Condor Camp Hackathon Problem 9.60

Juliana Carvalho de Souza

Problem 9.60 - *Studying Learned Features in Language Models*

“Try doing dimensionality reduction over neuron activations across a bunch of text, and see how interpretable the resulting directions are.”

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.

Analyses

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.

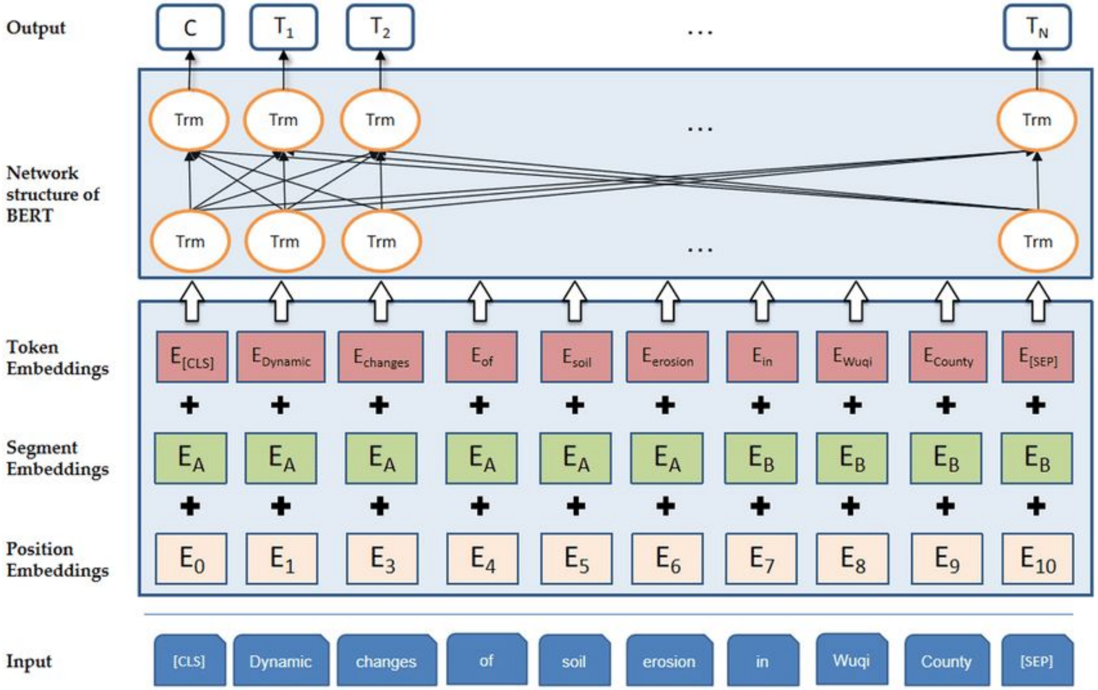
Plotting the [CLS] of the text through the layers of the model we could see that the separation gets greater as the text pass through more layers.

Method

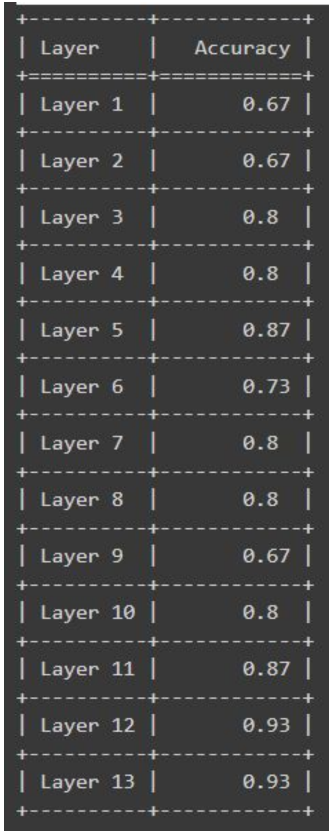
- Use real data from IBMD movie reviews compared with Toy data
- Selected BERT as a pretrained model
- Use PCA to reduce dimension and do the 2D plot
- Compared the accuracy through the layers - evaluated separability with logistic regression and cross-validation

Methods

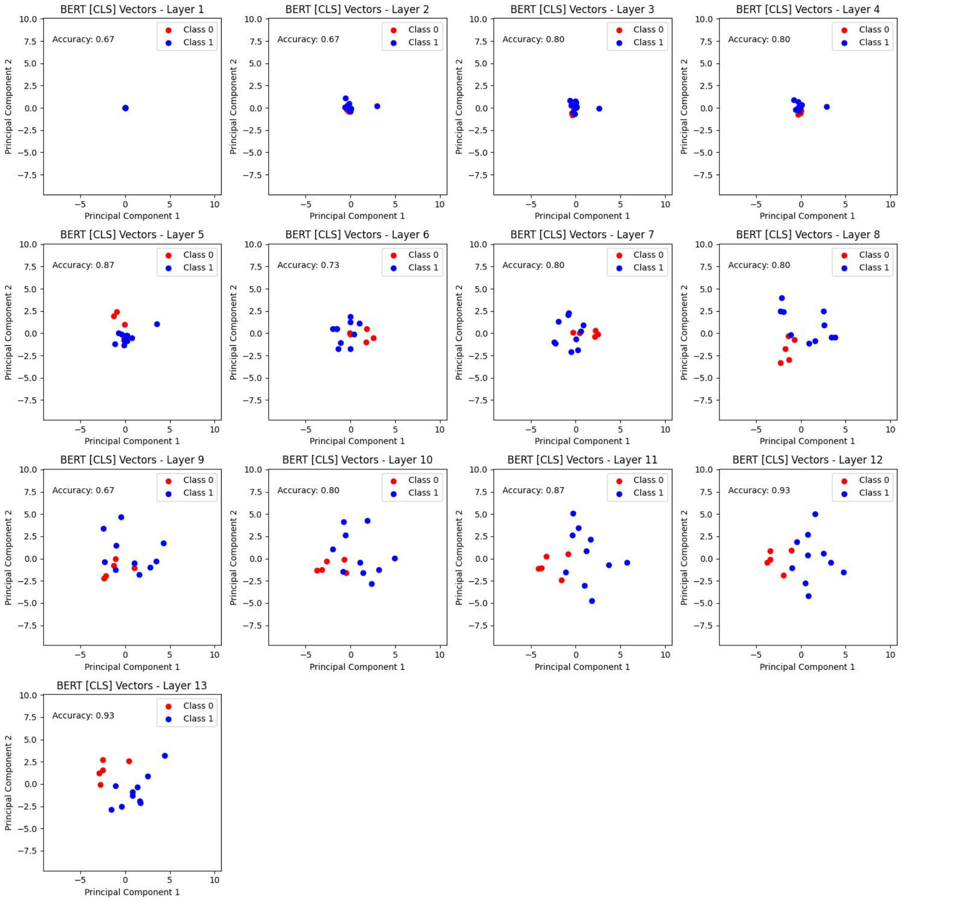
Use of BERT transformer on the texts embeddings, selecting the output from the layers



