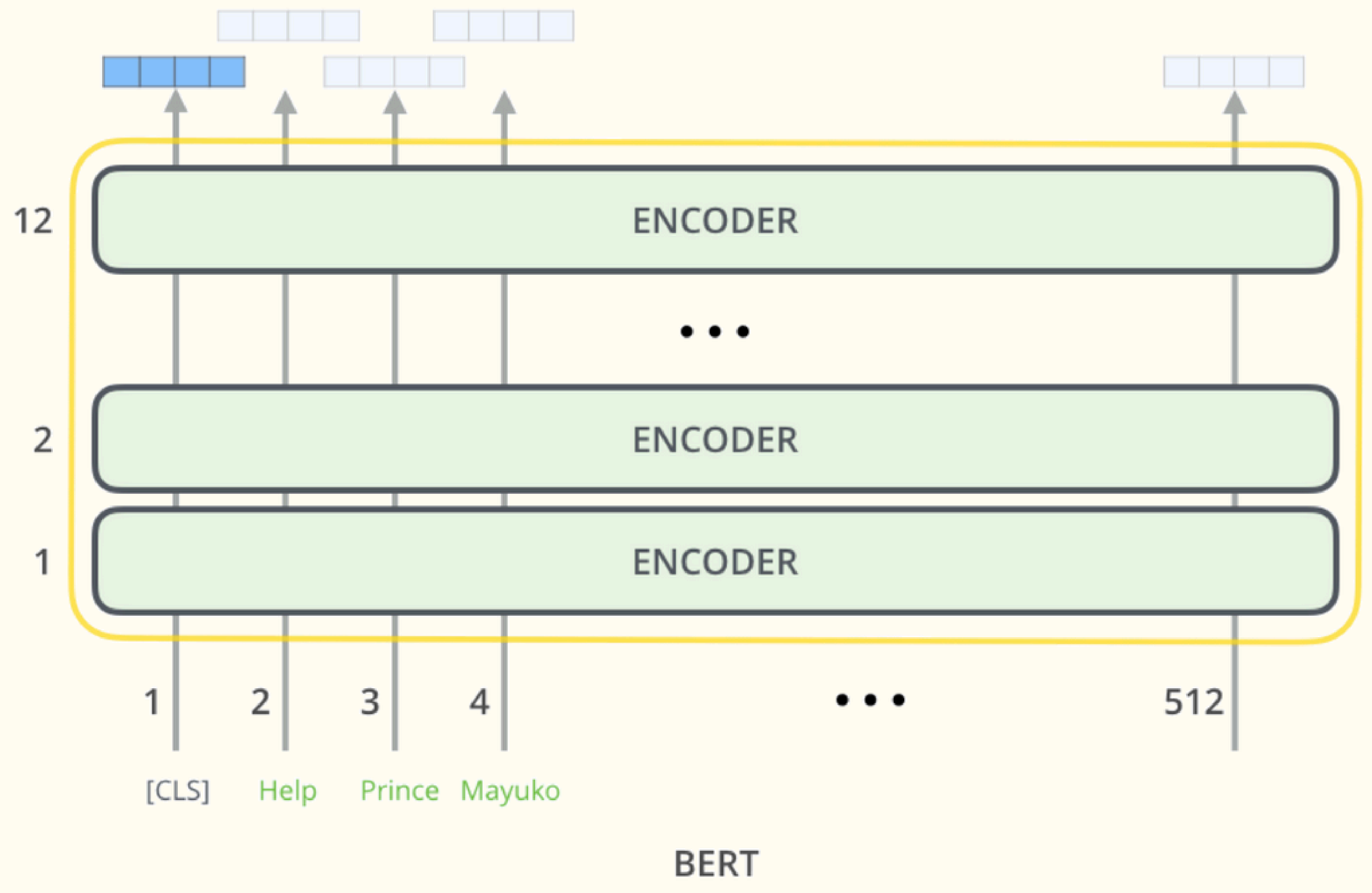
Testing multiple combinations of outputs from BERT’s encoders in BERTopic topic modeling

Author: Dominik Koterwa

Quick intro

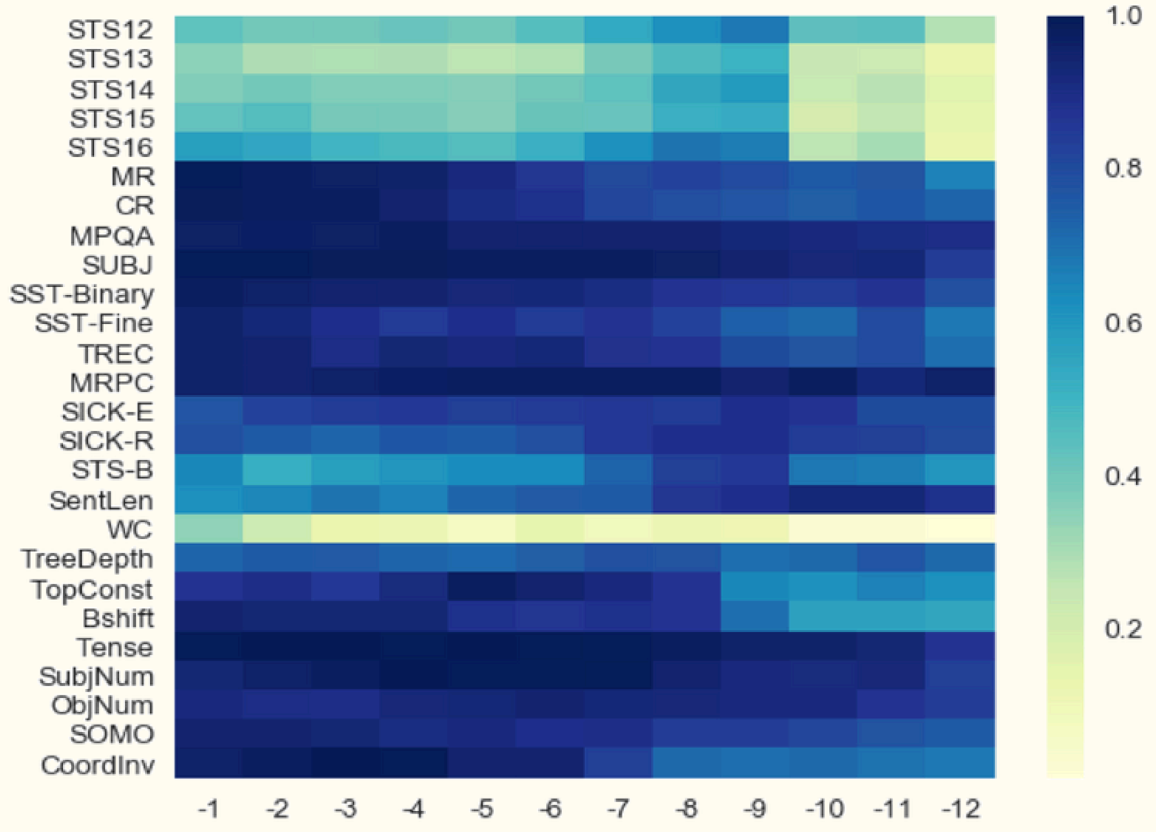
BERTopic is built using BERT, which is a model constructed from Encoder part of traditional Transformer. However, to provide a higher efficiency, **BERT is a stack of 12 Encoders**.

