Does BERT Pay Attention To Cyberbullying?

Fatma Elsafoury[1], Stamos Katsigiannis[2], Steven R. Wilson[3], and Naeem Ramzan[1]
[1]University of the West of Scotland, [2]Durham University, [3]University of Edinburgh

1. Cyberbullying detection

Cyberbullying is a form of spreading insults using mobile or internet technology

- Victims could suffer from **depression, anxiety, low self-esteem, self-harm**.
- **Automated cyberbullying detection** can prevent these risks by banning the bullies and providing support to the victims.
- Recent attention-based language models, like BERT, have improved cyberbullying detection but the **model's inner-workings have not been studied**.
- This work attempts to **explain BERT's performance** for cyberbullying detection.

2. BERT vs. RNNs for cyberbullying detection

We compared BERT's performance on cyberbullying-related datasets to RNN models. **Results show that BERT significantly improves cyberbullying detection**. The datasets contain comments collected from Kaggle, Twitter, and Wikipedia Talk Pages (WTP).

Dataset	No. Samples	No. Positive	LSTM	Bi-LSTM	BERT (Fine Tuned)
Kaggle-insults	7425	2578 (35%)	0.6420	0.653	**0.768**
Twitter-sexism	14742	3370 (23%)	0.6569	0.649	**0.760**
Twitter-racism	13349	1969 (15%)	0.6400	0.678	**0.757**
WTP-aggression	114649	14641 (13%)	0.7110	0.679	**0.753**
WTP-toxicity	157671	15221 (10%)	0.7230	0.737	**0.786**

Table 1: Dataset information and F1-scores achieved for each dataset

3. What is the role that attention weights play in BERT's performance?

We analysed BERT's attention weights to inspect if they are the reason behind its performance. We compared attention weights between BERT with and without fine-tuning on cyberbullying datasets. Then, we compared attention weights and feature importance scores measured using Integrated Gradients.

BERT's attention weights (Fine Tuned vs. Non Fine Tuned)

Figure 1 shows that attention weights' pattern **differs in BERT with and without fine-tuning**.

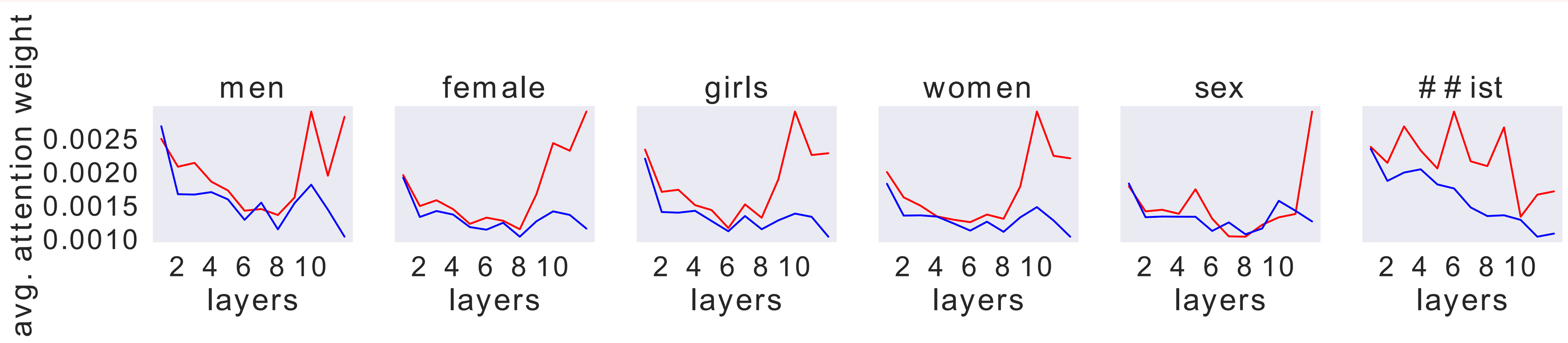

Figure 1: Mean attention weights of 12 heads per layer for fine-tuned BERT (red) and BERT without fine-tuning (blue), for the most important cyberbullying class-related tokens in the Twitter-sexism dataset according to Naive Bayes.

