

Al realizar un recuento de tweets con la aparición de alguna de estas palabras se obtiene para el conjunto de datos de Awareness que aparecen en **294 tweets** del total de 6496, y para el conjunto de Streaming aparecen en **787** tweets del total de 7540.

Por lo tanto, solo se puede observar aparición de síntomas de depresión en un **4.5 %** de tweets en Awareness y en un **10.4 %** de tweets en Streaming.

- overwhelmed (agobiado/a)
- exhausted (agotado/a)
- distressed (angustiado/a)
- anxiety (ansiedad)
- anxious (ansioso/a)
- tired (cansado/a)
- low (decaído/a)
- depression (depresión)
- depressed (deprimido/a o depresivo/a)
- discouraged (desanimado/a)
- desperate (desesperado/a)
- demotivated (desmotivado/a)
- insomnia (insomnio)
- cry (llorar)
- nervous (nervioso/a)
- worried (preocupado/a)
- lonely (solo/a)
- sad (triste)
- empty (vacío/a)

Figura 4.15. Lista de palabras más usadas por pacientes clínicos de depresión

4.4. Relación entre variables

Se han analizado por separado los resultados del análisis de sentimiento, de la extracción de temas y de la aparición de palabras de depresión en los tweets, pero un análisis conjunto de los diferentes resultados obtenidos ayudará a comprender mejor de qué forma se están comunicando las mujeres que han sufrido la pérdida del embarazo en las redes sociales.

Por lo tanto, en este apartado se analiza como el sentimiento, el tema detectado y la aparición de palabras de depresión se afectan entre ellos.

4.4.1. Correlación entre todas las variables

En primer lugar, para detectar rápidamente las posibles relaciones entre las variables disponibles, se ha generado una matriz de correlación entre todos los campos (Figura 4.16).

Algunas relaciones que se pueden observar en la Figura 4.16 son:

