Hierarchical Text Classification with Transformer-based Language Models

Jaco du Toit

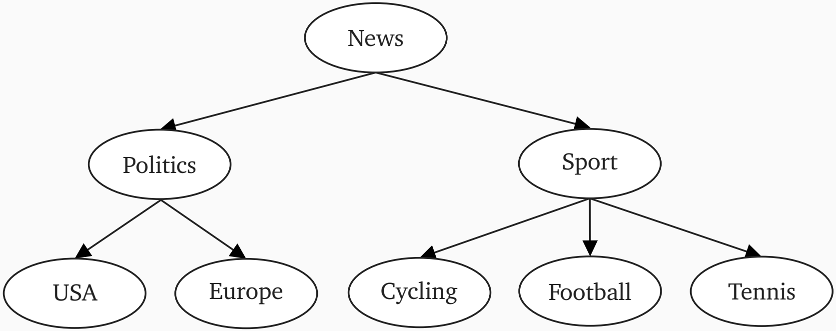
Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

forward together
sonke siya phambili
saam vorentoe

Supervisor: Marcel Dunaiski

Hierarchical Text Classification

- Objective: Classify text documents into a set of classes from a structured class hierarchy.



Motivation and Objectives

- Motivation:
  - Improves organisation and navigation of documents.
  - Allows users to select the level of granularity that they prefer.
  - What are the best approaches for incorporating the class hierarchy information?
  - Advancements in "flat" text classification have not been investigated for hierarchical text classification (HTC).
- Objectives:
  - Identify shortcomings of current approaches and promising unexplored areas of research.
  - Identify advancements in standard text classification approaches which have not been applied to HTC tasks.
  - Propose new HTC approaches from the identified unexplored areas of research.
  - Create new benchmark HTC datasets.

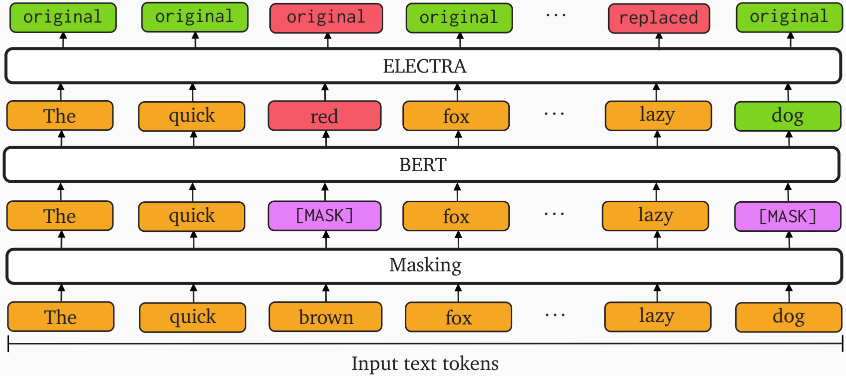
Hierarchical Text Classification Benchmark Datasets

- Web Of Science (WOS): Abstracts of research publications from Web of Science.
- Reuters Corpus Volume 1-Version 2 (RCV1-V2): News articles from Reuters.
- New York Times (NYT): News articles from New York Times.

Dataset	Levels	Classes	Avg. Classes	Train	Dev	Test
WOS	2	141	2.0	30,070	7,518	9,397
RCV1-V2	4	103	3.24	20,833	2,316	781,265
NYT	8	166	7.6	23,345	5,834	7,292

Transformer-based Language Models

- Trained through self-supervised learning tasks on large amounts of textual data.
- Attention mechanisms obtain contextually aware word embeddings.
- Self-supervised learning tasks:
  - Masked language modelling (BERT).
  - Replaced token detection (ELECTRA).



