Disease Identification From Literature

Project Proposal

Presented by:

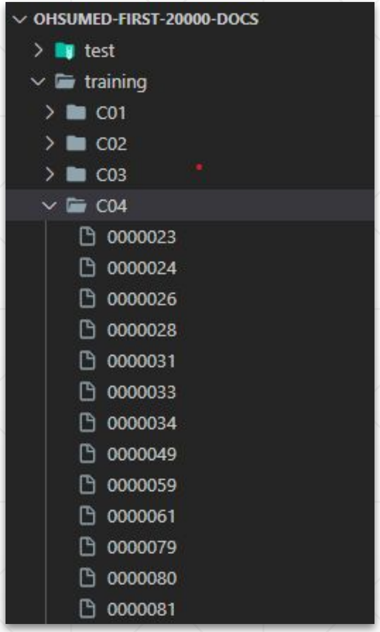
- Asif Shahriar # 1805040
- Debojit Pandit # 1805042

Project Idea

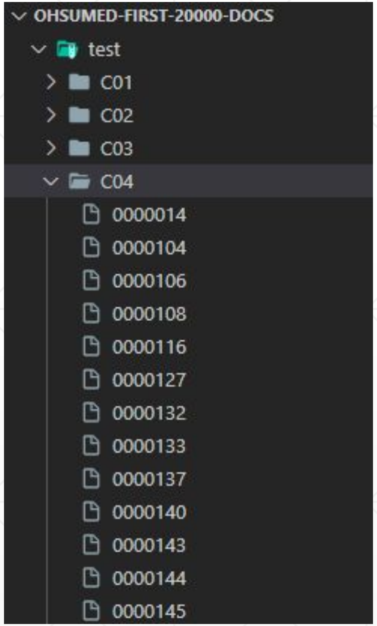
- There are 23 categories of diseases: disease-categories
- Task: Given a medical literature/document, attribute it to one or more of these categories
- NLP & Multi-label classification problem
- Kaggle link
- Papers:
  - GCN-paper-1
  - GCN-paper-2

Dataset

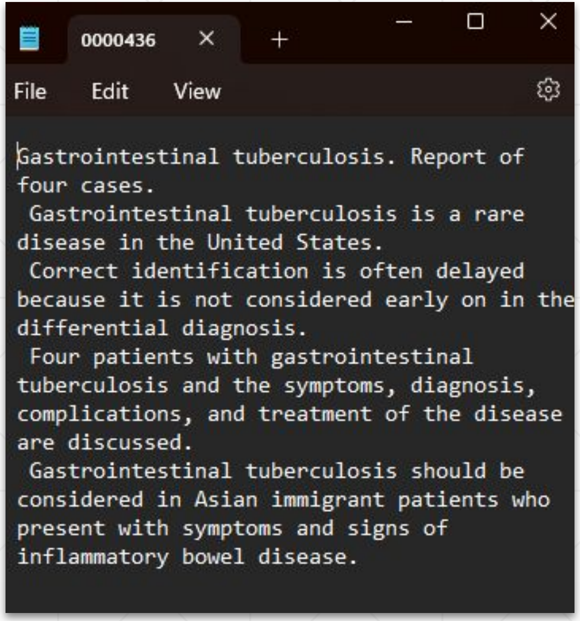
- **OHSUMED** ohsumed-dataset : consists of abstracts from medical journals



```
OHSUMED-FIRST-20000-DOCS
  test
  training
    C01
    C02
    C03
    C04
      0000023
      0000024
      0000026
      0000028
      0000031
      0000033
      0000034
      0000049
      0000059
      0000061
      0000079
      0000080
      0000081
```



```
OHSUMED-FIRST-20000-DOCS
  test
    C01
    C02
    C03
    C04
      0000014
      0000104
      0000106
      0000108
      0000116
      0000127
      0000132
      0000133
      0000137
      0000140
      0000143
      0000144
      0000145
```



Gastrointestinal tuberculosis. Report of four cases.

Gastrointestinal tuberculosis is a rare disease in the United States. Correct identification is often delayed because it is not considered early on in the differential diagnosis. Four patients with gastrointestinal tuberculosis and the symptoms, diagnosis, complications, and treatment of the disease are discussed. Gastrointestinal tuberculosis should be considered in Asian immigrant patients who present with symptoms and signs of inflammatory bowel disease.