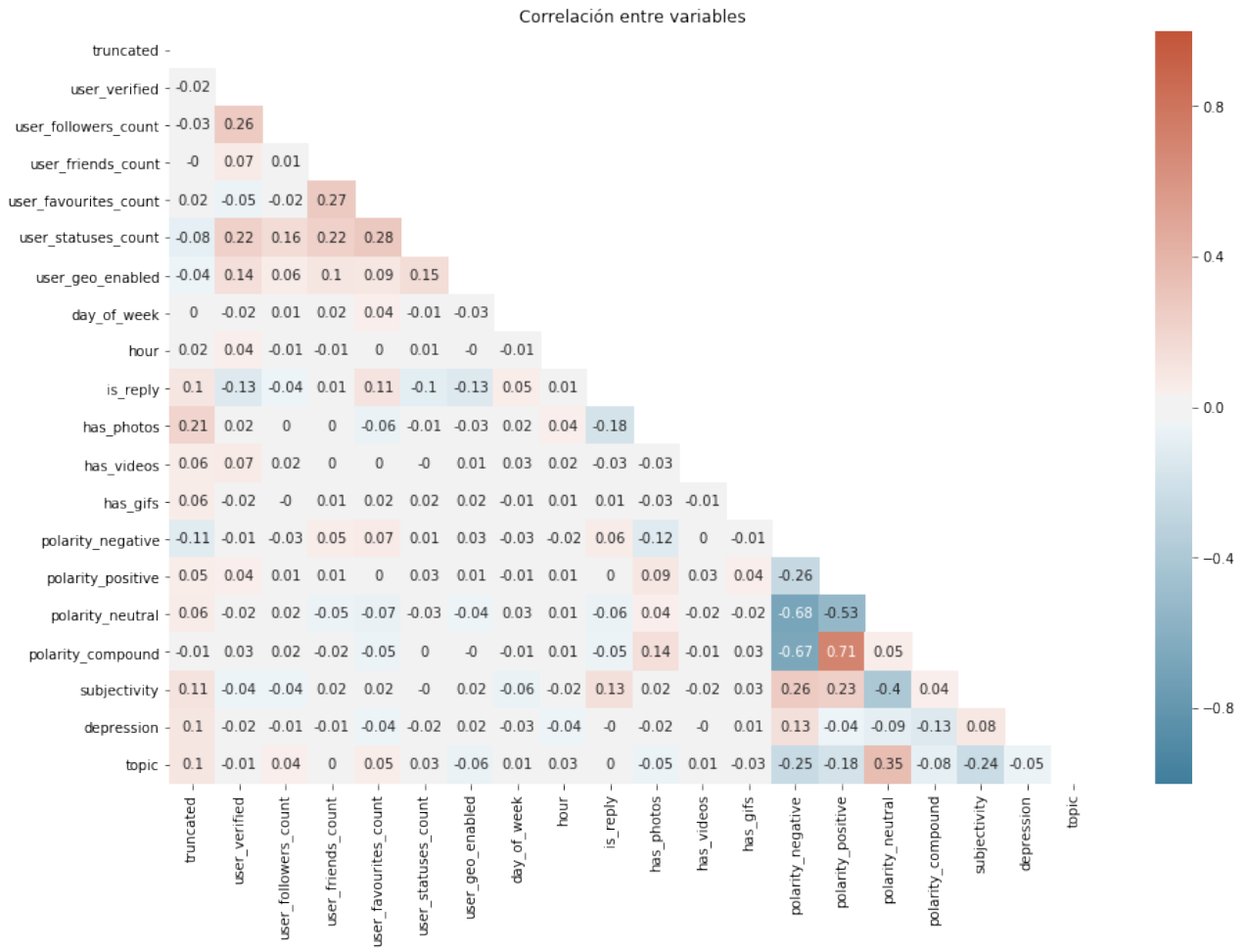


Figura 4.16. Matriz de correlación entre todas las variables numéricas

- El hecho de que un usuario esté verificado está positivamente correlacionado con tener un mayor número de seguidores.
- El hecho de que un usuario tenga un mayor recuento de publicaciones también implica que tenga un mayor número de tweets marcados como favorito.
- El hecho de que un tweet esté truncado y por lo tanto sea más largo, está positivamente correlacionado con el hecho de que contenga fotos, y está inversamente correlacionado con el hecho de ser una respuesta, así que los tweets más largos suelen contener fotos y no suelen ser respuestas.
- El hecho de que un tweet contenga una foto hace que la polaridad compuesta crezca ligeramente, así que los tweets con imágenes son ligeramente más positivos.

Se observan también correlaciones entre los campos obtenidos de los resultados, así que a continuación se van a explorar en profundidad la relación entre el tema detectado, la objetividad y subjetividad del tweet y la aparición de palabras de depresión.

4.4.2. Relación entre polaridad/subjetividad y depresión

Se ha calculado la polaridad y subjetividad medias mostrando la diferencia entre tweets que contienen palabras de depresión o no. En la Figura 4.17 se puede observar que los tweets que no contienen palabras de depresión de media muestran un sentimiento neutral, mientras que los tweets con presencia de estas palabras son claramente negativos. También se puede observar que la subjetividad es ligeramente más alta en el conjunto de datos con palabras de depresión en la Figura 4.18.

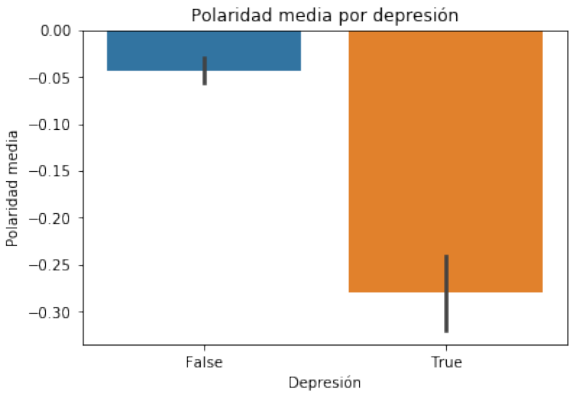


Figura 4.17. Polaridad media entre tweets con aparición de palabras de depresión

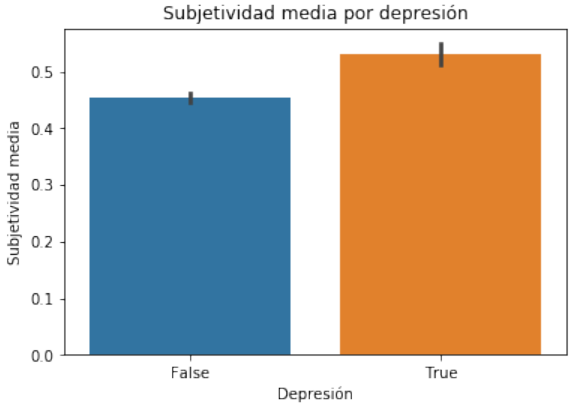


Figura 4.18. Subjetividad media entre tweets con aparición de palabras de depresión

4.4.3. Relación entre polaridad/subjetividad y tema identificado

Se ha calculado la media de polaridad y subjetividad para cada uno de los temas identificados en la extracción de temas.

Al observar en primer lugar los datos de Awareness en la Figura 4.19 se puede observar que el sentimiento es de media positivo para el tema awareness, que contiene mensajes de ánimo y concienciación hacia mujeres que hayan sufrido un aborto espontáneo. Para el tema que trata sobre el encarcelamiento en Oklahoma se puede observar que el sentimiento es de media negativo, ya que los tweets contenidos en este tema son en su mayoría quejas sobre este caso de encarcelamiento. Por último, el tema support, que habla sobre el apoyo necesario que necesita una mujer en una situación de pérdida de embarazo es de media neutral.

Por lo que incumbe a la subjetividad en la Figura 4.20, los tres temas son ligeramente subjetivos, siendo el último tema el que presenta datos más objetivos.

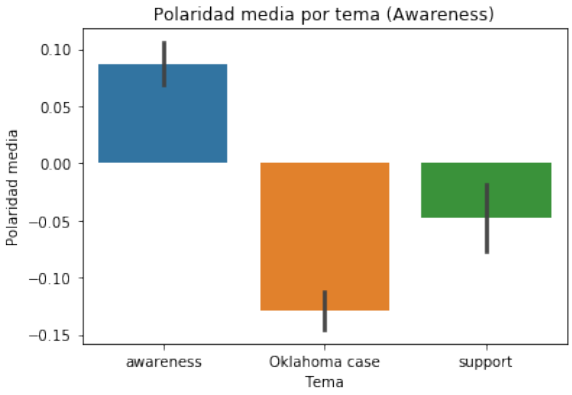


Figura 4.19. Polaridad media por tema (Awareness)

