Dimensionality reduction over neuron activations

Juliana Carvalho
FGV

Abstract

Dimensionality reduction can be applied to the activations of neurons across a bunch of texts to explore the interpretability of the resulting directions or patterns. In this work, we perform an PCA reduction to over posittive/negative analysis of comments. The experiments with real and toy data suggest that BERT is able to differentiate distinct concepts and that this differentiation becomes more sophisticated through its layers.

Keywords:Mechanistic interpretability, AI safety, Transformers, PCA

1. Introduction

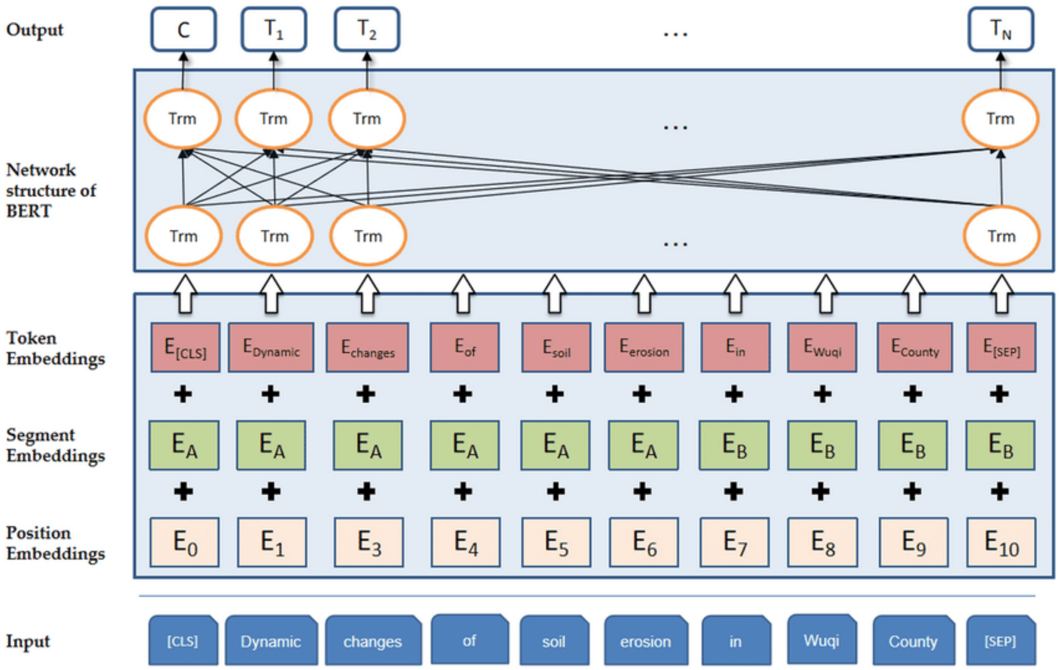
Our hypothesis is if the model improve though the activation layers. We would verify if BERT is able to differentiate distinct concepts and that this differentiation becomes more sophisticated through its layers.

2. Methods

We did dimensionality reduction over neuron activations across texts. Then, we verified how interpretable the resulting directions are.

The idea is to separate positive (1) and negative (0) comments in the vector space – the better the model, the better is the separation. Then, we would visusalize using a dimension reduction (PCA) of the vectors in 2 dimensions.

We selected BERT as a pretrained model.



Code:
https://colab.research.google.com/drive/13tj-COrH6yXzhC5Hw75oaDkEhS_0n7HY#scrollTo=uH0jCTsRECUQ

Toy data: generated by human volunteers regarding Wyttam Abbey evaluation's.

IMBD: filtered data by number of chars.

