# Clasificación automática de textos sobre Trastornos de Conducta Alimentaria (TCA) obtenidos de Twitter

**Autor: José Alberto Benítez-Andrades**
**Trabajo de Fin de Máster**
**Máster Universitario en Ingeniería y Ciencia de Datos**
**Curso 2020/2021**

**Directores:**
**Dr. D. Rafael Pastor Vargas**
**Dra. Dª. María Teresa García Ordás**

# Estructura de la presentación

1. Introducción y objetivos
2. Estado del arte
3. Metodología
4. Experimentos y resultados
5. Discusión y conclusiones
6. Planificación y estimación de costes
7. Aportaciones científicas realizadas

# 1. Introducción y objetivos

# 1. Introducción y objetivos

## Antecedentes y contexto

- Importancia al aspecto físico (Harris & Carr, 2001; Izquierdo et al. 2019; Urdapilleta et al. 2019; Lou & Tse, 2020).
- Problemática: sobrepeso y obesidad en la infancia (Lobstein et al., 2015; OMS, 2020).
- Trastornos de Conducta Alimentaria (TCA): enfermedades potencialmente mortales de naturaleza psicológica y física (Griffen et al. 2018; Hoek, 2016).
- Los medios de comunicación social en la investigación (Ackland, 2009; Carceller-Maicas, 2016; Lopez-Castroman et al. 2020; Timmins eta l. 2018).
- Twitter: versátil, facilidad para recopilar y procesar datos (Timmins et al. 2018).

# 1. Introducción y objetivos

## Antecedentes y contexto

- Redes, TCA e IA:
  - Detección de tuits a favor y en contra de los TCA (Fettach & Benhiba, 2019; Lewis & Arbuthnott, 2012; Oksanen et al. 2015).
  - Detección de comunidades de personas con TCA (Wang et al. 2018).

- No se han encontrado estudios que hagan uso del conjunto de datos para crear un clasificador que detecte:
  - Tuits escritos por personas que padecen o han padecido TCA.
  - Tuits que fomentan el padecer TCA.
  - Tuits informativos o no informativos.
  - Tuits de carácter científico.

# 1. Introducción y objetivos

## Objetivos generales y específicos

- **Pregunta 1:**
  - ¿Es posible conseguir modelos de aprendizaje automático o aprendizaje profundo capaces de clasificar de forma precisa tuits sobre TCA en las 4 categorías mencionadas?

- **Pregunta 2:**
  - ¿Qué modelos de aprendizaje automático o aprendizaje profundo obtienen mejores resultados a la pregunta 1?

# 1. Introducción y objetivos

## Objetivos generales y específicos

- **Objetivo 1:**
  - Aplicar técnicas de minería de datos que permitan generar un conjunto de datos para su posterior etiquetado y preprocesamiento.
- **Objetivo 2:**
  - Generar modelos de aprendizaje automático y aprendizaje profundo capaces de clasificar textos sobre TCA con alto grado de exactitud.
- **Objetivo 3:**
  - Comparar entre los diferente smodelos cuál es el que mejores resultados ofrece en base a las categorías que se pretenden clasificar o predecir.

# 2. Estado del arte

# 2. Estado del arte

## Medios sociales

- **Medios sociales:**
  - Aplicaciones basadas en la Web 2.0 de Internet que fomentan la participación de los usuarios.
  - El contenido generado por el usuario es esencial.
  - Los individuos generan perfiles específicos.
  - (Boyd & Ellison, 2007; Obar & Wildman, 2015; Sohota, 2020).
- Más de 2.500 millones de usuarios en todo el mundo (Statista, 2021).
- Redes de intereses específicos.
- Twitter
  - 280 caracteres, debate público amplio, usado en ciencias sociales y de la salud, de fácil acceso gracias a API.
  - La mayoría de estudios TCA y redes usa Twitter.