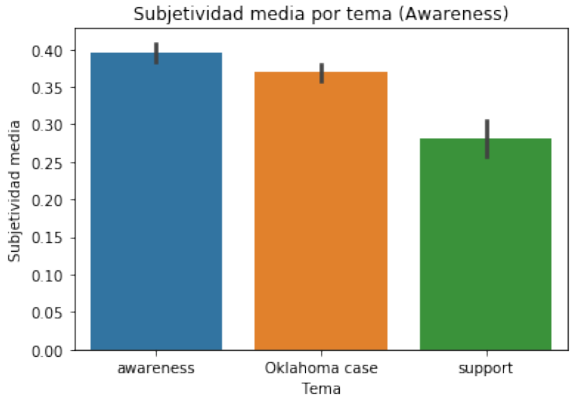


Figura 4.20. Subjetividad media por tema (Awareness)

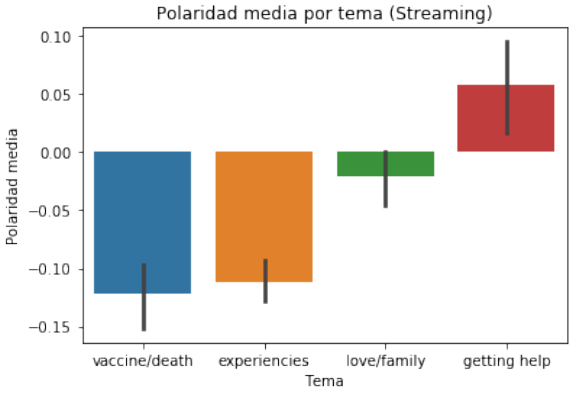


Figura 4.21. Polaridad media por tema (Streaming)

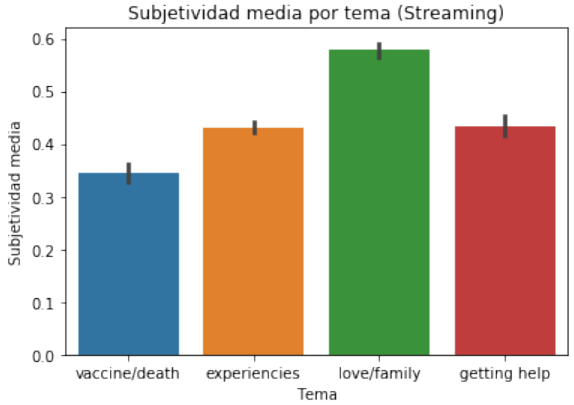


Figura 4.22. Subjetividad media por tema (Streaming)

Al observar los datos de Streaming en la Figura 4.21 se observa que tanto los temas que tratan sobre la vacunación y las experiencias reales de mujeres sobre abortos espontáneos son de media negativos. El segundo tema, que trata sobre las experiencias reales de las mujeres contiene una gran cantidad de tweets donde se pueden identificar problemas reales que afectan negativamente a la salud mental de las mujeres en estas situaciones, así que no es de extrañar que el sentimiento medio del tema sea negativo. El tema que trata sobre el amor y la familia es de media neutral, y el tema que da información sobre como conseguir ayuda es de media positivo.

Al observar la subjetividad en la Figura 4.22, una vez más se identifica que los diferentes temas son ligeramente subjetivos, excepto el que habla del amor y la familia, que es el más subjetivo. El tema que trata sobre la vacunación y el COVID-19 es el que muestra datos más objetivos.

4.4.4. Relación entre tema y depresión

Por último, se ha realizado un recuento de tweets que contengan palabras de depresión y de los que no para cada uno de los temas identificados.

En el conjunto de datos de Awareness (Figura 4.23) se puede observar que el tema con mayor presencia de palabras de depresión es el relacionado con el caso de encarcelamiento en Oklahoma, mientras que en el conjunto de datos de Streaming (Figura 4.24), tanto el tema de experiencias reales de aborto espotnáneo como el tema que habla sobre el amor y la familia son los que presentan mayor presencia de palabras de depresión.

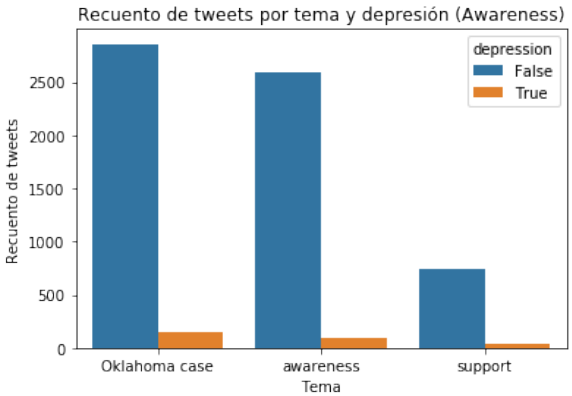


Figura 4.23. Recuento de tweets por tema y por aparición de depresión (Awareness)

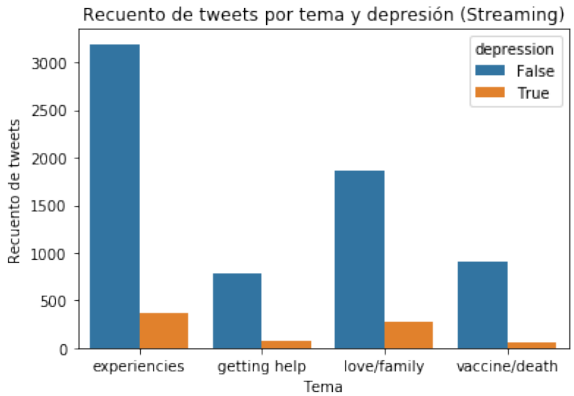


Figura 4.24. Recuento de tweets por tema y por aparición de depresión (Streaming)

Capítulo 5

Conclusiones

El objetivo principal de este proyecto ha sido identificar la aparición de problemas de salud mental, como ansiedad o depresión, en los usuarios de Twitter que hablan sobre el aborto espontáneo, de forma que se pueda caracterizar de qué hablan estos usuarios para profundizar y comprender los problemas reales alrededor de la pérdida del embarazo.