Attention weights vs. importance scores

- Table 2 shows that **no positive correlation was found between the tokens' attention weights and importance scores**.
- Consequently, **attention weights do not play a direct role in BERT's performance.**

Dataset	No. tokens	PCC (attention vs importance)
Kaggle-insults	4452	0.171
Twitter-sexism	3878	0.108
Twitter-racism	3991	0.056
WTP-aggression	4457	0.125
WTP-toxicity	4524	0.163

Table 2: Pearson's correlation between mean attention weights of fine-tuned BERT and mean absolute feature importance

4. What are the features that BERT relies on for its performance?

We analysed BERT's importance scores for the part-of-speech (POS) tags in the datasets. We hypothesised that BERT assigns the highest importance scores to informative POS tags for the task of cyberbullying detection, e.g. **Nouns**, and **Adjectives**. Results show that **the most important POS** according to BERT are **Auxiliaries** and **Punctuation**. This suggests that BERT does not rely on semantic features related to cyberbullying but instead **relies on syntactic biases** in the datasets.

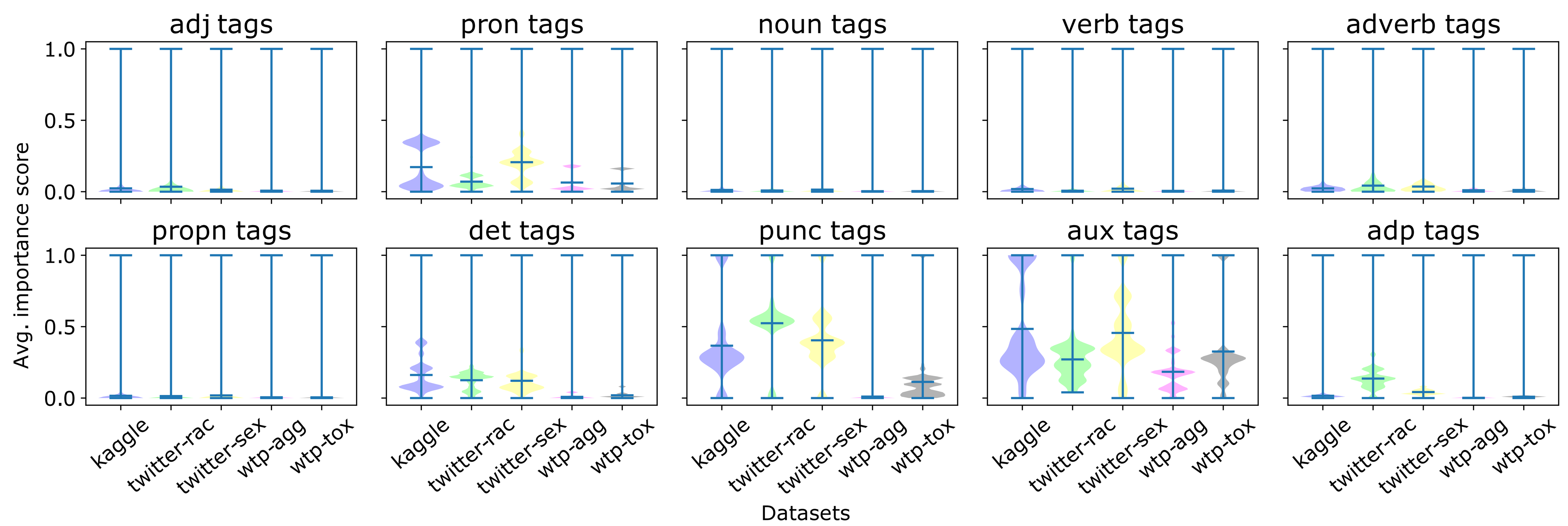

Figure 2: Mean normalised importance scores assigned by fine-tuned BERT to POS tags in the datasets

5. Take away messages

- **BERT performs significantly better than RNNs** on cyberbullying detection tasks.
- Although the pattern of attention weights changes when fine-tuning BERT, we found that **attention weights do not play a role in BERT's performance**.
- Results suggest that BERT does not rely on semantic features related to the task at hand, but **BERT relies on syntactical biases** in the datasets to achieve the high performance.