UNED

# 2

## Medic

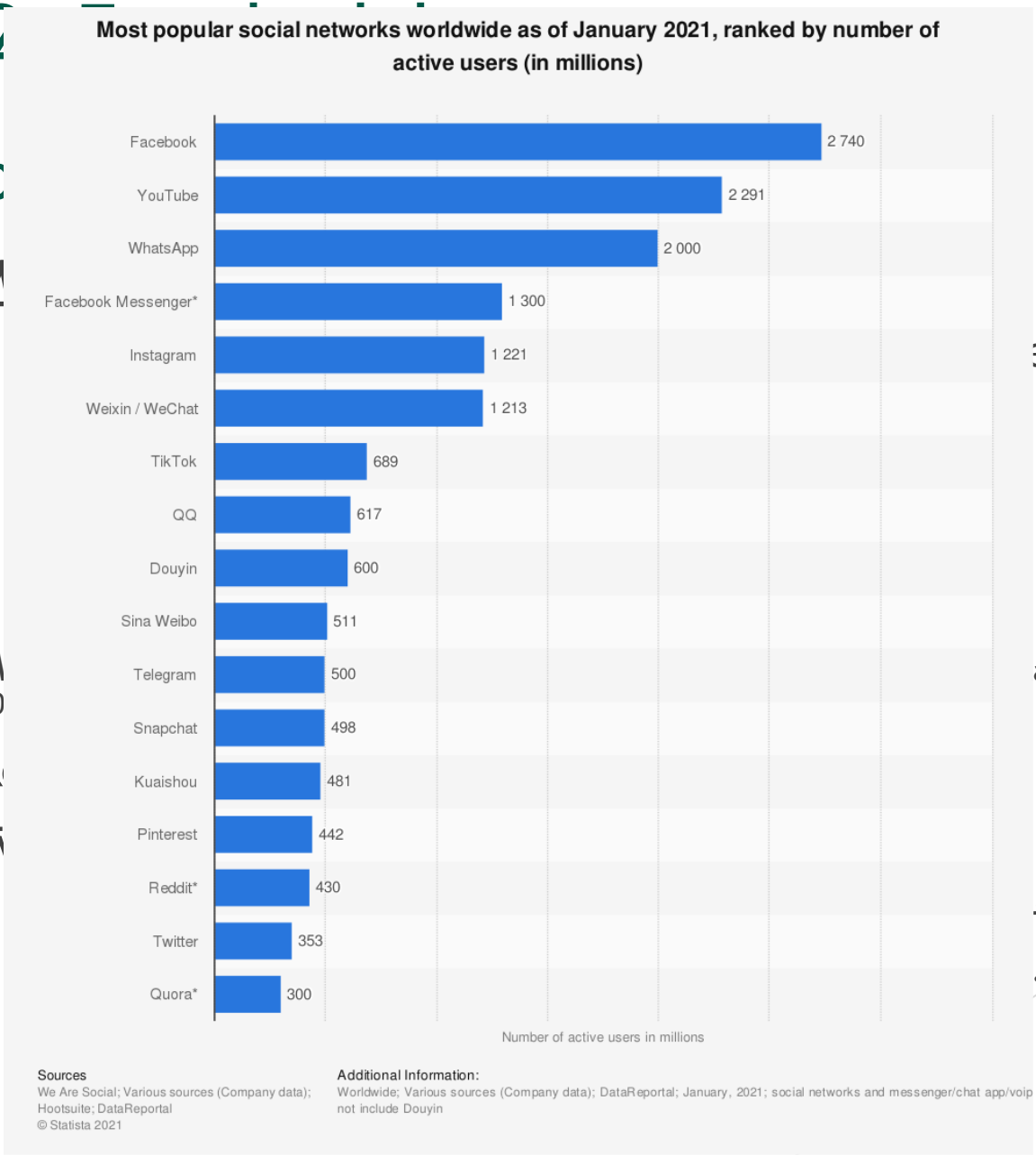
- M
- M
  20
- R
- Tv

e

atista,

iencias

.



# 2. Estado del arte

## Medios sociales

- Los estudios de redes digitales deben tomarse con precaución a la hora de generalizar los resultados.
- Los usuarios no representan necesariamente a toda la población.
  - Blank, 2017 determinó que los usuarios procedentes de Reino Unido suelen ser más jóvenes, ricos y educados en comparación con otros usuarios de Internet de Reino Unido.
  - Los usuarios de Estados Unidos también son más jóvenes y más ricos, pero no mejor educados.
  - Hay un sesgo por el hecho de que es más común que las personas más adineradas hagan uso de ciertos medios sociales.

# 2. Estado del arte

## Informática sanitaria y medios sociales

- Tres áreas principales de la informática sanitaria que hacen uso de Twitter (Zhang & Ahmed, 2019).
  - Ayuda a predecir eventos futuros.
    - Vigilancia sindrómica en epidemias, farmacovigilancia…
  - Usuarios y de su comportamiento.
    - Demografía de los usuarios, estructura de la red, frecuencia de la comunicación...
  - Impacto de los contenidos compartidos.
    - Twitter como apoyo emocional o como herramienta para difundir concienciación.

