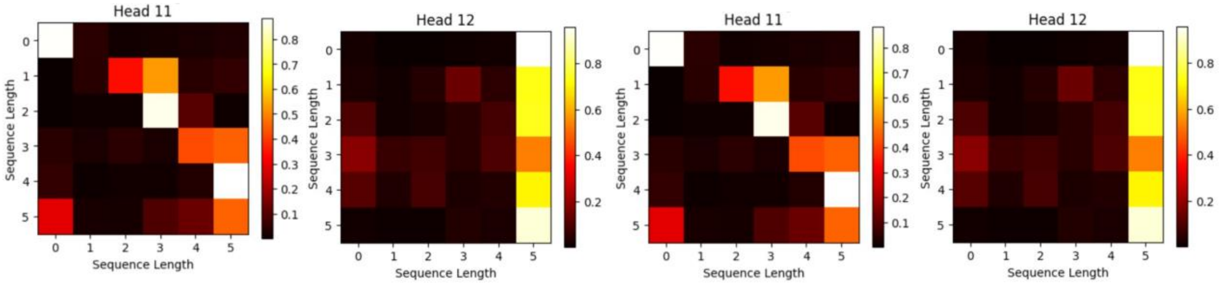
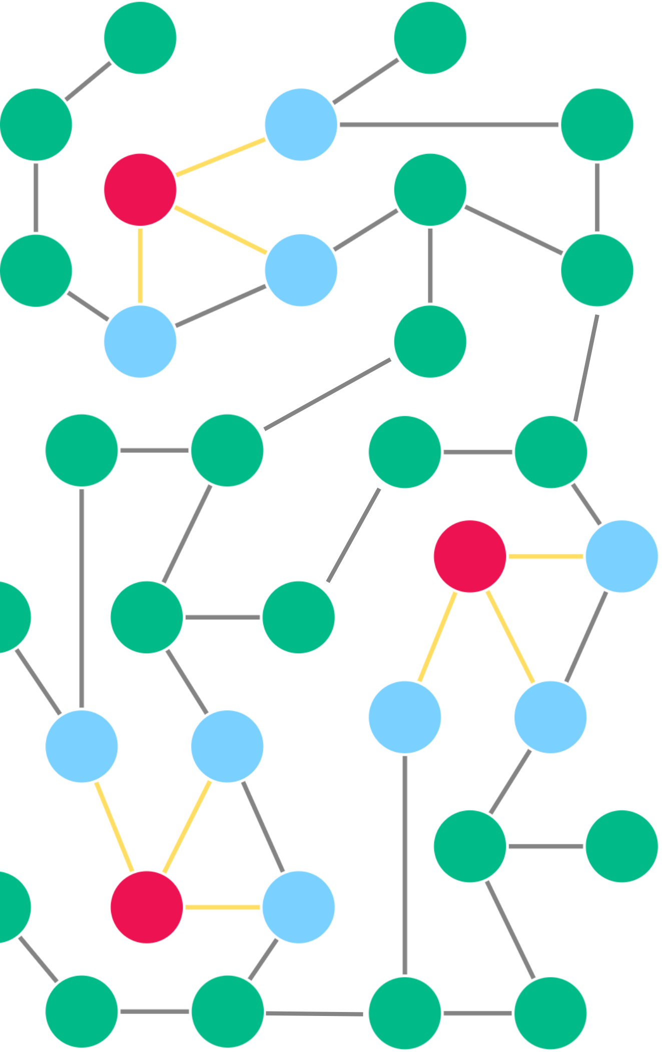


Course Project
27/05/2024

Team Project on the course “Selected Topics in Data Science”

Exploring Directed vs. Undirected Graph-Based Topological Data Analysis of Transformer Attention Maps

Kamil Garifullin

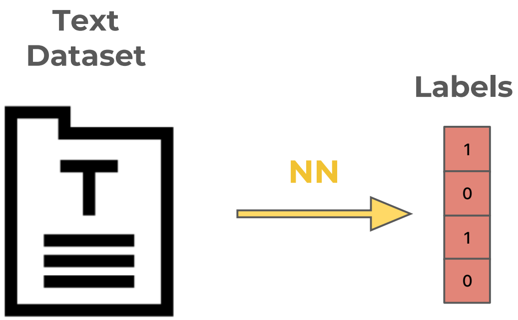


Introduction

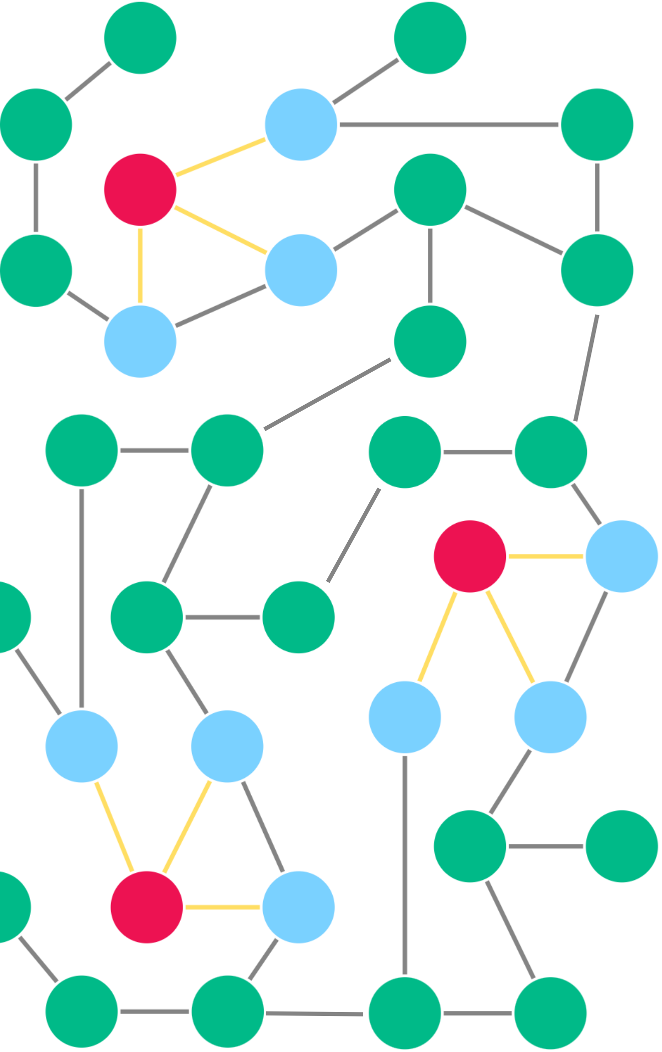
Motivation for research

Motivation

The main goal of this project is to systematically **compare** the performance and efficiency of persistent homology-based feature extraction from attention maps represented as **directed** graphs with their **non-directed** counterparts.



This study will solve the problem of **binary classification** of movie reviews: positive or negative.

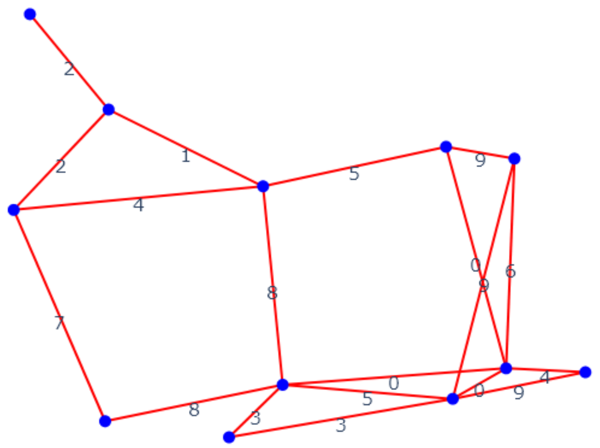


Methods

Topological Feature Extraction from Graphs

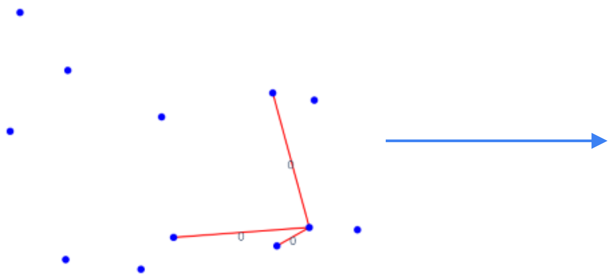
Topological Feature Extraction from **Undirected** Graphs

The process, known as **Vietoris-Rips** persistent homology, records the topological evolution of the graph across all edge weights.



