OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
TU DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Attention-based Gender-Stereotype Detection

Manuela Jeyaraj and **Sarah Jane Delany**

Technological University Dublin

ML Labs
Centre for Research Training
Science Foundation Ireland
Sfi For what's next

DEFINING GENDER - STEREOTYPE

- Perceptions about the typical physical, emotional, and social characteristics of individuals [1].
  - Gender-conforming (Stereotype)
  - Gender-non-conforming (Anti-Stereotype)
- Gender Stereotypes (GS) are NOT ALWAYS negative or harmful as opposed to Gender Bias.

AIMS and OBJECTIVES

- To effectively classify text containing GS from anti-stereotype text.
- To identify the type of language that perpetuates GS using attention.

METHODOLOGY

Dataset	Number of Samples	Class balance (Anti-Stereotype : Stereotype	Minimum length of text (in characters)	Maximum length of text (in characters)
StereoSet [2]	1,986	50:50	14	165
CrowS-Pairs [3]	524	40:60	13	183
Cryan et.al."Content" [4]	4,550	50:50	14	45,242
Cryan et.al."Reason" [4]	4,530	50:50	7	889

- An end-to-end approach : Fine-tuning a transformer model ⇨ Use attention ⇨ identifying features learned for gender-stereotype class.
- Then we observed the top feature attention scores learned by the model for correctly predicted stereotype instances in each dataset.

