UE19CS345
Network Analysis and Mining
Course Project

Interpretability of a Graph Classification problem: Toxic Conversations Dataset

PES1UG19CS007 Abdul Mannan
PES1UG19CS190 Ishita Chaudhary
PES1UG19CS321 Palak Kothari
PES1UG19CS387 Rithika Shankar

Topic and its uniqueness

PROBLEM STATEMENT

To predict if a conversation will turn toxic given the reply graph and further look into the interpretibility of the model.

RELATION TO LAB PROGRAMS

Insight into Pytorch Geometric models which helped build our understanding

UNIQUENESS

Graph classification, instead of node classification

Interpretability of a graph classification model

Dataset

Source: Harvard Dataverse

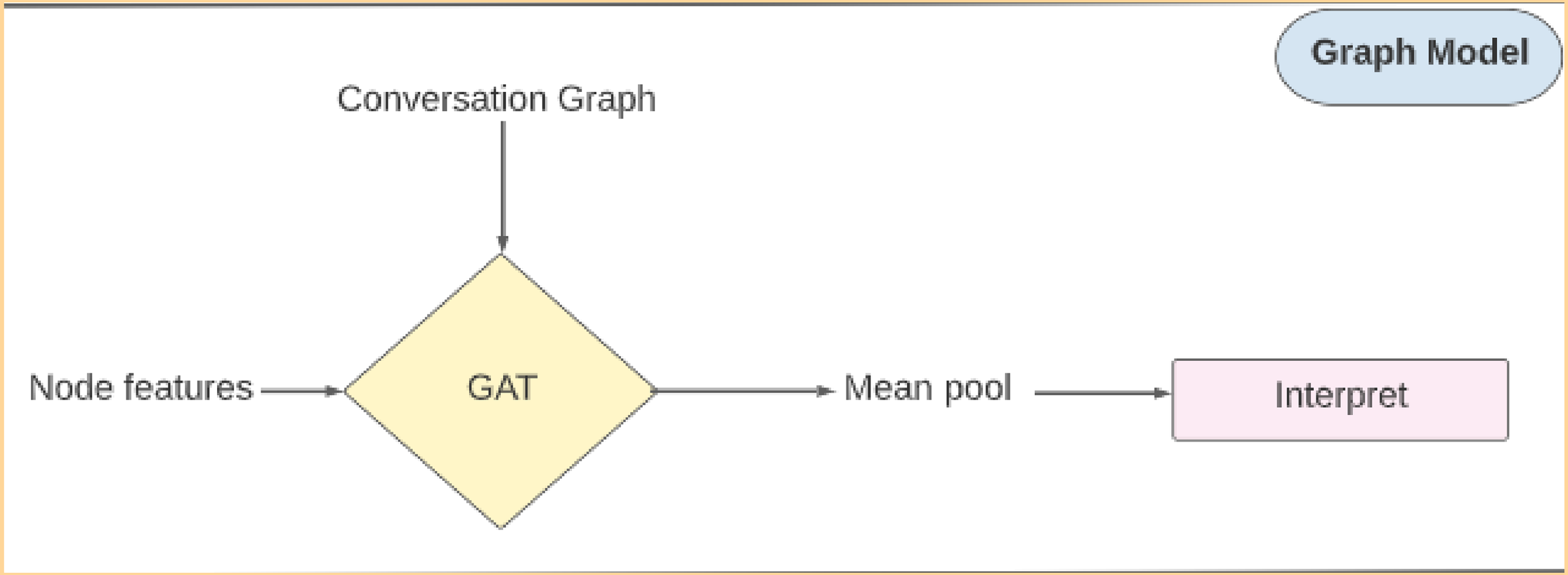
Size: 900 graphs

ATTRIBUTES
root_tweet_type
tweets
reply_tree
alignment_scores
toxicity_scores

Conversation Graph

Node Features

Design


Graph Model
Conversation Graph
Node features
GAT
Mean pool
Interpret

Final Results

Model	Accuracy
Graph Convolution Network	63%
Graph Attention Network	64%

Quantity and quality of work

No	Code Functionality	% Complete	Runs without problem
1	Creating labelled data	100%	Yes
2	Creating the reply graph	100%	Yes
3	Implementing the Graph classification models- GCN, GAT	100%	Yes
4	Implementing the interpretable model- GNN Explainer	100%	Yes

Remaining Portions

- Explore more graph classification models and compare the metrics
- Find the best interpretation of our results
- Predict whether the next reply in a thread will be toxic or not
- Improve upon existing classifiers

Top few learnings

- We observe that due to the attention mechanism in GAT it outperforms GCN, for the graph classification
- Looking into the interpretability helps us trust the model
- GNN Explainer provides a logical insight into the labels predicted (toxic or not toxic)
- It helps us find the crucial features or structures in the graph that affect the decisions of the classifiers

References

Martin Saveski, Brandon Roy, and Deb Roy. 2021. The Structure of Toxic Conversations on Twitter. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 1086–1097. https://doi.org/10.1145/3442381.3449861