Empowering Interdisciplinary Research with BERT-Based Models: An Approach Through SciBERT-CNN with Topic Modeling

266 Natural Language Processing, April 16, 2024

Darya Likhareva, Hamsini Sankaran, Siva Thiyagarajan

Outline

- Introduction
- Data
- Exploratory Data Analysis
- Primary model Architecture
- Key Model Results
- Discussion
- Conclusion

Team

Introduction

- **Problem**:
  - Growing Volume of Academic Publications
- **Limitations of Current Models:**
  - Poor or inadequate classification of diverse and interdisciplinary research
  - Improved classification methods are essential to ensure all research is recognized and utilized effectively
- **Our Approach:**
  - Multi-Label Text Classification with a SciBERT-CNN with Topic Modeling

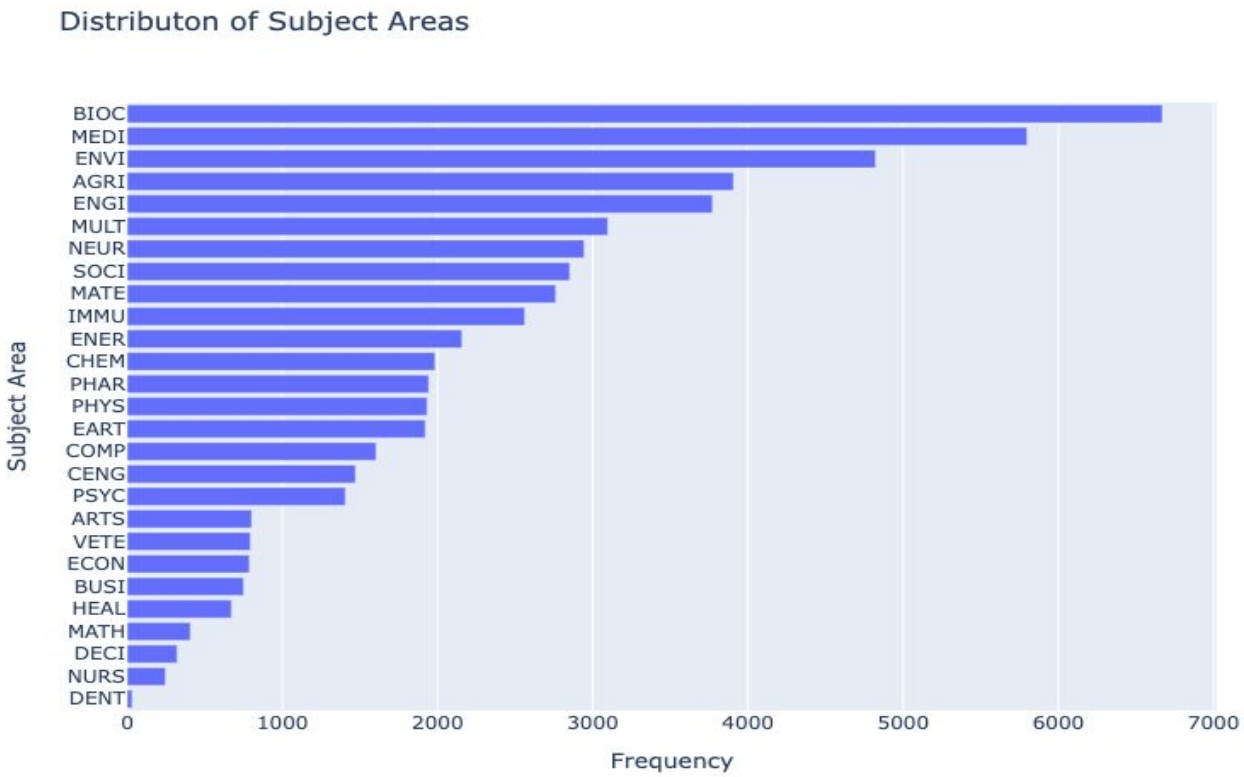
Data

Explore Our Data: [orieg/elsevier-oa-cc-by · Datasets at Hugging Face]

