Review of the project "Exploring Directed vs. Undirected Graph-Based Topological Data Analysis of Transformer Attention Maps"

The project explores the application of directed graph analysis to transformer attention maps, comparing it with traditional undirected graph methods. It focuses on using persistent homology to evaluate the impact of topological features of attention maps on downstream classification tasks.

The problem statement is clearly articulated, identifying the gap in current research regarding the analysis of transformer attention maps as undirected graphs, which overlooks the directional nature of the information flow within these models. The authors also effectively highlight the importance of this study by introducing the concept of directed graph analysis and its potential advantages. **The main idea** is thoroughly described, focusing on comparing persistent homology-based feature extraction from attention maps represented as directed versus undirected graphs. There are thorough **comparisons with relevant methods.** It reviews significant studies on topological data analysis (TDA) in NLP and for analyzing attention heads in linguistic acceptability, speech processing, and artificial text detection. It helps to establish the context for the current research and highlight its novelty.

Each part of the report is described clearly, with detailed explanations provided for the methodology, experimental setup, and results. The necessary information for understanding the study is largely present, though it could benefit from additional visual aids and details in the results part. The experiment protocol appears reasonable and well-structured. The report details the selection of the IMDb Movie Reviews dataset for binary sentiment classification and outlines the steps for feature extraction and comparison. The intermediate results are clearly reported, showing a comparative analysis between the baseline (BERT with a linear layer), undirected TDA, and directed TDA approaches. The results are reasonable and indicate that directed graphs provide valuable insights into the flow of information within transformer models.

The report is well-formatted and written with high quality. However, minor improvements in consistency, proofreading, and the addition of the results, conclusion and discussion would enhance its readability. In a discussion section it would be nice to mention the interpretation of the results in a broader research context. But I belief that it will be added in the next version.

The provided code runs without errors and implements the methodologies described in the report. It allows for reproducibility and further experimentation but would be better to update README with more details.

The problem statement could benefit from a brief mention of the specific impacts this gap in research might have on practical NLP applications, thereby emphasizing the significance of

addressing it. The illustration of the experimental pipeline and the process of feature extraction using TDA (like diagram) would enhance the clarity of the methodological description. The results section could be strengthened by including more statistical analysis, such as confidence intervals or significance testing, to provide a deeper understanding of the performance differences observed. Consider adding more diverse features beyond TDA to enhance the model's performance further.

Overall, the project report is a well-executed study that makes a deep exploring directed graph-based TDA of transformer attention maps. With minor enhancements in clarity, completeness, and presentation, the report could serve as an excellent reference for future research in this domain.