Topological Feature Extraction from Undirected Graphs

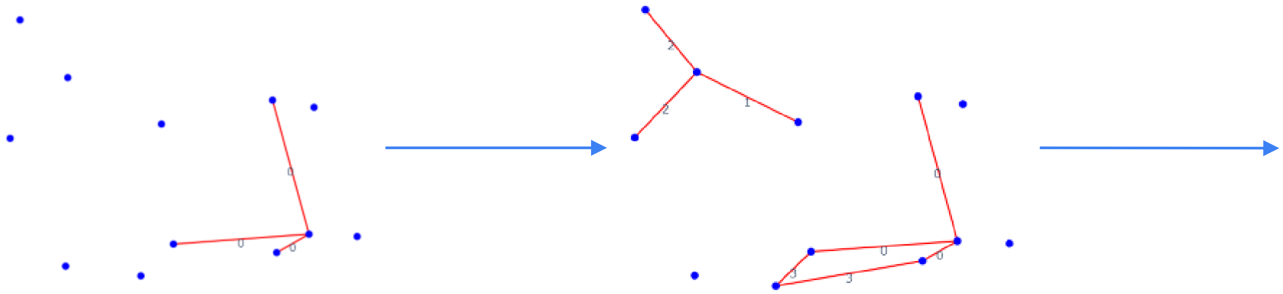
$\varepsilon = 0$



There are 9 connected components, and nothing much else.

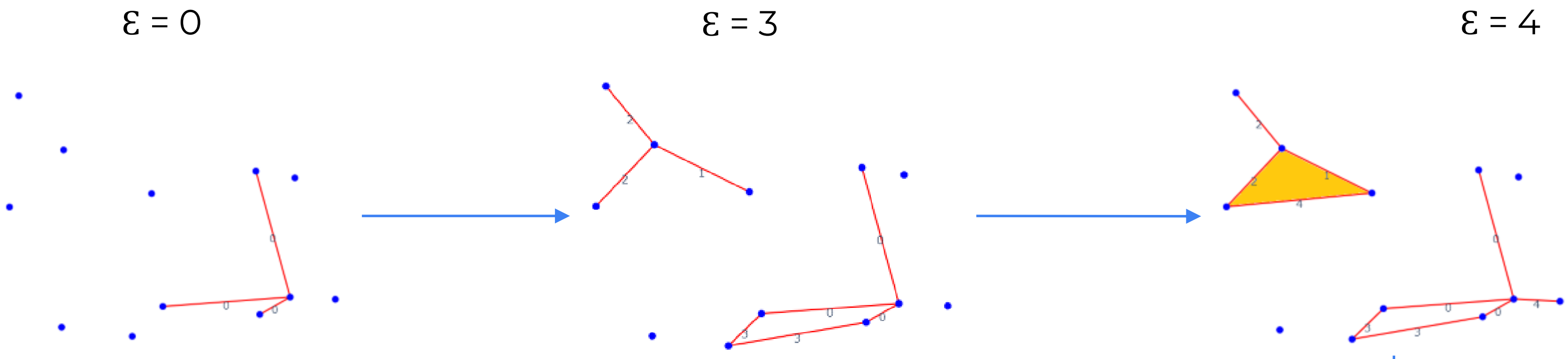
Topological Feature Extraction from Undirected Graphs

ε = 0 ε = 3



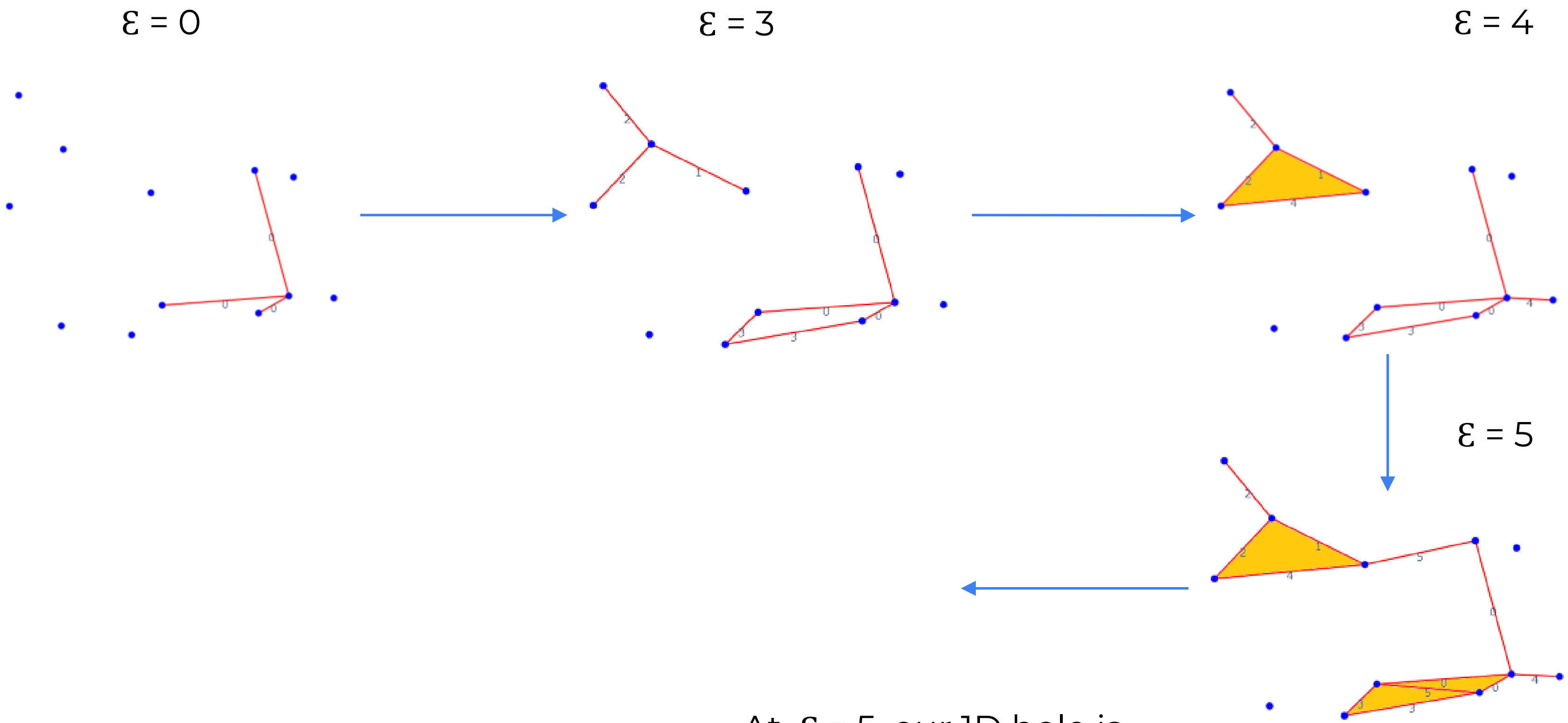
The newly arrived edges reduce the number of connected components further, but more interestingly they create a 1D hole!

