Detección de noticias falsas usando Procesado de Lenguaje Natural

Miguel Ricardo Claramunt Argote

Escola Tècnica Superior d'Enginyeria – Universitat de València

11 de julio de 2023

Esquema de la presentación

1.1. Introducción

Las *fake news* son piezas de texto con estilo pseudoperiodístico que tienen como objetivo desinformar a los lectores.

Motivos principales que incentivan la difusión de *fake news*:

- Económico: publicidad en la página web.
- Ideológico: apoyo de una agenda política.[1]

[1] (**AlcottGentzkow2017**; **Sydell2016**)

1.2. Motivación

Fact checking:

- Exhaustivo, laborioso.
- Mayoritariamente manual.
- Grandes beneficios si se automatiza: *LLMs*.

Otros beneficios: reducción de prejuicios en colectivos minorizados, mejora de la imagen del periodismo/periodista, etc.

1.3. Objetivos

Objetivo principal:

- Desarrollar una solución que permita detectar noticias falsas.
  - Clasificación a partir de características estilísticas.
  - Estudio comparativo: cantidad/calidad de información, arquitectura/tamaño de los modelos.

Objetivos transversales:

- Definición del término de *fake news* y discutir las implicaciones en nuestro trabajo.
- Conocer en profundidad cómo funcionan los modelos y los sesgos implícitos.
- Aplicación de técnicas de *Explanable AI* (XAI) para intentar entender el 'razonamiento' de los modelos.

2. Definiciones y área de estudio

> *Disinformation (Desinformación). Contenido proposicional de signos que tergiversa el estado del mundo con la intención de engañar.* — *(**Khan2021**)*

Problema complejo, límite del area de aplicación a noticias:

- Recopilables: gran disponibilidad de BBDD.
- Categorizables: aprendizaje supervisado.

3. Antecedentes y tecnologías

Se han desarrollado dos enfoques para abordar este problema:

- Basado en patrones: estilo o sintaxis, métricas de RRSS, minería de datos enfocada a emociones.
- Basado en evidencias. similaridad semántica entre *claim* y evidencia.

Trabajos relevantes:

- DeClarE (**Popat2018**)
- HAN (**Ma2019**)
- GET (**Xu2022**)

4.1. *Claim* y evidencia

- *Claim*: equivalente al titular de una noticia, pieza de información a verificar.
- Evidencia: hecho, dato o información que apoya o refuta la *claim*.

4.2. Dataset: Politifact[2] y Snopes[3]

Un *claim* y varias evidencias por noticia.

Creación de dos datasets, P-S_{One} y P-S_{All}:

- P-S_{One}: una *claim*, una evidencia elegida aleatoriamente.
- P-S_{All}: una *claim*, todas las evidencias.

Dataset	Etiqueta	#	#
PolitiFact	True	186	356
	False	170	
Snopes	True	116	433
	False	317	

Estadístico	Valor
μ	7,43
σ	6,34
Mín.	1
Q_1	2
Q_2	5
Q_3	11
Máx.	27

Cuadro: Conteo de muestras y estadísticos de longitud de noticias.

[2] (**Popat2017**)

[3] (**Vlachos2014**)

4.3. Dataset: News[4]

- Artículos periodísticos completos de diversas fuentes.
- Limpieza adicional para las noticias verdaderas: autor, aclaraciones.

Etiqueta	#
True	21417
False	23481

Cuadro: Conteo de muestras.

[4] (**Ahmed2017**)

4.4. Modelos y clasificadores utilizados

Modelos estadísticos (`scikit-learn`):

- *Bag of Words* y TF-IDF:
  - Regresión logística (LR)
  - Naïve Bayes (NB)
  - Support Vector Machine (SVM)
  - Stochastic Gradient Descent (SGD)
  - Random Forest (RF)

Modelos[5] basados en *transformers* (`transformers` + `pytorch`):

- BERT
- DistilBERT
- RoBERTa
- DeBERTa

[5] (**Devlin2018**; **Sanh2019**; **Liu2019**; **He2020**)

5.1. Resultados obtenidos: precisión (+)

Métricas utilizadas: precisión (+, −); *specificity* (−); *F1-score* (−).