- The output of each Encoder can be used as a representation of a sequence
- Usually, people use the output for the CLS token from the last Encoder
- BERTopic operates on the output from last Encoder with mean pooling and then uses it for topic modeling



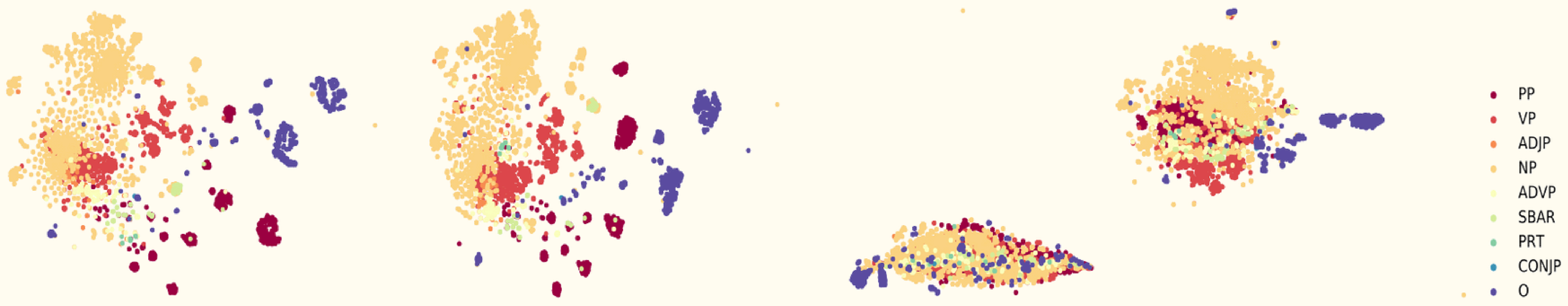
Source: http://jalammar.github.io/illustrated-bert/

Research about output of Encoders



Source: Ma et al., 2019

A large portion of research has been made in order to investigate the linguistic properties of BERT's activations from different layers. It has been stated that **in the first layers model concentrates knowledge about structural/syntactic features of the input. However, in next layers, this knowledge is replaced with higher concentration on semantics/task specific features.**