Topological Feature Extraction from Undirected Graphs



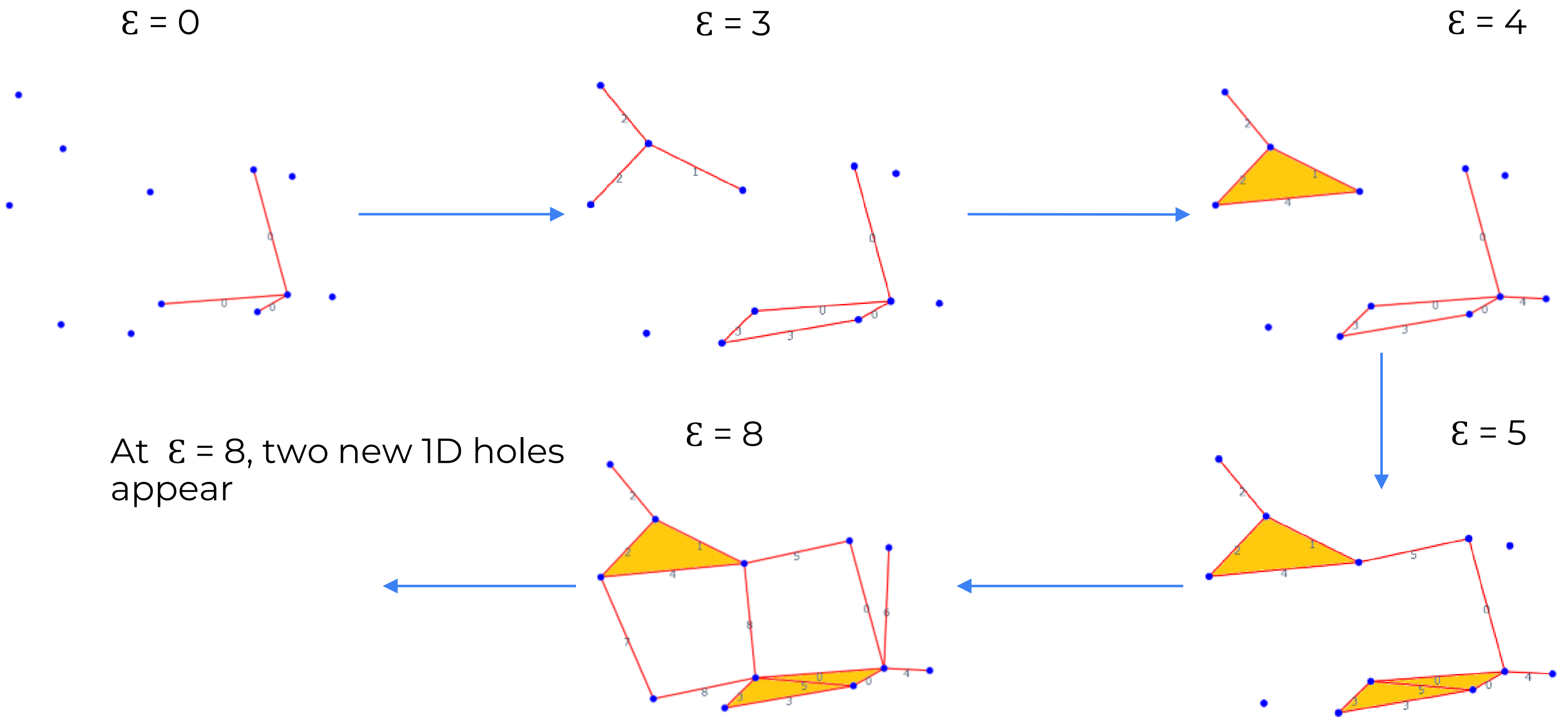
As an example of a “higher”-simplex, at ℇ = 4 we get our first triangle

Topological Feature Extraction from Undirected Graphs



At Ɛ = 5, our 1D hole is filled

Topological Feature Extraction from Undirected Graphs

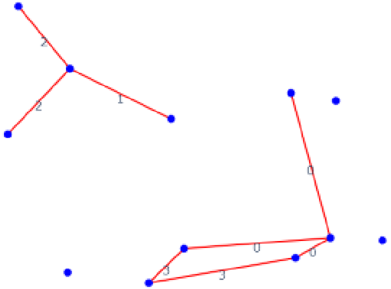


At ε = 8, two new 1D holes appear

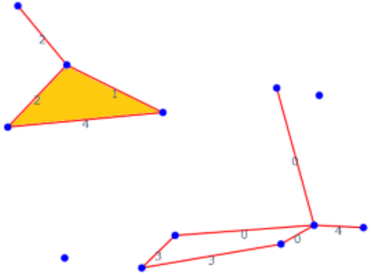
Topological Feature Extraction from Undirected Graphs

Finally, at Ɛ = 9, some more connected components merge, but no new voids are either created or destroyed

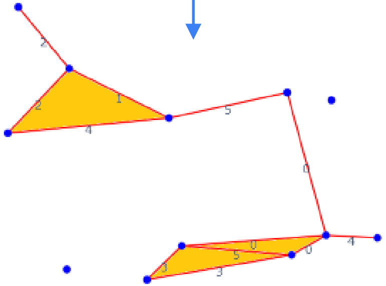
Ɛ = 3



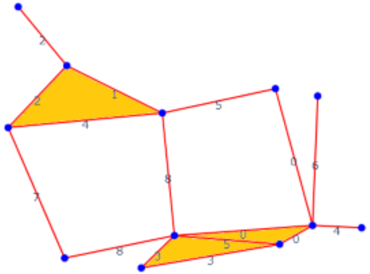
Ɛ = 4



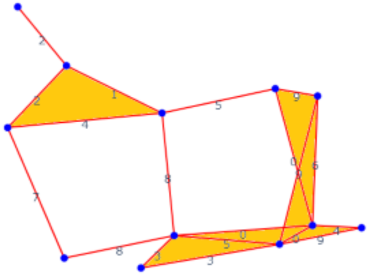
Ɛ = 5



Ɛ = 8



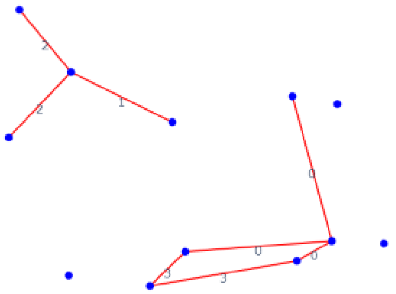
Ɛ = 9



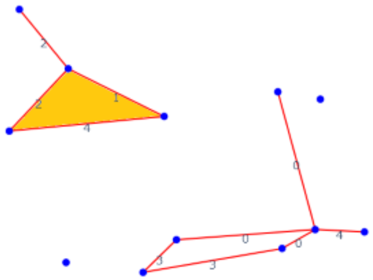
Topological Feature Extraction from Undirected Graphs

We can stop as we have reached the maximum value of ε, beyond which nothing will change: there is only one connected component left, but there are also two 1D holes which will never get filled.

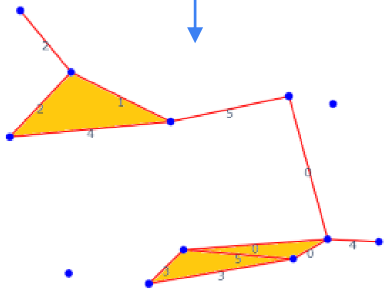
ε = 3



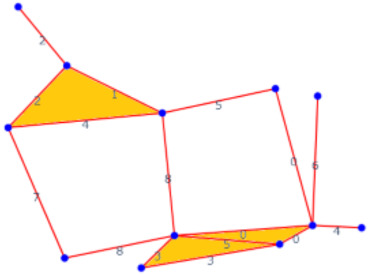
ε = 4



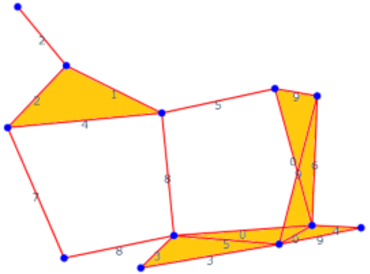
ε = 5



ε = 8

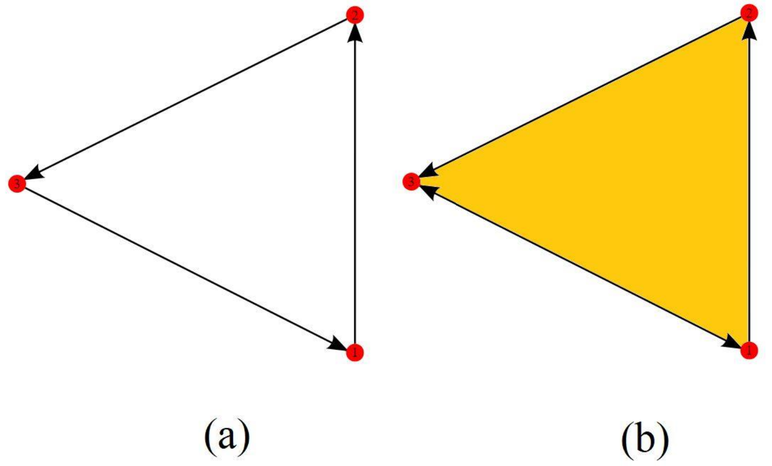


ε = 9



Topological Feature Extraction from Directed Graphs

The ideas and constructions underlying the algorithm in this case are very similar to the ones described above for the undirected case. Again, we threshold the graph and its directed edges according to an ever-increasing parameter and the edge weights.