RESULTS and DISCUSSION

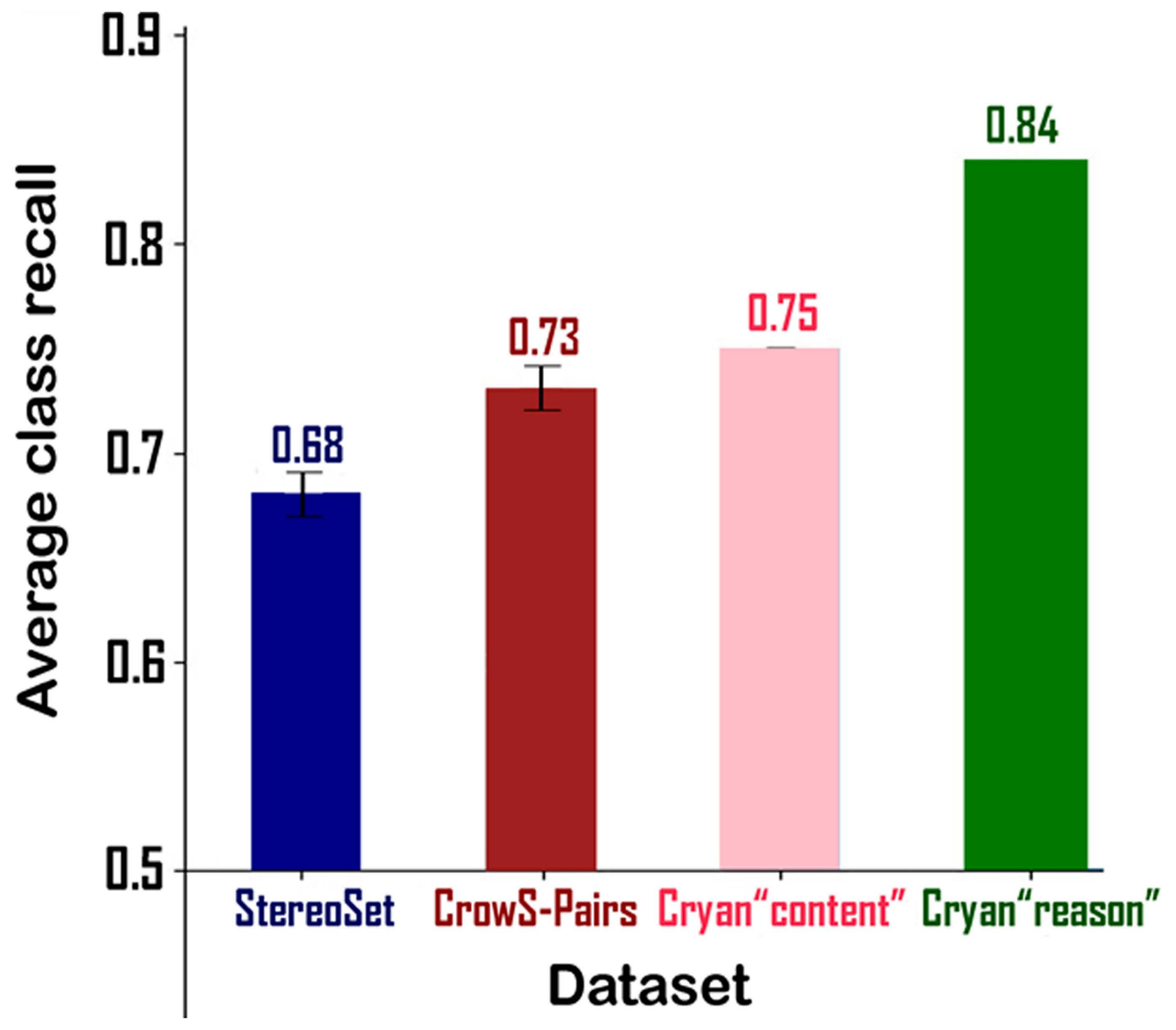

Figure 1. Model's performance (average class recall) across datasets

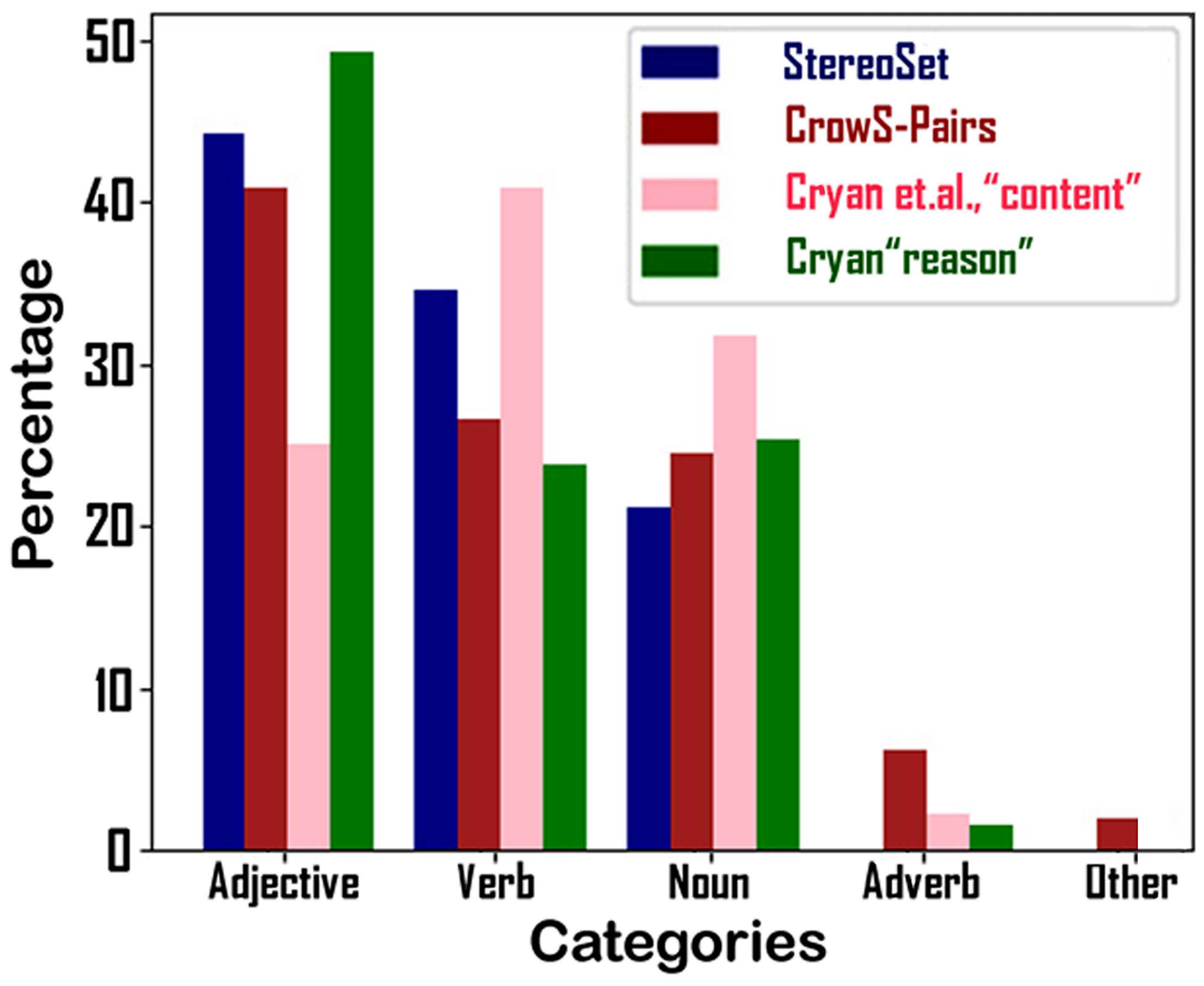

Figure 2. POS analysis of top features for a stereotype class based on the learned attention

- Adjectives and verbs are mainly learned by the model to predict a GS text.

