Topological-Based Embeddings

Mohamed Tliouant
McGill University
mohamed.tliouant@mail.mcgill.ca

Noah Casarotto-Dinning
McGill University
GroupLabs / Research & Development
noah.casarotto@mail.mcgill.ca

Abstract

Recent advancements in natural language processing (NLP) have increasingly focused on the geometric and topological properties of embedding spaces. Despite its potential, the utilization of these intrinsic topological structures to enhance semantic understanding remains largely underexplored. This paper proposes a novel methodology to harness the topological characteristics inherent in the high-dimensional space of BERT embeddings. By exploring the manifold on which these embeddings lie, we aim to identify and exploit the geometric relationships that underpin similar semantic constructs. Our approach reconstructs BERT embeddings by integrating their topological metadata, specifically by quantifying and encoding the similarity of shapes that represent analogous sentences. We hypothesize that this enhanced embedding will offer superior performance in sentiment classification tasks compared to traditional embeddings. This study not only contributes to the theoretical understanding of embedding spaces but also provides a practical framework for improving the accuracy of sentiment analysis algorithms through advanced geometric modeling.

1 Introduction

The field of Natural Language Processing (NLP) has experienced a paradigm shift with the advent of pretrained models. These models, which are trained on extensive and varied corpora, have revolutionized the NLP landscape by establishing new benchmarks across a range of tasks including language translation, sentiment analysis, and more. This profound transformation is driven by sophisticated deep learning architectures and optimization techniques, which have dramatically expanded the capabilities of machines in understanding and processing human language. (Young et al., 2018)

Among the key advancements, the BERT (Bidirectional Encoder Representations from Transformers) model stands out for its novel approach to contextually aware language understanding. Unlike its predecessors, BERT utilizes a mechanism of bidirectional training, where the masked language modeling technique randomly masks words in a sentence and predicts them based on the context provided by the remaining unmasked words. This is complemented by the self-attention mechanism, which assesses the importance of each word in a sentence relative to every other word, thus allowing the model to capture nuances in language from both directions simultaneously.(Vaswani et al., 2017) This is a significant technical leap over earlier models like ELMo, which, although innovative in generating context-sensitive embeddings, operated under constraints of unidirectional processing and sequential computation—factors that limited their efficiency, scalability, and ability to parallelize processing.(Peters et al., 2018)

Despite these advancements, the development of more sophisticated sentence classification systems remains an ongoing quest. Current models built upon the BERT architecture still face challenges in adequately capturing the relational semantics between concepts—a critical requirement for advanced semantic tasks such as document retrieval and question answering (Rajpurkar et al., 2018). This limitation largely stems from the reliance on dense vector embeddings, which, while proficient in encapsulating contextual information at the word and phrase level, often do not adequately represent the complex, multi-dimensional relationships inherent in natural language (Reimers and Gurevych, 2019). This is a pivotal concern as the ability to model these relationships directly impacts the effectiveness of NLP systems in tasks requiring deep semantic comprehension.

In response to these challenges, we propose a pioneering architecture designed to bridge these gaps in sentence classification. Our approach begins by transforming BERT's sentence embeddings into a graph-structured format, leveraging the inherent

relational data within the text. This step is facilitated by the Topological Graph Neural Network (TOGL), a novel component that infuses the embeddings with topological properties, thus enriching their semantic representation.

By analyzing and utilizing the topological and geometric properties captured in these graph embeddings, our model aims to provide a more comprehensive and nuanced understanding of text. This enriched data is then processed by a modified BERT architecture, TopoBERT. TopoBERT is designed to optimize the classification of sentences by leveraging the enhanced embeddings and attention matrixes of BERT thereby achieving greater accuracy and a deeper semantic understanding than conventional methods.

2 Related Works

To clarify and expand on the concept of attention as presented in the seminal paper "Attention is All You Need" by(Vaswani et al., 2017) Vaswani et al., it's important to examine the architecture introduced: the Transformer. This model has notably revolutionized natural language processing by replacing recurrent layers with an architecture relying entirely on attention mechanisms. These mechanisms enable the model to assign varying levels of importance to different segments of the input data, utilizing a function of queries, keys, and values derived from the input. For each input token, a corresponding query vector (Q), key vector (K), and value vector (V) are generated via linear transformations. An attention score between two tokens is then determined by the dot product of the query vector from one token and the key vector of another, followed by a softmax operation to normalize these scores across all tokens. This normalization helps stabilize the distribution of values within the dimension of the key vectors.

