# A topic in the combination of Large Language Model and Topological Data Analysis

Zeyang Ding

May 29, 2025

# Table of contents

# Outline

# BERT Model

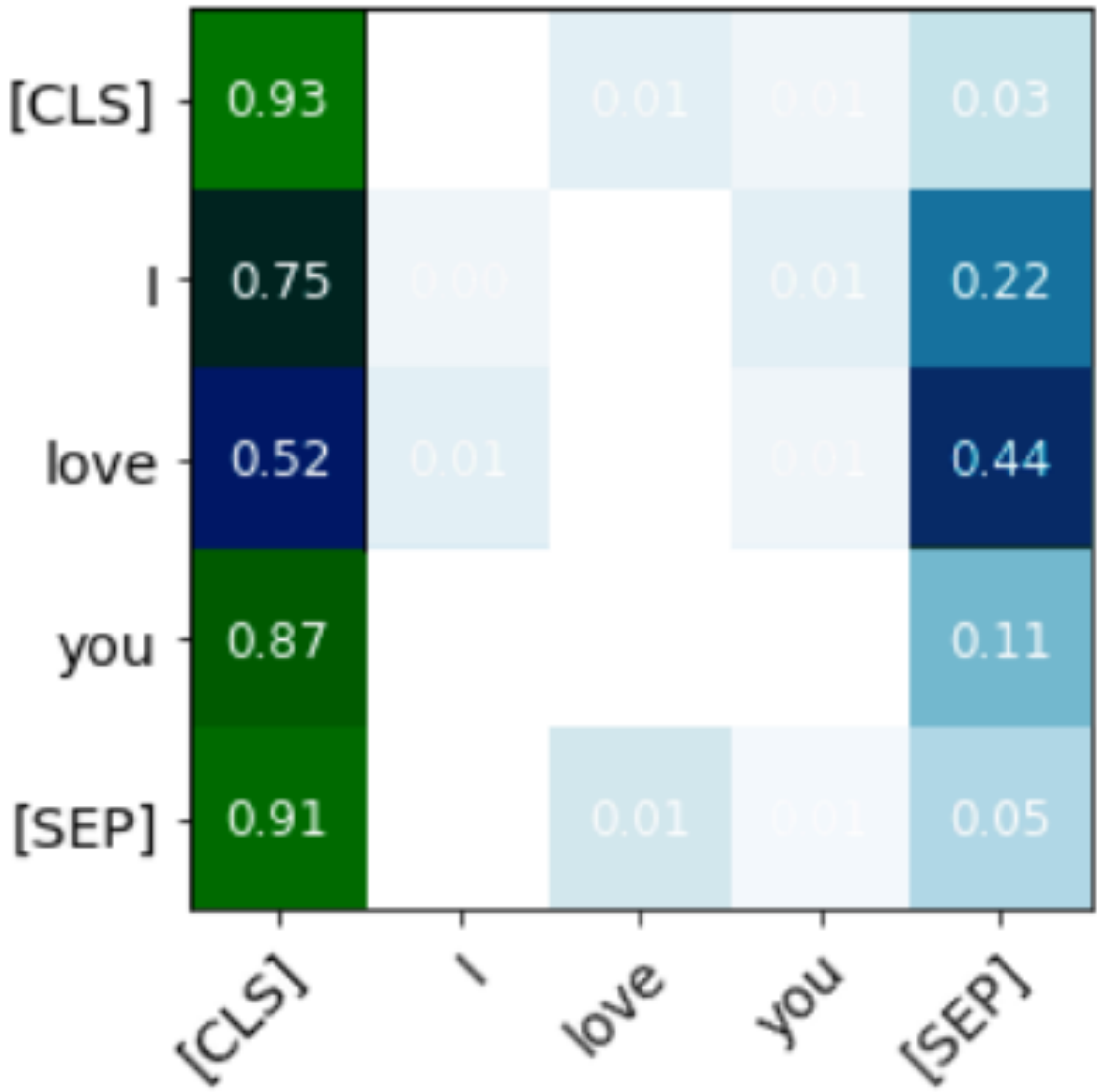
The BERT architecture comprises $L$ encoder layers with $H$ attention heads in each layer. The input of each attention head is a matrix $X$ consisting of the $d$-dimensional representations (row-wise) of $m$ tokens, so that $X$ is of shape $m \times d$. The head outputs an updated representation matrix $X^{\mathrm{out}}$:

$$\begin{aligned} X^{\mathrm{out}} &= W^{\mathrm{attn}}(XW^{\mathrm{V}}) \\ \text{with } W^{\mathrm{attn}} &= \mathrm{softmax}\left(\frac{(XW^{\mathrm{Q}})(XW^{\mathrm{K}})^{\mathrm{T}}}{\sqrt{d}}\right), \end{aligned} \tag{1}$$