# 2. Estado del arte

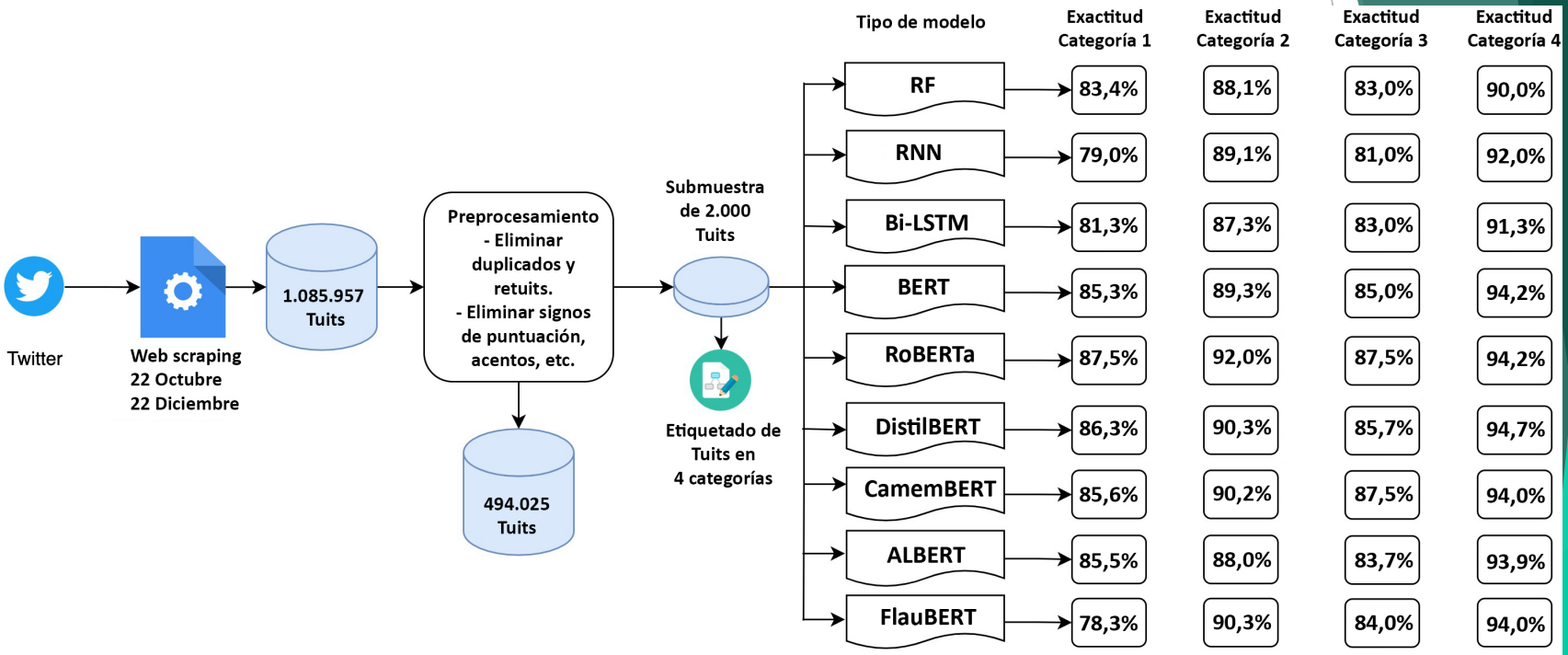
## Métodos de clasificación e informática sanitaria

- Aprendizaje automático supervisado.
- La clasificación es un método eficiente para categorizar grandes conjuntos de datos (Oussous et al. 2018).
  - Mayor precisión de las predicciones en comparación con el trabajo humano manual (Geirhos et al. 2020; Van Der Walt & Eloff, 2018; Youyou et al. 2015).
- Diversos métodos posibles:
  - Naïve Bayes, Support Vector Machine (SVM), Random Forest (RF), Decision Trees (DT), Gradient Boosting Trees (GBT), Gradient Boosted Regression Trees (GBRT) o Logistic Regression (LG).
  - Naïve Bayes para clasificar cuatro estados de salud: gripe, depresión, embarazo, TCA (Prieto et al. 2014).
  - SVM para pronosticar brotes de enfermedades. Fiebre hemorrágica del dengue (Kesorn et al. 2015). Incluso COVID-19 (Dixit et al. 2021; Singh et al. 2020).
  - GBRT para predecir usuarios que tuitean sobre cigarrillos electrónicos (Kim et al. 2017).

