Does BERT Pay Attention To Cyberbullying?

Fatma Elsafoury, Stamos Katsigiannis, Steven R. Wilson, and Naeem Ramzan

sigir21

UNIVERSITY OF THE WEST of SCOTLAND UWS

Durham University

THE UNIVERSITY of EDINBURGH

Cyberbullying Detection

- What is cyberbullying?
  - Spreading insults using an electronic medium.

Cyberbullying Detection

- What is cyberbullying?
  - Spreading insults using an electronic medium.
- Why detect cyberbullying?
  - Support victims, warn/block bullies.

Cyberbullying Detection

- What is cyberbullying?
  - Spreading insults using an electronic medium.
- Why detect cyberbullying?
  - Support victims, warn/block bullies.
- How to improve the detection of cyberbullying?
  - Attention-based pre-trained language models → BERT.

What is BERT's performance on different cyberbullying-related datasets?

Dataset information and binary F1-scores achieved for each dataset. We used bert-base-uncased.

Dataset	No. Samples	No. Positive	LSTM	Bi-LSTM	BERT (Fine-Tuned)
Kaggle (Insults)	7425	2578 (35%)	0.642	0.653	**0.768**
Twitter (Sexism)	14742	3370 (23%)	0.656	0.649	**0.760**
Twitter (Racism)	13349	1969 (15%)	0.640	0.678	**0.757**
WTP*(Aggression)	114649	14641 (13%)	0.711	0.679	**0.753**
WTP* (Toxicity)	157671	15221 (10%)	0.723	0.737	**0.786**

Answer: BERT performs significantly better than RNNs on cyberbullying detection tasks.

***WTP: Wikipedia Talk Pages**

What is BERT's performance on different cyberbullying-related datasets?

Dataset information and binary F1-scores achieved for each dataset. We used bert-base-uncased.

Dataset	No. Samples	No. Positive	LSTM	Bi-LSTM	BERT (Fine-Tuned)
Kaggle (Insults)	7425	2578 (35%)	0.642	0.653	**0.768**
Twitter (Sexism)	14742	3370 (23%)	0.656	0.649	**0.760**
Twitter (Racism)	13349	1969 (15%)	0.640	0.678	**0.757**
WTP*(Aggression)	114649	14641 (13%)	0.711	0.679	**0.753**
WTP* (Toxicity)	157671	15221 (10%)	0.723	0.737	**0.786**

Answer: BERT performs significantly better than RNNs on cyberbullying detection tasks. **Why?**

***WTP: Wikipedia Talk Pages**

What is the role that attention weights play in BERT’s performance?

BERT's Attention weights vs. Importance scores

Pearson's correlation coefficient between mean attention weights of fine-tuned BERT, mean absolute feature importance scores.

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

Answer: Attention weights do not play an important role in BERT’s performance.

What is the role that attention weights play in BERT’s performance?

BERT's Attention weights vs. Importance scores

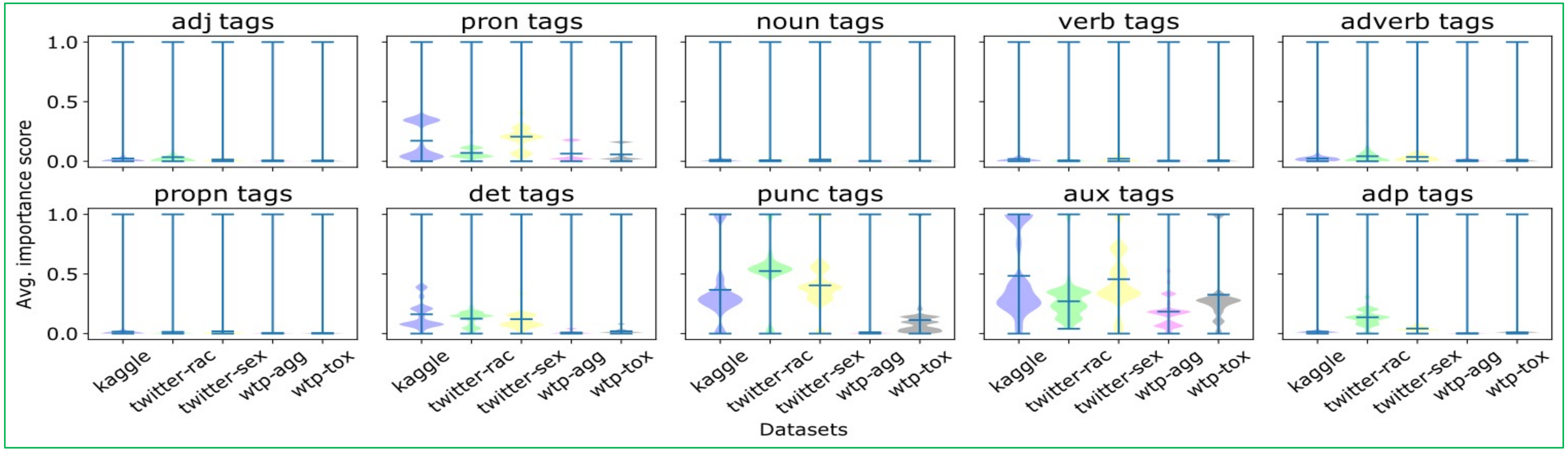
Pearson's correlation coefficient between mean attention weights of fine-tuned BERT, mean absolute feature importance scores.