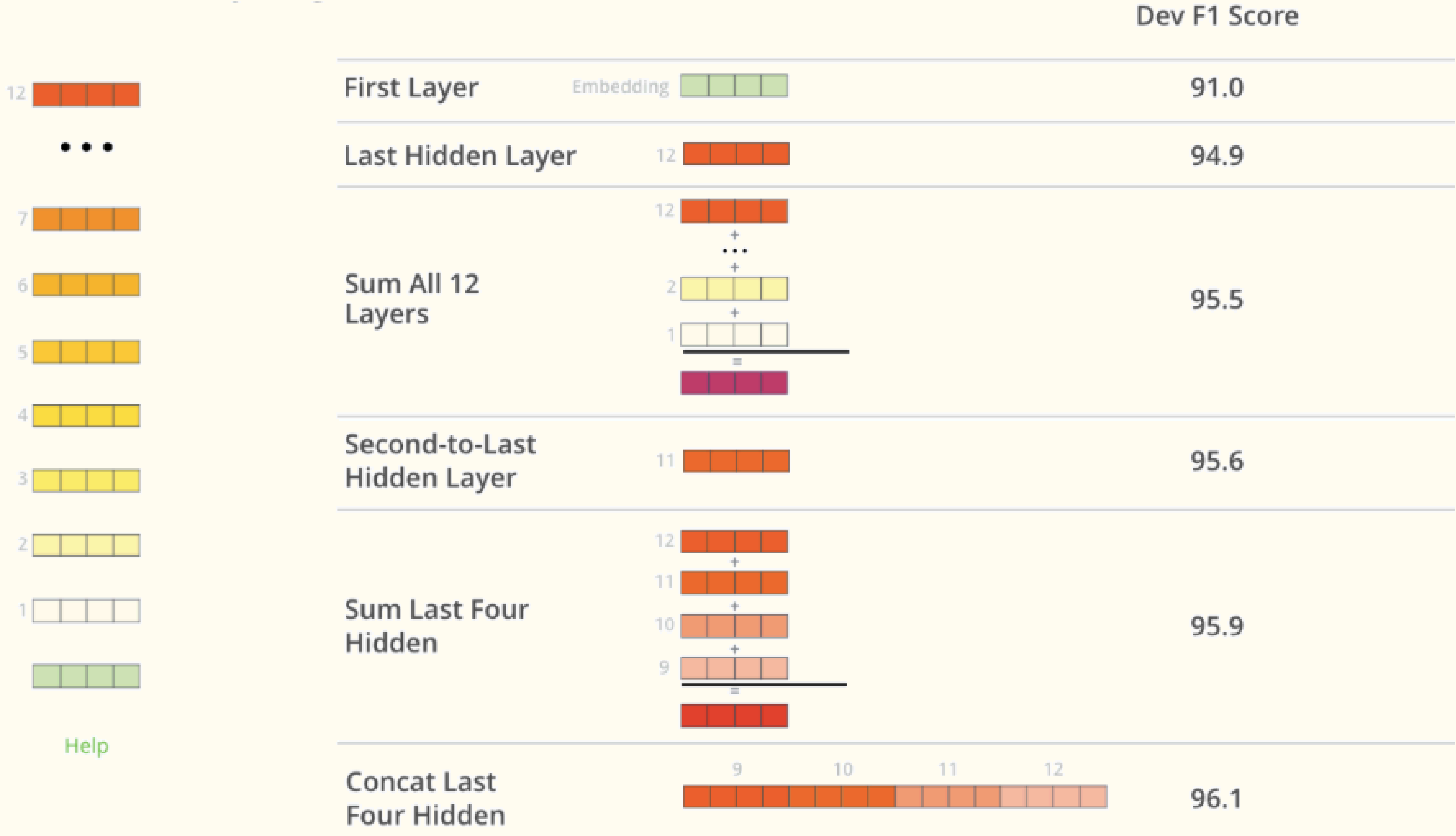


(a) Layer 1 (b) Layer 2 (c) Layer 11 (d) Layer 12

Source: Jawahar et al., 2019

BERT’s authors also analyzed that

Results on CoNLL-2003 Named Entity Recognition task



12
• • •
7
6
5
4
3
2
1
Help

		Dev F1 Score
First Layer	Embedding	91.0
Last Hidden Layer	12	94.9
Sum All 12 Layers	12 + ••• + 2 + 1 =	95.5
Second-to-Last Hidden Layer	11	95.6
Sum Last Four Hidden	12 + 11 + 10 + 9 =	95.9
Concat Last Four Hidden	9 10 11 12	96.1

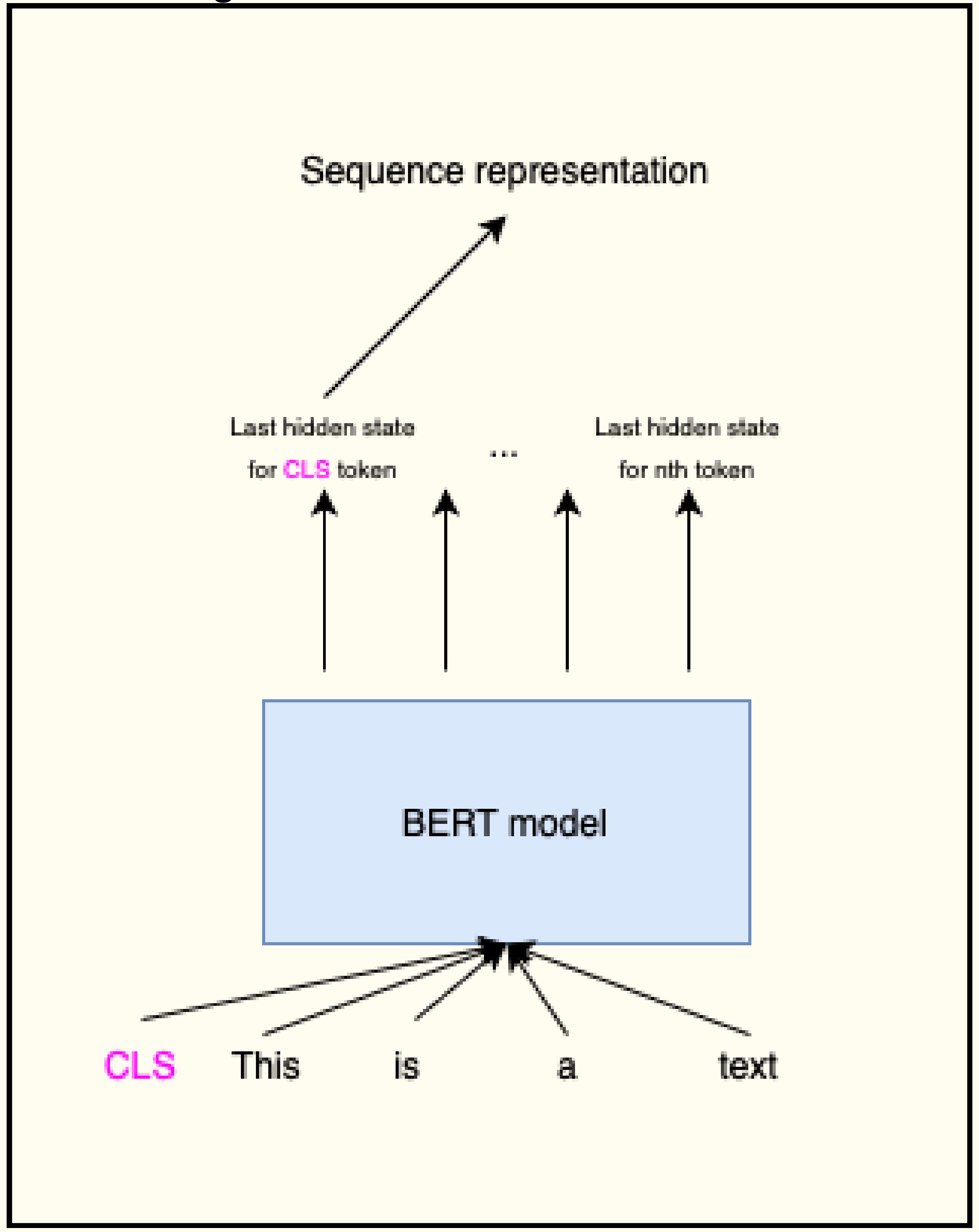
Source: http://jalammar.github.io/illustrated-bert/

In BERTopic, Mean Pooling from last layer is being used as a representation of input sequence.