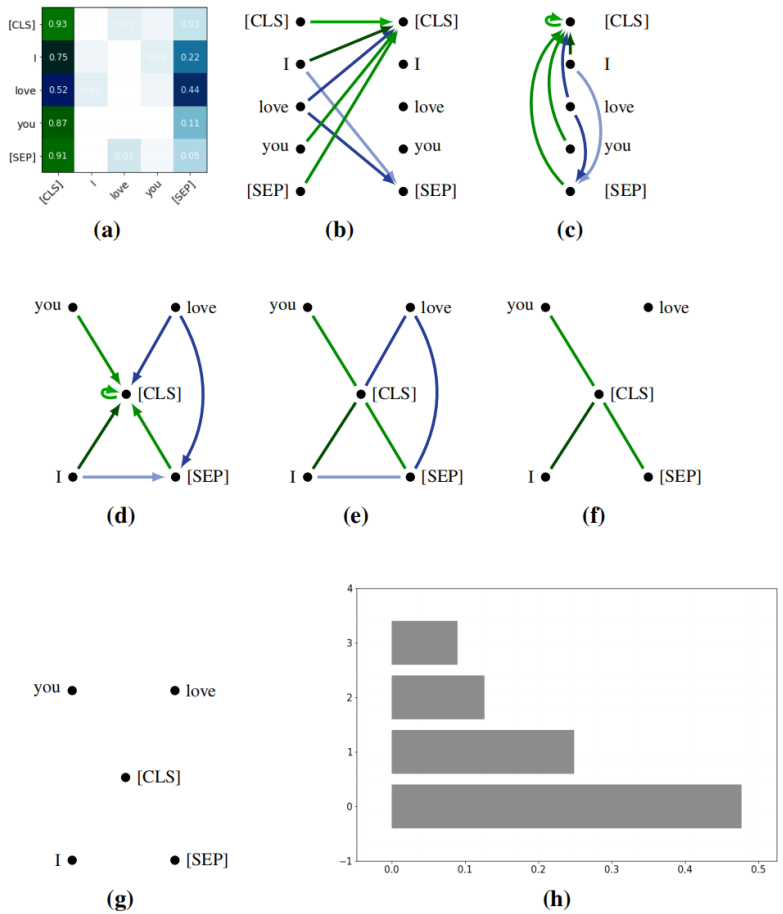
where $W^{\mathrm{Q}}$, $W^{\mathrm{K}}$, $W^{\mathrm{V}}$ are trained projection matrices of shape $d \times d$ and $W^{\mathrm{attn}}$ is of shape $m \times m$ matrix of attention weights. Each element $w_{ij}^{\mathrm{attn}}$ can be interpreted as a weight of the $j$-th input's *relation* to the $i$-th output: larger weights mean stronger connection between the two tokens.

# Attention Map and Attention Graph

An *attention map* displays an attention matrix $W^{attn}$ in form of a heat map, where the color of the cell $(i, j)$ represents the *relation* of the $i$-th token to the output representation of the $j$-th token. The attention matrix is considered to be a weighted graph with the vertices representing tokens and the edges connecting pairs of tokens with strong enough mutual relation (the higher the weight, the stronger the relation). The construction of such graph appears to be quite problematic: a threshold needs to be set to distinguish between weak and strong relations. This leads to instability of the graph's structure: changing the threshold affects the graph properties such as the number of edges, connected components, cycles. The choice of the optimal thresholds is essential to define which edges remain in the graph. TDA methods allow extracting the overall graph's properties which describe the development of the graph with respect to changes in the threshold.


[CLS]
I
love
you
[SEP]
0.93
0.01
0.01
0.03
0.75
0.00
0.01
0.22
0.52
0.01
0.01
0.44
0.87
0.11
0.91
0.01
0.01
0.05
[CLS]
I
love
you
[SEP]

# Representing attention map with a weighted graph



# Outline

# Topological Features

First, Fix a set of thresholds $T = \{t_i\}_{i=1}^{k}, 0 < t_1 < ... < t_k < 1$. Consider an attention head $h$ and corresponding weights $W^{\text{attn}} = (w_{i,j}^{attn})$. Given a text sample $s$, for each threshold level $t \in T$ we define the weighted directed graph $\Gamma_s^h(t)$ with edges $\{j \to i \mid w_{ij}^{\text{attn}} \geq t\}$ and its undirected variant $\overline{\Gamma_s^h(t)}$ by setting an undirected edge $v_i v_j$ for each pair of vertices $v_i$ and $v_j$ which are connected by an edge in at least one direction in the graph $\Gamma_s^h(t)$.
Consider the following features of the graphs:

- the first two Betti numbers of the undirected graph $\overline{\Gamma_s^h(t)}$.
- the number of edges ($\mathbf{e}$), the number of strongly connected components ($\mathbf{s}$) and the amount of simple directed cycles ($\mathbf{c}$) in the directed graph $\Gamma_s^h(t)$.