Con tal de analizar en profundidad la conversación alrededor del aborto espontáneo y de comprender como este afecta a la salud mental de las mujeres, se han analizado los datos extraídos a partir de diferentes parámetros.

En primer lugar, en los dos conjuntos de datos extraídos se han identificado tweets que contienen palabras normalmente usadas por pacientes de **depresión** clínica, por lo que se han podido identificar usuarios con posibles problemas de salud mental. Alrededor un **5 %** de los tweets de Awareness y un **10 %** de los tweets de Streaming presentan palabras de este conjunto.

Los resultados también muestran que los tweets de usuarios que usan este tipo de palabras son significativamente más **negativos** y más **subjetivos**.

En segundo lugar, para comprender de qué trata la conversación alrededor del aborto espontáneo en Twitter, se han analizado en profundidad el sentimiento y el tema que se puede extraer de los tweets obtenidos.

Al analizar el recuento del sentimiento se ha observado que **los tweets negativos predominan en los dos conjuntos de datos**, estando cerca el número de tweets positivos. En cambio, el número de tweets neutrales ha sido mucho menor que los positivos y negativos. Por lo tanto, **los tweets que tratan sobre el aborto espontáneo son generalmente positivos o negativos, pero no neutrales**.

Respecto a la subjetividad, no se han observado patrones claros, pero se ha podido observar que esta se encuentra relacionada con la polaridad. Como más positivo o negativo sea un tweet, también más subjetivo será este. Por lo tanto, **los tweets muy positivos o negativos suelen ser subjetivos, mientras que los tweets neutrales suelen ser más objetivos**.

Analizando los patrones horarios y el sentimiento se ha observado que **la polaridad disminuye en general durante las horas de la noche, mientras que la subjetividad aumenta ligeramente**.

En el conjunto de datos de Awareness se han identificado 3 temas: uno hablando de **concienciación** sobre el aborto espontáneo (awareness), otro hablando sobre el **apoyo** necesario para superar la pérdida del embarazo (support), y por último un tema centrado en el caso de **encarcelamiento** por asesinato de una mujer que sufrió un aborto espontáneo a causa del uso del drogas en Oklahoma (Oklahoma case).

Se puede destacar que los tweets dentro del tema de la concienciación sobre el aborto son de media **positivos**, mientras que los tweets comentando el caso de encarcelamiento en Oklahoma son de media **negativos**.

En el conjunto de datos de Streaming se han identificado 4 temas diferentes. En primer lugar se ha identificado un tema que trata sobre las **experiencias reales de aborto espontáneo** vividas por mujeres (miscarriage experiencies), dentro del cual se han identificado un seguido de problemas que afectan negativamente a la salud física y mental de las mujeres en esta situación. Por otro lado, se han identificado también los temas que hablan del **amor y la famila** (love/family), sobre la **vacunación y el COVID-19** (vaccine/death) y sobre cómo **conseguir ayuda** habiendo vivido un aborto espontáneo (getting help).

En este caso, tanto los tweets dentro del tema que habla de experiencias reales como el que habla sobre la vacunación y el COVID-19 son de media **negativos**. El único tema detectado como **positivo** de media en este conjunto de datos es el que habla sobre como conseguir ayuda habiendo sufrido un aborto espontáneo.

Por último, al cruzar el modelado de temas con la identificación de palabras con relevancia clínica, se ha contextualizado el uso de estas palabras. Se ha podido observar que las palabras que indican depresión aparecen sobretodo en el tema del caso del encarcelamiento de Oklahoma, y en los temas de experiencias reales y amor/familia. La aparición de estas palabras en temas que hablan de experiencias reales de aborto espontáneo hace saltar la alarma en busca de **posibles problemas de salud mental** causados por el suceso.

Como objetivos secundarios, se han **extraído** de forma satisfactoria los tweets durante dos períodos en el tiempo usando un proceso de streaming a través de la API de Twitter y se han encontrado las **palabras clave** más adecuadas para realizar la extracción de tweets. Durante el procesamiento de los datos, se han aplicado transformaciones en diversas columnas, siendo especialmente importantes las **transformaciones en el texto** de los tweets. Usando los datos extraídos se han podido aplicar el **análisis de sentimiento** y la **extracción de temas** satisfactoriamente, que han aportado los resultados necesarios mostrados en las **visualizaciones**, para llegar a los objetivos de este proyecto.

Capítulo 6

Líneas de trabajo futuro

Con la realización de este proyecto se abren numerosas líneas de trabajo futuro.

En primer lugar, la primera idea que surge a continuación de este proyecto es el **aumentar el conjunto de datos** obtenido recolectando más tweets con el código desarrollado con tal de analizar en más profundidad la conversación alrededor de la pérdida del embarazo. Por las características del proyecto, el tiempo de recolección de tweets no puedo ser muy grande, pero se podría ampliar el análisis realizado en este proyecto con una cantidad mayor de datos. El código realizado es reutilizable para poder realizar el mismo análisis con cualquier conjunto de tweets sobre el aborto que se obtenga.