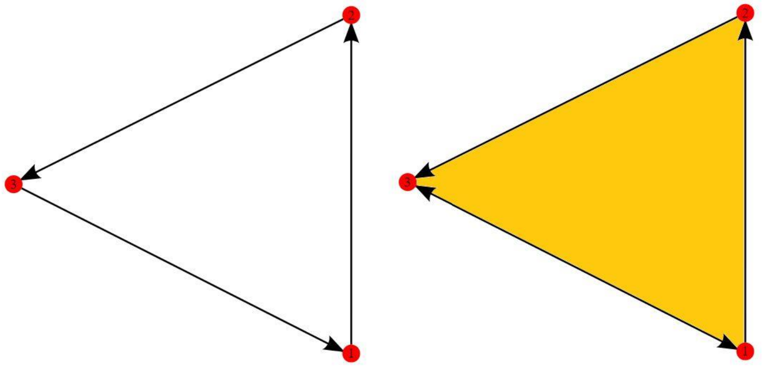


(1, 2), (2, 3) and (3,1) **form a 1D hole**

(1, 2), (2, 3) and (1, 3) form the boundary of (1, 2, 3) – **not a 1D hole**

Topological Feature Extraction from Directed Graphs

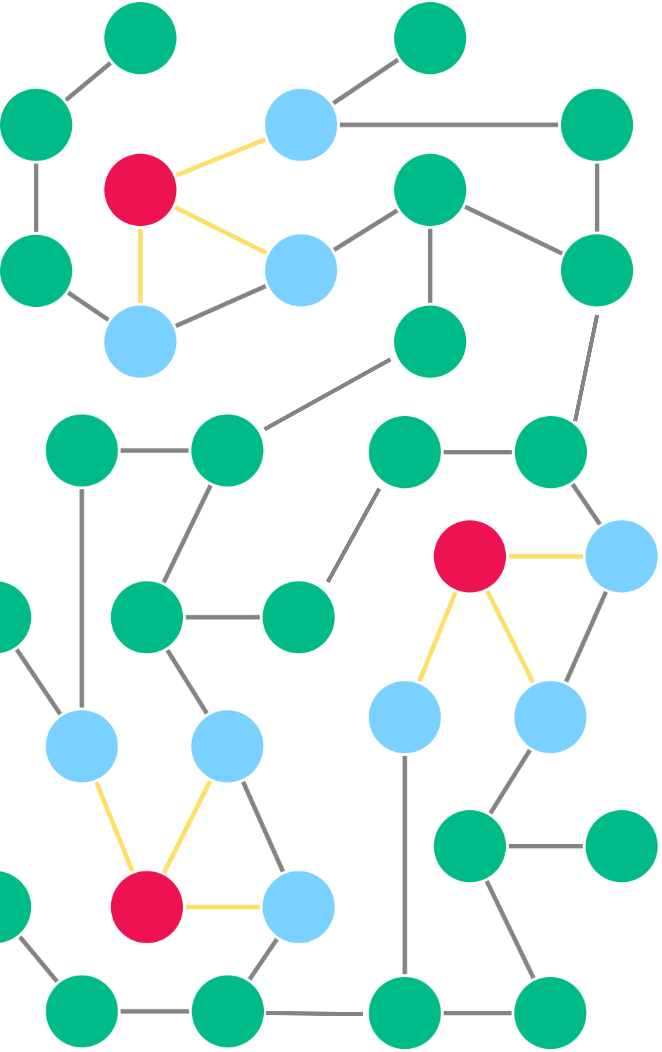
But! Because the graphs are directed, there are interesting consequences that distinguish these methods from methods for extracting features from undirected graphs.



(a) (b)

(1, 2), (2, 3) and (3,1) **form a 1D hole**

(1, 2), (2, 3) and (1, 3) form the boundary of (1, 2, 3) – **not a 1D hole**

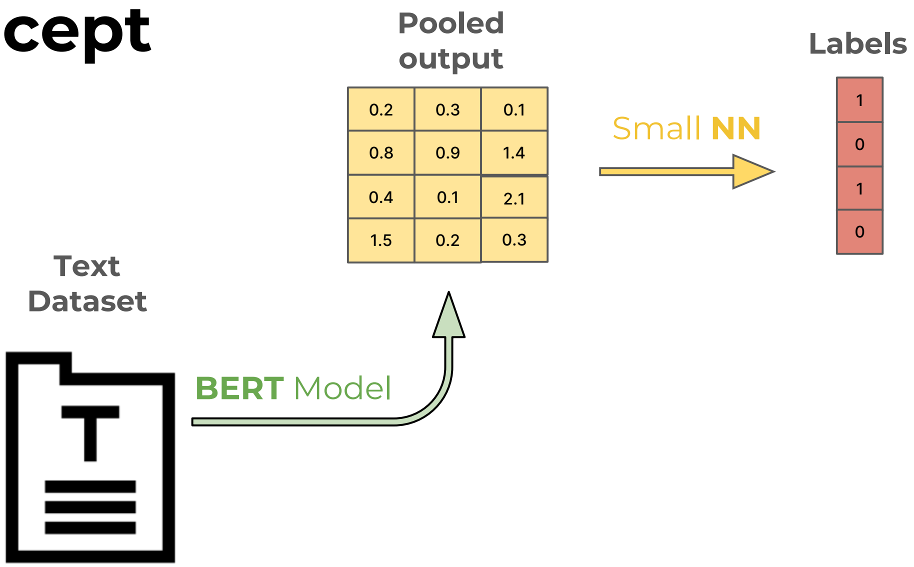


Methods

general idea

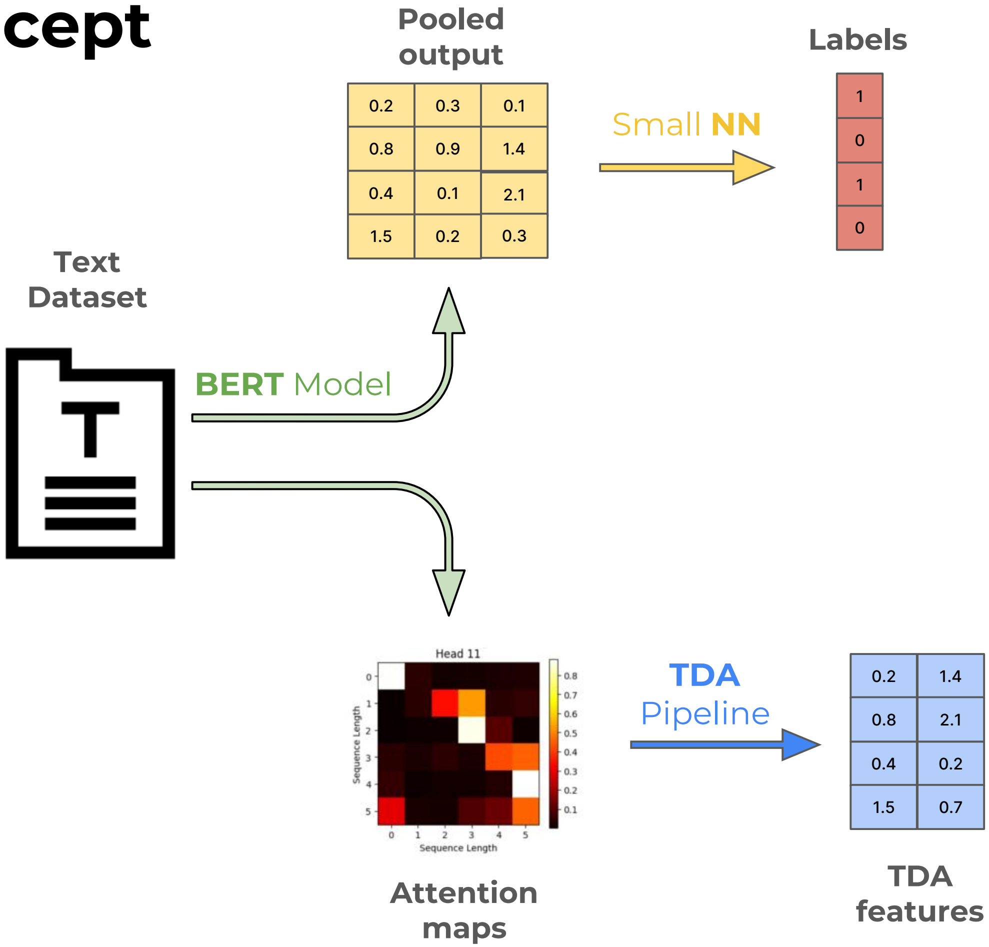
idea behind our method and pipeline of the work