# 3. Metodología

# 3. Metodología

## Flujo de trabajo




| Tipo de modelo | Exactitud Categoría 1 | Exactitud Categoría 2 | Exactitud Categoría 3 | Exactitud Categoría 4 |
|---|---|---|---|---|
| RF | 83,4% | 88,1% | 83,0% | 90,0% |
| RNN | 79,0% | 89,1% | 81,0% | 92,0% |
| Bi-LSTM | 81,3% | 87,3% | 83,0% | 91,3% |
| BERT | 85,3% | 89,3% | 85,0% | 94,2% |
| RoBERTa | 87,5% | 92,0% | 87,5% | 94,2% |
| DistilBERT | 86,3% | 90,3% | 85,7% | 94,7% |
| CamemBERT | 85,6% | 90,2% | 87,5% | 94,0% |
| ALBERT | 85,5% | 88,0% | 83,7% | 93,9% |
| FlauBERT | 78,3% | 90,3% | 84,0% | 94,0% |

# 3. Metodología

## Recopilación

- 22 de octubre de 2021 a 22 de diciembre de 2021.
- T-Hoarder para recopilar tuits (Congosto et al. 2017).
  - 3 experimentos que contenían un total de 23 términos relacionados con TCA, en inglés.
- Etiquetado manual de un subconjunto clasificándolos en 4 categorizaciones:
  - Categorización 1: tuits escritos por personas que padecen TCA.
  - Categorización 2: tuits que fomentan padecer un TCA.
  - Categorización 3: tuits informativos o de opinión.
  - Categorización 4: tuits de carácter científico.

# 3. Metodología

R(

**Tabla 3.2**: *Ejemplos de tuits categorizados.*

| Categoría | Tema | Tuit |
|---|---|---|
| Tipo 1 | Escrito por alguien que padece TCA | i was stressed and ate a whole bowl of pasta, where's my badge for being the worst anorexic #edtwt |
| Tipo 1 | Escrito por alguien que no padece TCA | Is your #teenager not eating or eating a lot less than normal? She might be suffering from #anorexia. We can help; please come see us https://t.co/GfStM1IVGz #weightloss #losingweight https://t.co/z5NK0tjNIt |
| Tipo 2 | Fomenta los TCA | Currently feeling like the best anorexic #EDtwt https://t.co/1BZPMs8bGU |
| Tipo 2 | No fomenta los TCA | Higher-calorie diets could lead to a speedier recovery in patients with anorexia nervosa, study shows https://t.co/mipX3nrhHN #mentalhealth #diet #anorexia |
| Tipo 3 | Informativo | #AnorexiaNervosa – A Father and Daughter Perspective - Highlights from RCPsychIC 2019 #EatingDisorders #mentalhealth https://t.co/iq3GH5ce6C |
| Tipo 3 | De opinión | Binge eating makes me sad :( #eatingdisorder #bingeeating https://t.co/0jjf7YrVyc |
| Tipo 4 | Científico | The problem extends to Food and Drug Administration and National Institutes of Health datasets used in a recent study appearing in Reproductive Toxicology. #ai #technology #BigData #ML https://t.co/DFvh6gNA38 |
| Tipo 4 | No científico | Do not waste time thinking about what you could have done differently. Keep your eyes on the road ahead and do it differently now. #anorexia #eatingdisorder #recovery #nevergiveup #alwayskeepfighting https://t.co/YalYzclBDM |

# 3. Metodología

## Métodos de clasificación utilizados

- Métodos utilizados:
  - Random Forest (RF).
    - Se utilizó una validación cruzada con k = 5.
  - Recurrent Neural Network (RNN).
  - Redes Bi-LSTM.
  - Modelos BERT.
    - Se hizo uso de *ClassificationModel* de la biblioteca *simpletransformers*.
    - Se utilizaron 6 modelos pre-entrenados (BERT, RoBERTa, DistilBERT, CamemBERT, ALBERT y FlauBERT).
- Se usó scikit-learn para dividir el conjunto de datos y obtener las métricas *f1 score* y *accuracy*.
- En las redes neuronales se aplicaron 5 iteraciones.

# 3. Metodología

## Random Forest, RNN y Bi-LSTM

- Random Forest:
  - Construye múltiples árboles de decisión y los fusiona para obtener una predicción más precisa y estable.
- RNN:
  - Conexiones entre nodos forman un grafo dirigido a lo largo de una secuencia temporal.
  - Frecuente para reconocimiento de escritura o habla.
- Bi-LSTM:
  - LSTM Bidireccionales: extensión de las LSTM que pueden mejorar el rendimiento en problemas de clasificación de secuencias.