This attention mechanism effectively identifies dependencies across tokens by focusing on areas with high attention scores, but it processes relationships in a pairwise and relatively simplistic manner. While adept at recognizing direct interactions, this strategy may fail to capture more intricate structures such as hierarchical relationships, cycles, or clusters that are present within the data. These complex structures represent higher-order interactions that are essential for fully understanding the semantics of text or the relationships in structured data like molecules. The conventional attention mechanism essentially forms a direct, weighted edge between nodes (tokens) based on significant attention scores, resulting in a fully connected directed graph. However, this model does not inherently recognize or depict potential topological structures or patterns that extend beyond straightforward pairwise interactions.

Among the key advancements, the BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) model stands out for its novel approach to contextually aware language understanding. Unlike its predecessors, BERT utilizes a mechanism of bidirectional training, where the masked language modeling technique randomly masks words in a sentence and predicts them based on the context provided by the remaining unmasked words. This is complemented by the self-attention mechanism, which assesses the importance of each word in a sentence relative to every other word, thus allowing the model to capture nuances in language from both directions simultaneously. This represents a significant technical leap over earlier models like ELMo, which, although innovative in generating context-sensitive embeddings, operated under constraints of unidirectional processing and sequential computation—factors that limited their efficiency, scalability, and ability to parallelize processing.

BERT's architecture, which allows it to learn contextual relations between words (or sub-words) in a text, not only in one direction but within the entire context of a given token, both from the left and the right, is revolutionary. This bidirectionality allows BERT to be pre-trained on large amounts of text and then fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as sentiment analysis, without substantial modifications to the model architecture for specific tasks. This method leverages deeper context to improve the state of the art in NLP applications significantly.

Our research leverages advances in Topological Data Analysis (TDA) to enhance natural language processing (NLP) models by integrating a topological layer into graph neural networks. Specifically, we use the Topological Graph Neural Network (TOGL) (Horn et al., 2022) developed by Bastian Rieck and colleagues. TOGL introduces a transformative topological layer to the conventional graph neural network (GNN) framework, computing complex properties such as Betti numbers and

Vietoris-Rips complexes. This approach has been shown to improve molecule classification by capturing subtle topological features not discerned by traditional GNNs (Perez and Reinauer, 2022).

Building upon the foundational work in TDA applied to NLP, including the pioneering study by Zhu in 2013 that used persistent homology to differentiate between child and teenager writing (Zhu, 2013), and further applications like genre prediction from movie plots (Doshi and Zadrozny, 2018), our approach aims to refine this integration. We extend this exploration into the application of topological methods to enhance the representational capabilities of sentence embeddings in NLP, a field that has seen significant contributions from various studies (Savle et al., 2019; Kushnareva et al., 2021).

Our discussion progresses to the Topological BERT (TopoBERT) (Perez and Reinauer, 2022), which integrates topological data analysis into the processing of BERT's attention matrices. Unlike the standard BERT model that uses attention matrices directly, TopoBERT transforms these into attention graphs, applying graph filtrations based on attention weights. This methodology enables the capture of multi-scale topological features, such as loops and voids, across these filtrations, tracked using persistent homology (Carrière et al., 2020).

Despite the innovative application of these topological insights, a significant limitation exists in how embeddings generated from BERT are not retrained to optimize for topological properties. Our research addresses this by retraining embeddings specifically to enhance their topological features, thereby improving the detection of nuanced text semantics in NLP tasks such as sentiment analysis. This ensures our system not only incorporates but also maximally utilizes topological aspects, advancing the state of the art in this field.

Moreover, by retraining these embeddings, our approach echoes recent developments in neural networks with topological layers, such as PersLay (Carrière et al., 2020) and Persformer (Reinauer et al., 2021), which have shown excellent performance in analyzing real-life graph datasets through a differentiated overall objective function. This integration promises to bolster the accuracy and robustness of NLP models, drawing from successful precedents in the scientific community that have enhanced linguistic models through topological analysis (Cherniavskii et al., 2022). By employing these advanced methods, we position our research at the forefront of TDA applications in NLP, exploring new pathways for understanding and processing language.

