Universidad Rey Juan Carlos

PRÁCTICA OBLIGATORIA EVALUABLE

CLASIFICACIÓN DE IDEOLOGÍAS EN TUITS

Objetivo

El objetivo es poner en práctica diferentes conceptos aprendidos durante el curso relacionados con la Minería de Texto y Procesamiento del Lenguaje Natural a través de un caso de estudio real.

Normativa de entrega

La entrega de esta práctica es ***obligatoria*** y puede realizarse ***individualmente o por parejas***. A continuación, se detallan otros datos de interés relacionados con la normativa de entrega:

- La fecha límite de entrega de la práctica será el día ***19 de mayo de 2025, a las 23:59h***.
- La entrega de la práctica se hará a través del formulario habilitado para ello en aula virtual.
- Se deberá enviar un único archivo comprimido con todo el código fuente que se haya implementado (ficheros .py o notebooks), junto con los *requirements* de ejecución necesarios para ejecutar la práctica (versiones de las librerías utilizadas, versión de Python empleada, ...).
- **Además**, en el mismo fichero comprimido se deberá adjuntar
    - Una memoria explicando lo que se ha hecho.
    - El resultado de las predicciones sobre un conjunto de test que se proporcionará más adelante.

Enunciado

Twitter (ahora X) es una plataforma de comunicación ampliamente utilizada, que tiene naturaleza de red social pero donde las relaciones son asimétricas. Es decir, un usuario de Twitter decide a quien seguir, pero la persona a la que sigue no necesariamente tiene que seguirle a él. Cuando se sigue a alguien se ven sus tuits en la cronología o timeline (conjunto ordenado de todos los mensajes que llegan de la gente a la que se sigue).

El formato breve de publicación es algo positivo que ha posibilitado la agilidad y eficiencia característica de Twitter. Ha fomentado una cultura de aprovechar al máximo el espacio disponible, por lo que cuando se publica algo se va al grano. En Twitter se puede encontrar un poco de todo, desde personas de diferentes ámbitos que resulten de interés para seguir, noticias de última hora, tendencias en tecnología o en moda, recomendaciones de todo tipo de cosas, opiniones, etc. Sin duda Twitter constituye una gran fuente de información.

El **objetivo** de esta práctica es poder clasificar diferentes tuits según su ideología política, en particular en cuatro posibles categorías:

- moderado de derechas (*moderate_right)*
- moderado de izquierdas (*moderate_left*)
- de derechas (*right)*
- de izquierdas *(*left*).*

Corpus

Se trabajará con una colección de tuits en español procedente de la tarea de investigación PoliticEs 2022 (García-Díaz, 2022), dentro del Workshop IberLef 2022 (*Iberian Languages Evaluation Forum*). En esta práctica se proporcionará una colección de tuits de entrenamiento (fichero *training.csv*) y de desarrollo (fichero *development.csv*).

En concreto, el corpus que se proporciona para la práctica se compone de 28 061 tuits en la parte de entrenamiento y 4 677 tuits en la parte de desarrollo. Para todos ellos, se disponen de diversas anotaciones relacionadas tanto con el contenido textual del tuit como con diversas características del usuario que lo realizó (género, profesión e ideología política). La estructura de los ficheros que componen dicho corpus se muestra en la Figura 1.

Sin embargo, para esta práctica se desea **categorizar el contenido textual del tuit** analizado, correspondiente a la columna *tweet* (columna número 7), **en función de la ideología política** indicada en la columna *ideology_multiclass* (columna número 6).