Siguiendo con una idea similar, en este proyecto solamente se han recolectado datos de Twitter, pero se podrían conseguir resultados muy interesantes extrayendo **datos de otras redes sociales**, como Instagram o TikTok, para realizar un análisis conjunto de cómo es la conversación alrededor del aborto espontáneo en las redes sociales en general.

En segundo lugar, sería de gran interés mantener y visualizar en **gráficos interactivos los resultados de este trabajo en tiempo real**, para comprender como evoluciona la conversación sobre el aborto espontáneo. Esta idea se podría llevar a cabo creando por ejemplo una aplicación web con un dashboard que contenga gráficos mostrando el sentimiento, el tema y la aparición de palabras de depresión, que de fondo fuese actualizando el conjunto de tweets extraídos automáticamente con un proceso de streaming.

Una aplicación de estas características podría desarrollarse con la tecnología Shiny de R, o con la tecnología Dash de Python, por ejemplo.

Otra aplicación que se podría desarrollar a partir del análisis de depresión de este proyecto podría ser un **chatbot** que pueda detectar usuarios que hayan publicado recientemente sobre haber sufrido un aborto espontáneo y que estén mostrando signos de depresión a partir de las palabras que están compartiendo. Detectando estos usuarios, este chatbot podría ofrecer ayuda psicológica o derivar a esta persona a los profesionales sanitarios necesarios, o simplemente hacer

saltar la alarma sobre la seguridad y bienestar de esta persona. Este tipo de chatbots están actualmente en auge, ya que existen varios de ellos donde los usuarios pueden hacer preguntas sobre su salud mental [48].

Por otro lado, aún sin ser uno de los objetivos de este trabajo, se han identificado **problemas claros en las experiencias que han compartido las mujeres que han sufrido un aborto espontáneo** en Twitter. Esto demuestra que las redes sociales son una fuente muy valiosa para identificar problemas que suceden en la vida real. En este proyecto estos problemas han sido detectados manualmente, así que una línea de investigación futura muy interesante sería desarrollar un proyecto que pudiese buscar e identificar tweets de estas características, para poder identificar los problemas directamente de las personas que los han vivido. Existen diversos estudios que afirman que las mujeres tienen más probabilidades de no sentirse escuchadas ni comprendidas por los profesionales sanitarios [49][50], por lo que es muy importante escuchar estas opiniones para intentar cerrar la brecha de género que existe en la sanidad actualmente.

Glosario

aborto involuntario pérdida espontánea de un feto antes de la semana 20 del embarazo. La pérdida del embarazo después de 20 semanas se llama muerte fetal. Un aborto espontáneo es un suceso que ocurre naturalmente, a diferencia de los abortos médicos o abortos quirúrgicos.

API acrónimo de Application Programming Interfaces, que en español significa interfaz de programación de aplicaciones. Se trata de un conjunto de definiciones y protocolos que permite la comunicación entre dos aplicaciones de software a través de un conjunto de reglas.

aprendizaje no supervisado método de aprendizaje automático que trata de encontrar patrones no conocidos en un conjunto de datos sin información previa.

aprendizaje supervisado método de aprendizaje automático que trata de encontrar patrones en datos desconocidos aprendiento datos de entrenamiento previamente etiquetados.

CRISP-DM acrónimo de Cross Industry Standard Process for Data Mining.

código abierto software cuyo código fuente se ha puesto a disposición de todo el mundo de manera gratuita y otorgado con licencias que facilita su reutilización o adaptación a contextos diferentes.

hashtag Conjunto de caracteres precedidos por una almohadilla (#) que sirve para identificar o etiquetar un mensaje en las webs de microblogs.

LDA acrónimo de Latent Dirichlet Analysis.

machine learning subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan a partir de transformación estadísticas lineales.

NLP acrónimo de Natural Language Processing (procesamiento del lenguaje natural).

Python lenguaje de programación de alto nivel diseñado para ser fácil de leer y sencillo de implementar, además de ser de código abierto, lo que significa que su uso es gratuito, incluso para aplicaciones comerciales.

retweet consiste en publicar nuevamente un Tweet propio o de otra persona.

streaming tecnología que permite transmitir archivos de audio y vídeo en un flujo continuo a través de una conexión.

tweet mensaje publicado en Twitter que contiene texto, fotos, GIF o video.

tweet favorito modo que se usa en la red social Twitter para guardar los tweets que un usuario considera sus favoritos y poder leerlos más tarde.

Twitter red social y servicio de microblogging usado para la comunicación en tiempo real utilizado por millones de personas y organizaciones.

VADER acrónimo de Valence Aware Dictionary and sEntiment Reasoner.

Bibliografía

[1] Women's Brain Project. Women's Brain Project Site; 2022. Accessed: 2022-01-01. Available from: https://www.womensbrainproject.com/.

[2] March of Dimes. Miscarriage; 2021. Accessed: 2021-09-26. https://www.marchofdimes.org/complications/miscarriage.aspx.

[3] World Health Organization. Why we need to talk about losing a baby; 2021. Accessed: 2021-09-26. https://www.who.int/news-room/spotlight/why-we-need-to-talk-about-losing-a-baby.

[4] Kolte A, Olsen L, Mikkelsen E, Christiansen O, Nielsen H. Depression and emotional stress is highly prevalent among women with recurrent pregnancy loss. Human reproduction. 2015;30(4):777–782.