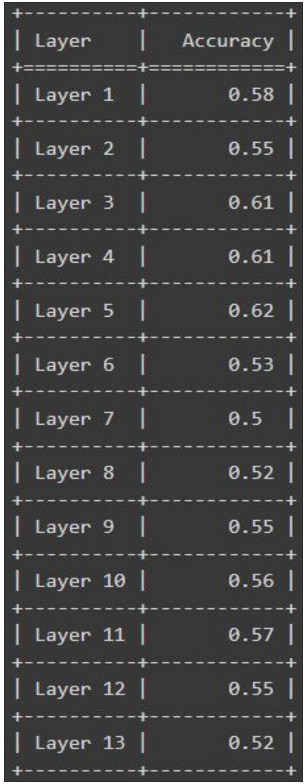
Results - Toy data (human comments about Wytham Abbey)



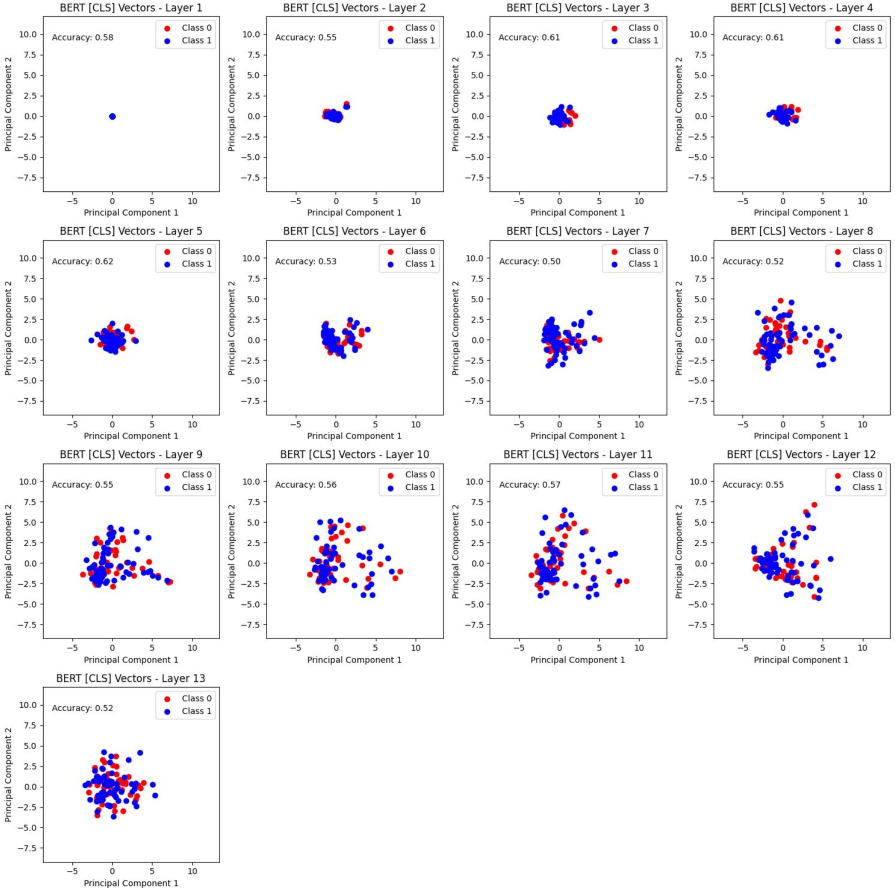
```
+----------+------------+
| Layer    |   Accuracy |
+==========+============+
| Layer 1  |       0.67 |
+----------+------------+
| Layer 2  |       0.67 |
+----------+------------+
| Layer 3  |       0.8  |
+----------+------------+
| Layer 4  |       0.8  |
+----------+------------+
| Layer 5  |       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 |
+----------+------------+
```



Results: real data - ratings from IMBD



```
+----------+------------+
| 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 |
+----------+------------+
```



Conclusions

- In toy data, there is a good separation as layers increase
- In IMBD data, the separation increases and then decreases:
    - As it’s raw a data not cleaned it can contain grammar mistakes or other problems that the model is not capable of capturing the separation at the last layers
    - Possibility of overfitting

In conclusion: we could see good direction separation in the plot. It’s almost clear the separation direction of positive/negative review in Toy data (last layer). The experiments suggest that BERT is able to differentiate distinct concepts and that this differentiation becomes more sophisticated through its layers, as semmantics