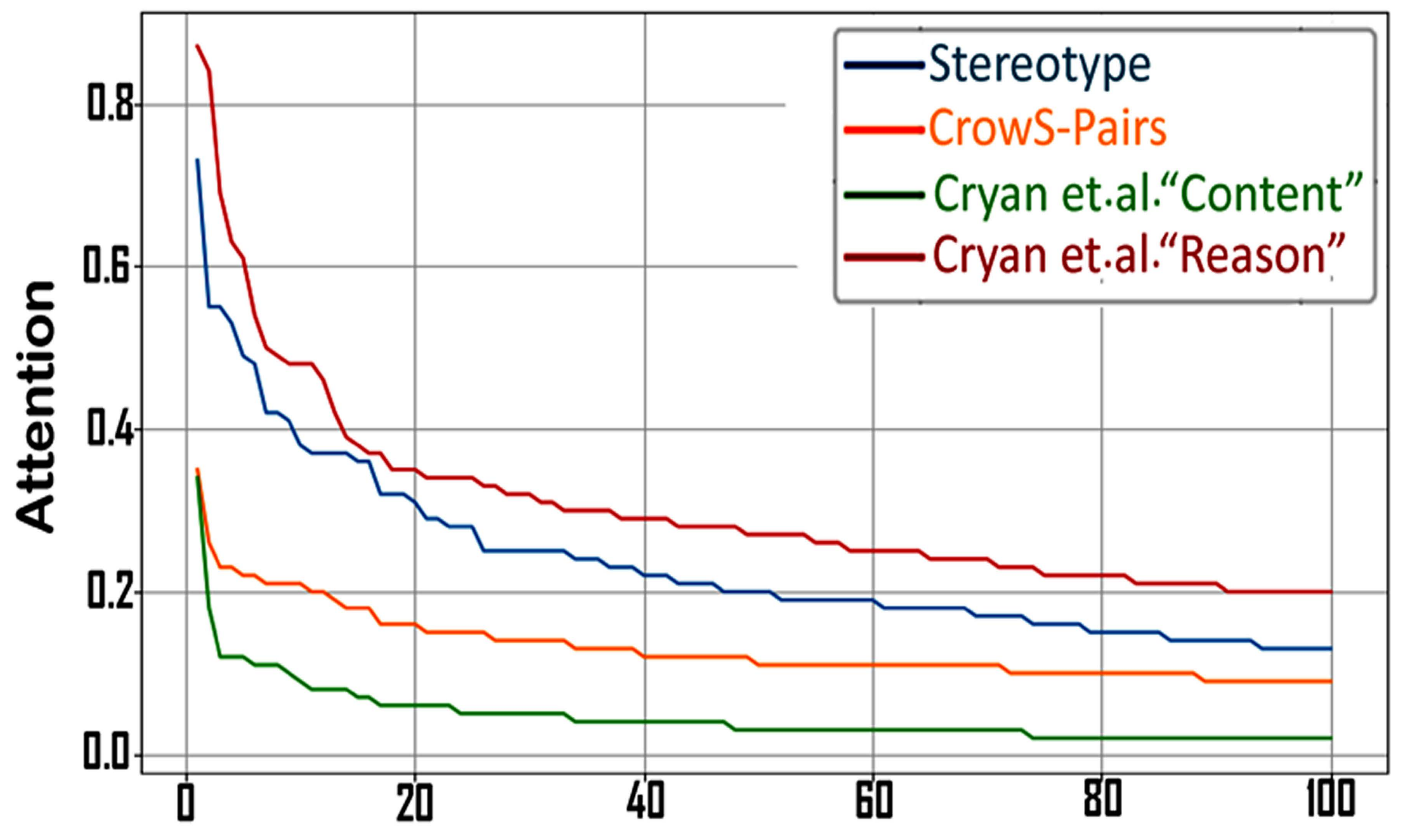

Figure 3. Attention distribution of learned features for a stereotype class

- Though StereoSet dataset has the lowest accuracy of the 4, it has learned features with high attention scores.
- This shows that the model's prediction of a stereotype is dependent on context words.

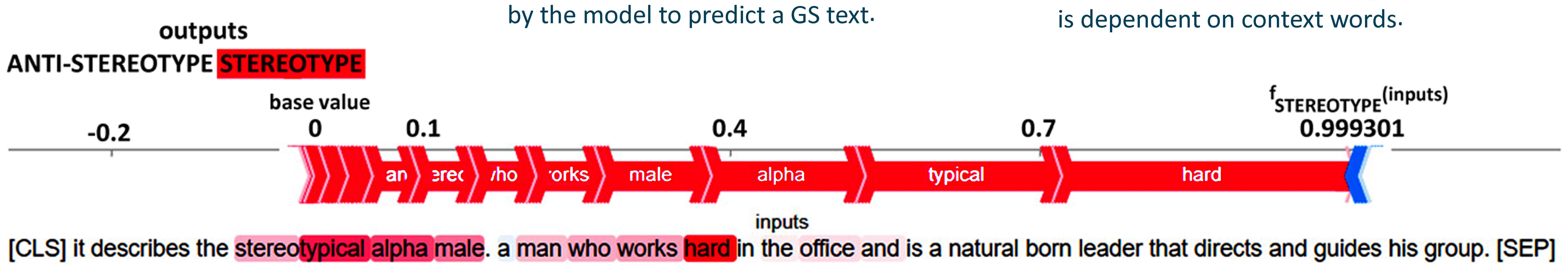

Figure 4. Visualization of attention for a sample text

- Each bar represents the attention learned for the respective feature on the scale.
- However, this feature alone is not representative of the model's overall decision.
- Feature "hard" contributes more to the models decision of that sentence being classified as a GS with the highest normalized attention of 0.3.
- It also uses the positive attention of context words ("typical", "alpha", etc.)

CONCLUSION

- Using learned attention of a model is an effective end-to-end approach to understand language features that prompt a GS text.
- Use of stereoSet, CrowS-Pairs and Cryan et.al.'s "reason" datasets is very useful and valid for GS studies considering the lack of labelled datasets.

ACKNOWLEDGEMENT

This work was funded by Science Foundation Ieland through the SFI Centre for Research Training in Machine Learning (18/CRT/6183).

REFERENCES

[1] Ellemers, N. (2018). Gender stereotypes. *Annual Review of Psychology,* 69, pp. 275-298.

[2] Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet : Measuring Stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing.* (Volume 1 : Long papers). pp. 5356-5371

[3] Nangia, N., Vania, C., Bhalerao, R., & Bowman, S.R. (2020). CrowS-Pairs : A Challenge dataset for measuring social biases in masked language models. In *2020 Conference on Empirical Methods in Natural Language Processing EMNLP 2020*. pp. 1953-1967.

[4] Cryan, J., Tang, S., Zhang, X., Metzger, M., Zheng, H., & Zhao, B.Y. (2020). Detecting gender stereotypes : lexicon vs. supervised learning methods. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1-11.*