3. Results

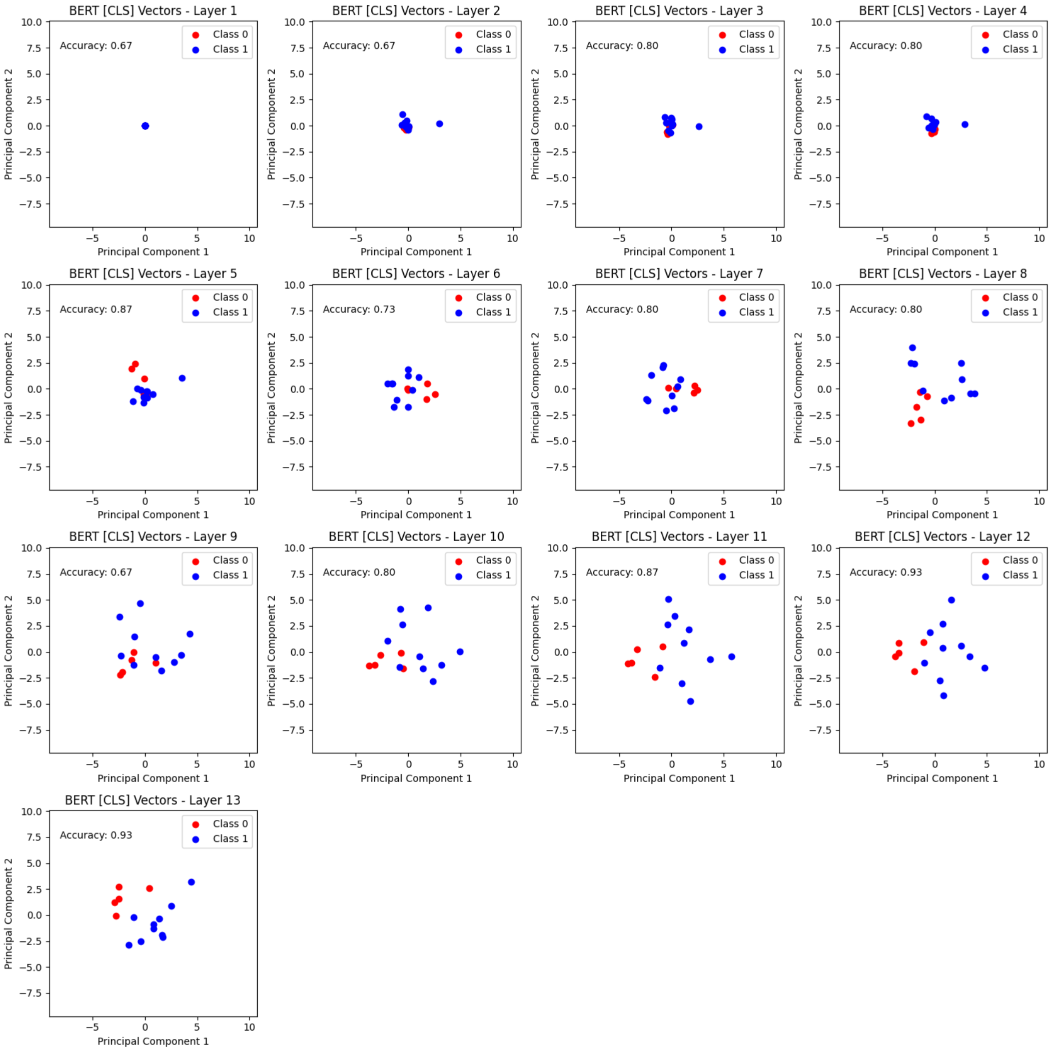
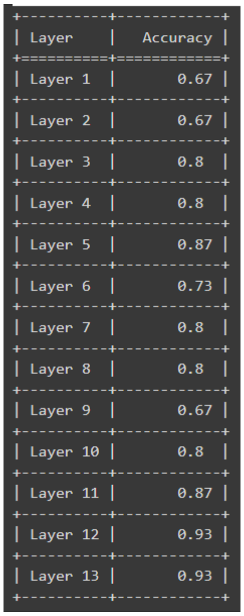


Figure 1 – Toy data (PCA)



```
+----------+-----------+
| Layer    |  Accuracy |
+==========+===========+
| Layer  1 |      0.67 |
+----------+-----------+
| Layer  2 |      0.67 |
+----------+-----------+
| Layer  3 |      0.8  |
+----------+-----------+
| Layer  4 |      0.8  |
+----------+-----------+
| Layer  5 |      0.87 |
+----------+-----------+
| Layer  6 |      0.73 |
+----------+-----------+
| Layer  7 |      0.8  |
+----------+-----------+
| Layer  8 |      0.8  |
+----------+-----------+
| Layer  9 |      0.67 |
+----------+-----------+
| Layer 10 |      0.8  |
+----------+-----------+
| Layer 11 |      0.87 |
+----------+-----------+
| Layer 12 |      0.93 |
+----------+-----------+
| Layer 13 |      0.93 |
+----------+-----------+
```