Concept



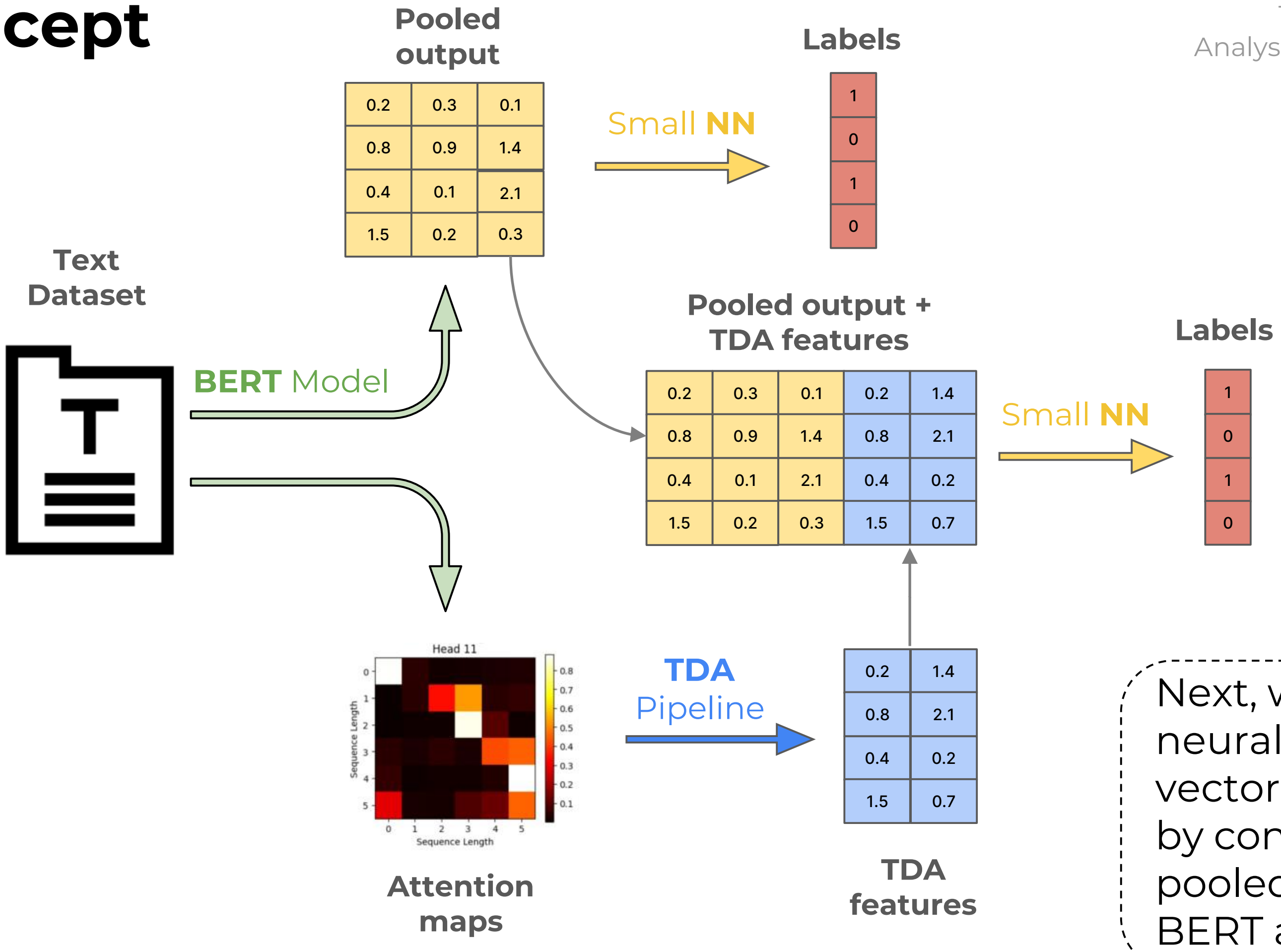
For the **baseline** in this work, a combination of **BERT-base-uncased** and one linear layer was used.

Concept



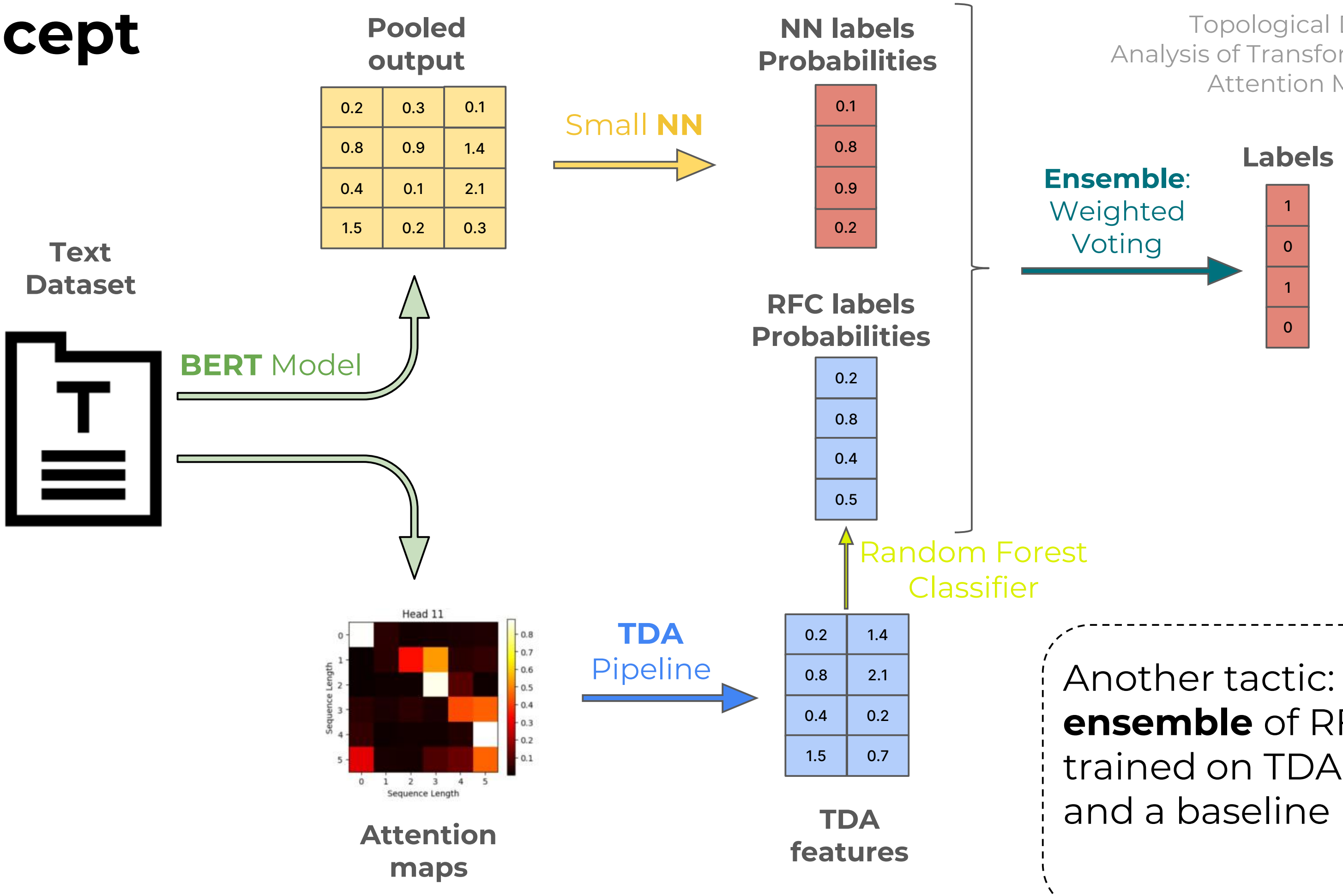
We can obtain **TDA-features** by considering attention maps as undirected or directed graphs

