Why Are They Smarter? Comparing Human and Machine Attention Maps

M. Garzoni di Adorgnano, O.A. Kirschstein Schafer, G. Kooij, R.E. Lucas, M. Pieke

Utrecht University, Utrecht, Nederland

Abstract

Machine learning systems are known for outperforming humans on a variety of different tasks. It is not always clear whether the decision-making process is similar between such systems and humans. In some cases, having an understanding of such (dis)similarities can be advantageous in order to improve human decision making. This paper determines whether BERT uses similar or different words to humans to classify a restaurant review is positive or negative.

Introduction

Replication of previous work by Cansu et al.

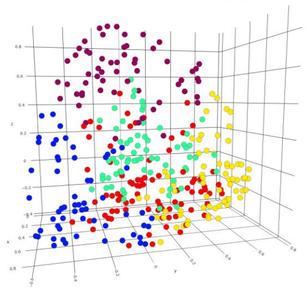
Example sentence: *"what happened food is normally good both our meals were less than satisfactory hard to screw up bar food but they managed i also dont get the 32 oz happy hour beer i guess if you like warm beer it is cheap service was good probably will not go back"*

Is this review positive or negative?

Methods

Methods cont.

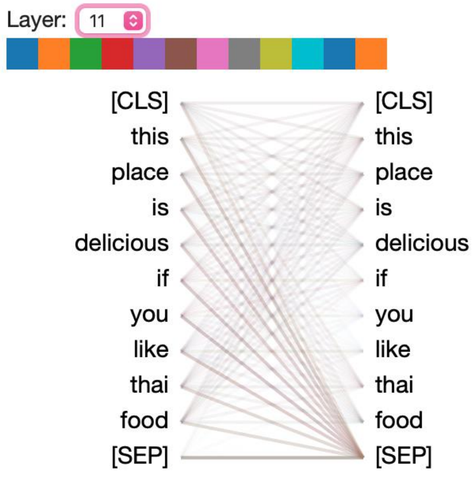
We trained BERT on word embeddings derived from restaurant reviews from Yelp. We used the existing human attention maps and generated new ones for BERT.



We used k-Means clustering on the reviews. We will use McNemer tests to find out in which cluster BERT performs significantly better than the humans.

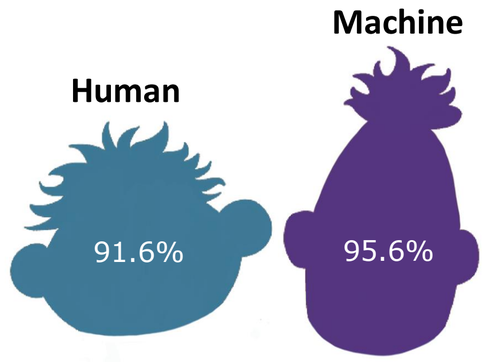
Results

We created attention maps for every review in the dataset using BERT.



Results cont.

We calculated the accuracy for both human and BERT on the test data. The accuracy was 91.6% for humans and 95.6% for BERT.



We found that BERT outperforms humans in every cluster.

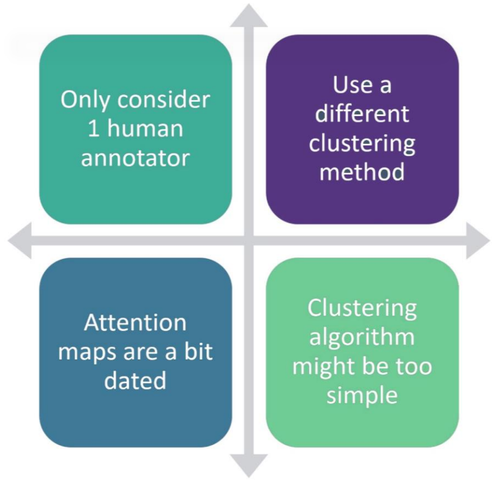
Cluster	P-Value	Statistics
0	0.00	11.0
1	0.00	11.0
2	0.00	9.0
3	0.00	6.0
4	0.00	4.0

Conclusion

While this research remains exploratory, it opens the door to future decision makers with new insights into

- how they can improve the quality of their decisions in the future or
- how to use machine learning systems to augment these decision making processes.

Limitations



References

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. 2020. Human attention maps for text classification: Do humans and neural networks focus on the same words?. In Proceedings of the 58th annual meeting of the association for computational linguistics. 4596–4608.

Acknowledgements

We would like to thank Dong Nguyen, Yupei Du, Heysem Kaya and Gizem Sogancioglu for their guidance and support during the course of this project.