To get the whole set of topological features for the given text sample $s$ and the attention head $h$, concatenate the features for all the thresholds, starting from $T$.

# Features Derived from Barcodes

For each text sample we calculate barcodes of the first two persistent homology groups (denoted as $H_0$ and $H_1$) on each attention head of the BERT model. Compute the following characteristics of these barcodes:

- The sum of lengths of bars;
- The mean of lengths of bars;
- The variance of lengths of bars;
- The number of bars with time of birth/death greater/lower than threshold;
- The time of birth/death of the longest bar (excluding infinite);
- The overall number of bars;
- The entropy of the barcode.

# Features Based on Distance to Patterns

Consider distances from the given graph to attention patterns as the graph features $d_i(\Gamma) = d(\Gamma, \Gamma_i)$:

- Attention to the previous token. $\Gamma_{feature} : E = (i+1, i),\ i = \overline{1, n-1}$.
- Attention to the next token. $\Gamma_{feature} : E = (i, i+1),\ i = \overline{1, n-1}$.
- Attention to [CLS]-token. [CLS]-token corresponds to the vertex 1 of the set $V = [1, n]$ as it denotes the beginning of the text. $\Gamma_{feature} : E = (i, 1),\ i = \overline{1, n}$.
- Attention to [SEP]-token. Suppose $i_1, \ldots, i_k$ are the indices of [SEP]-tokens. Then $\Gamma_{feature} : E = (i, i_t),\ i = \overline{1, n},\ t = \overline{1, k}$.
- Attention to punctuation marks. Let $i_1, \ldots, i_k$ be the indices of the tokens which correspond to commas and periods. $\Gamma_{feature} : E = (i, i_t),\ i = \overline{1, n},\ t = \overline{1, k}$. Note that this pattern can be potentially divided into Attention to commas and Attention to periods.

# Outline

# Data

| Text Source | | Train | | Validation | | Test | | \|Vocab\| | | Length | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **H** | **M** | **H** | **M** | **H** | **M** | **H** | **M** | **H** | **M** |
| WebText | GPT-2 Small; pure sampling | 20K | 20K | 2.5K | 2.5K | 2.5K | 2.5K | 220K | 532K | $593 \pm 177$ | $515 \pm 322$ |
| Amazon Review | GPT-2 XL pure sampling | 5K | 5K | 1K | 1K | 4K | 4K | 47K | 49K | $179 \pm 170$ | $177 \pm 171$ |
| RealNews | GROVER top-$p$ sampling | 5K | 5K | 1K | 1K | 4K | 4K | 98K | 75K | $721 \pm 636$ | $519 \pm 203$ |

Table 1: Statistics for the datasets used in the experiments on the artificial text detection task. **H**=Human; **M**=Machine.

# Results

| Model | WebText & GPT-2 Small | Amazon Reviews & GPT-2 XL | RealNews & GROVER |
|---|---|---|---|
| **TF-IDF, N-grams** | 68.1 | 54.2 | 56.9 |
| **BERT [CLS trained]** | 77.4 | 54.4 | 53.8 |
| **BERT [Fully trained]** | **88.7** | **60.1** | **62.9** |
| **BERT [SLOR]** | 78.8 | 59.3 | 53.0 |
| **Topological features** | 86.9 | 59.6 | 63.0 |
| **Barcode features** | 84.2 | 60.3 | 61.5 |
| **Distance to patterns** | 85.4 | 61.0 | 62.3 |
| **All features** | **87.7** | **61.1** | **63.6** |

# Robustness towards Unseen Models

This setting tests the robustness of the artificial text detection methods towards unseen TGMs on the **WebText & GPT-2** dataset. The baselines and models are trained on texts from the GPT-2 small model and further used to detect texts generated by unseen GPT-style models with pure sampling: GPT-2 Medium (345M), GPT-2 Large (762M) and GPT-2 XL (1542M). Note that such a setting is the most challenging as it requires the transfer from the smallest model to that of the higher number of parameters.