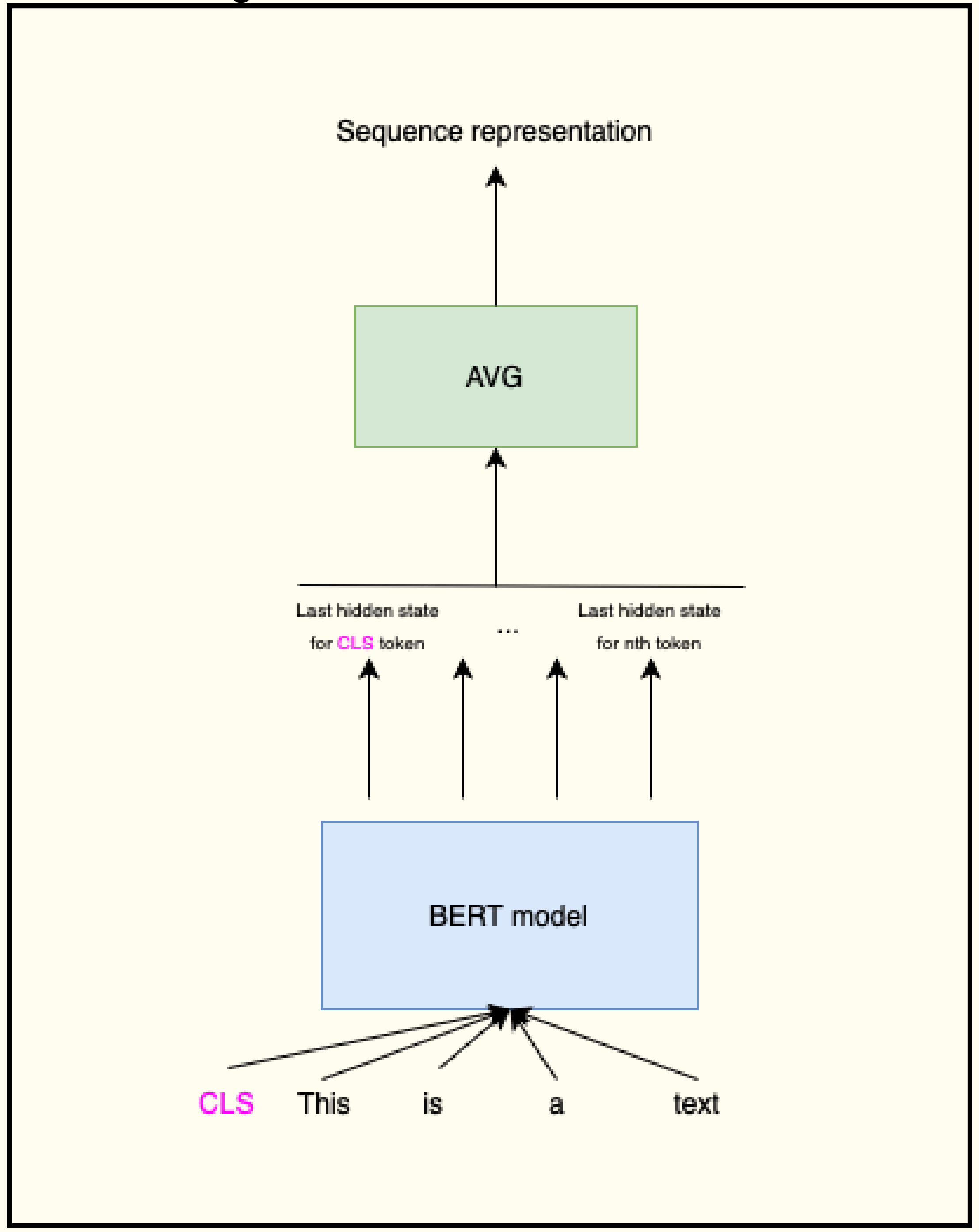
The question is **how different types of representations affect the quality of BERTopic?** In this research I test 18 of them.

