Text Classification Using TextGCN

Mayank Gupta (2022H1030099P)
Padshah Rohan Chirag (2022H1030121P)

Problem Statement:

- The objective of the assignment is to identify the medical research proposals which involve studying the SARS-CoV-2 model.
- Develop a machine learning model which can classify the research proposals on the basis of submitted title, keywords and abstract information (AKT)

Why TextGCN:

- TextGCN adapts the GCN framework for text data analysis.
- Nodes represent words, and edges capture co-occurrence relationships between words and documents.
- Word embeddings are integrated into the graph convolutional layers.
- The architecture consists of input representations, graph construction, graph convolutional and prediction layer.
- Training process includes labeled and unlabeled data.
- Labeled data provides supervision for the initial model.
- Unlabeled data helps propagate information and improve model predictions.
- Optimization algorithms minimize the loss function.

Advantages of Semi-Supervised Learning with TextGCN

- Leveraging both labeled and unlabeled data provides significant benefits.
- Improved model performance by utilizing abundant unlabeled data.
- Cost-effective solution by reducing the reliance on labeled data.
- Captures latent semantic relationships through unsupervised learning

Results with TextGCN

innovate achieve lead

Split	Training Accuracy	Test Accuracy	Doc. Combination Used
80:20	0.98	0.58	Title+Keyword+Abstract
70:30	0.96	0.45	Title+Keyword+Abstract

TextGCN to BERT GCN

- We transitioned from TextGCN to BERTGCN to enhance our text analysis capabilities.As TextGCN used Identity matrix as feature matrix, BertGCN takes embeddings from pretrained BERT model and uses them as feature matrix. This can improve the Training of GCN model.
- For prediction it uses linear interpolation where it tries to consider prediction of pre trained BERT model or GCN model trained with BERTs embeddings with a factor lamda.

Results with BertGCN

Split	Test Accuracy	Doc. Combination Used
80:20	0.71	Title+Keyword+Abstract

Thank You !