Concept

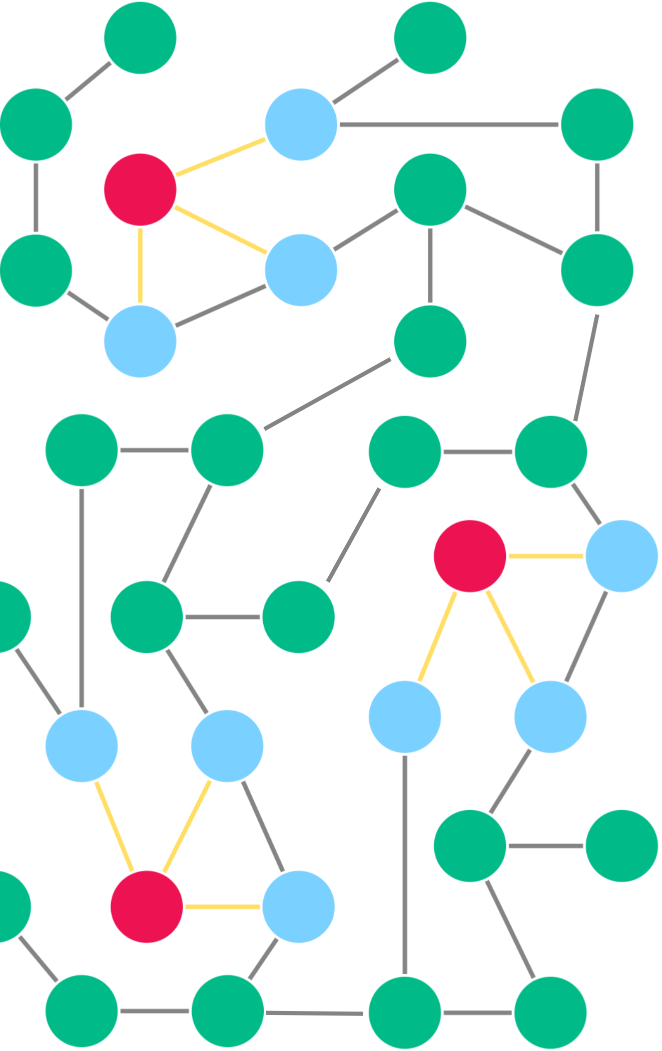


Next, we trained linear neural network on vectors obtained by concatenating pooled outputs from BERT and TDA-features

Concept



Another tactic: to use an **ensemble** of RFC model trained on TDA-features and a baseline



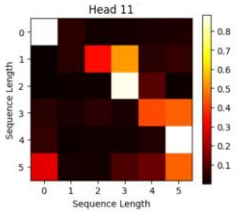
Methods

feature calculation pipeline

Concept

Text Dataset

BERT Model



Attention map

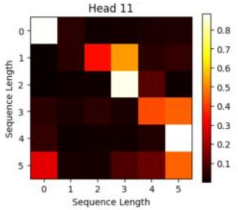
After obtaining the attention map, we can consider it as a directed or undirected graph.

Concept

Undirected graphs:
VietorisRips
Persistence

**Text
Dataset**

BERT Model



Attention map

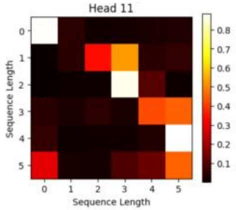
Directed graphs:
Flagser
Persistence

We can extract topological features from ***directed*** graphs via the **Flagser Persistence** and from ***undirected*** graphs via the **Vietoris Persistence**

Concept

Text Dataset

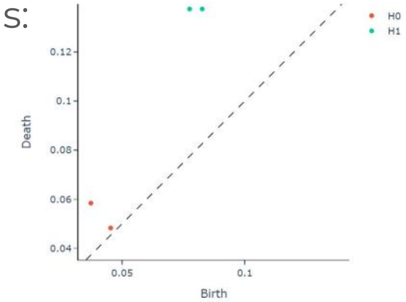
BERT Model



Attention map

Undirected graphs: VietorisRips Persistence

Pers. diagram



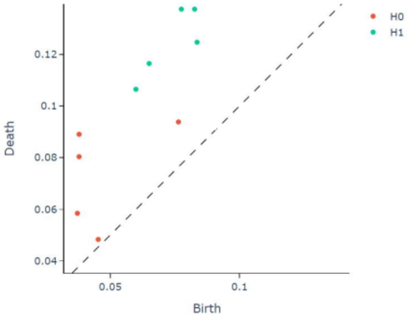
features calculation

0.2	1.4
0.8	2.1
0.4	0.2
1.5	0.7

TDA features from undirected graphs

Directed graphs: Flagser Persistence

Pers. diagram



features calculation

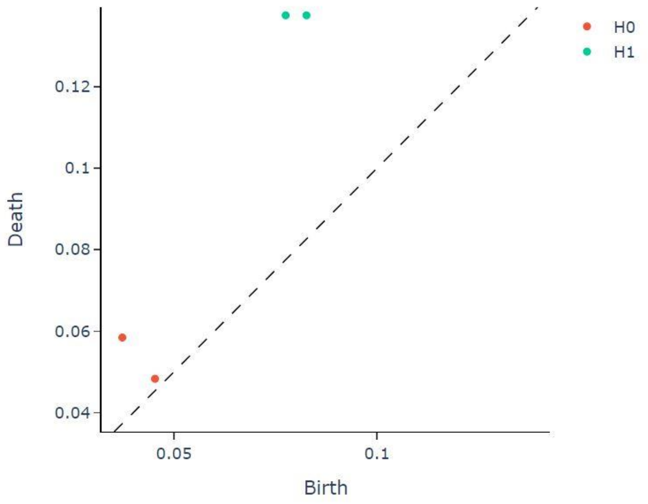
1.1	0.9
0.2	0.6
0.8	0.1
1.2	0.8

TDA features from directed graphs

After that we obtain Persistence Diagrams and calculate TDA features

How to get features from diagrams?

Pers. diagram



H0 Homology group

H1 Homology group

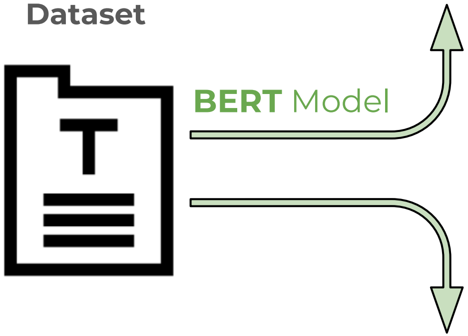
- Mean length of bars
- Birth time of the longest bar
- Death time of the longest bar
- Total aggregate length of bars
- Bars lengths variance
- Overall count of bars present
- Barcode's entropy assessment