# 3. Metodología

## BERT

- Bidirectional Encoder Representations from Transformers (BERT).
- Novedad en procesamiento del lenguaje natural PLN.
- Innovación técnica: entrenamiento bidireccional del *Transformer*.
- En este trabajo se aplican 6 modelos preentrenados BERT a 4 categorizaciones.
- Funcionamiento:
  - Los *Transformer* disponen de dos mecanismos, uno lee el texto de entrada, codificador y otro realiza una predicción, decodificador.
  - Lee toda la secuencia de palabras a la vez, de izquierda a derecha y de derecha a izquierda.
  - Hacen uso de *Masked LM y Next Sentence Prediction*.

# 3. Metodología

## Configuración

- Jupyter lab y Python 3.6.
- Pytorch para los modelos BERT.
- Tensorflow para RF, RNN y Bi-LSTM.
- Se dividieron los datos en un 70% para entrenamiento y 30% para validación.

# 4. Experimentación y resultados

# 4. Experimentación y resultados

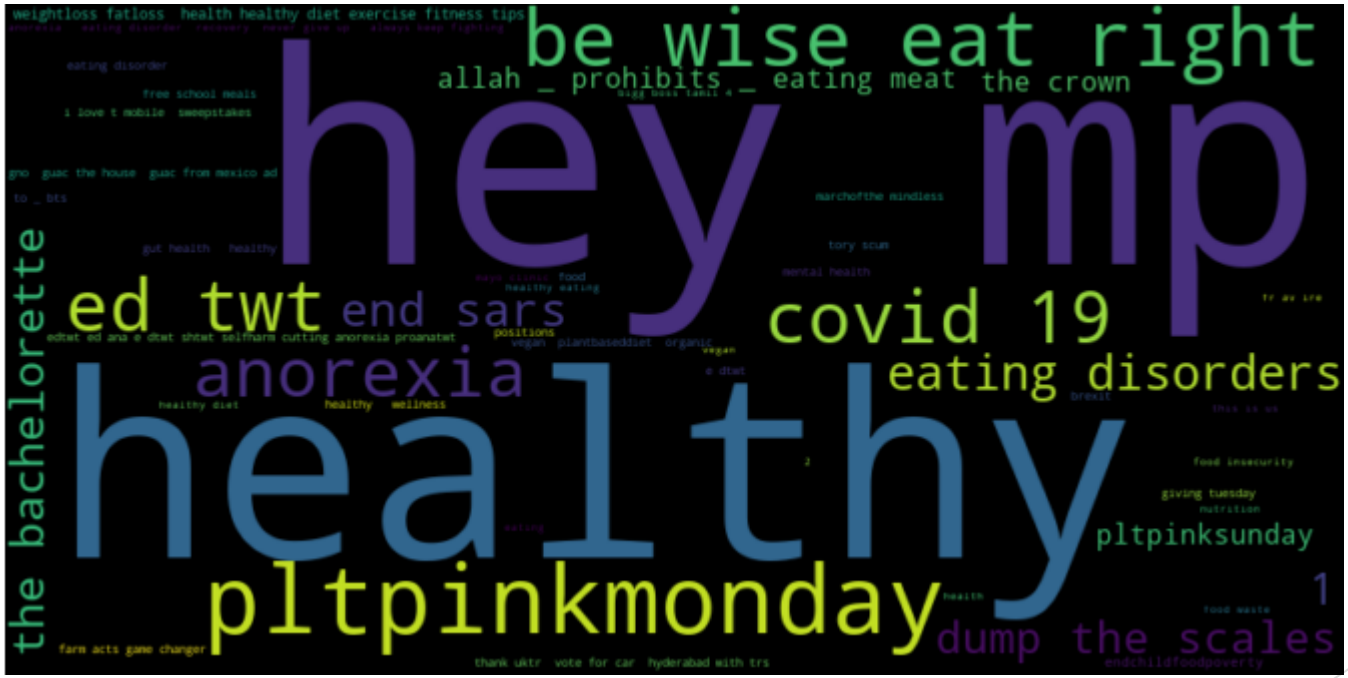
## Resultados de la recopilación y preprocesamiento

- Datos recopilados entre el 22 de octubre de 2020 y el 22 de diciembre de 2020.
- 1.085.957 tuits recopilados.
- Tras el preprocesamiento: 494.025 tuits.