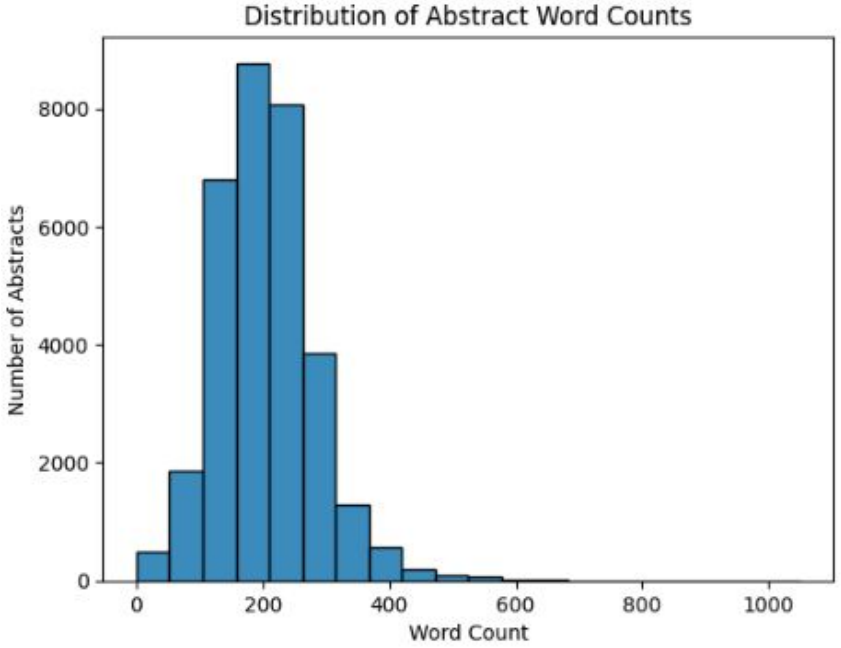
Elsevier OA CC-BY Corpus Attributes
Title
Abstract
Subject Areas
Keywords
ASJC codes
Body Text
Author Highlights

Elsevier OA CC-BY Corpus Dataset Sizes and Structures	
Train Data	(32072, 7)
Validation Data	(4009, 7)
Test Data	(4008, 7)

