Figure 8: The columns represent attention paid to: (a) modified chunk on funny sentence, (b) non-modified chunk on funny sentences, (c) modified chunk on serious sentences, and (d) non-modified chunk on serious sentences. The first row is distilBERT-1S, the second row is RoBERTa-1S Lighter color represents higher average attention.

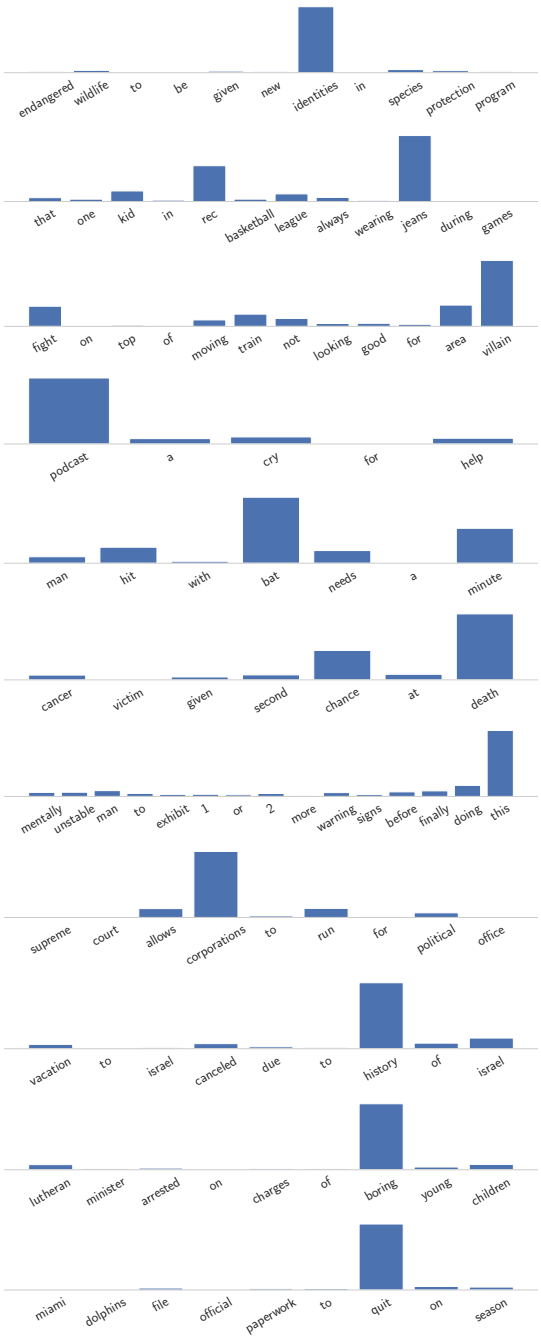

Figure 9: Examples of attention distributions on funny sentences by the head 10-6 of BERT.