# 4. Experimentación y resultados

## Resultados de la recopilación y preprocesamiento

- Palabras más repetidas en los 494.025 tuits



# 4. Experimentación y resultados

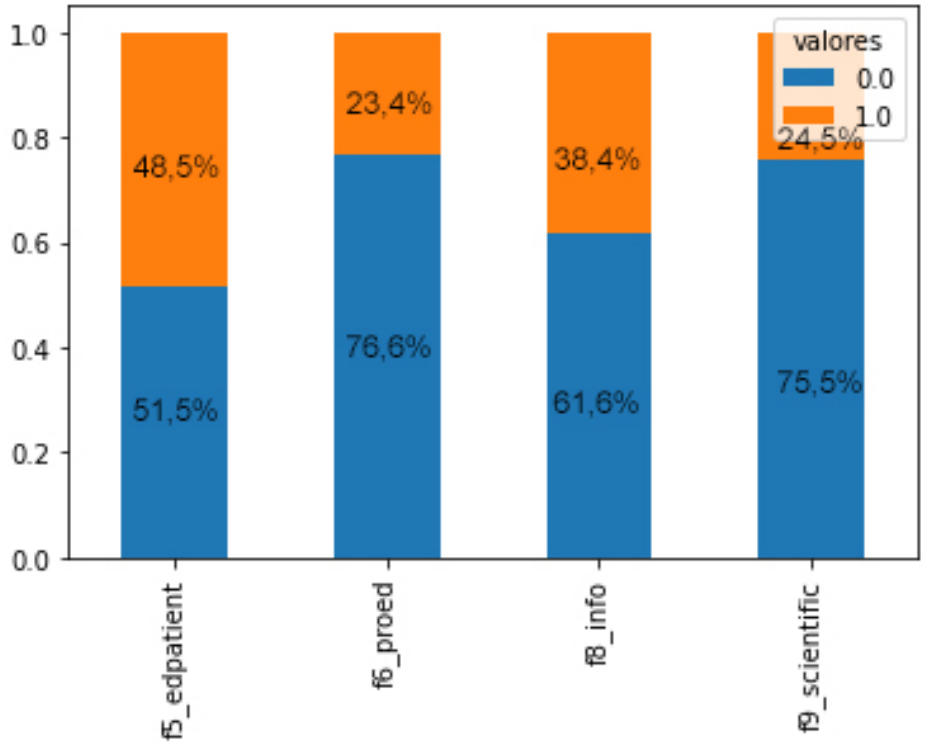
## Resultados de la recopilación y preprocesamiento

- Palabras más repetidas en la submuestra de 2.000 tuits.

# 4. Experimentación y resultados

## Resultados de la recopilación y preprocesamiento

- Categorizaciones realizadas y balance de datos:



# 4. Experimentación y resultados

## Resultados

**Tabla 4.1**: *Resultados obtenidos con los diferentes modelos para la clasificación de la categorización 1.*

| Tipo de modelo | Modelo preentrenado | f1 | acc | tiempo |
|---|---|---|---|---|
| Random Forest | | 0,831 | 83,4 % | 94,3s |
| RNN | | 0,770 | 79,0 % | 152,1s |
| Bi-LSTM | | 0,780 | 81,3 % | 163,2s |
| BERT | bert-based-multilingual-cased | 0,848 | 85,3 % | 1257,4s |
| RoBERTa | roberta-base | 0,868 | 87,5 % | 1116,2s |
| DistilBERT | distilbert-base-cased | 0,856 | 86,3 % | 1343,3s |
| CamemBERT | camembert-base | 0,838 | 85,6 % | 1472,3s |
| ALBERT | albert-base-v1 | 0,849 | 85,5 % | 1372,7s |
| FlauBERT | flaubert_base_cased | 0,757 | 78,3 % | 1203,9s |

**Tabla 4.2**: *Resultados obtenidos con los diferentes modelos para la clasificación de la categorización 2.*

| Tipo de modelo | Modelo preentrenado | f1 | acc | tiempo |
|---|---|---|---|---|
| Random Forest | | 0,711 | 88,1 % | 99,2s |
| RNN | | 0,800 | 89,1 % | 163,1s |
| Bi-LSTM | | 0,780 | 87,3 % | 175,3s |
| BERT | bert-based-multilingual-cased | 0,746 | 89,3 % | 1232,1s |
| RoBERTa | roberta-base | 0,818 | 92,0 % | 1158,8s |
| DistilBERT | distilbert-base-cased | 0,780 | 90,3 % | 1327,8s |
| CamemBERT | camembert-base | 0,763 | 90,2 % | 1457,5s |
| ALBERT | albert-base-v1 | 0,714 | 88,0 % | 1352,3s |
| FlauBERT | flaubert_base_cased | 0,784 | 90,3 % | 1207,1s |

