

Trabajo de Sentiment Analysis.

En este trabajo se debe desarrollar un sistema de Sentiment Analysis para Twitter basado en aprendizaje automático, por ejemplo usando el clasificador SVM de sklearn o utilizando transformers, o ambas aproximaciones.

Se proporciona un corpus que tiene tres ficheros, dos etiquetados y otro que se debe etiquetar:

- training.txt, es el fichero etiquetado para entrenamiento. Contienen el identificador del tweet, la etiqueta y el texto del tweet. Cada campo está separado por tabulador.
- development.txt, tiene el mismo formato que el fichero anterior y se debe utilizar elegir modelo y ajustar parámetros. Si queréis también podéis utilizarlo como corpus de entrenamiento.
- test.txt, este fichero tiene el mismo formato que los anteriores pero todos los tweets tienen como etiqueta UNK.

A cada tweet se le debe asignar una única etiqueta entre 4 posibles (P, N, NONE, NEU).

- La etiqueta P indica que el tweet transmite una opinión/sentimiento general "positivo".
- La etiqueta N indica que el tweet transmite una opinión/sentimiento general "negativo".
- La etiqueta NONE indica que el tweet no transmite globalmente ninguna opinión o sentimiento.
- La etiqueta NEU indica que el tweet transmite sentimiento tanto positivo como negativo y de alguna manera esos sentimientos se neutralizan.

Qué hay que realizar en el trabajo:

1) Utilizar los ficheros de entrenamiento y validación para aprender un clasificador utilizando un modelo de aprendizaje automático. Para conseguir esto antes debéis elegir un método de tokenización y vectorización de los tweets.

2) Con el clasificador aprendido, etiquetar todos los tweets de test y generar un nuevo fichero con los resultados. El formato del nuevo formato debe ser:

```
identificador1 (tabulador) etiqueta1  
identificador2 (tabulador) etiqueta2  
identificador3 (tabulador) etiqueta3
```

Es decir, no se debe incluir el texto del tweet en el fichero de resultados.

- 3) Para mejorar los modelos se puede utilizar el diccionario de polaridad proporcionado (ElhPolar_esV1.lex).
- 4) Subir a la actividad de PoliformaT todos los fichero requeridos (resultados, código y informe)
- 5) Opcionalmente, o alternativamente, se pueden utilizar Transformers. Se recomienda utilizar un modelo preentrenado y hacer finetuning con los datos del corpus seleccionado.

A la actividad de PoliformaT se deben subir tres ficheros, con tres nombres concretos:

- resultado.txt. Fichero con los resultado del etiquetado del test. Como se ha indicado anteriormente el formato del fichero debe ser (identificador \t polaridad). Si aprendéis más de un clasificador y generáis más de un fichero de resultados los podéis subir todos.
- codigo.zip. Un fichero comprimido en formato zip con todo el código desarrollado. No es necesario incluir las librerías python que se pueden instalar con "pip"
- informe.pdf. Un fichero pdf con un informe del trabajo realizado: que modelos habéis utilizado, que preproceso, parámetros, resultados en validación, y en general toda la información que consideréis importante.

Una orientación de cómo desarrollarlo se puede consultar en el siguiente enlace (token classification).

<https://huggingface.co/learn/nlp-course/en/chapter7/2?fw=pt>

También puede ser interesante consultar el libro Jurafsky. En este capítulo se describe como hacer el finetuning.

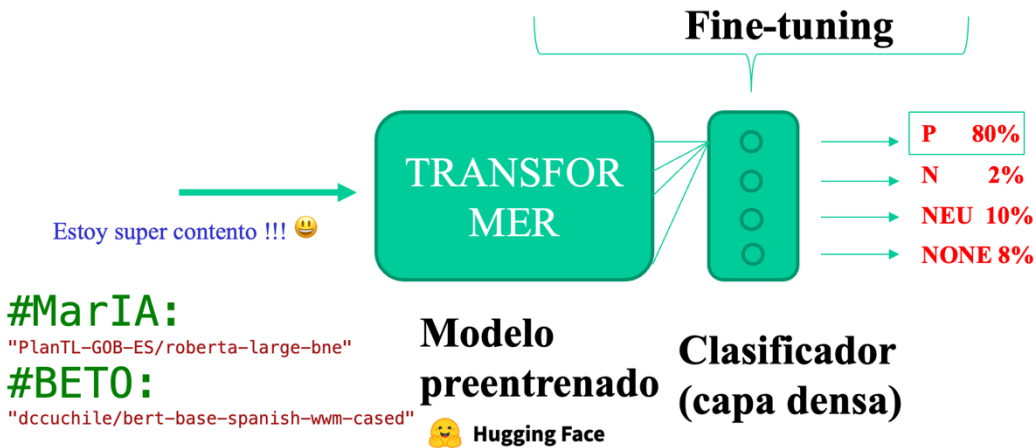
<https://web.stanford.edu/~jurafsky/slp3/11.pdf>

Algunos modelos de Hugging Face preentrenados para el castellano con los que se puede entrenar pueden ser:

#MarIA: "PlanTL-GOB-ES/roberta-large-bne"

#Beto: "dccuchile/bert-base-spanish-wwm-cased"

Esquema de como abordar el problema con transformers



#Usar los datos de la tarea para hacer el Fine-tuning