[5] Farren J, Jalmbrant M, Ameye L, Joash K, Mitchell-Jones N, Tapp S, et al. Post-traumatic stress, anxiety and depression following miscarriage or ectopic pregnancy: a prospective cohort study. BMJ open. 2016;6(11).

[6] World Health Organization. Defining competent maternal and newborn health professionals; 2018. Accessed: 2021-09-26. https://apps.who.int/iris/bitstream/handle/10665/272817/9789241514200-eng.pdf?ua=1.

[7] Bardos J, Hercz D, Friedenthal J, Missmer SA, Williams Z. A national survey on public perceptions of miscarriage. Obstetrics and gynecology. 2015;125(6).

[8] Kuchinskaya O, Parker L. 'Recurrent losers unite': Online forums, evidence-based activism, and pregnancy loss. Social science & medicine (1982). 2018;216:74–80.

[9] Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, et al. CRISP-DM 1.0 Step-by-step data mining guide. The CRISP-DM consortium; 2000. Available from: https://www.the-modeling-agency.com/crisp-dm.pdf.

[10] Indurkhya N, Damerau FJ. Handbook of Natural Language Processing. 2nd ed. Chapman & Hall/CRC; 2010.

[11] Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Sanz Lamora N, Álvarez A, González-Rodríguez A, et al. Characterization of Anorexia Nervosa on Social Media: Textual, Visual, Relational, Behavioral, and Demographical Analysis. J Med Internet Res. 2021 Jul;23(7):e25925. Available from: https://www.jmir.org/2021/7/e25925.

[12] Leis A, Ronzano F, Mayer MA, Furlong LI, Sanz F. Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. J Med Internet Res. 2019 Jun;21(6):e14199. Available from: http://www.jmir.org/2019/6/e14199/.

[13] Leis A, Ronzano F, Mayer MA, Furlong LI, Sanz F. Evaluating Behavioral and Linguistic Changes During Drug Treatment for Depression Using Tweets in Spanish: Pairwise Comparison Study. J Med Internet Res. 2020 Dec;22(12):e20920. Available from: http://www.jmir.org/2020/12/e20920/.

[14] Ramírez-Cifuentes D, Freire A, Baeza-Yates R, Puntí J, Medina-Bravo P, Velazquez DA, et al. Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis. J Med Internet Res. 2020 Jul;22(7):e17758. Available from: https://www.jmir.org/2020/7/e17758.

[15] Ferrara E, Yang Z. Measuring Emotional Contagion in Social Media. PLoS ONE. 2015;10.

[16] Graells-Garrido E, Baeza-Yates R, Lalmas M. Representativeness of Abortion Legislation Debate on Twitter: A Case Study in Argentina and Chile. Companion Proceedings of the Web Conference 2020. 2020.

[17] Mercier RJ, Senter K, Webster R, Riley AH. Instagram Users' Experiences of Miscarriage. Obstetrics & Gynecology. 2019.

[18] Cesare N, Oladeji O, Ferryman K, Wijaya D, Hendricks-Muñoz KD, Ward A, et al. Discussions of miscarriage and preterm births on Twitter. Paediatric and Perinatal Epidemiology. 2020;34(5):544-52. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ppe.12622.

[19] Klein AZ, Cai H, Weissenbacher D, Levine LD, Gonzalez-Hernandez G. A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes. Journal of Biomedical Informatics. 2020;112:100076. Articles initially published in Journal of Biomedical Informatics: X 5-8, 2020. Available from: https://www.sciencedirect.com/science/article/pii/S2590177X2030010X.

[20] Twitter Developer Platform. Twitter API Tools & Libraries; 2022. Accessed: 2021-10-17. Available from: https://developer.twitter.com/en/docs/twitter-api/tools-and-libraries.

[21] Twitter Developer Platform. Twitter API Tools Documentation; 2022. Accessed: 2022-01-01. Available from: https://developer.twitter.com/en/docs/twitter-api.

[22] Tweepy. Tweepy Documentation; 2022. Accessed: 2022-01-01. Available from: https://docs.tweepy.org/en/stable/.

[23] Twitter Developer Platform. Twitter Developer Platform; 2022. Accessed: 2022-01-01. Available from: https://developer.twitter.com/en.

[24] Twitter Developer Platform. Rate limits; 2022. Accessed: 2022-01-01. Available from: https://developer.twitter.com/en/docs/twitter-api/rate-limits.

[25] Twitter Developer Platform. Data Dictionary: Tweet Object; 2022. Accessed: 2022-01-01. Available from: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet.

[26] JSON. Introducing JSON; 2022. Accessed: 2022-01-01. Available from: https://www.json.org/json-en.html.

[27] Wikipedia. Pregnancy and Infant Loss Remeberance Day; 2022. Accessed: 2022-01-01. Available from: https://en.wikipedia.org/wiki/Pregnancy_and_Infant_Loss_Remembrance_Day.

[28] Awareness Days. Global Wave of Ligth 2021; 2021. Accessed: 2022-01-01. Available from: https://www.awarenessdays.com/awareness-days-calendar/global-wave-of-light-2021/.

[29] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press; 2008. Available from: http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

[30] Hutto CJ, Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text; 2015. .

[31] GitHub. VADER Sentiment Analysis; 2014. Accessed: 2022-01-01. Available from: https://github.com/cjhutto/vaderSentiment.

[32] TextBlob. TextBlob: Simplified Text Process; 2020. Accessed: 2022-01-01. Available from: https://textblob.readthedocs.io/en/dev/.