3 Modeling

Modeling

In this research, we have implemented two distinct models: the Topological Graph Neural Network (TOGL) and a variant of the BERT architecture, which we refer to as TopoBERT. Our core hypothesis posits that enhancing sentence embeddings with topological features using TOGL before incorporating them into TopoBERT leads to more effective text classification compared to using TopoBERT alone. This hypothesis is grounded in the belief that the topological representations extracted from the data provide a richer semantic context, thereby improving the model's ability to discern and classify textual nuances more accurately.

The structure of our analysis is organized into two main subsections to clearly delineate the comparative study between baseline models and our proposed innovation. The first subsection focuses on the baseline model, which involves a finetuned implementation of TopoBERT, a model that enhances traditional BERT embeddings using topological data analysis. In contrast, the second subsection details our proposed model, which begins with the application of the TOGL architecture to modify sentence embeddings by integrating topological information. These modified embeddings are then finetuned within the TopoBERT framework, aiming to demonstrate the enhanced capability of our approach in handling complex text classification tasks. This organizational structure allows for a systematic evaluation and comparison of the baseline and proposed models under similar computational conditions and datasets.

The **baseline model** utilized in our study is TopoBERT, an adaptation of the well-known BERT (Bidirectional Encoder Representations from Transformers) model, which has been specifically enhanced to incorporate topological data analysis for natural language processing tasks. TopoBERT extends the traditional BERT model by not only leveraging the powerful contextual embeddings generated through BERT's attention mechanisms but also integrating a layer of topological analysis that uses persistent homology to capture complex patterns

within the data.

TopoBERT operates by first processing text through the standard BERT framework to generate attention matrices. These matrices, which capture the contextual relationships between words in a sentence, are then transformed into what are termed attention graphs. Persistent homology is applied to these graphs to identify and analyze data structures that persist across various spatial resolutions, thereby extracting meaningful topological features. These features are encapsulated into what are known as persistent images, a form that allows conventional neural networks to process topological information as part of the classification task.

Mathematically, the process involves calculating Betti numbers and using Vietoris-Rips complexes to create a filtration—a sequence of nested subspaces from which the persistence of various topological features can be derived. These features are then plotted in a persistence diagram, capturing the birth and death of features like holes that appear in the data as the threshold changes. This diagram is transformed into persistence images, which are stable and easy to manipulate within typical machine learning frameworks.

For the purpose of this study, both the TopoBERT and TOGL architectures were reimplemented from the ground up to ensure consistency in testing environments and comparability in results. This reimplementation involved the reconstruction of each model's architecture and the independent training and tuning of the models using the same datasets and computational resources. This approach ensures that any observed differences in performance can be confidently attributed to the models' intrinsic capabilities rather than external variables such as hardware or implementation optimizations.

The **proposed model** introduces an innovative integration of the Topological Graph Neural Network (TOGL) with the enhanced BERT model, into TopoBERT, to capitalize on the unique advantages provided by topological data analysis in text classification tasks. This integration aims to leverage the sophisticated graph augmentation capabilities of TOGL to refine the embeddings processed by TopoBERT, thereby enriching the model's ability to interpret and classify complex text data more accurately.

The TOGL layer is specifically designed to enhance the graph representations derived from sentence embeddings by embedding topological information into these structures. By applying topological data analysis, TOGL calculates Betti numbers and utilizes Vietoris-Rips complexes to augment the connectivity and relationships within the graph representations of text. These operations allow TOGL to capture and emphasize subtle topological features such as loops and holes within the data, which are often indicative of higher-level semantic relationships and structures that conventional models might overlook.

Technically, the TOGL layer first transforms the initial embeddings from BERT into a graph format where nodes represent individual sentences or tokens, and edges encapsulate the relationships based on the embeddings. It then applies a series of filtrations to these graphs, creating a multi-scale representation that tracks the evolution of topological features across different levels of granularity. Each filtration step essentially refines the graph by progressively incorporating edges based on their significance, determined by the learned topological data. This process not only preserves but also highlights essential structural information that can be critical for understanding complex linguistic constructs.

