## Topic

- Topological Analysis of Text Embeddings

## Motivation

- Text embeddings (e.g., from Word2Vec, GloVe, BERT, or GPT models) encode semantic information in high-dimensional vector spaces. However, interpreting these embeddings is challenging due to their dimensionality. Topological Data Analysis (TDA) provides tools (e.g., persistent homology) to extract meaningful structures (e.g., clusters, loops, hierarchies) from high-dimensional data.
    - Apply TDA to text embeddings to uncover latent semantic structures.
    - Compare topological features across embedding models (e.g., static vs. contextual embeddings).
    - Explore applications in text classification, topic modeling, or bias detection.

## Methodology

- Data & Embeddings
    - Datasets: AG News, 20 Newsgroups, or custom text corpora.
    - Embedding Models:
        - Static: Word2Vec, GloVe
        - Contextual: BERT, RoBERTa, GPT
- Topological Tools
    - Persistent Homology (using giotto-tda, ripser) to identify clusters and holes.
    - Mapper Algorithm (using Kepler-Mapper) to visualize high-dimensional structures.
    - Dimensionality Reduction: UMAP/t-SNE for visualization.
- Evaluation
    - Compare topological summaries across models.
    - Assess whether topological features correlate with semantic meaning.