# 4. Experimentación y resultados

## Resultados

**Tabla 4.3**: *Resultados obtenidos con los diferentes modelos para la clasificación de la categorización 3.*

| Tipo de modelo | Modelo preentrenado | f1 | acc | tiempo |
|---|---|---|---|---|
| Random Forest | | 0,790 | 83,0 % | 95,3s |
| RNN | | 0,780 | 81,0 % | 151,5s |
| Bi-LSTM | | 0,793 | 83,0 % | 164,8s |
| BERT | bert-based-multilingual-cased | 0,811 | 85,0 % | 1292,7s |
| RoBERTa | roberta-base | 0,840 | 87,5 % | 1142,5s |
| DistilBERT | distilbert-base-cased | 0,809 | 85,7 % | 1332,0s |
| CamemBERT | camembert-base | 0,847 | 87,5 % | 1462,0s |
| ALBERT | albert-base-v1 | 0,791 | 83,7 % | 1331,3s |
| FlauBERT | flaubert_base_cased | 0,797 | 84,0 % | 1202,1s |

**Tabla 4.4**: *Resultados obtenidos con los diferentes modelos para la clasificación de la categorización 4.*

| Tipo de modelo | Modelo preentrenado | f1 | acc | tiempo |
|---|---|---|---|---|
| Random Forest | | 0,83 | 90,0 % | 93,4s |
| RNN | | 0,860 | 92,0 % | 148,4s |
| Bi-LSTM | | 0,853 | 91,3 % | 154,3s |
| BERT | bert-based-multilingual-cased | 0,888 | 94,2 % | 1272,4s |
| RoBERTa | roberta-base | 0,880 | 94,2 % | 1149,1s |
| DistilBERT | distilbert-base-cased | 0,890 | 94,7 % | 1328,5s |
| CamemBERT | camembert-base | 0,885 | 94,0 % | 1403,6s |
| ALBERT | albert-base-v1 | 0,876 | 93,9 % | 1302,9s |
| FlauBERT | flaubert_base_cased | 0,875 | 94,0 % | 1227,1s |

# 4. Experimentación y resultados

## Resultados de la recopilación y preprocesamiento

- Porcentajes de mejora entre el mejor modelo BERT y el mejor modelo de entre RF, RNN y Bi-LST:
  - Categorización 1:
    - RoBERTa 87,5% vs RF 83,4% = 4,92% de mejora
  - Categorización 2:
    - RoBERTa 92,0% vs RNN 89,1% = 3,25% de mejora
  - Categorización 3:
    - RoBERTa 87,5% vs RF 83,0% = 5,42% de mejora
  - Categorización 4:
    - DistilBERT 94,7% vs RNN 92,0% = 2,94% de mejora

# 5. Discusión y conclusiones

# 5. Discusión y conclusiones

## Discusión

- A pesar de que los modelos BERT obtuvieron mejores resultados de exactitud (*accuracy*) el coste computacional fue muy elevado.
- Incluso en los modelos más simples se obtuvieron resultados prometedores no excediendo un 5,42% de mejora entre los modelos BERT los más simples.

# 5. Discusión y conclusiones

## Conclusiones

- Respuesta a la pregunta 1:
  - Ha sido posible conseguir modelos de aprendizaje automático y aprendizaje profundo capaces de clasificar de forma precisa los tuits sobre TCA en las cuatro categorías planteadas en este estudio.
- Respuesta a la pregunta 2:
  - Los modelos de aprendizaje automático que mejor resultado obtuvieron fueron los BERT.

# 5. Discusión y conclusiones

## Líneas futuras

- Aumentar el conjunto de datos de entrenamiento y validación con un mayor número de tuits etiquetados.
- Aplicar técnicas de PLN que hagan uso de ontologías para plantear automatizaciones y razonamientos lógicos.
- Integrar modelos predictivos en un proyecto de desarrollo real como, por ejemplo, un bot de Twitter capaz de detectar si los tuits están siendo escritos por personas que padecen o han padecido TCA.
- Investigar en mayor profundidad los efectos de los TCA.
- Hacer uso de otras plataformas de medios sociales.

# 6. Planificación y estimación de costes

# 6. Planificación y estimación de costes

## Planificación del trabajo

**Tabla 6.1**: *Estimación de trabajo y coste.*

| Tipo de tarea | Duración |
|---|---|
| Revisión bibliográfica de la literatura | 6 días |
| Instalación del entorno de recopilación de tuits | 5 días |
| Instalación de programas | 2 días |
| Tratamiento de los datos en crudo | 8 días |
| Etiquetado de la submuestra de 2.000 tuits | 10 días |
| Programación de los algoritmos de aprendizaje automático | 13 días |