Quick introduction to best-known pooling methods

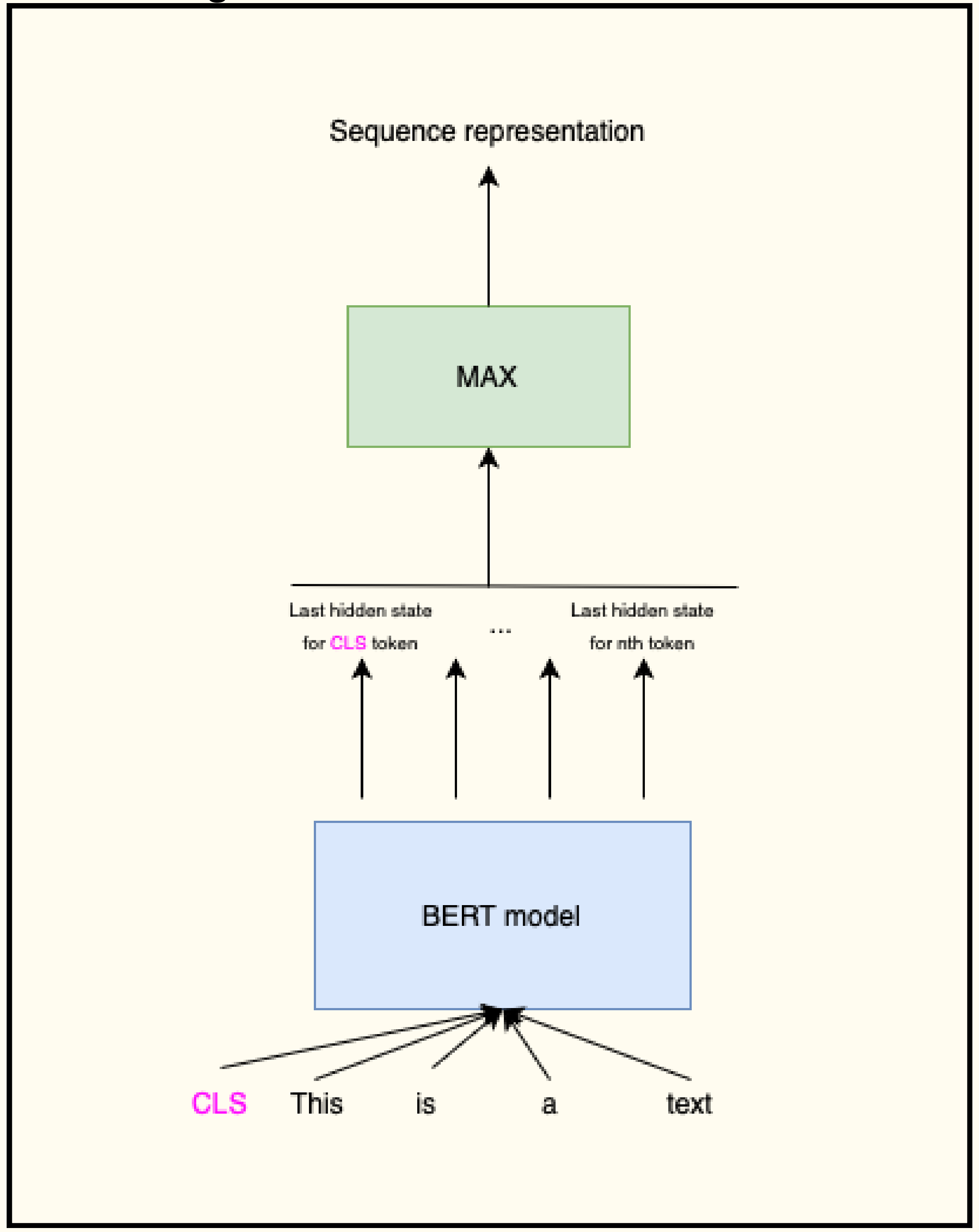
CLS Pooling


Sequence representation
Last hidden state for CLS token
...
Last hidden state for nth token
BERT model
CLS
This
is
a
text

Mean Pooling


Sequence representation
AVG
Last hidden state for CLS token
...
Last hidden state for nth token
BERT model
CLS
This
is
a
text

Max Pooling

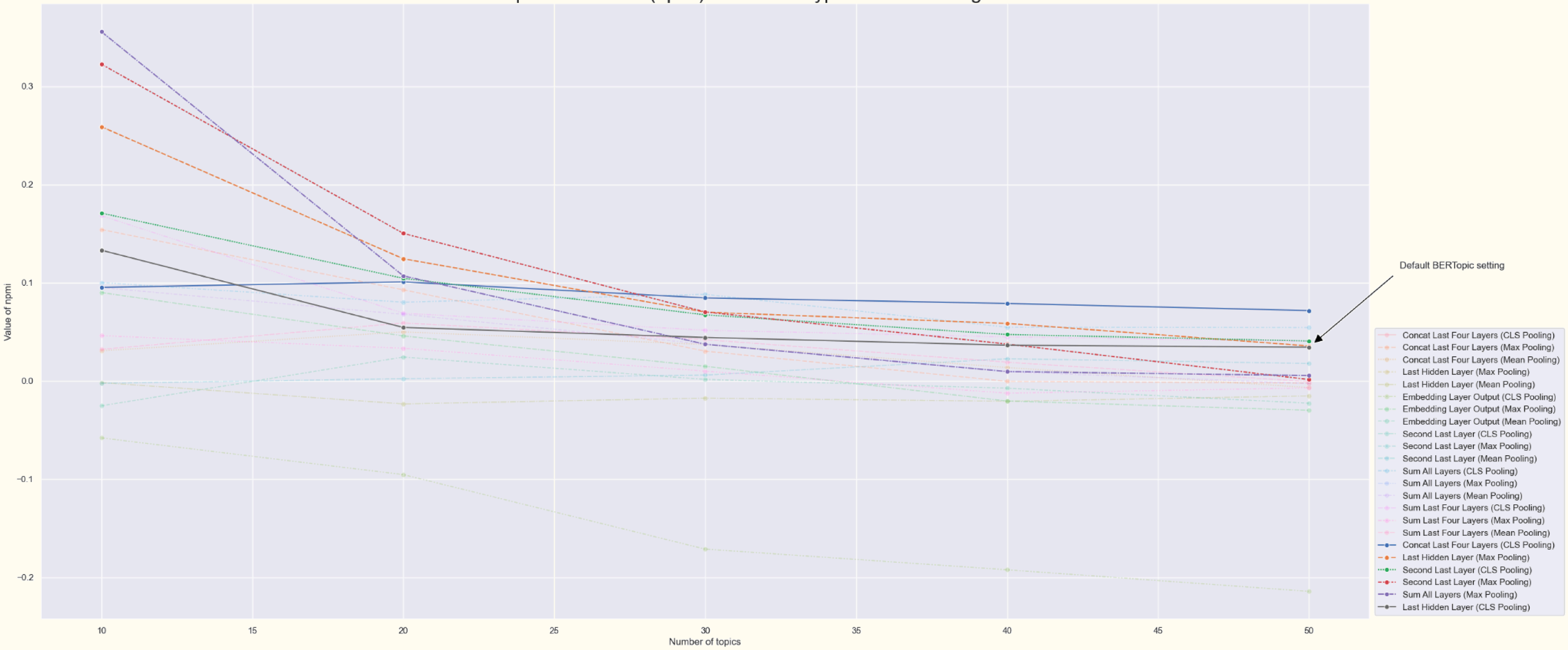

Sequence representation
MAX
Last hidden state for CLS token
...
Last hidden state for nth token
BERT model
CLS
This
is
a
text

Description of experiments

- **I chose the dataset of Trump’s tweets**, because it is a common benchmark for topic modeling and was used in BERTopic paper.
- Used “all-MiniLM-L12-v2" Sentence Transformer to produce embeddings.
- 18 different types of embeddings (6 types of aggregation * 3 types of pooling) have been collected.
- Then, I used Trainer provided by the BERTopic’s author to conduct evaluation.
- **Topic Coherence and Topic Diversity measures** have been used to compare the quality of vaious options.