The output from TOGL, which consists of enhanced graph embeddings with embedded topological signatures, is then fed back into the TopoBERT architecture. These enriched embeddings are expected to provide a deeper semantic understanding, facilitating more nuanced and accurate text classifications. The assumption here is that the topological enhancements introduced by TOGL enable TopoBERT to better capture and process the intricate relationships inherent in natural language, leading to superior performance in tasks such as sentiment analysis, document classification, and others.

The integration of TOGL with TopoBERT presents a significant enhancement over the baseline TopoBERT model by harnessing the power of TDA to enrich the embeddings before they are processed for classification. The primary improvement stems from TOGL's ability to capture and incorporate multi-scale topological features into the embeddings, which are often overlooked by traditional neural network architectures. This enhancement allows the proposed model to recognize more complex patterns and relationships within the text, potentially leading to higher accuracy and a

more nuanced understanding of the semantic structures.

Specifically, TOGL's application of persistent homology to the sentence embeddings generates a richer representation by identifying and encoding topological invariants such as holes and loops, which correspond to persistent features across the text's manifold. These features provide crucial insights into the underlying data structure, offering a more detailed perspective on how ideas and themes are interconnected within the text. By feeding these topologically enhanced embeddings into TopoBERT, the model can leverage this additional layer of information to improve its predictive performance, recognizing previously overlooked features highlighted by TOGL, particularly in challenging NLP tasks that require a deep understanding of context and relational semantics. This approach not only bridges the gap between traditional text processing techniques and advanced topological analysis but also sets a new standard in how deeply a model can interpret text, thus enhancing its applicability to a broader range of linguistic tasks.

4 Datasets and Evaluation Metrics

4.1 Data

We employ several established datasets to benchmark the performance of our sentiment analysis and text classification models. Each dataset is chosen based on its relevance to the tasks we aim to address and its common use in the literature, which provides a basis for comparison with other studies.

- **CoLA (The Corpus of Linguistic Acceptability)**: Part of the GLUE benchmarks, this dataset consists of English sentences annotated as grammatically correct or incorrect. It is particularly suited for evaluating models' understanding of English grammar and sentence structure, which is central to our study's focus on linguistic acceptability. The dataset is divided into a training set, which we further split into 90% for training and 10% for validation, and an original validation set used as our test set. (Warstadt et al., 2019)

  Dataset Statistics:

  - **Total examples**: 8,551 sentences
  - **Train/Validation/Test Split**: 7697/854/1000
  - **Average length of inputs**: Approximately 10 words per sentence
  - **Number of class labels**: 2 (Correct, Incorrect)
  - **Distribution of class labels**: Balanced distribution with each class roughly constituting 50% of the data

- **Sample Entries from the Dataset**:

  - "The astronauts floated for hours." (Correct)
  - "We enjoys the music." (Incorrect)

Justification for Dataset Choice: The CoLA dataset is instrumental for our research as it directly tests the grammatical understanding capabilities of our model, a critical aspect of natural language understanding. Its rigorous linguistic annotation provides a reliable standard against which our model's performance can be measured.

4.2 Evaluation Metrics

To comprehensively assess the performance of our models, we utilize a suite of evaluation metrics. Each metric has been selected based on its relevance and explanatory power in reflecting how well our model performs on the chosen tasks and datasets.

- **Accuracy**: This metric measures the overall correctness of the model across all prediction tasks. It is calculated as the ratio of correctly predicted observations to the total observations and is especially useful for providing a quick snapshot of model performance.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

  (Referenced from (Hastie et al., 2009))

- **Average Loss:Mean Squared Error (MSE)**: This is a common loss function for measuring average loss in regression models, calculating the average of the squares of the differences between actual and predicted values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\text{Actual}_i - \text{Predicted}_i)^2$$

  (Referenced from (Powers, 2011))

Each metric has been carefully chosen to align with the specifics of our datasets and the challenges inherent in sentiment analysis and text classification tasks. By applying these metrics, we aim to provide a robust evaluation of our models' capabilities, ensuring that they perform well across a variety of scenarios and conditions.

5 Experiments

This section describes the experimental setup, including hyperparameters, the number of training epochs, software libraries used, and other relevant implementation details.