## Planificación del trabajo

**Tabla 6.2**: *Salario medio según el rol de los trabajadores.*

| Recursos | Salario/año | Salario/mes | Salario/día | Salario/hora |
|---|---|---|---|---|
| Ingeniero en informática | 26.404,00 € | 2.200,33 € | 110,02 € | 13,75 € |
| Ingeniero de datos | 36.001,00 € | 3.000,08 € | 150,00 € | 18,75 € |
| Científico de datos | 32.937,00 € | 2.744,75 € | 137,24 € | 17,15 € |

**Tabla 6.3**: *Estimación de trabajo y coste.*

| Recurso | Trabajo (horas) | Coste (€) |
|---|---|---|
| Ingeniero en Informática | 38 | 522,50 |
| Ingeniero de datos | 147 | 2.756,25 |
| Científico de datos | 174 | 2.984,10 |
| **Total** | **359** | **6.262,85** |

# 6. Planificación y estimación de costes

## Planificación del trabajo

**Tabla 6.4**: *Estimación de costes de Hardware.*

| Nombre | Coste unitario | Unidades | Coste total |
|---|---|---|---|
| Ordenador sobremesa | 2.450,00 € | 1 | 2.450,00 € |
| Ratón | 12,00 € | 1 | 12,00 € |
| Monitor | 165,00 € | 2 | 330,00 € |
| Servidor VPS | 400,00 € | 1 | 400,00 € |
| **Total hardware** | **-** | **5** | **3.192,00 €** |

**Tabla 6.5**: *Estimación de costes de Software.*

| Nombre | Coste unitario | Unidades | Coste total |
|---|---|---|---|
| Licencia de Microsoft Office | 579,00 € | 1 | 579,00 € |
| Licencia de Microsoft Project | 849,00 € | 1 | 849,00 € |
| Licencia de Windows 10 | 116,98 € | 1 | 116,98 € |
| **Total Software** | **-** | **3** | **1.544,98 €** |

# 6. Planificación y estimación de costes

## Planificación del trabajo

**Tabla 6.6**: *Presupuesto detallado para realizar.*

| Tipo de coste | Coste | Porcentaje |
|---|---|---|
| Recursos humanos | 6.262,85 € | 56,94 % |
| Hardware | 3.192,00 € | 29,02 % |
| Software | 1.544,98 € | 14,05 % |
| **Total** | **10.999,83 €** | **100 %** |

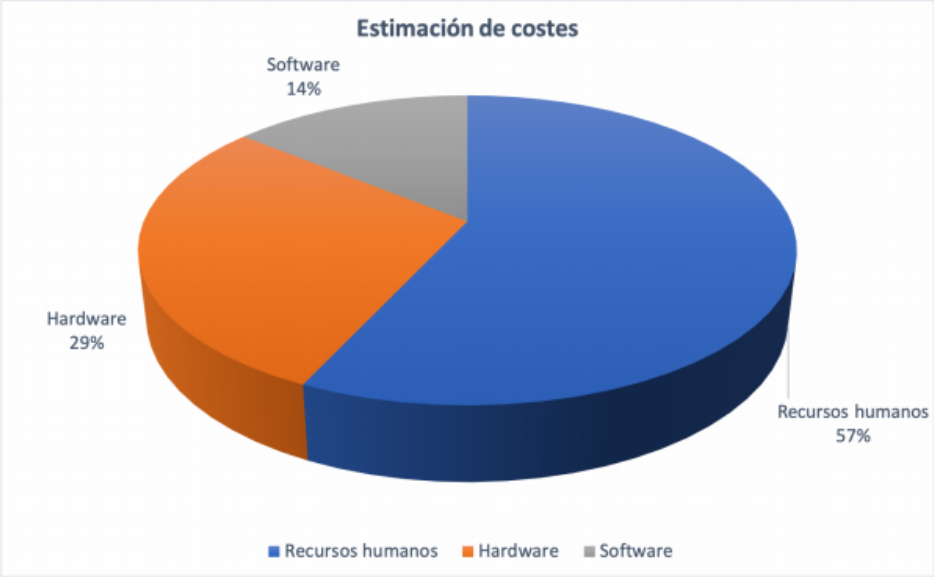



**Figura 6.1**: *Porcentaje estimado de costes según el tipo de recurso.*

# 7. Aportaciones científicas

# 7. Aportaciones científicas

- Comunicación oral en el congreso ““34th IEEE CBMS International Symposium on Computer-Based Medical Systems”
- **Título:** BERT Model-Based Approach For Detecting Categories of Tweets in the Field of Eating Disorders (ED)”
- **Autores:** José Alberto Benítez-Andrades, José Manuel Alija-Pérez, Isaías García-Rodríguez, Carmen Benavides, Héctor Alaiz-Moretón, Rafael Pastor-Vargas, María Teresa García-Ordás.


thank
you