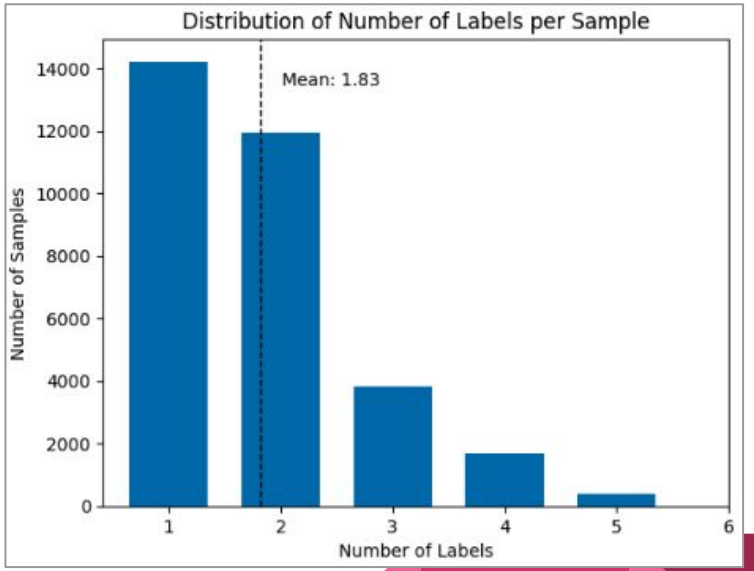
Exploratory Data Analysis



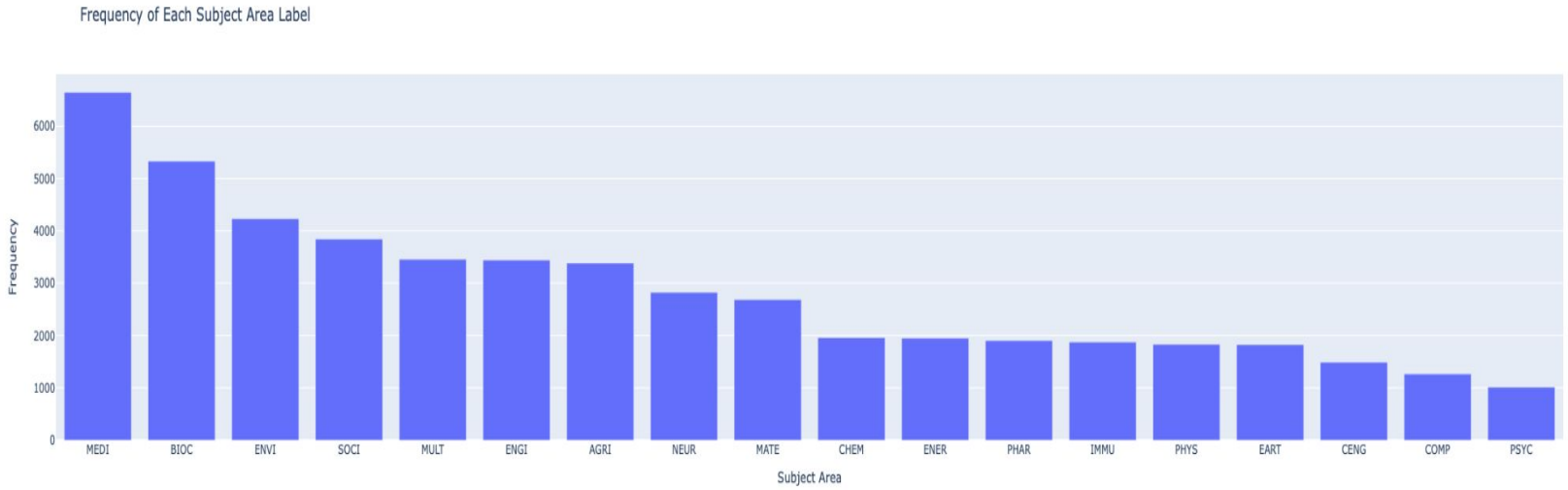
Distribution Of Abstract



Distribution Of Number of Labels Per Sample



After Class Rebalancing



Baseline:

