# Clasificación de texto basada en Transformers: comparación de representaciones y reducción de dimensiones

Francisco Fernández Condado

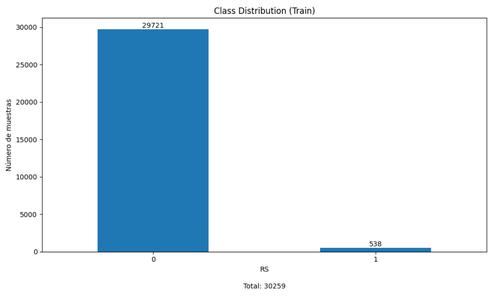
Universidad del País Vasco / Euskal Herriko Unibertsitatea

## Objetivos

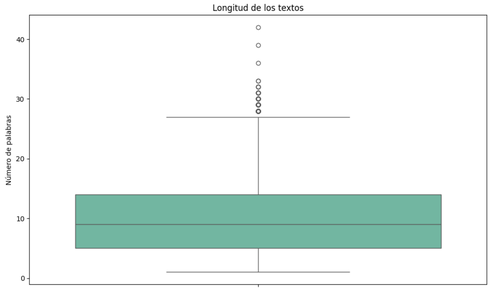
- Objetivo: dado un texto corto, predecir la clase objetivo tratando de maximizar F-Score en la clase minoritaria.
- Research questions:
  - RQ1 ¿Qué modelo Transformer ofrece la mejor representación para estos datos?
    1. RoBERTa
    2. BERTweet
  - RQ2 ¿Mejora los resultados aplicar una técnica de reducción de dimensionalidad?
    1. PCA
    2. UMAP

## Tarea y datos

- **Tarea:** clasificación binaria de textos con distribución de clases altamente desbalanceada.
- **Distribución de clase RS:**




- **Longitud de los textos:**




- **Pre-Proceso:** tokenización estándar y uso de embeddings basados en Transformers, descartando instancias inválidas.

## Clasificador 1: RoBERTa

RoBERTa es un modelo de Transformer preentrenado de una forma muy similar a BERT, con mejoras interesantes en lo que respecta al impacto de grandes cantidades de hiperparámetros clave y el tamaño de los datos de entrenamiento.

## Algoritmo

Las pruebas han sido realizadas utilizando Random Forest aplicando SMOTE [Chawla et al., 2002] como técnica de Over-Sampling para reducir el desbalanceo.

## Resultados experimentales: RQ1

Para comparar el mejor clasificador, se han realizado pruebas experimentales usando el mismo conjunto de datos, aplicando SMOTE con los mismos parámetros y empleando las distintas vectorizaciones que propone tanto un modelo como otro.

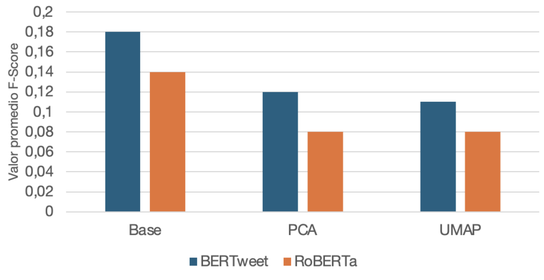



Figura 1:Valores promedios de F-Score para la clase minoritaria usando BERTweet y RoBERTa

- **Configuraciones:**
  - RoBERTa: Fine-tuning con learning rate $1 \times 10^{-5}$ y batch size 16.
  - BERTweet: Fine-tuning con learning rate $3 \times 10^{-5}$ y batch size 32.
  - 5 iteraciones por cada prueba
- Los resultados muestran mejores valores para F-Score, Accuracy y Recall utilizando BERTweet frente a RoBERTa, independientemente de la reducción de dimensionalidad.

## Clasificador 2: BERTweet

BERTweet es otro modelo de Transformer entrenado de forma similar a BERT, pero en este caso optimizado para textos cortos. Se ha elegido por sus buenos resultados en pruebas experimentales [Guo et al., 2020] con datasets similares al utilizado, que se corresponden con textos de poca longitud con características típicas del lenguaje informal como el uso de emoticonos. Más adelante se aplicará con reducciones de dimensionalidad:

- PCA
- UMAP

| Clasificador | Método | Resultados (mean) |
|---|---|---|
| BERTweet | Base | Accuracy: 0.96<br>F1-Score: 0.18<br>Recall: 0.21 |
| BERTweet | PCA | Accuracy: 0.94<br>F1-Score: 0.12<br>Recall: 0.22 |
| BERTweet | UMAP | Accuracy: 0.92<br>F1-Score: 0.11<br>Recall: 0.25 |
| RoBERTa | Base | Accuracy: 0.96<br>F1-Score: 0.14<br>Recall: 0.16 |
| RoBERTa | PCA | Accuracy: 0.93<br>F1-Score: 0.08<br>Recall: 0.17 |
| RoBERTa | UMAP | Accuracy: 0.97<br>F1-Score: 0.08<br>Recall: 0.08 |

Cuadro 1:Resultados medios de las experimentaciones. BERTweet ofrece una mayor precisión y exhaustividad en todas las pruebas realizadas

## Resultados experimentales: RQ2

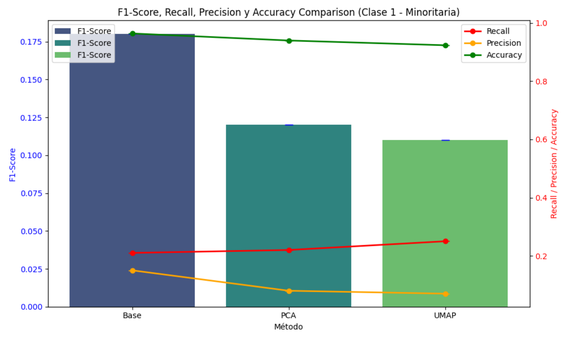



Figura 2:Comparativa de métricas usando BERTweet

- Los resultados muestran una mayor precisión al no aplicar reducción de dimensionalidad.
- UMAP mejora ligeramente la exhaustividad.
- RF da mejor resultado con más parámetros.

## Conclusiones y trabajo futuro

- Resumen: se han explorado 2 clasificadores + 2 análisis para ver cuál ofrece mejores resultados.
- Para estos datos, la mejor opción en cuanto a precisión es utilizar BERTweet sin reducción, UMAP mejora tiempos.
- Convendría realizar las mismas pruebas con otros datos. Probablemente en textos largos no se obtengan los mismos resultados, así como probar otras técnicas y embeddings.

## Bibliografía

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002).
Smote: synthetic minority over-sampling technique.
*Journal of artificial intelligence research*, 16:321–357.

[Guo et al., 2020] Guo, Y., Dong, X., Al-Garadi, M. A., Sarker, A., Paris, C., and Aliod, D. M. (2020).
Benchmarking of transformer-based pre-trained models on social media text classification datasets.
In Kim, M., Beck, D., and Mistica, M., editors, *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 86–91, Virtual Workshop. Australasian Language Technology Association.