[33] GitHub. VADER Sentiment Analysis: About the scoring; 2014. Accessed: 2022-01-01. Available from: https://github.com/cjhutto/vaderSentiment#about-the-scoring.

[34] Bafna P, Pramod D, Vaidya A. Document clustering: TF-IDF approach. In: 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT); 2016. p. 61-6.

[35] Wang YX, Zhang YJ. Nonnegative Matrix Factorization: A Comprehensive Review. IEEE Transactions on Knowledge and Data Engineering. 2013;25(6):1336-53.

[36] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2003;3:993-1022. Available from: http://portal.acm.org/citation.cfm?id=944937.

[37] Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics. 2010 12;1:43-52.

[38] Pleplé Q. Perplexity To Evaluate Topic Models; 2013. Accessed: 2022-01-01. Available from: http://qpleple.com/perplexity-to-evaluate-topic-models/.

[39] Chang J, Boyd-Graber JL, Gerrish S, Wang C, Blei DM. Reading Tea Leaves: How Humans Interpret Topic Models. In: Neural Information Processing Systems. vol. 22; 2009. p. 288-96.

[40] Mifrah S, Benlahmar EH. Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus. International Journal of Advanced Trends in Computer Science and Engineering. 2020 08.

[41] Röder M, Both A, Hinneburg A. Exploring the Space of Topic Coherence Measures. WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining. 2015 02:399-408.

[42] Gensim. Gensim Documentation: Latent Dirichlet Allocation; 2022. Accessed: 2022-01-01. Available from: https://radimrehurek.com/gensim/models/ldamodel.html.

[43] Hancock J. In: Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient); 2004. .

[44] Gilbert AC. After miscarriage, woman is convicted of manslaughter. The 'fetus was not viable,' advocates say. USA TODAY. 2021. Available from: https://eu.usatoday.com/story/news/nation/2021/10/21/oklahoma-woman-convicted-of-manslaughter-miscarriage/6104281001/.

[45] Thompson P, Cruz AT. How an Oklahoma women's miscarriage put a spotlight on racial disparities in prosecutions. NBC News. 2021. Available from: https://www.nbcnews.com/news/us-news/woman-prosecuted-miscarriage-highlights-racial-disparity-similar-cases-rcna4583.

[46] Levinson-King R. US women are being jailed for having miscarriages. BBC News. 2021. Available from: https://www.bbc.com/news/world-us-canada-59214544.

[47] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. American Psychiatric Publishing; 2013.

[48] Denecke K, Abd-alrazaq A, Househ M. In: Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges; 2021. p. 115-28.

[49] Northwell Health. Gaslighting in women's health: No, it's not just in your head; 2020. Accessed: 2022-01-01. Available from: https://www.northwell.edu/katz-institute-for-womens-health/articles/gaslighting-in-womens-health.

[50] TODAY. Dismissed: There's a gender gap at the doctor's office every woman needs to know about. Meet the doctors fighting for change; 2019. Accessed: 2022-01-01. Available from: https://www.today.com/health/dismissed-health-risk-being-woman-t153804.

Apéndice A

Consulta del código del proyecto

El código generado para este proyecto se encuentra alojado en el siguiente repositorio de Github:

