Detección de Fake News mediante Modelos de NLP: Clasificación de Noticias como Verdaderas o Falsas usando TinyBERT y DistilBERT

Diana Cordero

Abril 2025

1. Introducción

En la actualidad, las noticias falsas (fake news) son un fenómeno que afecta la credibilidad de los medios de comunicación, influyendo en la opinión pública y las decisiones políticas. El término Fake News se refiere a la propagación deliberada de información falsa, generando un peligroso ciclo de desinformación que distorsiona la percepción de la realidad.

Las redes sociales, en particular, han revolucionado la forma en que consumimos información, permitiendo que cualquier usuario sea a la vez productor y difusor de contenido, con lo quela propagación de información errónea se ha acelerado. Debido a esto, se genera la necesidad de contar con sistemas automáticos capaces de identificar y clasificar las noticias como verdaderas o falsas.

Este artículo presenta una aproximación utilizando técnicas de procesamiento de lenguaje natural (NLP) para la clasificación de noticias. Se utilizan modelos preentrenados como TinyBERT y DistilBERT para detectar la veracidad de las noticias en el dataset LIAR, una base de datos de declaraciones políticas que contiene ejemplos de noticias etiquetadas como verdaderas o falsas.

2. Conjunto de datos

El dataset utilizado en este proyecto es el dataset LIAR, un conjunto de datos que contiene 12,800 afirmaciones de políticos de diferentes contextos, categorizadas como verdaderas, mayormente verdaderas, mitad y mitad, mayormente falsas y falsas. Cada afirmación está etiquetada con una categoría que indica su veracidad.

3. Antecedentes

La detección de noticias falsas ha sido objeto de estudio durante los últimos años, con varios enfoques que incluyen el análisis de texto, la detección de patrones semánticos y la clasificación basada en modelos de machine learning. Algunos trabajos destacados en este ámbito son:

- En “Detección de Noticias Falsas (Fake News) en Internet Utilizando Deep Learning”, Centeno et al (2022), utilizó modelos de regresión logística y modelos de clasficiacion con transformadores, como ALBERT, Bert, ELECTRA, RoBERTa y XLM-Roberta, que utilizan arqquitecturas de atención para comprender el contexto y significado del texto, y posteriormente, hacer la clasificación.
- Bhardwaj et al (2024) proponen un modelo de análisis de sentimientos basado en emociones para detectar cuentas falsas, bots y campañas maliciosas. Su enfoque utiliza algoritmos de machine learning, reforzando la viabilidad de los sistemas automatizados para combatir la desinformación, especialmente en entornos donde el contenido se viraliza rápidamente.

4. Metodología

4.1. TinyBERT

TinyBERT es un modelo diseñado explícitamente para casos de uso donde la velocidad de inferencia y el uso reducido de memoria son prioridades clave. Al igual que DistilBERT, TinyBERT también utiliza técnicas de distilación de conocimiento, pero se enfoca principalmente en tareas específicas, aplicando la distilación tanto a nivel de logits (salidas no normalizadas del modelo antes de la activación) como a nivel de capas intermedias del modelo original (BERT).

Lo que distingue a TinyBERT es su enfoque de *two-stage distillation* (distillación en dos etapas): primero se realiza una distilación general durante el preentrenamiento, y luego una segunda fase específica para la tarea (*task-specific distillation*) durante el ajuste fino. Esto permite que el modelo mantenga una representación útil del lenguaje a pesar de contar con muchas menos capas y parámetros.

El proceso comenzó con la implementación del modelo preentrenado TinyBERT, optimizado para tareas de Procesamiento de Lenguaje Natural (NLP). Las etapas fueron:

1. **Tokenización**: Utilizamos el tokenizador de TinyBERT para transformar el texto en secuencias de tokens, preservando la estructura semántica.
2. **Entrenamiento Base**: Configuración inicial de 3 épocas con los parámetros por defecto.

Los resultados del entrenamiento inicial se muestran en la Tabla 1:

Cuadro 1: Resultados del entrenamiento con 3 épocas

Época	Pérdida Entrenamiento	Pérdida Validación	Exactitud
1	1.6959	1.7369	0.2399
2	1.6391	1.7007	0.2578
3	1.5902	1.7184	0.2523

El modelo alcanzó una pérdida final de entrenamiento de 1.685 y una pérdida de validación de 1.718. La exactitud del modelo es baja, lo que indica que el modelo no está clasificando bien.

4.1.1. Extensión a 5 Épocas

Para mejorar el rendimiento, se extendió el entrenamiento a 5 épocas. Los resultados se muestran en la Tabla 2.

Cuadro 2: Resultados con 5 épocas de entrenamiento

Época	Pérdida Entrenamiento	Pérdida Validación	Exactitud
1	1.7309	1.7205	0.2453
2	1.6409	1.6963	0.2812
3	1.5548	1.7331	0.2679
4	1.5170	1.7832	0.2492
5	1.4559	1.8250	0.2461

Aunque la pérdida de entrenamiento disminuyó a 1.586, se observa sobreajuste, evidenciado por el aumento en la pérdida de validación.