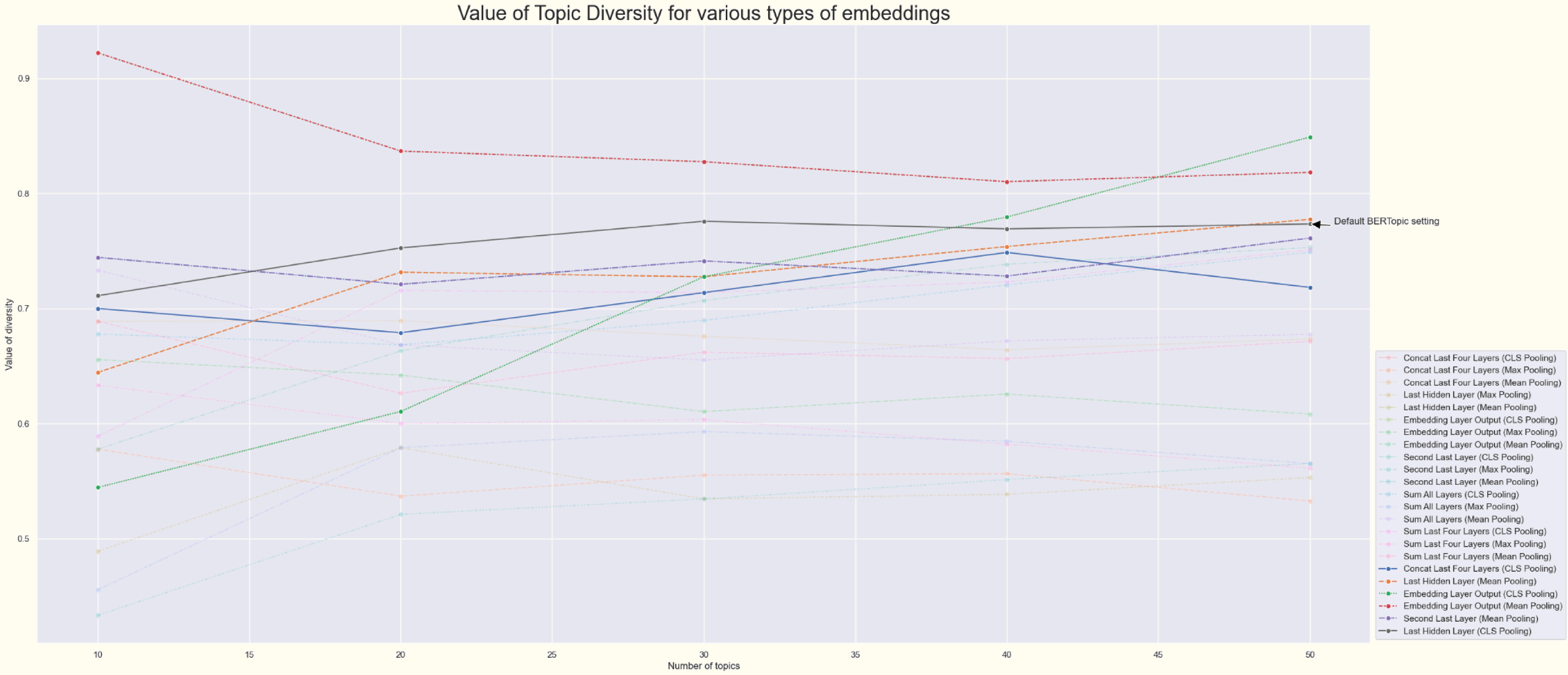
Results on Trump Tweets dataset (1)

