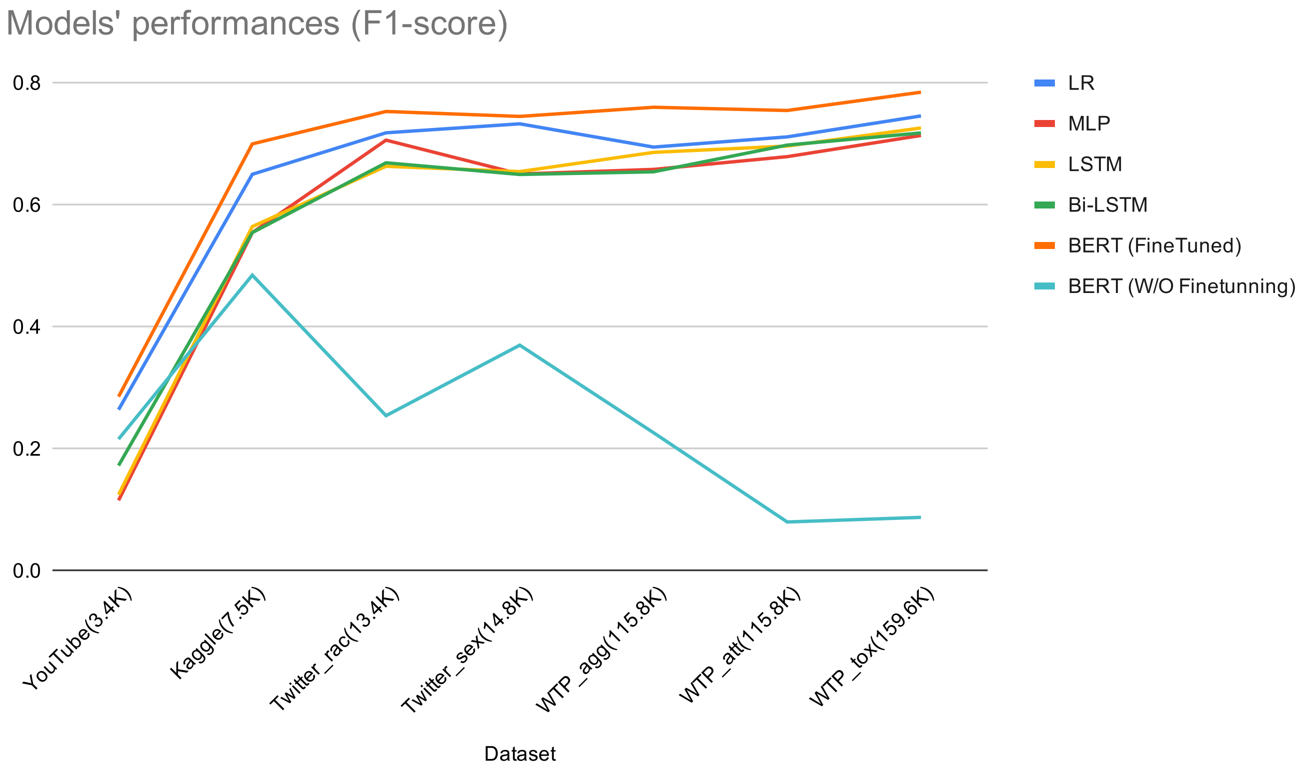
DOES BERT PAY ATTENTION TO ATTRIBUTION

Fatma Elsafoury

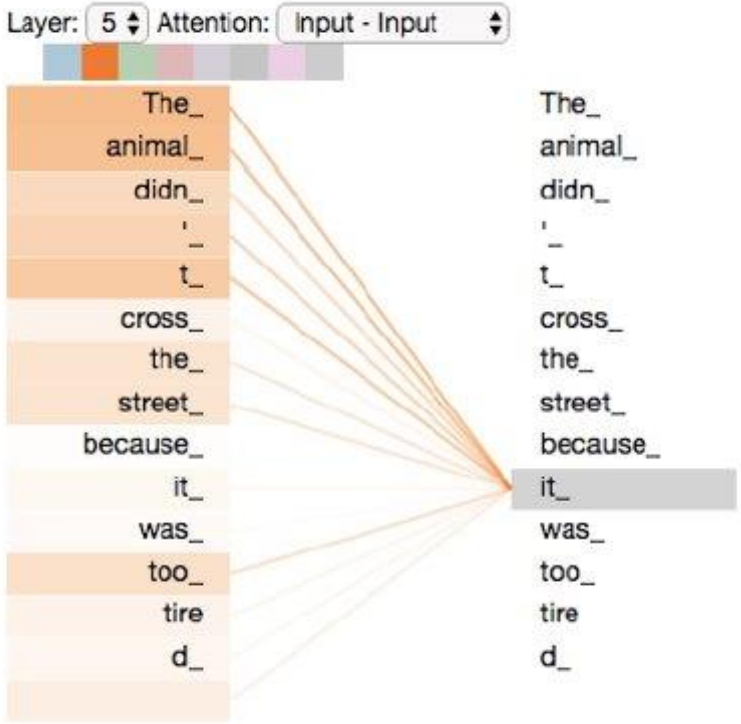
CYBERBULLYING DETECTION

- **Cyberbullying:** Using electronic forms to abuse another person or a group of people [1].
- **Detection:** Using Natural Language Processing and Machine Learning models.

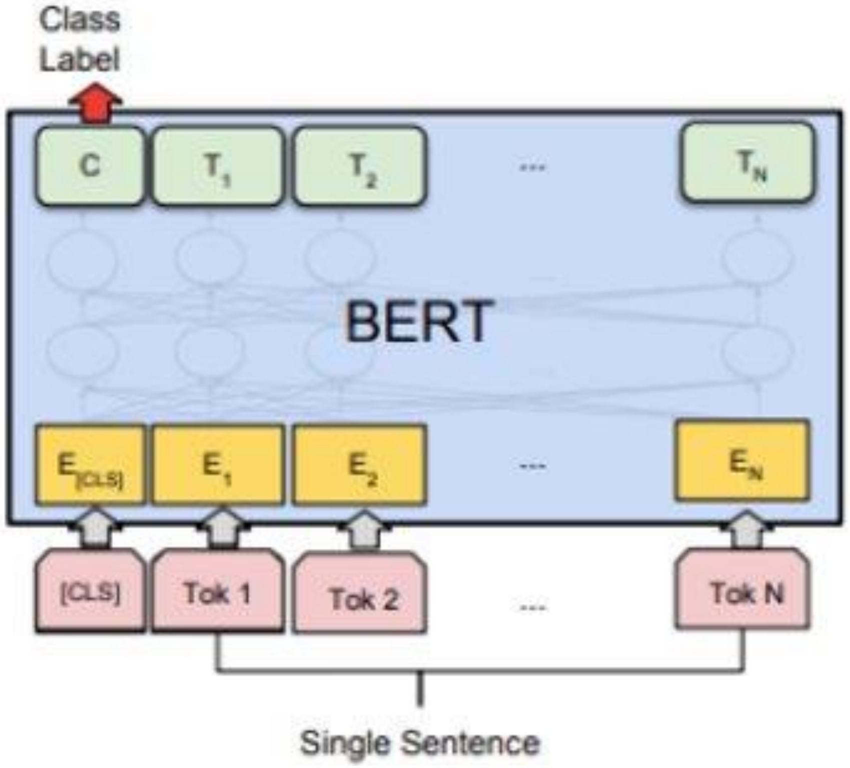


Bidirectional Encoder Representations from Transformers (BERT) [2]

- Uses attention mechanism which assigns attention weight of each word in a sentence.

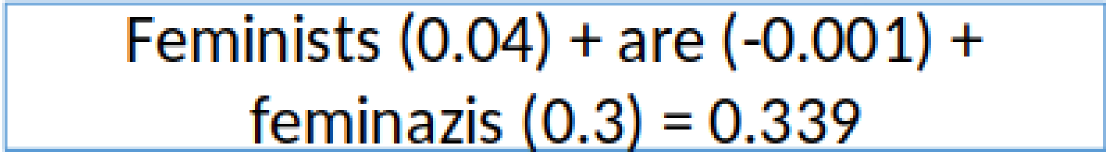


Layer: 5 Attention: Input - Input



Do attention weights play a role in the model's outcome?

- Collect the attention weights (Fine-tuned BERT) of the unique words in our dataset.
- Use Integrated gradients algorithm to get the feature(unique word) attribution(importance) score [3].



Feminists (0.04) + are (-0.001) + feminazis (0.3) = 0.339

Do attention weights play a role in the model's outcome?