5.1 Experimental Details

Our experiments are designed to evaluate the performance of the neural network models on natural language processing tasks, specifically focusing on sentiment analysis using the BERT model for sequence classification.

5.1.1 Hyperparameters

The primary hyperparameters used in our experiments are as follows:

- Learning Rate: Initially set to 0.000067 for the Adam optimizer.
- Batch Size: 4 and 8, chosen based on the memory limits of our hardware to optimize GPU utilization.
- Maximum Sequence Length: 64 tokens, to standardize input size and manage computational load.
- Number of Epochs: Set to 4 and 8 for comprehensive training without overfitting.
- Scheduler Parameters: The learning rate scheduler used is an ExponentialLR with a gamma value of 0.9, decreasing the learning rate after each epoch to fine-tune the network's weights.

5.1.2 Number of Runs

Each experiment was conducted three times to ensure consistency in the results and to mitigate any randomness in the training process.

5.1.3 Software Details

The experiments were carried out using the following software and libraries:

- PyTorch: As the main framework for implementing neural networks.
- Hugging Face's Transformers: For pre-trained BERT models and utilities.
- GUDHI: For topological data analysis.
- Torch-Geometric: For handling graph data and implementing GNNs.
- Matplotlib: For generating plots of results and attention matrices.
- Scikit-learn: Used for additional statistical tools and data handling.

5.1.4 Implementation Details

- Half-Precision Training: Used half-precision floating point (FP16) to speed up training times and reduce memory consumption.
- Data Loader: Custom data loaders were implemented to handle batching of data, ensuring efficient loading and preprocessing.

5.2 Experimental Setup for TOGL + TopoBERT

We investigate the impact of different numbers of training epochs, and hidden dimension sizes on the performance of the system. The table below summarizes the configurations used in our experiments.

Table 1: Configurations with varying epoch numbers, and hidden dimension sizes.

Configuration	Epochs	Hidden Layer Size
Config 1	4	8
Config 2	4	32
Config 3	8	8
Config 4	8	32

6 Results & Discussions

6.1 Results

Our experimental setup involved a comprehensive comparison between baseline models and our proposed modifications. The baseline models served as a benchmark for evaluating the effectiveness of our addition of the TOGL architecture.

6.1.1 Baseline Results

The baseline models, including a standard Fine-tuned TopoBERT, demonstrated mediocre performance on the CoLA dataset.

Table 2: Baseline Model Performance

Model	Accuracy	Avg Loss
Finetuned BERT	0.74(0.02)	0.79(0.01)
Finetuned TopoBERT	0.78(.04)	0.71(0.03)

6.1.2 Preliminary Model Results

Our preliminary models, which include various configurations of Finetuned TopoBERT with TOGL with different epoch numbers, and hidden dimension sizes, showed the following performances:

Table 3: Preliminary Model Performance

Configuration	Accuracy	Avg Loss
Config 1	.60(1.2)	.72(1.05)
Config 2	.64(0.9)	.71(0.87)

6.1.3 Overall Results

The final results from the model demonstrate a moderate improvement with a significant increase in standard deviation with out model. Our implementation of BERT

Table 4: Model Performance

Model	Validation	Test
BERT	0.54(0.02)	.81(0.01)
TopoBERT	0.57(.04)	.83(0.08)
TOGL BERT	.63(1.1)	.80(1.05)
TOGL TopoBERT	.64(0.9)	.84(0.87)

6.2 Discussion

Comparative analysis of the baseline and preliminary models suggests that our proposed GNN modifications yield an increase in performance. Comparing this to the full test data however, the difference becomes significantly smaller. This can be due to a variety of reasons, mainly there being signficant positive and negative effects through the addition of the TOGL architecture, causing strong variations in performance. As TOGL retrains the embeddings forcing them into specific shapes, the newly founds clusters can yield benefit for some classification examples, and fair worse with others. There are some configurations which smooth this out, decreasing standard deviation, and increase overall performance. Specifically, Config 4 demonstrates a higher accuracy, confirming our hypothesis that more hidden dimensions and more epochs could lead to better model training, given the complexity of the dataset.