- Mejores modelos:
  - P-S_{One}: BoW + SGD
  - P-S_{All}: BoW + NB
  - News: TF-IDF + SGD
- BoW > TF-IDF > LARGE > BASE
- CASED, ML > UNCASED
- P-S_{One} $\approx$ P-S_{All} > News

5.2. Resultados obtenidos: precisión ($-$)

- Mejores modelos:
  - P-S_{One}: BoW + SGD
  - P-S_{All}: BoW + NB
  - News: TF-IDF + NB
- BoW $>$ TF-IDF $>$ LARGE $>$ BASE
- Pocas diferencias entre variantes
- P-S_{One} $>$ P-S_{All} $>$ News

5.3. Resultados obtenidos: *specificity* $(-)$

- Mejores modelos:
  - LARGE [1,00]
  - DistilBERT$_{\text{B, C \& ML}}$ y BERT$_{\text{B, C}}$ obtienen resultados similares.
- P-S$_{\text{One}} \approx$ P-S$_{\text{All}} \approx$ News
- Otros [0,49, 0,71]
  - P-S$_{\text{One}} \approx$ P-S$_{\text{All}} >$ News

5.4. Resultados obtenidos: *F1-Score* (−)

- Mejores modelos:
  - P-S_{One}: $BERT_{B,\,C}$, $DeBERTa_{B}$, M_{L}
  - P-S_{All}: TF-IDF + RF, M_{L} obtienen resultados similares
  - News: TF-IDF + SGD
- LARGE > BASE
- CASED, ML > UNCASED
- P-S_{One} ≈ P-S_{All} > News

5.5. Resultados generales

- LARGE $>$ BASE
- CASED, ML $>$ UNCASED
- P-S_{One} $>$ P-S_{All} $>$ News
- No hay indicios de sobreajuste.

5.6. Interpretabilidad de los modelos: SHAP

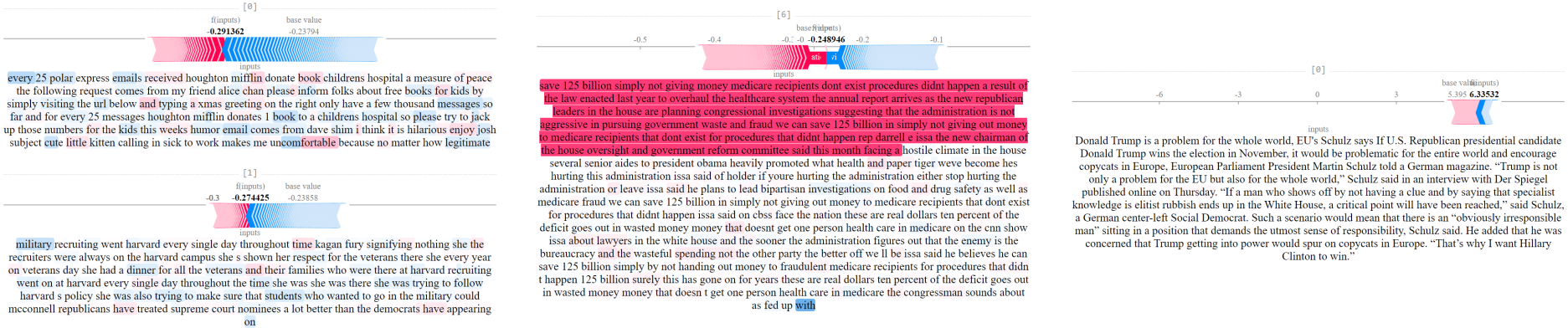

Figura: Valores Shapley de DistilBERT$_{B,\ C,\ ML}$ para los tres *datasets*.

- LARGE, CASED, ML *vs.* otros[6].
- Rendimiento general: P-S$_{All}$ > P-S$_{One}$ > News.

[6] Existen algunas excepciones según *dataset* y modelo

5.7. Evaluación

- LARGE ofrece resultados más consistentes.
- Alto valor de *specificity*: gran capacidad de detectar noticias verdaderas.
- Análisis cuantitativo *vs.* cualitativo: resultados generalmente poco concluyentes; distinto a lo esperado (*datasets*, modelos...).

6. Discusión

- SHAP: limitaciones en la interpretación (alta variabilidad, escasa justificación matemática, etc.).
- Limitaciones de los modelos BERT: conocimiento sintáctico, semántico, sentido común, etc.
- Posibles sesgos:
  - *Word Embeddings*: sesgos implícitos, diferencia de criterios estático *vs.* contextuales.
  - Tokenización: *cased vs. uncased*.
  - Destilado de modelos: $\uparrow$ sesgos, no relacionado con tamaño del modelo.
  - Modelos aumentados: mayor capacidad de adquirir sesgos.

7. Conclusiones

Objetivos propuestos y grado de cumplimiento:

- Definir *fake news*, delimitar área de estudio.
- Desarrollar un clasificador de *fake news*.
- Estudio comparativo de rendimiento: *datasets*, modelos.
- Estudio del rendimiento con XAI.
- Estudiar los sesgos involucrados en el estudio.

Referencias I