Table 1: toy data table

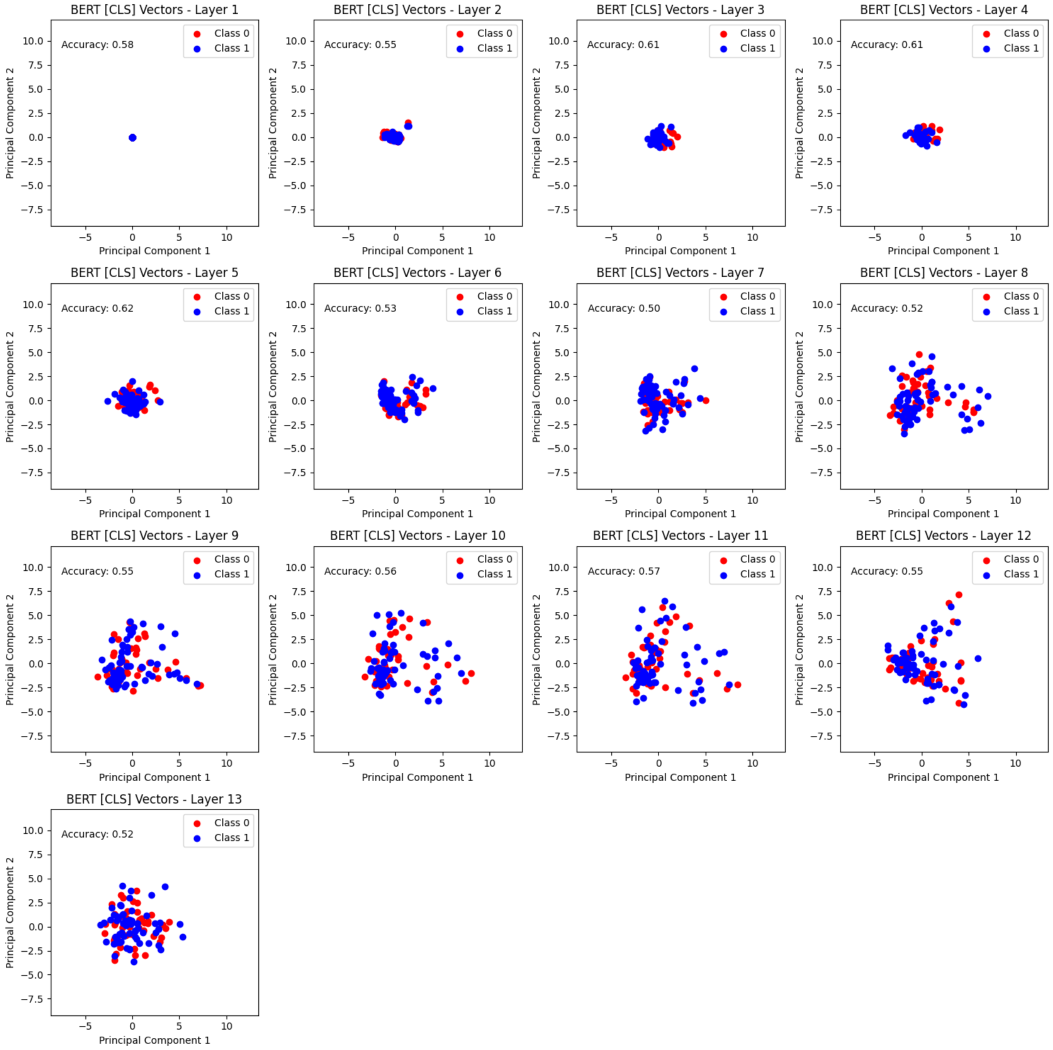
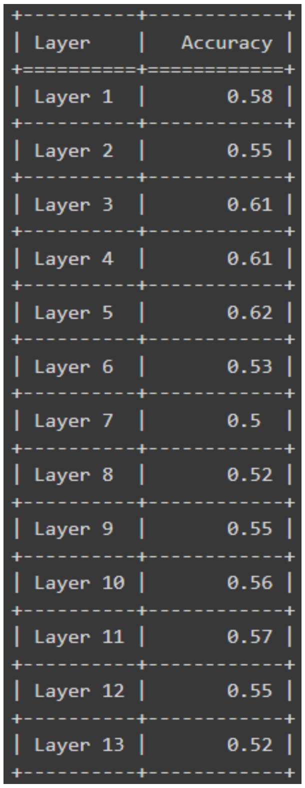


Figure 2 – IMBD data (PCA)



```
+----------+------------+
| Layer    |   Accuracy |
+==========+============+
| Layer 1  |       0.58 |
+----------+------------+
| Layer 2  |       0.55 |
+----------+------------+
| Layer 3  |       0.61 |
+----------+------------+
| Layer 4  |       0.61 |
+----------+------------+
| Layer 5  |       0.62 |
+----------+------------+
| Layer 6  |       0.53 |
+----------+------------+
| Layer 7  |       0.5  |
+----------+------------+
| Layer 8  |       0.52 |
+----------+------------+
| Layer 9  |       0.55 |
+----------+------------+
| Layer 10 |       0.56 |
+----------+------------+
| Layer 11 |       0.57 |
+----------+------------+
| Layer 12 |       0.55 |
+----------+------------+
| Layer 13 |       0.52 |
+----------+------------+
```

Table 2: Real data table

4. Discussion and Conclusion

The experiments with real and toy data suggest that BERT is able to differentiate distinct concepts and that this differentiation becomes more sophisticated through its layers. It is like if the model is learning more complex language context and phenomenon as the vectors passes through layers. Intuitively, the first layers would be responsible for simple language concepts, like identifying words and punctuation, and the following layers would be responsible to merge these concepts to create a more sophisticated understanding of natural language, including context and semantics. Literature in this area reinforce this conclusion (Yun et al., 2021). The experiments are not conclusive but they show a clear direction of research in how to interpret the reasoning inside Transformer-based language models.

5. References

Z. Yun, Y. Chen, B. Olshausen, and Y. LeCun, “Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors,” in Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Online: Association for Computational Linguistics, Jun. 2021, pp. 1–10. doi: 10.18653/v1/2021.deelio-1.1.