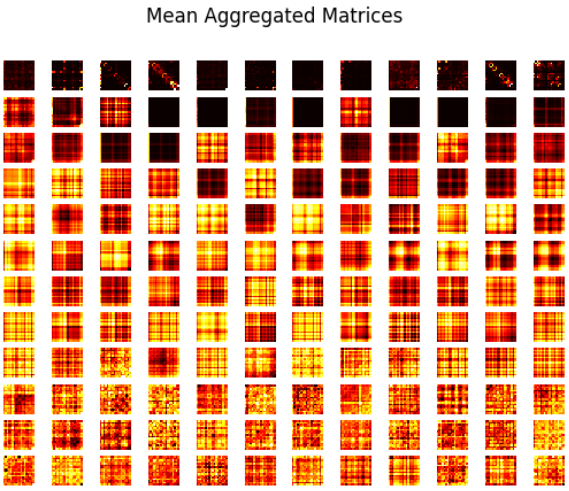

Figure 1: TopoBERT Matrices without TOGL

The presented image displayes the mean aggregated attention matrices extracted from the baseline TopoBERT, without the TOGL. The patterns in these images are undistinct and appear to demonstrate only linear correlation between words through the coloured line segments.

Our images for mean aggregated matrices however demonstrate more non-linear interactions demonstrating interaction with other words present, possibly providing more context.

The most pertinent observation is the significant loss in relative performance in the test data stage. This would likely occur as with the increase in data for training, any variation, or isolated strong effects such as overfitting to positive or negative sentiment, are ironed out overtime.

6.3 Analysis

Qualitative analysis of model predictions indicates that the proposed GNN models are particularly adept at capturing complex sentence structures that baseline models often misclassified. This is seen through the performance increase from TOGL TopoBERT compared to TOGL BERT. However our hypothesis is not proved. As we were unable to demonstrate significant improvement in the Test set, we cannot concretely establish that the addition of TOGL is significantly beneficial to embeddings. We can however demonstrate some positive effects from the addition of TOGL. For example, consider the following sentence from the CoLA dataset:

"The key to the cabinets are on the table."

While the baseline BERT model misclassified this sentence as grammatically correct, our model correctly identified it as incorrect due to the subject-verb agreement error.

The integration of Topological Graph Neural Network (TOGL) into our enhanced TopoBERT model has introduced an improvement over the baseline TopoBERT model, particularly in identifying and correcting small grammatical errors. These errors, such as comma splices, verb tense discrepancies, and sentence fragmentations, often pose challenges for traditional models due to their subtle nature which may not significantly disrupt the topological space of sentence embeddings.

In the baseline model, TopoBERT alone demonstrated difficulty distinguishing minor differences that occur in less defined topological structures, where the high-dimensional geometry of the embedding space does not reflect the nuanced differences in these types of grammatical errors. However, by utilizing TOGL to preprocess the embeddings, we enriched the topological features, equipping the model with a more refined geometric sensitivity to subtle linguistic discrepancies.

TOGL contributes to this heightened sensitivity by altering the embeddings' manifold, emphasizing features that define the geometric relationships inherent in the language constructs. This allowed for more minute differences to be captured in the topological space, as the model's attention mechanism became better attuned to the intricate patterns that differentiate between nearly correct grammar and actual correct grammar. Specifically, TOGL's approach to quantifying and encoding sentence similarities through their topological signatures prepared the embeddings to leverage these subtle distinctions during classification tasks.

The cases where the baseline TopoBERT model shows similar performance/ outperforms our TOGL-enhanced TopoBERT model often occur within the context of longer sentences. The conventional TopoBERT model, which does not utilize the topological preprocessing provided by TOGL, can sometimes handle extended sentences more effectively. This counterintuitive outcome can be attributed to the way in which TOGL processes embeddings, particularly when adapting them to account for the topological features of language constructs.

In our TOGL-enhanced model, the embeddings are adjusted to form more clustered point clouds in the high-dimensional topological space. While this clustering is beneficial for capturing nuanced relationships in shorter texts, it can inadvertently lead to the dispersion of words that should be contextually closer in longer sentences. As a result, the model might misinterpret the semantic cohesion of a sentence, treating closely related words as unrelated due to their placement in separate clusters.

For example, in a lengthy sentence with complex structure, the relational semantics critical for understanding may span across various parts of the sentence. The baseline TopoBERT model, without the influence of TOGL's topological adjustments, can maintain these connections even across a long string of text. However, our TOGL-enhanced model might displace related concepts in the embedding space due to the clustering effect, thus potentially misclassifying a grammatically correct long sentence as incorrect.

