Kriti 2024

Automated Research Paper Categorisation

Manas Hostel

The Problem Statement

- The problem statement involved classifying research papers into one or more of the 57 given categories based on their title and abstract
- Basically, it was a **“Multi-label text classification”** problem with Macro F1 score as the judging criteria
- The given training dataset had 51210 samples with columns : “Id”, “Title”, “Abstract”, “Categories"

Data Cleaning

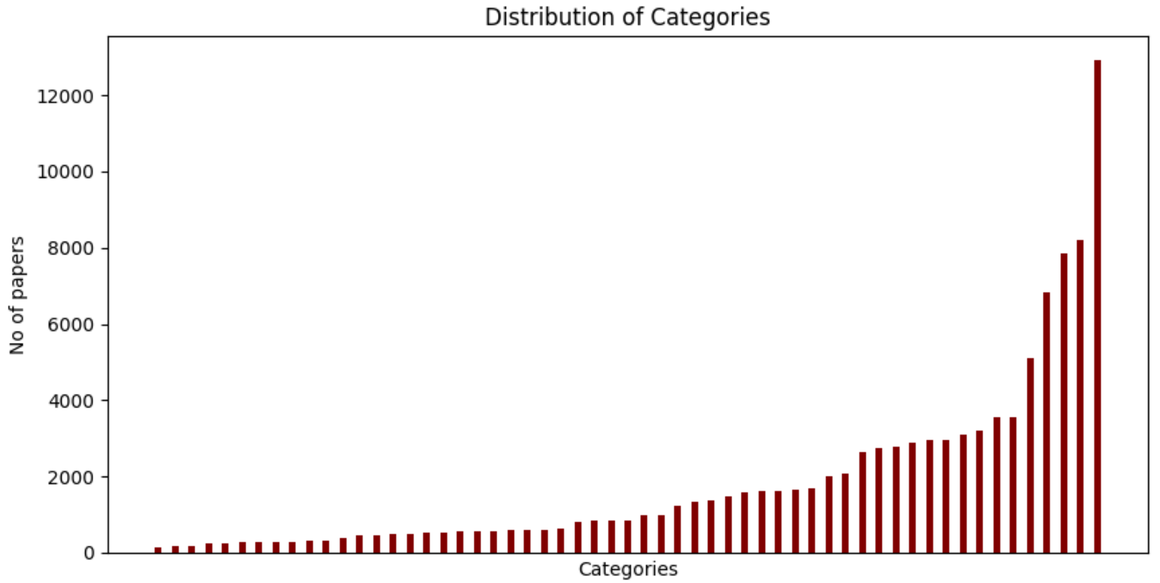
- Parsed the “Categories” column from str to list data type
- Created a “Text” column by joining “Title” and “Abstract” and cleaned it:
  - Parsed LATEX to text using **pylatexenc** library
  - Removed redundant spaces and punctuations
  - Converted it to lower case

```
On dense orbits in the boundary
of a Coxeter system In this
paper, we study the minimality of
the boundary of a Coxeter system.
We show that for a Coxeter system
$(W,S)$ if there exist a maximal
spherical
subset $T$ of $S$ and an element
$s_0\in S$ such that $m(s_0,t)\ge
3$ for each
$t\in T$ and $m(s_0,t_0)=\infty$
for some $t_0\in T$, then every
orbit
$W\alpha$ is dense in the
boundary $\partial\Sigma(W,S)$ of
the Coxeter system
$(W,S)$, hence
$\partial\Sigma(W,S)$ is minimal,
where $m(s_0,t)$ is the order
of $s_0t$ in $W$.
```

```
on dense orbits in the boundary
of a coxeter system in this
paper, we study the minimality of
the boundary of a coxeter system.
we show that for a coxeter system
(w,s) if there exist a maximal
spherical subset t of s and an
element s_0 s such that m(s_0,t)≥
3 for each t t and m(s_0,t_0)=
for some t_0 t, then every orbit
w is dense in the boundary (w,s)
of the coxeter system (w,s),
hence (w,s) is minimal, where
m(s_0,t) is the order of s_0t in
w.
```

Data Analysis

The key insight that we got from data analysis was that the categories had a long tailed distribution/class imbalance



Approaches

BERT-based Transformers

bert-base-uncased & scibert-scivocab-uncased

T5 Transformer

Solving class imbalance

SciBERT with a special loss function for long-tailed distributions

$$L_{DB} = \begin{cases} -\hat{r}_{DB}(1-q_i^k)^{\gamma} log(q_i^k) & \text{if } y_i^k = 1 \\ -\hat{r}_{DB}\frac{1}{\lambda}(q_i^k)^{\gamma} log(1-q_i^k) & \text{otherwise.} \end{cases}$$

Results

- bert-base-uncased [2 Epochs] : 0.49
- **scibert_scivocab_uncased [6 Epochs] = 0.65**
- scibert_scivocab_uncased (Sp. Loss Function) = 0.65
- T5 Transformer [2 Epochs] = 0.60