Project Proposal: Graph Convolutional Networks for Text Classification
Team12 – Networkx

Background
This project considers the problem of classifying given sentence into the defined category efficiently by constructing the heterogeneous graph using the preprocessed text corpus. Most of the existing models (RNN and Statistical models) consider pre-trained word embeddings of the words or sentences and apply the classifier on top of it to classify the sentence/word into one of the category. To improve the efficacy of the model, GCN (Graph Convolution Network) have been considered for the encoding sentences in better representation.

Paper Summary :
A unique graph convolutional network approach for text categorization is proposed in this research. The authors attempted to construct a heterogeneous graph from a text corpus in order to feed it into the model and concurrently train word and document embeddings with graph neural networks. The proposed technique outperforms state-of-the-art text classification algorithms on many benchmark datasets, without the need of pre-trained word embeddings or external information. This approach also automatically learns predictive word and document embeddings.

Project Description :
Text classification is a challenging task in standard NLP era. As semantic representations closely related to syntactic ones. We will work on graph convolution network based technique which uses GCN as sentence encoder. The dataset present in this project do not have much annotation to train the GCN for classification as the model is semi-supervised. Hence, we can able to extend this work to Indic languages by crawling the news articles with different categories as labels for each article and train the baseline and proposed models on it for significant improvements over the baseline model. Also, in traditional NLP era, the transformer models can learn the graph structure of the given input text for better representation of text to produce the state-of-art results. We will compare the performance of the proposed model with the BERT for Text classification model and discuss the results with detailed analysis.

Main goal :
Aim of this paper is to build the heterogenous graph from the preprocessed corpus and learn the word and document embeddings to classify the given text into one of the categories.

Objectives:
- To highlight the effectiveness of GCN‘s in NLP.
- Using GCN two layer network, train the heterogenous graph build on text to get the word or documents embeddings.
- Once the embeddings generated classifier will classify the text into one of the category present in the heterogenous graph.

Datasets :
- The 20NG dataset contains 18,846 documents evenly categorized into 20 different categories.
- The Ohsumed corpus is from the MEDLINE database, contains 23 disease categories.
- R52 and R8 are two subsets of the Reuters 21578 dataset and contains 52, 8 categories respectively.
- MR is a movie review dataset for binary sentiment classification. It contains 5,331 positive and 5,331 negative reviews.
- [text_gcn/data at master · yao8839836/text_gcn (github.com)]

Baseline & Evaluation of results:
In this project, we'll use the GCN two layer network as the baseline to train the heterogenous graph built on the text data. The objective of this model is to build the embeddings for the nodes present in the heterogenous graph and classify the text to one of the category present in the heterogenous graph. Also, we would like to compare the results with Bert based text classifications finetune models (as these models are considered to be performed better than RNN

models). As this project mainly consider this as classification task, we'll use the very well-known metrics (categorical cross entropy) as evaluation metrics for the models.

Timeframe + Work Distribution (For one months)

	Task	Work Distribution	Start and End Dates
Phase One	Report preparation	Madan, Prateek & Pavan	(April 6 – April 9)
	Understanding GCN	Madan, Prateek & Pavan	
	Baseline setup	Prateek & Pavan	
Phase Two	Reproduce baseline results	Prateek & Pavan	(April 9 - April 13)
	Methodology	Prateek, Madan & Pavan	
	Description	Prateek, Madan & Pavan	
	Multilingual data collection	Prateek & Pavan	
Phase Three	Setup Transformers	Prateek, Madan & Pavan	(April 14 – April 24)
	Comparing with transformer models	Prateek, Madan & Pavan	
	Reporting results	Prateek, Madan & Pavan	
	Error analysis	Prateek, Madan & Pavan	
	Hyper parameter tuning	Prateek, Madan & Pavan	
	Training multilingual data	Prateek, Madan & Pavan	
Phase Four	Future Works	Prateek, Madan & Pavan	(April 25 – April 30)
	Conclusions	Prateek, Madan & Pavan	
	References	Prateek, Madan & Pavan	

References :

- [Yao, L., Mao, C. and Luo, Y., 2019, July. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 7370-7377).]
- [Thomas N. Kipf and Max Welling. 2017. Semi supervised classification with graph convolutional networks. In Proceedings of ICLR.]
- [Sun, C., Qiu, X., Xu, Y. and Huang, X., 2019, October. How to fine-tune bert for text classification?. In China national conference on Chinese computational linguistics (pp. 194-206). Springer, Cham.]
- [https://medium.com/analytics-vidhya/tf-idf-term-frequency-technique-easiest-explanation-for-text-classification-in-nlp-with-code-8ca3912e58c3]
- [Text Classification Algorithms: A Survey | by Kamran Kowsari | Text Classification Algorithms | Medium]