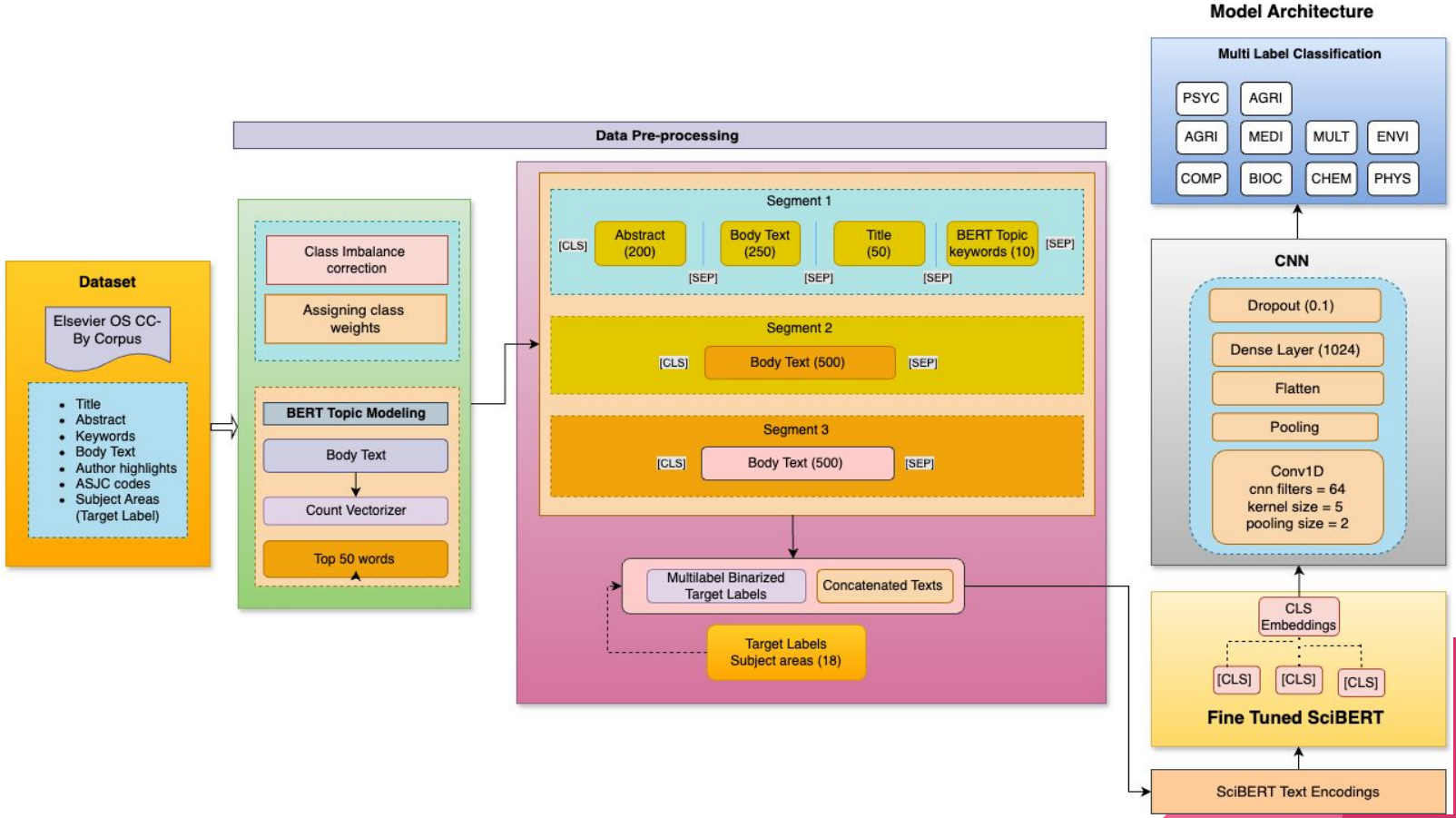
- BERT with Abstract with standard transformer architecture
- **Baseline Weighted Average: 0.59**
- Poor performance due to class imbalance and underrepresented labels

	precision	recall	f1-score	support
AGRI	0.73	0.58	0.64	453
ARTS	0.70	0.14	0.23	102
BIOC	0.80	0.54	0.64	842
BUSI	0.61	0.24	0.34	83
CENG	0.69	0.11	0.19	177
CHEM	0.60	0.21	0.31	234
COMP	0.78	0.28	0.41	224
DECI	1.00	0.00	0.00	37
DENT	1.00	0.00	0.00	4
EART	0.81	0.64	0.71	227
ECON	0.89	0.23	0.37	107
ENER	0.72	0.60	0.65	299
ENGI	0.68	0.50	0.58	463
ENVI	0.64	0.70	0.67	597
HEAL	0.89	0.39	0.54	105
IMMU	0.80	0.51	0.62	314
MATE	0.78	0.55	0.65	340
MATH	1.00	0.00	0.00	64
MEDI	0.81	0.62	0.71	737
MULT	0.99	0.65	0.79	386
NEUR	0.76	0.87	0.81	383
NURS	1.00	0.03	0.05	37
PHAR	0.69	0.35	0.47	231
PHYS	0.78	0.25	0.38	254
PSYC	0.59	0.43	0.50	182
SOCI	0.76	0.39	0.52	367
VETE	0.71	0.70	0.70	76
micro avg	0.75	0.51	0.61	7325
macro avg	0.79	0.39	0.46	7325
weighted avg	0.76	0.51	0.59	7325
samples avg	0.79	0.58	0.60	7325