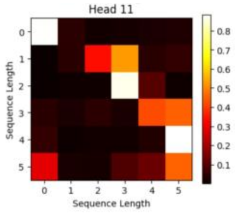
14 features from 1 pers. diagram

Concept

Text Dataset

BERT Model





Attention map from last layer and last head

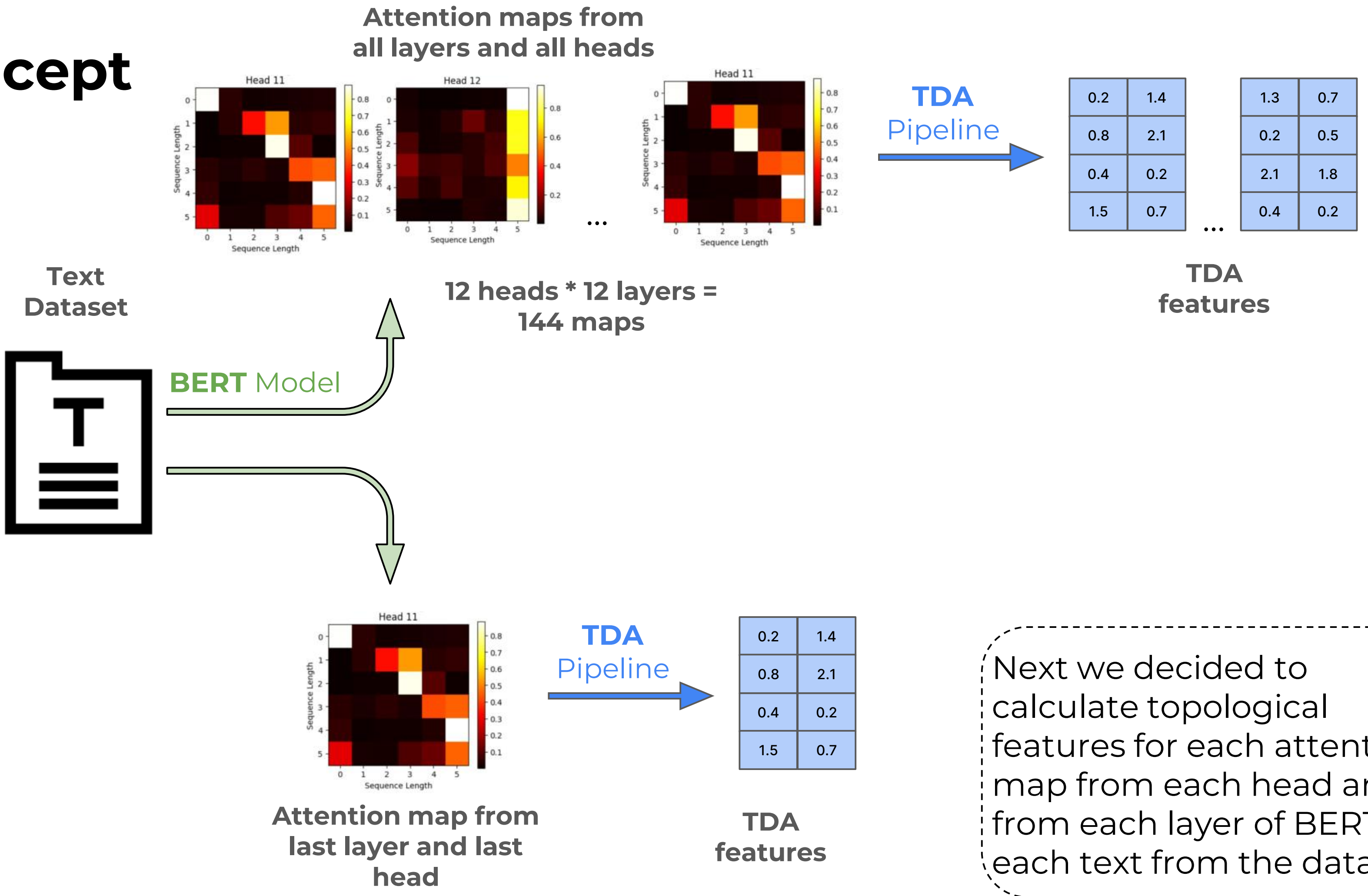
TDA Pipeline

0.2	1.4
0.8	2.1
0.4	0.2
1.5	0.7

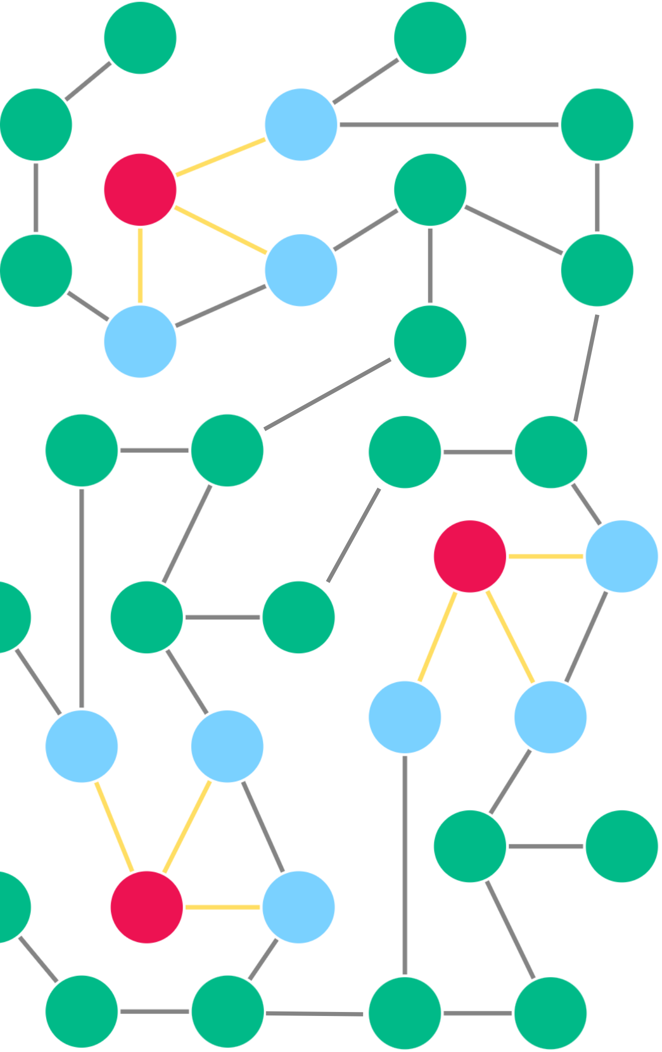
TDA features

First, we extracted features only for the 12th attention map of the last layer of BERT for each text from the dataset.

Concept



Next we decided to calculate topological features for each attention map from each head and from each layer of BERT for each text from the dataset.



Dataset

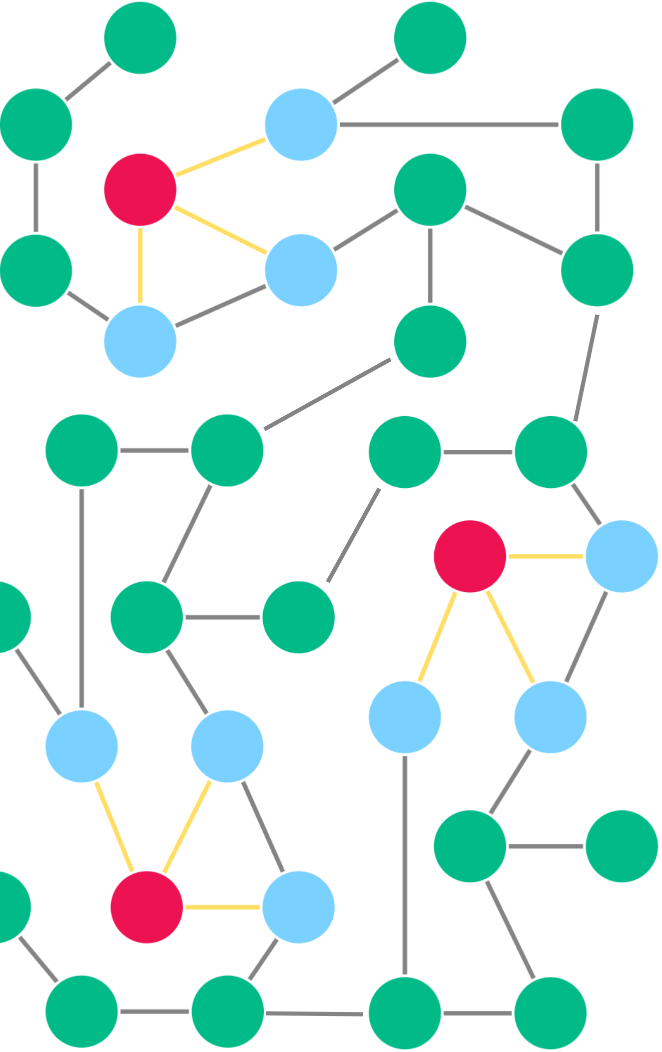
description of the data used in the following research