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

Answer: Attention weights do not play an important role in BERT’s performance. **What about linguistic features?**

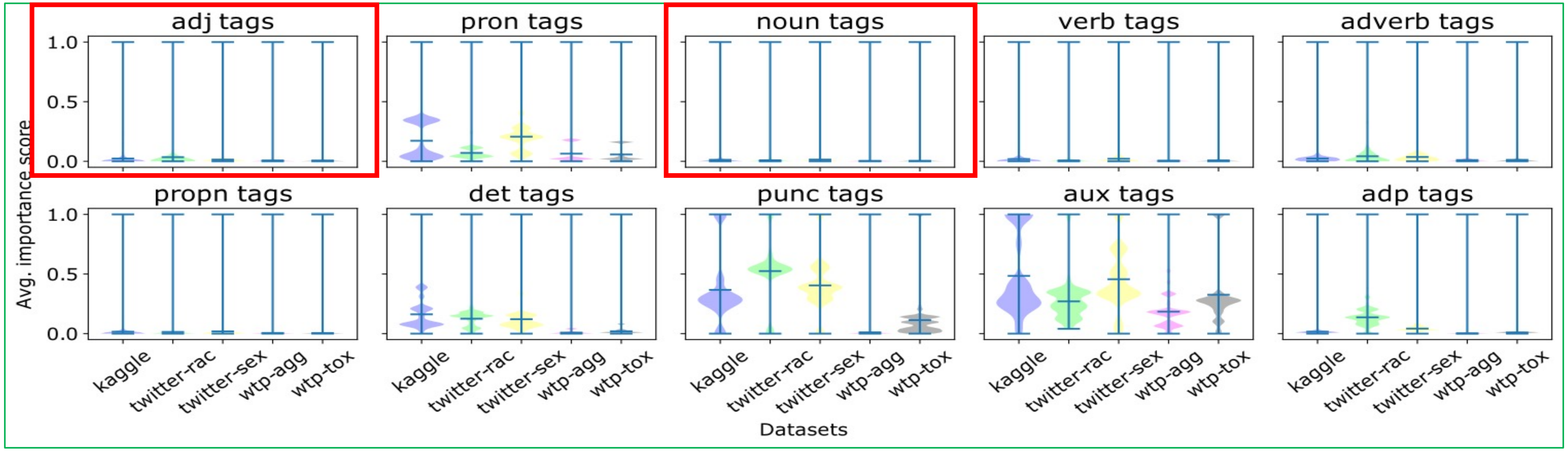
What are the features that BERT relies on for its performance?

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



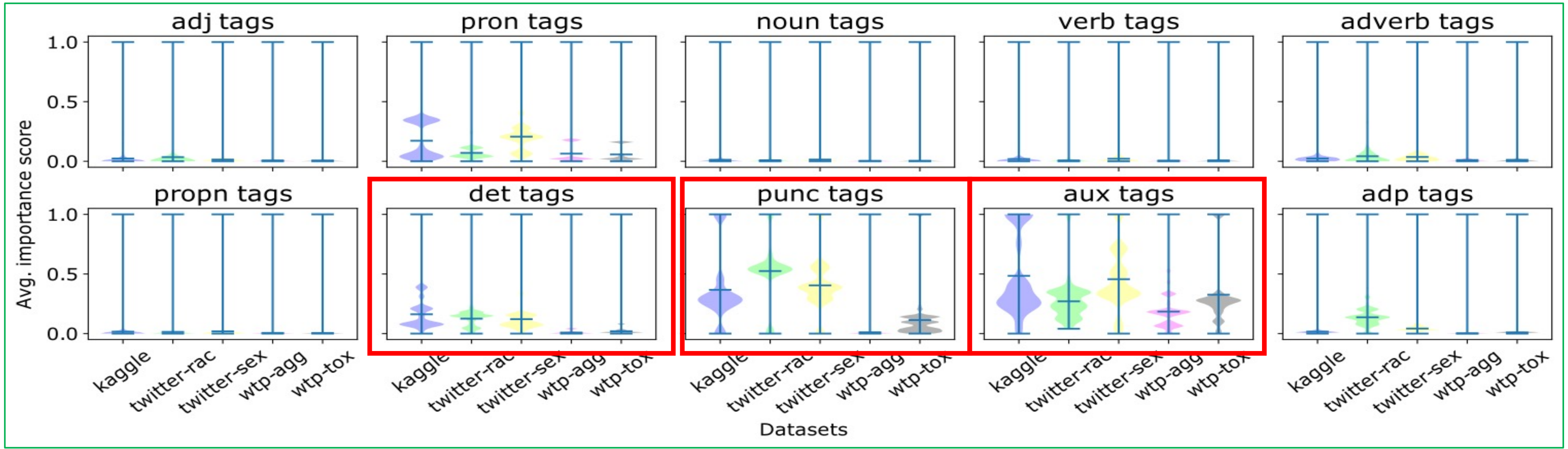
What are the features that BERT relies on for its performance?

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



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

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



Answer: BERT does not rely on linguistic features related to cyberbullying but instead it relies on syntactical biases.

Take away messages

- BERT performs significantly better than RNNs on cyberbullying detection tasks.
- Attention weights do not play a role in BERT’s performance.
- Results suggest that BERT relies on syntactical biases in the datasets to achieve its high performance.

Thank you

✉ fatma.elsafoury@uws.ac.uk
@fatmaelsafoury