Value of Topic Coherence (npmi) for various types of embeddings

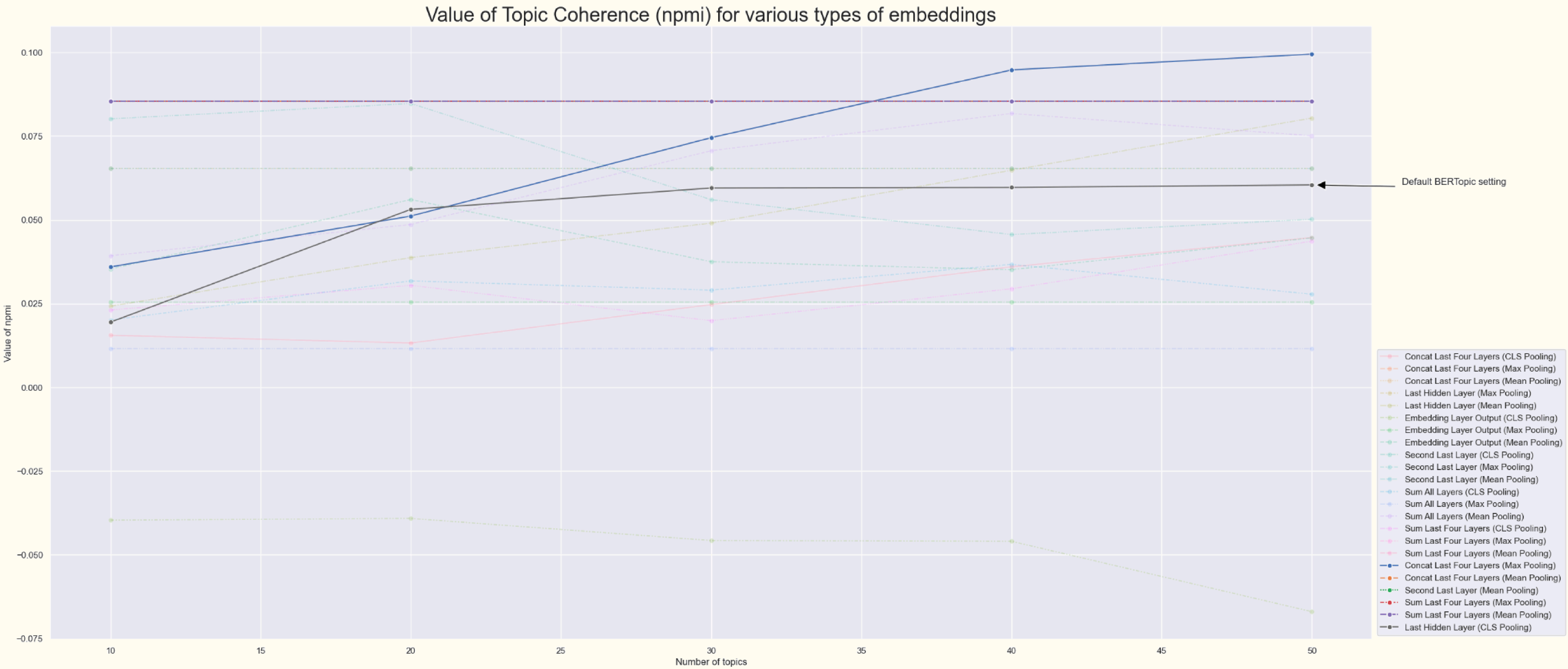


Default BERTopic setting

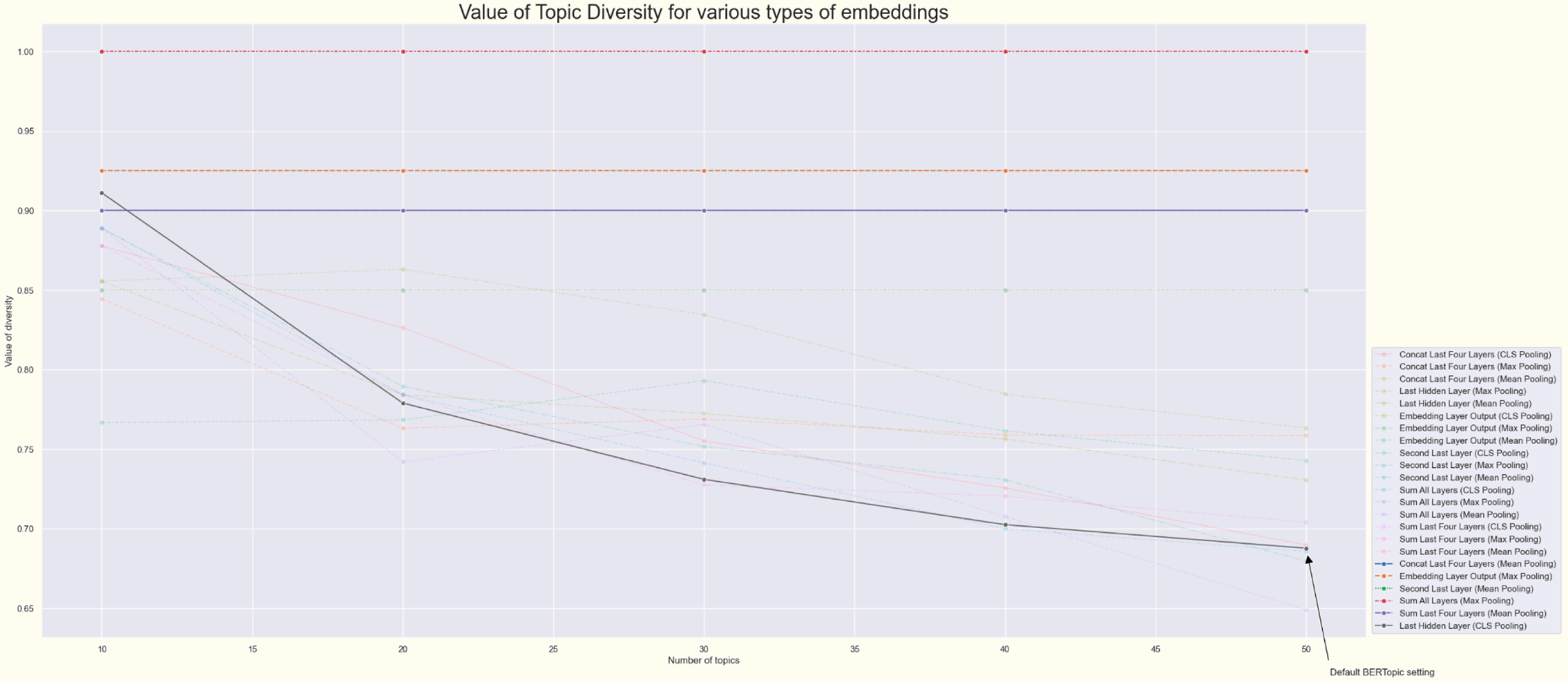
Results on Trump Tweets dataset (2)



Results on 20 News Groups dataset (1)