Model Experiments

Model	Embeddings	Feature Modifications
BERT [Baseline]	BERT	abs only
RoBERTa	RoBERTa	abs only
Longformer	Longformer	abs only
SciBERT	SciBERT	abs only
SciBERT	SciBERT	abs + body_text
SciBERT	SciBERT	abs + body_text, CLS embeddings
SciBERT	SciBERT	abs + body_text + title + keywords, CLS embeddings + CNN
SciBERT	SciBERT	abs + body_text + title + top 10 important words from body_text, Keybert
SciBERT [Final Model]	SciBERT	abs + body_text + title + top 10 important words from body_text, BERT topic modeling

Primary Model Architecture



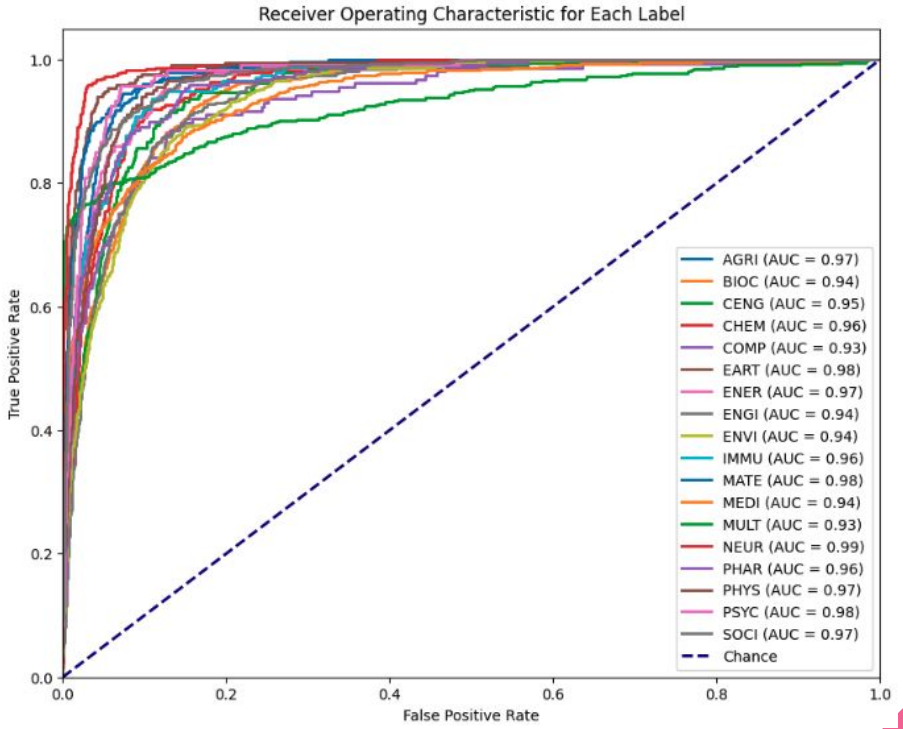
** Weighted Pr, Recall & F1 score

Final Model

- Significant improvement in the performance of our SciBERT-CNN with BERT topic modeling.
- Fine-tuning implemented in the SciBERT-CNN model: CNN's convolutional and max-pooling layers, dropout strategy, dense layer, and classification layer with a sigmoid function
- **SciBERT-CNN Weighted Average: 0.70**

Label	Baseline F1-Score	Best Model			
		Precision	Recall	F1-Score	Support
AGRI	0.64	0.71	0.78	0.74	413
BIOC	0.34	0.69	0.69	0.69	653
CENG	0.19	0.41	0.68	0.51	189
CHEM	0.31	0.55	0.66	0.60	248
COMP	0.41	0.47	0.57	0.51	157
EART	0.71	0.57	0.93	0.71	217
ENER	0.65	0.64	0.63	0.64	235
ENGI	0.58	0.56	0.76	0.64	444
ENVI	0.67	0.67	0.60	0.63	512
IMMU	0.62	0.67	0.59	0.63	238
MATE	0.65	0.70	0.89	0.78	331
MEDI	0.71	0.82	0.71	0.76	851
MULT	0.79	0.98	0.68	0.80	467
NEUR	0.81	0.86	0.86	0.86	394
PHAR	0.47	0.51	0.75	0.61	219
PHYS	0.38	0.49	0.75	0.59	203
PSYC	0.50	0.46	0.81	0.58	120
SOCI	0.52	0.87	0.69	0.77	470
micro avg	0.61	0.67	0.72	0.69	6361
macro avg	0.46	0.65	0.72	0.67	6361
weighted avg	0.59	0.70	0.72	0.70	6361
samples avg	0.60	0.72	0.74	0.70	6361