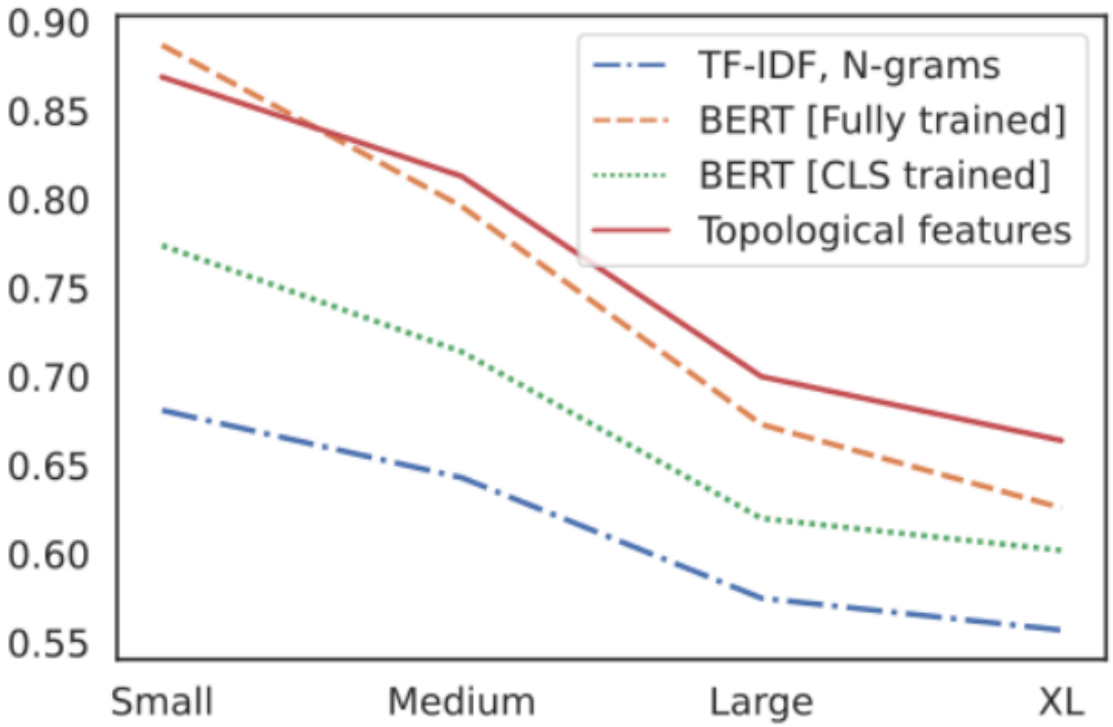
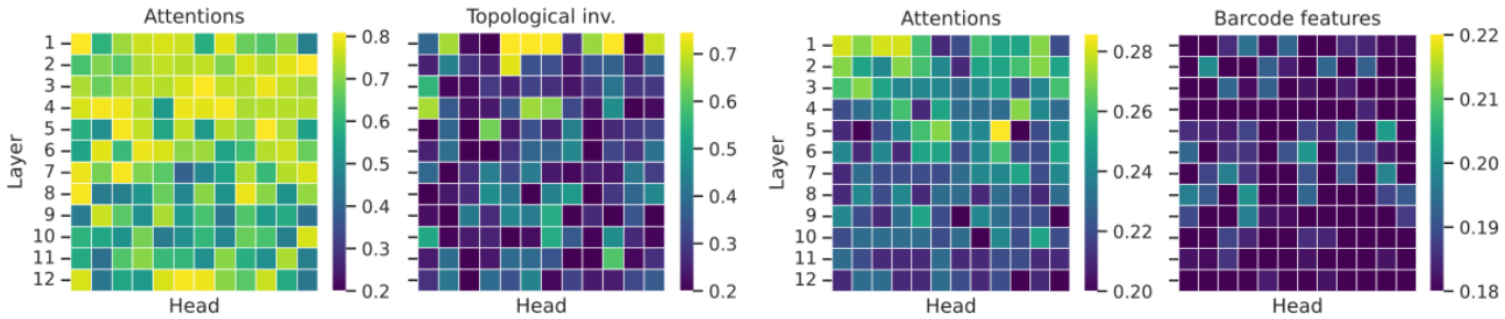



Figure 2: The results of the robustness experiments. X-axis=GPT-2 model size. Y-axis=Accuracy score.