De este modo, a través de técnicas y modelos de procesamiento del lenguaje natural, se deberá clasificar el contenido textual de los tuits (columna *tweet*) en las categorías *moderate_right, moderate_left, right* y *left* (columna *ideology_multiclass*).

```
code,label,gender,profession,ideology_binary,ideology_multiclass,tweet
36617,@user10,male,journalist,right,moderate_right,"EE UU y China: Los dos grandes pelea
11991,@user10,male,journalist,right,moderate_right,"Sensación Previsible a esta hora: A
40804,@user10,male,journalist,right,moderate_right,"No te salves. no te quedes inmóvil a
48101,@user10,male,journalist,right,moderate_right,"Al menos 25 militares venezolanos, 
27627,@user10,male,journalist,right,moderate_right,"Rivera que , con Sanchez ,da una may
45149,@user10,male,journalist,right,moderate_right,"Pablo Iglesias que , en este momento
26192,@user10,male,journalist,right,moderate_right,"El Presidente del Gobierno en funci
13931,@user10,male,journalist,right,moderate_right,"El [POLITICAL_PARTY] supera el 40% d
```

Figura 1. Fragmento del subconjunto de entrenamiento del corpus de la colección PoliticEs.

Métricas de evaluación

Se empleará la métrica *macro f1* para medir la calidad predictiva de las técnicas y modelos de clasificación desarrollados. Esta métrica calcula la media aritmética de los valores *f1* de las clases *moderate_right, moderate_left, right* y *left* analizadas, otorgándoles la misma relevancia en la clasificación. La siguiente ecuación muestra matemáticamente cómo calcular el valor *macro f1*, donde *n* es el número de clases analizadas (en este caso, cuatro):

$$\text{Macro-F1} = \frac{1}{n}\sum_{i=1}^{n} \text{F1}_i$$

Como recordatorio, el valor de *f1* de cada clase se obtiene haciendo una media armónica entre sus valores de *precision* y *recall*. En particular, el valor de *f1* es calculado por medio de las siguientes ecuaciones:

$$\text{F1}_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \qquad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}$$

Donde TP son los verdaderos positivos (*True Positive),* TN los verdaderos negativos (*True Negatives),* FP los falsos positivos (*False Positives)* y, por último, FN los falsos negativos (*False Negatives)*.

Referencias

José Antonio García-Díaz, Salud María Jiménez-Zafra, María-Teresa Martín Valdivia, Francisco García-Sánchez, L. Alfonso Ureña-López, Rafael Valencia-García. "Overview of PoliticEs 2022: Spanish Author Profiling for Political Ideology". Procesamiento del Lenguaje Natural 2022, 69: 265-272.

Requisitos mínimos que hay que cumplir

Para la realización de la práctica se podrán utilizar libremente todas las librerías vistas en clase. Por otra parte, **es importante no aplicar un único enfoque** para resolver el problema. Se requiere desarrollar varias propuestas y analizar los resultados de cada una indicando cuál puede ser más apropiada.

De este modo, es crucial no limitar la memoria únicamente al modelo que obtenga la mejor clasificación, sino evaluar y comparar el rendimiento de las diversas técnicas y enfoques abordados en la asignatura. Esto incluye, entre otros aspectos:

- Aplicación de diversos preprocesamientos textuales (como lematización, filtrado por categoría gramatical, …).
- Representación a través de vectorizaciones basadas en bolsas de palabras (como pesado binario, TF-IDF, …).
- Empleo de embeddings estáticos (como Word2Vec, Gensim, FastText, ...)
- Empleo de embeddings contextuales (como BERT, ELMO, ...)
- Utilización de modelos "clásicos" de aprendizaje automático (como Naive Bayes, Regresión Logística, SVM, …)
- Utilización de modelos de aprendizaje profundo (como CNNs, RNNs, LSTMs, fine-tuning de modelos pre-entrenados, ...)

Requisitos mínimos que debe cumplir la memoria escrita

Esta sección del enunciado pretende aportar algunas directrices para la redacción de la memoria:

1) La memoria debe escribirse en español, con un mínimo de 15 páginas sin contar la portada y el índice en caso de que se incluya.

2) Debe tener un formato estructurado con secciones y subsecciones.

3) Las imágenes no deberían de ser mayores de un tercio de una página.

4) El esqueleto debe seguir una estructura aproximada a:

    1. Portada

    2. Índice (opcional)

    3. Introducción

    4. Propuestas con sus experimentos y sus resultados

    5. Conclusiones