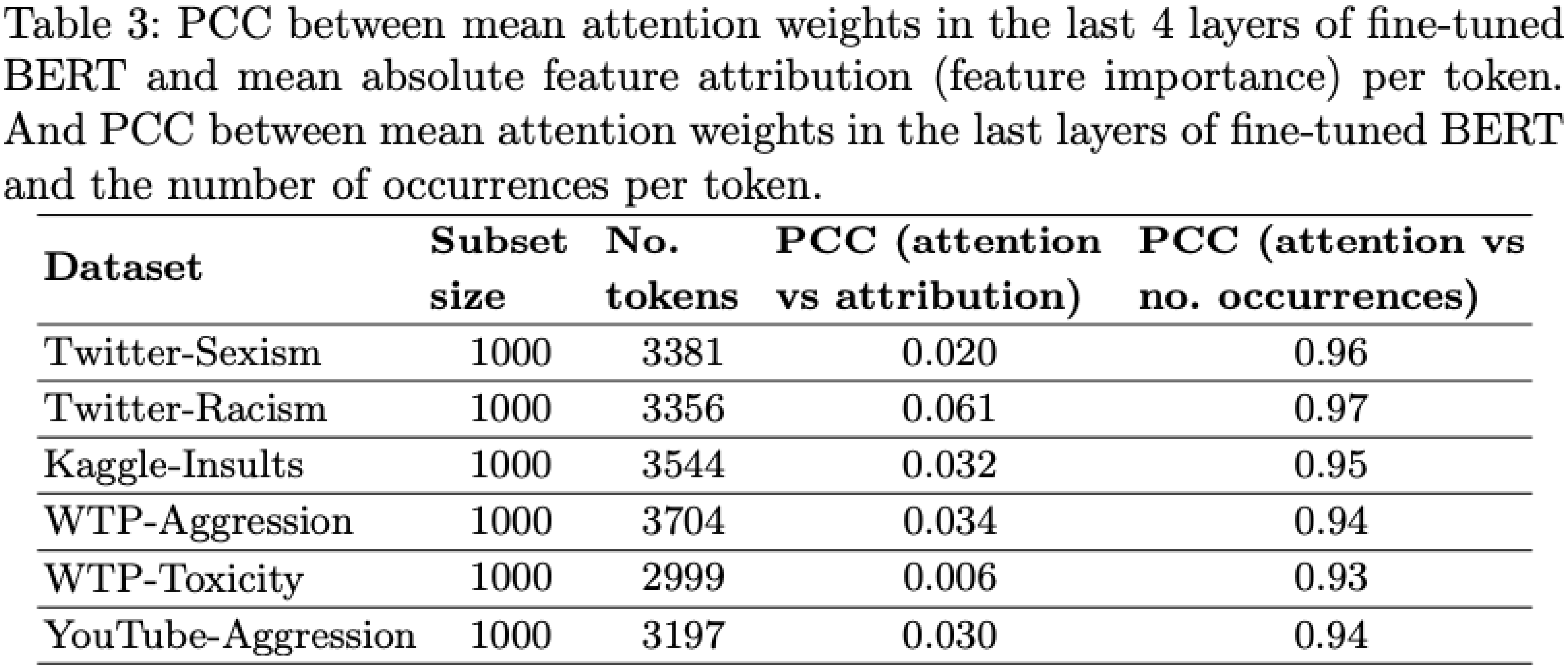


Table 3: PCC between mean attention weights in the last 4 layers of fine-tuned BERT and mean absolute feature attribution (feature importance) per token. And PCC between mean attention weights in the last layers of fine-tuned BERT and the number of occurrences per token.

Dataset	**Subset size**	**No. tokens**	**PCC (attention vs attribution)**	**PCC (attention vs no. occurrences)**
Twitter-Sexism	1000	3381	0.020	0.96
Twitter-Racism	1000	3356	0.061	0.97
Kaggle-Insults	1000	3544	0.032	0.95
WTP-Aggression	1000	3704	0.034	0.94
WTP-Toxicity	1000	2999	0.006	0.93
YouTube-Aggression	1000	3197	0.030	0.94

Thank you

References

- [1] Tibor Bosse and Sven Stam. 2011. A normative agent system to prevent cyberbullying. InProceedings of the 2011 IEEE/WIC/ACM InternationalConferences on Web Intelligence and Intelligent Agent Technology-Volume 02. IEEE Computer Society, 425–430.
- [2] Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert:Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
- [3]Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep net-works. In: ICML. Proceedings of Machine Learning Research, vol. 70, pp.3319–3328. PMLR (2017)