ROC CURVE


Receiver Operating Characteristic for Each Label
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
AGRI (AUC = 0.97)
BIOC (AUC = 0.94)
CENG (AUC = 0.95)
CHEM (AUC = 0.96)
COMP (AUC = 0.93)
EART (AUC = 0.98)
ENER (AUC = 0.97)
ENGI (AUC = 0.94)
ENVI (AUC = 0.94)
IMMU (AUC = 0.96)
MATE (AUC = 0.98)
MEDI (AUC = 0.94)
MULT (AUC = 0.93)
NEUR (AUC = 0.99)
PHAR (AUC = 0.96)
PHYS (AUC = 0.97)
PSYC (AUC = 0.98)
SOCI (AUC = 0.97)
Chance

Challenges, Conclusion & Future work

- Challenges: Data Augmentation, Compute Resources

- Tackled the challenging task of multi-label text classification
- Significant performance improvements by combining SciBERT-CNN with BERT topic modeling. By focusing on abstracts, body text segments, titles, and a selection of the top 10 words, we surpassed other BERT-based model benchmarks.

- Future efforts will explore data augmentation and the integration of domain-specific keywords to bolster underrepresented classes and refine overall label performance.

Thank you!

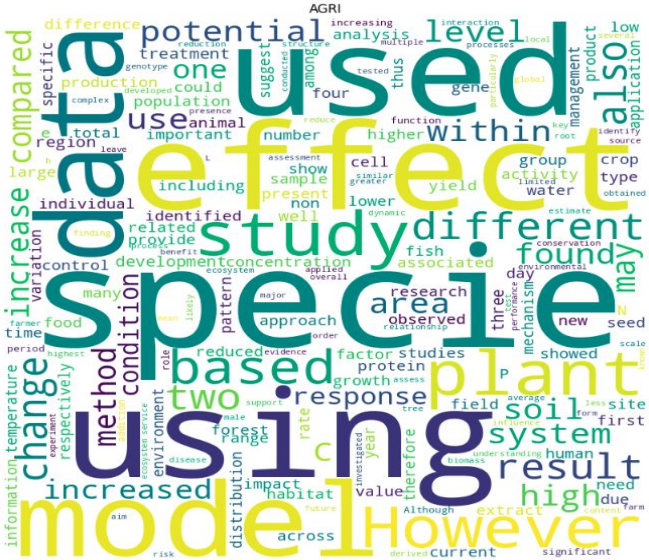
Questions? Comments? Concerns?

BASELINE MODEL (BERT) INFERENCE

Labels	Precision	Recall	F1-Score	Support
AGRI				
BIOC				
CENG				
CHEM				
COMP				
EART				
ENER				
ENGI				
ENVI				

IMMU				
IMMU				
MEDI				
MULT				
NEUR				
PHAR				
PHYS				
PSYC				
SOCI				

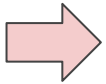
WORD CLOUD - COMMON WORDS IN SUBJECT AREAS



Weighted Average Inference Results

BASELINE - BERT

micro avg	0.67	0.72	0.69	6361
macro avg	0.65	0.72	0.67	6361
weighted avg	0.70	0.72	0.70	6361
samples avg	0.72	0.74	0.70	6361



FINE TUNED SciBERT-CNN

micro avg	0.67	0.72	0.69	6361
macro avg	0.65	0.72	0.67	6361
weighted avg	0.70	0.72	0.70	6361
samples avg	0.72	0.74	0.70	6361