4.1.2. Optimización con Early Stopping

Se implementaron dos técnicas de regularización:

- **Early Stopping**: Interrupción anticipada al no mejorar la validación
- **Submuestreo**: Entrenamiento con subconjuntos de datos (50 %)

Los resultados (Tabla 3) mostraron:
La pérdida final de entrenamiento fue de 1.458, con un tiempo reducido de 242.79 segundos (vs 3024.15 en 5 épocas completas). Sin embargo, la pérdida de validación sugirió la necesidad de:

Cuadro 3: Resultados con Early Stopping (4 épocas)

Época	Pérdida Entrenamiento	Pérdida Validación	Exactitud
1	1.6261	1.6768	0.2600
2	1.5173	1.6760	0.2750
3	1.4137	1.7342	0.2500
4	1.2765	1.8085	0.2600

Época	**Training Loss**	**Validation Loss**	**Accuracy**
1	No log	1.6978	0.258
2	1.6871	1.7403	0.277

Cuadro 4: Resultados del entrenamiento con DistilBERT base uncased

- Ajuste de hiperparámetros
- Aumento de datos
- Fine-tuning específico

4.2. DistilBERT base uncased

Como alternativa al modelo `TinyBERT`, se empleó el modelo `DistilBERT base uncased`. DistilBERT es una versión más ligera y rápida del modelo BERT, desarrollada con el objetivo de facilitar su uso en entornos con recursos computacionales limitados. Este modelo fue entrenado utilizando una técnica conocida como *knowledge distillation*, donde un modelo más pequeño (el estudiante) aprende a imitar el comportamiento de un modelo más grande y robusto (el maestro), en este caso BERT base.

A diferencia de otros enfoques donde la distilación se aplica en la etapa de ajuste fino (*fine-tuning*), DistilBERT aplica esta técnica durante la etapa de preentrenamiento. En este proceso, se emplea una pérdida compuesta que combina tres componentes: modelado del lenguaje, distilación y distancia coseno entre representaciones internas del modelo. Gracias a esta estrategia, DistilBERT logra reducir el tamaño de BERT en aproximadamente un 40 %, manteniendo un 97 % de sus capacidades de comprensión del lenguaje y aumentando su velocidad de inferencia en un 60 %.

El modelo utilizado fue `distilbert-base-uncased`, una versión que ignora las mayúsculas y minúsculas y ha demostrado un buen equilibrio entre rendimiento y eficiencia, siendo apto para tareas de clasificación de texto en tiempo real o sobre dispositivos con recursos limitados.

Se realizó la tokenización de los datos utilizando el tokenizador correspondiente al modelo, manteniendo el mismo preprocesamiento que con los modelos anteriores. Para el entrenamiento, se usó un conjunto reducido de datos compuesto por 5,000 ejemplos para entrenamiento y 1,000 para validación, y la siguiente configuración de entrenamiento:

- Número de épocas: 3
- Tamaño del batch: 8 por dispositivo
- `gradient_accumulation_steps=2` para emular un batch efectivo mayor
- Evaluación y guardado de modelo por época
- Carga automática del mejor modelo al finalizar
- Sin reportes a plataformas externas

El rendimiento obtenido superó ligeramente a TinyBERT en precisión.

5. Resultados y próximos pasos

5.1. Resultados obtenidos

Se entrenaron y evaluaron dos modelos de lenguaje preentrenados: **TinyBERT** y **DistilBERT base uncased**, utilizando un subconjunto reducido del conjunto de datos original (5,000 ejemplos para

Modelo	Training Loss	Validation Loss	Accuracy
TinyBERT	1.2648	1.7996	0.250
DistilBERT	1.6871	1.7403	0.277

Cuadro 5: Comparación de rendimiento entre modelos

entrenamiento y 1,000 para validación). Ambos modelos fueron entrenados durante tres épocas con los mismos parámetros de entrenamiento para garantizar una comparación justa.

Aunque TinyBERT presenta un menor *training loss*, su desempeño en validación fue ligeramente inferior en precisión comparado con DistilBERT. Esto podría indicar una ligera tendencia al sobreajuste por parte de TinyBERT. DistilBERT, en cambio, logró generalizar un poco mejor en el conjunto de validación, siendo así una opción más balanceada para el contexto actual.

5.2. Próximos pasos

Para continuar con esta línea de trabajo, se proponen los siguientes pasos:

- **Entrenamiento con el conjunto completo:** Ampliar la cantidad de datos de entrenamiento y validación para evaluar si los modelos mejoran su capacidad de generalización.
- **Optimización de hiperparámetros:** Ajustar parámetros como la tasa de aprendizaje, tamaño de batch y número de épocas para explorar su impacto en el rendimiento.
- **Exploración de otros modelos:** Considerar modelos adicionales como BERT base o RoBERTa, con el fin de establecer comparaciones más robustas.
- **Uso de técnicas de regularización y early stopping:** Para prevenir el sobreajuste y mejorar la estabilidad durante el entrenamiento.

Estos pasos permitirán refinar la estrategia de modelado y mejorar el rendimiento de los modelos en tareas de clasificación de texto.

6. Bibliografía

- ¿Qué son las Fake News?: guía para combatir la desinformación en la era de la posverdad / IFJ. (2018, 22 agosto). https://www.ifj.org/media-centre/reports/detail/que-son-las-fake-news-guia-para-combatir-la-desinformacion-en-la-era-de-la-posverdad/category/publications
- Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2019). TinyBERT: Distilling BERT for Natural Language Understanding. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1909.
- Mariana Esmeralda Centeno Reyes, Cristopher Jesús Chaire Rodríguez, Mario Isaac Fonseca Martínez, Diana Martínez Frías, Fernando Oviedo Paramo, & Juan Carlos Gómez Carranza. (2024). Detección de Noticias Falsas (Fake News) en Internet Utilizando Deep Learning. JÓVENES EN LA CIENCIA, 28. https://doi.org/10.15174/jc.2024.4598
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1910.01108