Data Preprocessing

- Current representation: document 0008272 is contained in two directories: C07 & C15 -> two labels
- Need to create a csv file , with entries like this:
  - 0008272 00000010000000100000000

Proposed Experiments

- We found papers using GCN models on Ohsumed dataset
- GCN-paper-1
- GCN-paper-2

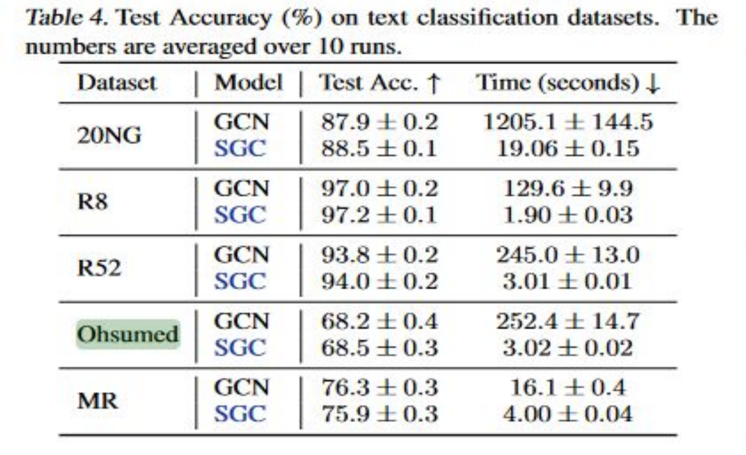
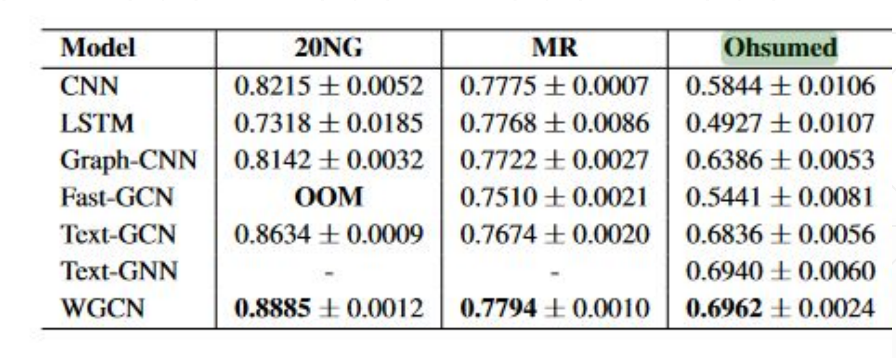


Table 4. Test Accuracy (%) on text classification datasets. The numbers are averaged over 10 runs.

Dataset	Model	Test Acc. ↑	Time (seconds) ↓
20NG	GCN	87.9 ± 0.2	1205.1 ± 144.5
	SGC	88.5 ± 0.1	19.06 ± 0.15
R8	GCN	97.0 ± 0.2	129.6 ± 9.9
	SGC	97.2 ± 0.1	1.90 ± 0.03
R52	GCN	93.8 ± 0.2	245.0 ± 13.0
	SGC	94.0 ± 0.2	3.01 ± 0.01
Ohsumed	GCN	68.2 ± 0.4	252.4 ± 14.7
	SGC	68.5 ± 0.3	3.02 ± 0.02
MR	GCN	76.3 ± 0.3	16.1 ± 0.4
	SGC	75.9 ± 0.3	4.00 ± 0.04



Model	20NG	MR	Ohsumed
CNN	0.8215 ± 0.0052	0.7775 ± 0.0007	0.5844 ± 0.0106
LSTM	0.7318 ± 0.0185	0.7768 ± 0.0086	0.4927 ± 0.0107
Graph-CNN	0.8142 ± 0.0032	0.7722 ± 0.0027	0.6386 ± 0.0053
Fast-GCN	OOM	0.7510 ± 0.0021	0.5441 ± 0.0081
Text-GCN	0.8634 ± 0.0009	0.7674 ± 0.0020	0.6836 ± 0.0056
Text-GNN	-	-	0.6940 ± 0.0060
WGCN	**0.8885** ± 0.0012	**0.7794** ± 0.0010	**0.6962** ± 0.0024

- We will use models like **LSTM-CNN** and **BERT** for better performance