https://github.com/LauraPlanas/MiscarriageTwitterAnalysis

Apéndice B

Lista de tweets que identifican problemas alrededor del aborto espontáneo

B.1. Problemas económicos

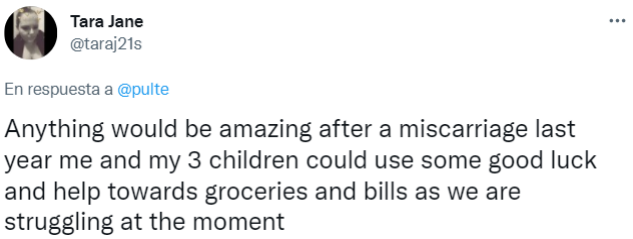


Figura B.1. Ejemplo de tweet que muestra problemas económicos

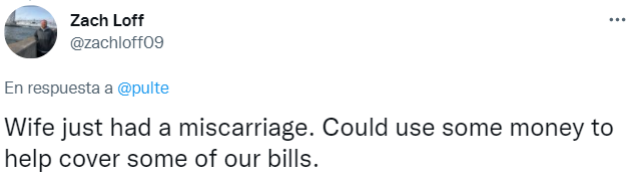


Figura B.2. Ejemplo de tweet que muestra problemas económicos

B.2. Problemas de estigma

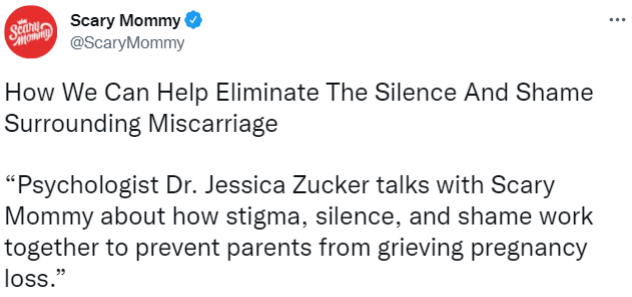


Figura B.3. Ejemplo de tweet que muestra problemas de estigma

Figura B.4. Ejemplo de tweet que muestra problemas de estigma

B.3. Problemas del sistema sanitario

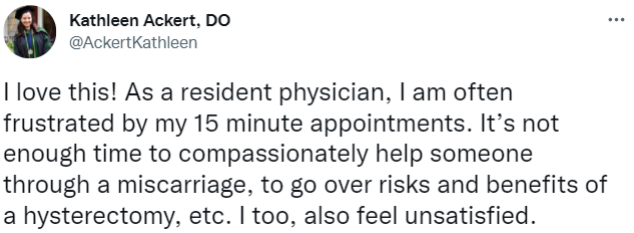


Figura B.5. Ejemplo de tweet que muestra problemas en el sistema sanitario

Figura B.6. Ejemplo de tweet que muestra problemas en el sistema sanitario

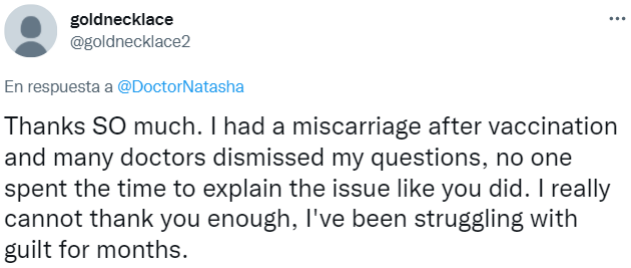


Figura B.7. Ejemplo de tweet que muestra problemas en el sistema sanitario

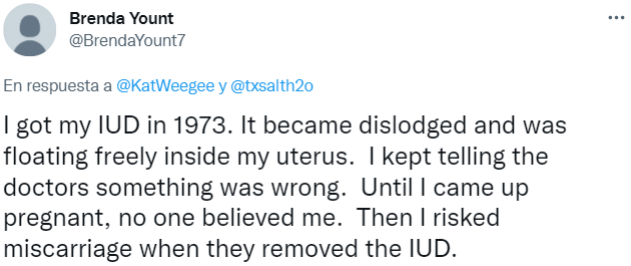


Figura B.8. Ejemplo de tweet que muestra problemas en el sistema sanitario

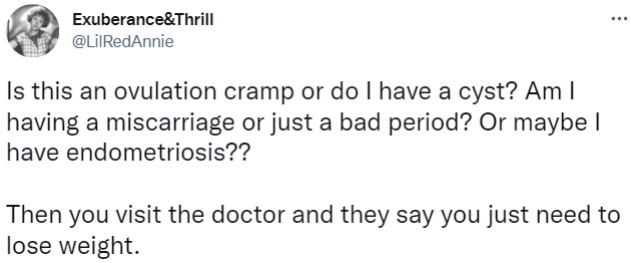


Figura B.9. Ejemplo de tweet que muestra problemas en el sistema sanitario

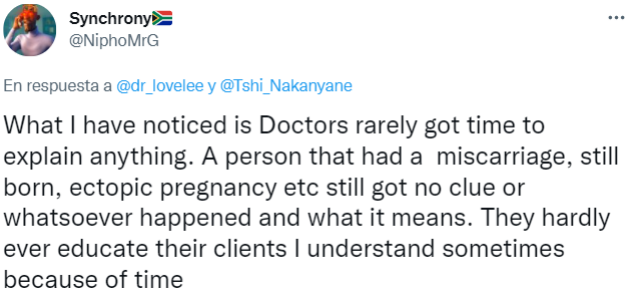


Figura B.10. Ejemplo de tweet que muestra problemas en el sistema sanitario

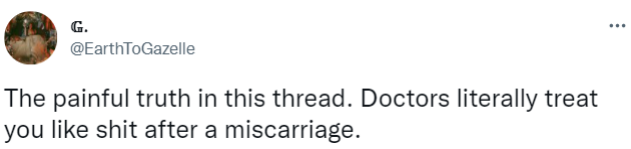


Figura B.11. Ejemplo de tweet que muestra problemas en el sistema sanitario

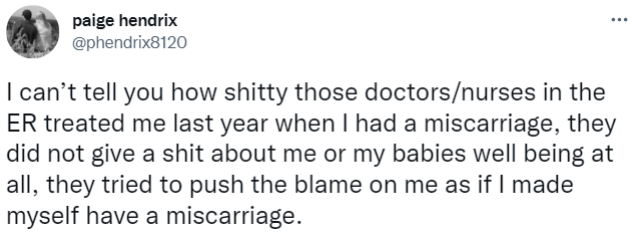


Figura B.12. Ejemplo de tweet que muestra problemas en el sistema sanitario

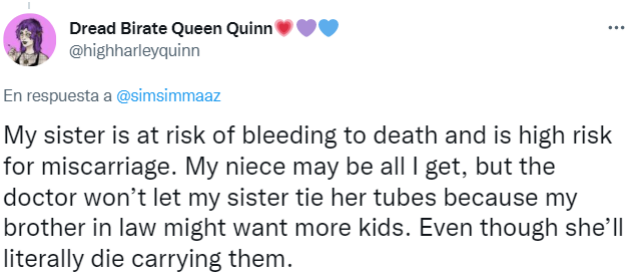


Figura B.13. Ejemplo de tweet que muestra problemas en el sistema sanitario

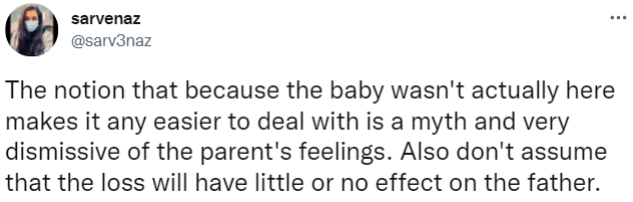


Figura B.14. Ejemplo de tweet que muestra problemas en el sistema sanitario