Datasets

Characteristic of the used data for training

IMDB Movie Reviews

- 50K movie reviews for binary sentiment classification
- Train: 25,000 movie reviews
- Test: 25,000



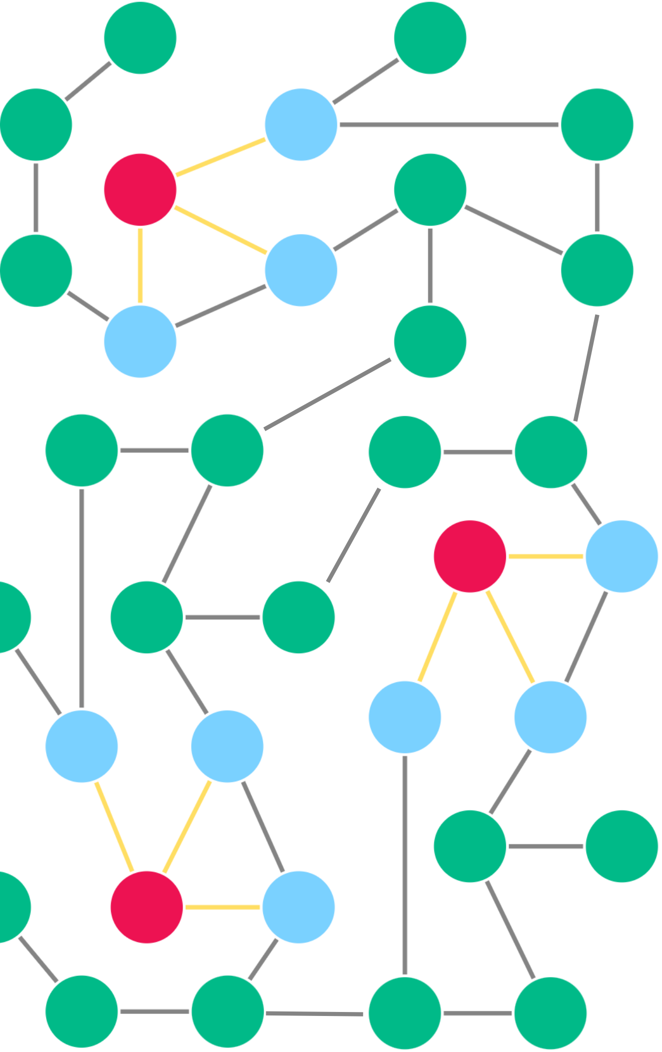
Results

Feature extraction from the **last layer and last head of attention**

Main Results

trainset size	En-BERT + linear layer(baseline)	Baseline + undirected TDA	Baseline + undirected TDA
Full dataset (25000 texts)	**73.9%**	73.4%	73.0%

The features obtained for the **last attention layer and the last head** turned out to be unrepresentative for both the case of directed graphs and the case of undirected graphs.
The accuracy of the **baseline** solution turned out to be **0.5-1% higher** compared to models trained on vectors that were formed by concatenating pooled outputs from BERT and TDA-features



Results

Feature extraction from the **all layers and all heads of attention**

Main Results

trainset size	En-BERT + linear layer(baseline)	Baseline + undirected TDA	Baseline + undirected TDA
A half of the dataset (12500 texts)	68.2%	**70.0%**	-

Due to the fact that feature extraction from attention maps obtained from all heads and layers is very computationally expensive, we used half of the texts from the dataset to train and validate the model

Main Results

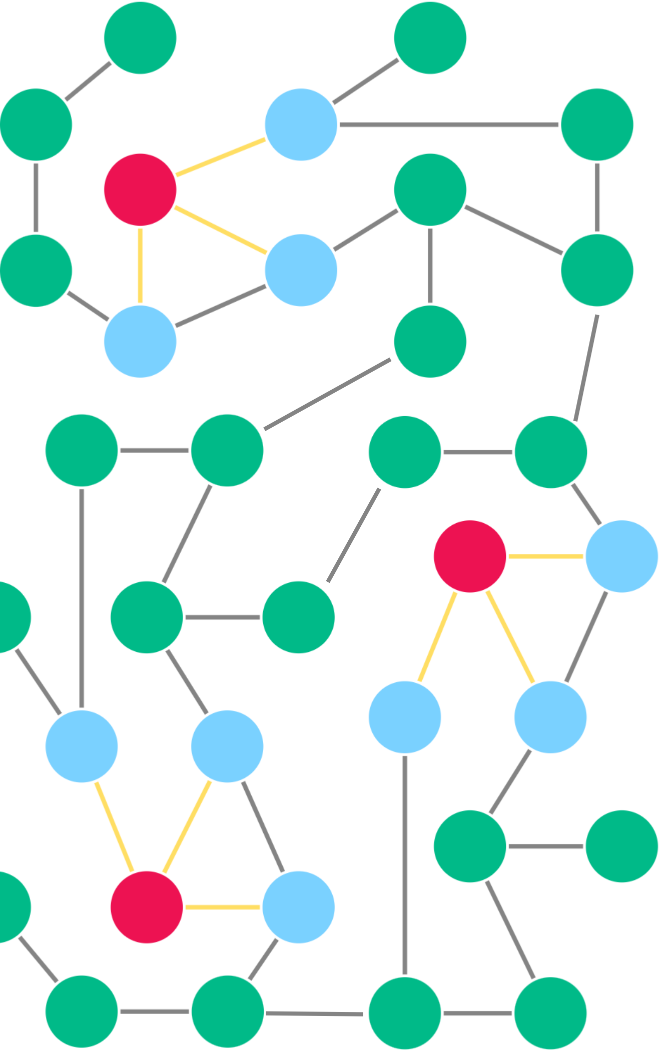
trainset size	En-BERT + linear layer(baseline)	Baseline + undirected TDA	Baseline + undirected TDA
A half of the dataset (12500 texts)	68.2%	**70.0%**	-

FlagserPersistence, which we used to calculate bar codes, is poorly parallelized and requires a lot of calculation time.
So, we could not calculate TDA-features, because even on half of the dataset it would take a huge amount of time and resources that we did not have.

Main Results

trainset size	En-BERT + linear layer(baseline)	Baseline + undirected TDA	Baseline + undirected TDA
A half of the dataset (12500 texts)	68.2%	**70.0%**	-

The features obtained **from all layers and heads of attention**, if we consider attention maps as undirected graphs, turned out to be representative and the model, which is **an ensemble of RFC** (random forest classifier) trained on these TDA-features and the baseline model, turned out to be **2% higher** than the baseline.



Conclusion

Main Results

Using of TDA features obtained by considering attention maps as **undirected** graphs **increased** the prediction accuracy by **2%**

Feature extraction from the **last layer and all last head of attention**

trainset size	En-BERT + linear layer(baseline)	Baseline + undirected TDA	Baseline + undirected TDA
Full dataset (25000 texts)	73.9%	73.4%	73.0%

Feature extraction from the **all layers and all heads of attention**

trainset size	En-BERT + linear layer(baseline)	Baseline + undirected TDA	Baseline + undirected TDA
A half of the dataset (12500 texts)	68.2%	70.0%	-

Questions?

Kamil Garifullin
Kamil.Garifullin@skoltech.ru
Data Science, Ms-1

References:

[1] **Acceptability judgements via examining the topology of attention maps.** Cherniavskii, D., Tulchinskii, E., Mikhailov, V., Proskurina, I., Kushnareva, L., Artemova, E., ... & Burnaev, E. (2022). arXiv preprint arXiv:2205.09630.

[2] **Topological data analysis for speech processing.** Tulchinskii, E., Kuznetsov, K., Kushnareva, L., Cherniavskii, D., Barannikov, S., Piontkovskaya, I., ... & Burnaev, E. (2022).. arXiv preprint arXiv:2211.17223.

[3] **Artificial text detection via examining the topology of attention maps.** Laida Kushnareva, Daniil Cherniavskii, Vladislav Mikhailov, Ekaterina Artemova, Serguei Barannikov, Alexander Bernstein, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2021

[4] https://giotto-ai.github.io/gtda-docs/latest/notebooks/persistent_homology_graphs.html