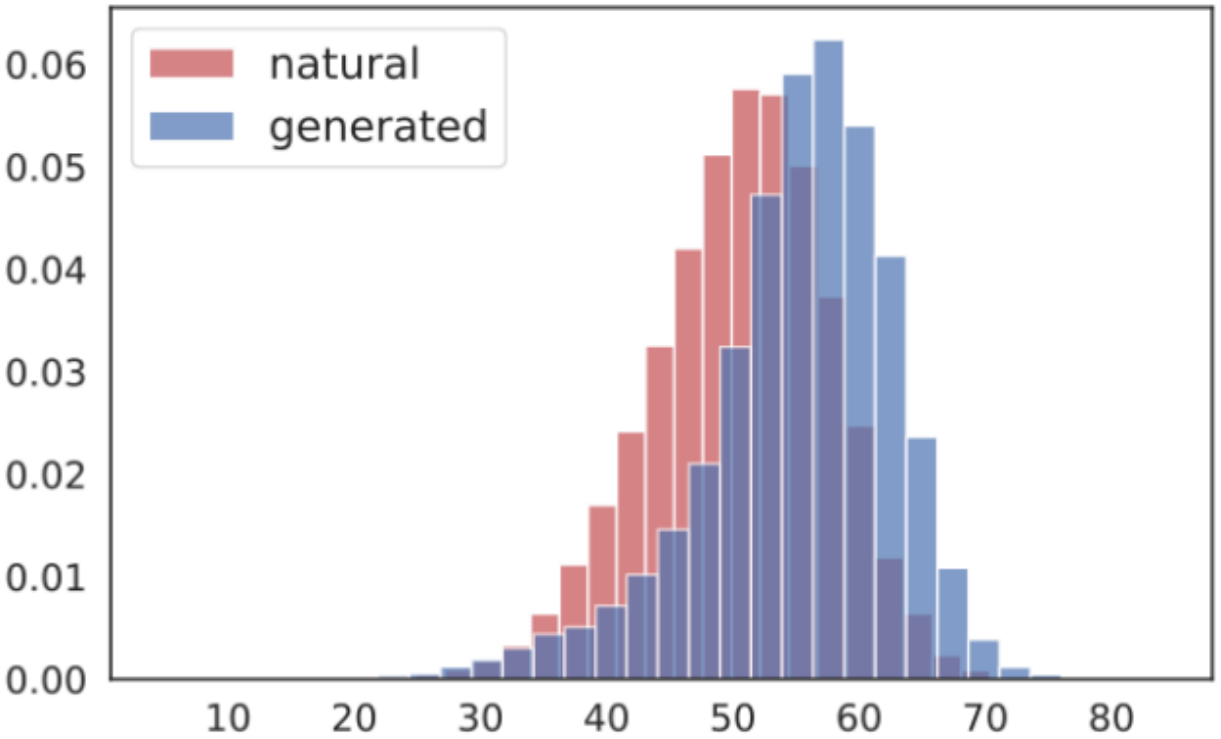
# Attention Head-wise Probing




(a) LENGTH (b) DEPTH

Figure 3: Heat maps of attention head-wise probing on LENGTH (Left) and DEPTH (Right) tasks. Attentions=Frozen attention weights. X-axis=Head index number. Y-axis=Layer index number. The brighter the color, the higher the accuracy score for the attention head.

# Structural Differences between Natural and Generated texts



# Outline

# Core Ideas to Borrow

1. **Multi-dimensional Topological Feature Extraction Framework**
   - In NLP's artificial text detection, three interpretable feature groups are extracted from attention maps: 0- and 1-dimensional barcodes (connectivity and loop counts), barcode summary statistics, and pattern-distance features comparing to canonical attention patterns.
   - Similarly, for HuBERT-based speech models one can compute 0-dimensional barcodes and summary statistics on each head's attention graph for downstream tasks.
2. **Head-wise Functional Visualization**
   - In NLP, TDA features allow probing at the attention-head level, revealing which heads best capture syntax, semantics, or particular pattern structures.
   - In speech, analogous analyses highlight heads that—without supervision—separate "voice vs voiceless", speaker identity, or emotion, suggesting specialization in silence/pause patterns or spectral structures.

# Main Challenges in Extending to Speech

1. **Sequence Length and Temporal Resolution** Speech transformers often process thousands of frames, making persistence computations on full attention graphs computationally intractable without downsampling or sliding-window segmentation.
2. **Noise and Acoustic Variability** Environmental noise, speaker timbre, and channel effects introduce spurious edges in thresholded attention graphs; robustification—e.g. fusing TDA with MFCC/PLP features—is required to mitigate sensitivity.
3. **Hierarchical Semantics vs. Pure Temporal Structure** Text has clear hierarchies (word→phrase→sentence); speech adds layers of phonemes, syllables, words, and prosody. Multiscale extraction (frame→phoneme→word) is necessary to capture the full hierarchical structure.

# Thanks for Your Listening !
# &
# Question and Answer.