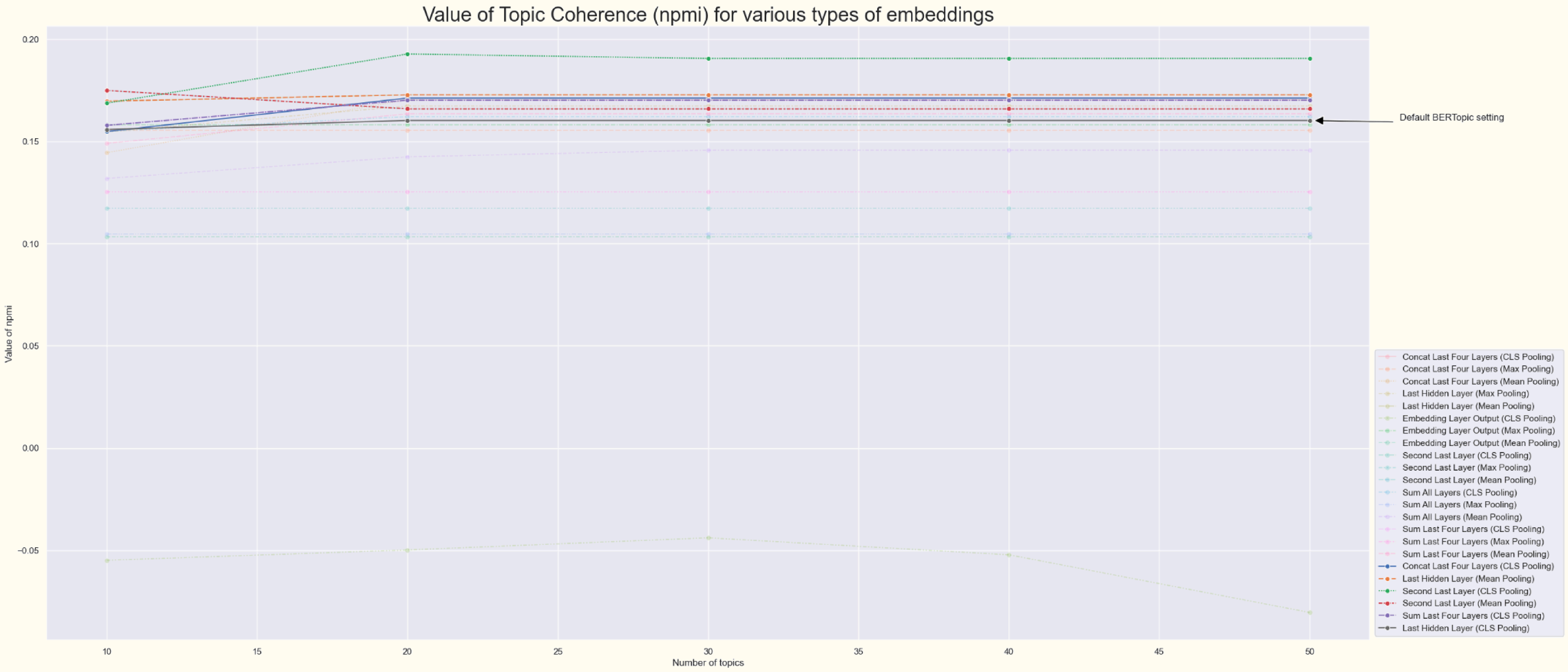
Results on 20 News Groups dataset (2)



Value of Topic Diversity for various types of embeddings

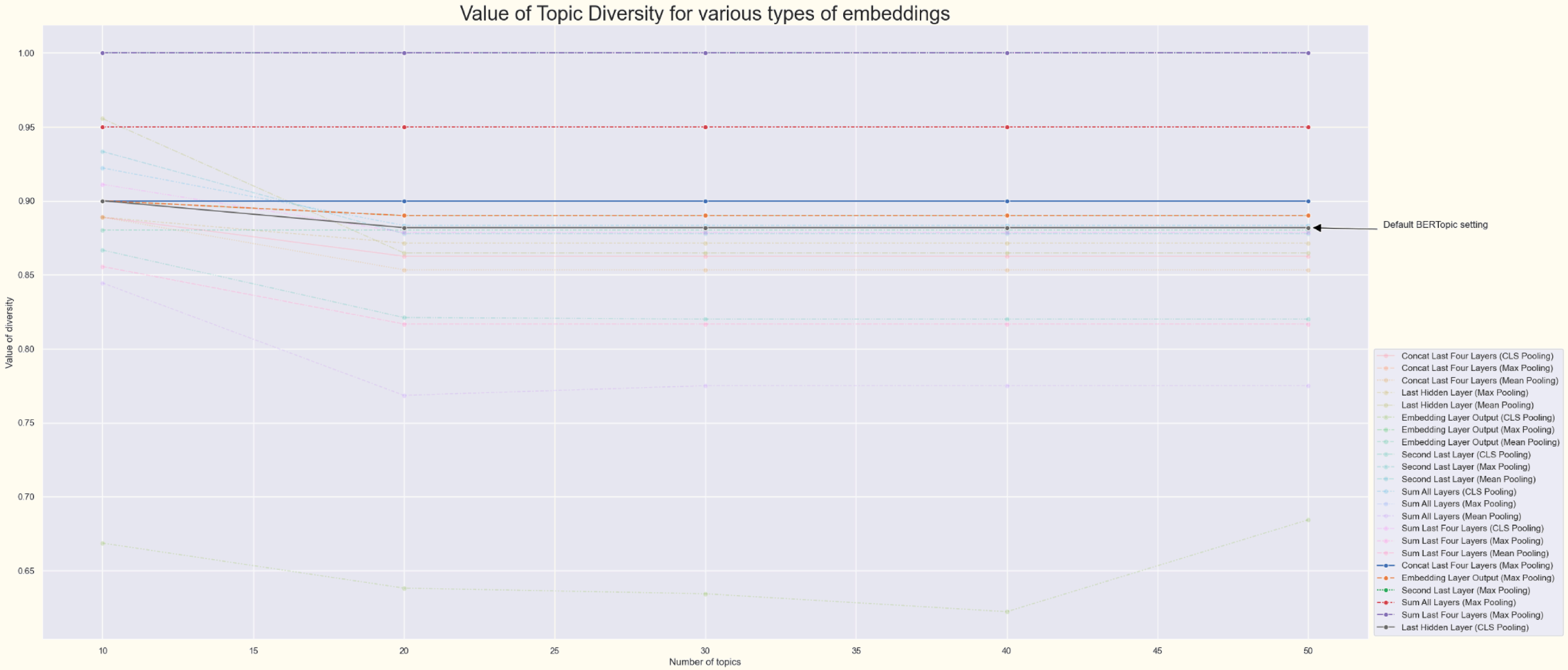
Default BERTopic setting

Results on BBC News dataset (1)



Default BERTopic setting

Results on BBC News dataset (2)



Conclusions

- **Default BERTopic configuration does not guarantee the best value of Coherence or Diversity** for any dataset and any number of topics.
- It is useful to explore more possibilities and options when in comes to topic modeling with models based on Transformers.
- However, **it is costly to explore larger search grid.**
- It would be beneficial to see how does different pooling strategy behave with different models producing the embeddings.
- Big shout-out to free Kaggle Notebooks for making this research possible.