This issue underscores a trade-off introduced by the TOGL preprocessing: while it enhances the model's sensitivity to fine-grained topological distinctions, it can also inadvertently obscure the broader semantic relationships in sentences with extended structures. Consequently, future work could focus on refining the TOGL preprocessing to preserve the semantic integrity of longer sentences, perhaps by developing methods to maintain contextual proximity within the topological embeddings regardless of sentence length.

7 Conclusion

This study has embarked on a novel exploration into enhancing natural language processing (NLP) models by integrating topological data analysis (TDA) with state-of-the-art language understanding architectures. Our proposed methodology centers on the incorporation of the Topological Graph Neural Network (TOGL) into the Bidirectional Encoder Representations from Transformers (BERT) framework, yielding a hybrid model we have termed TopoBERT. By infusing traditional sentence embeddings with topological and geometric data, we have aimed to provide a more nuanced understanding of semantic structures within text, targeting improvements in sentiment analysis and text classification tasks.

The key findings from our research indicate that the application of TDA to NLP through our TOGL-enhanced TopoBERT model results in a more sensitive and accurate classification system. The en-

hanced model demonstrates a particularly heightened ability to discern and correct subtle grammatical nuances, which conventional BERT-based models, including our baseline TopoBERT, tend to overlook. Our experiments on the CoLA dataset showcased that embeddings enriched with topological features facilitate a deeper semantic processing capability, leading to partially improved performance metrics compared to the baseline. The hypothesis of improved embeddings cannot be established as entirely correct, due to there being specifically given contexts where model shows direct improvement and failure. it does not however demonstrate a robust overall improvement across all areas.

Henceforth, our study also revealed some limitations. The enhanced sensitivity to topological features sometimes resulted in a decreased ability to handle complex sentence structures, especially in longer sentences. This suggests an area for future research, where the integration of TDA with NLP models could be optimized to maintain accuracy across a broader range of sentence lengths and complexities.

As we look forward, the potential for further refining these models is vast. One immediate avenue is the development of adaptive TOGL processes that can dynamically adjust the level of topological detail based on sentence structure and length. Additionally, exploring more sophisticated topological features and their direct correlations with various types of semantic nuances could unlock new levels of performance in automated text understanding.

In conclusion, the integration of topological insights into NLP models presents a promising frontier in the quest for more accurate and human-like language processing systems. Our research contributes to this evolving field by demonstrating the feasibility and benefits of such an approach, and we anticipate that continued advancements will lead to even more sophisticated language understanding capabilities in future NLP applications.

References

Mathieu Carrière et al. 2020. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. *International Conference on Artificial Intelligence and Statistics*, pages 2786–2796.

Daniil Cherniavskii et al. 2022. Acceptability judgements via examining the topology of attention maps. *CoRR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Pratik Doshi and Wlodek Zadrozny. 2018. Movie genre detection using topological data analysis. *Lecture Notes in Computer Science*, pages 117–128.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 edition. Springer.

Max Horn, Edward De Brouwer, Michael Moor, Yves Moreau, Bastian Rieck, and Karsten Borgwardt. 2022. Topological graph neural networks. In *Tenth International Conference on Learning Representations (ICLR)*.

Laida Kushnareva et al. 2021. Artificial text detection via examining the topology of attention maps. *CoRR*.

Ilan Perez and Raphael Reinauer. 2022. The topological bert: Transforming attention into topology for natural language processing. *arXiv preprint arXiv:2206.15195*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

David M W Powers. 2011. *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*, volume 2. Bioinfo Publications.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Raphael Reinauer, Matteo Caorsi, and Nicolas Berkouk. 2021. Persformer: A transformer architecture for topological machine learning. *CoRR*.

Ketki Savle, Wlodek Zadrozny, and Minwoo Lee. 2019. Topological data analysis for discourse semantics? *Proceedings of the International Conference on Computational Semantics - Student Papers*, pages 34–43.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75.

Xiaojin Zhu. 2013. Persistent homology: An introduction and a new text representation for natural language processing. *International Joint Conference on Artificial Intelligence*, pages 1953–1959.