Outline

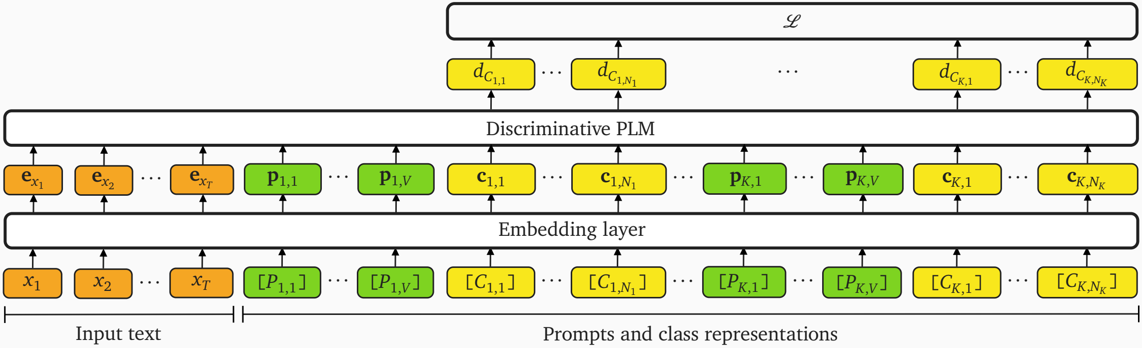
- Four chapters each structured as a paper:
  - Part 1 - Prompt Tuning Discriminative Language Models.
  - Part 2 - Language Models with Label-wise Attention Mechanisms.
  - Part 3 - Combining Language and Topic Models.
  - Part 4 - Introducing Three New Benchmark Datasets.

Part 1 - Prompt Tuning Discriminative Language Models

- Background:
  - Prompt tuning for text classification.
    - "**x** is about `[MASK]`"
  - Hierarchy-aware Prompt Tuning (HPT).
    - "**x** Level 1 class: `[MASK]` Level 2 class: `[MASK]`"
  - Prompt Tuning framework for Discriminative PLMs (DPT).
    - "**x** Class: Politics, $\cdots$, Sport"
- Objectives:
  - Combine the HPT and DPT approaches to investigate the efficacy of prompt tuning discriminative language models for hierarchical text classification tasks.
  - Propose improvements to DPT.

Model Architecture

- Our approach applies the prompt tuning paradigm to discriminative language models by appending prompts (green) and class representations (yellow) to the text token sequence (orange).



Positional embeddings

- Assign the same position IDs to all of the class tokens at a certain level.
- Allows the approach to scale to HTC tasks with much larger hierarchical class structures while maintaining many more input text tokens than DPT.

Dataset	Levels	Classes	DPT		HPTD		
			Tokens	%Tokens	Tokens	%Tokens	Additional tokens
WOS	2	141	369	72.07	508	99.21	+139
RCV1-V2	4	103	405	79.10	504	98.44	+99
NYT	8	166	338	66.02	496	96.88	+158
Ill. Ex.	2	50	460	89.84	508	99.21	+48
Ill. Ex.	2	200	310	60.54	508	99.21	+198
Ill. Ex.	2	800	0	0	508	99.21	+798
Ill. Ex.	8	50	454	88.67	496	96.88	+42
Ill. Ex.	8	200	304	59.38	496	96.88	+192
Ill. Ex.	8	800	0	0	496	96.88	+792

Main Results

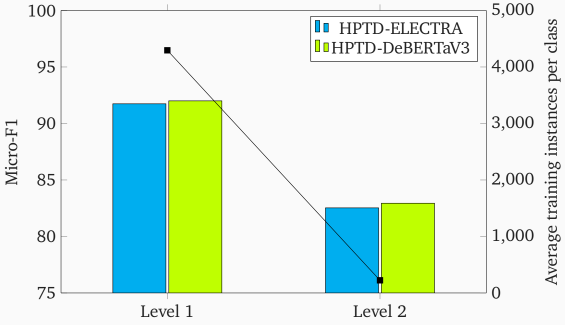
- We compare two discriminative language models: ELECTRA and DeBERTaV3.
- Using DeBERTaV3 model improves performance over ELECTRA on two datasets.
- Our approach outperforms previously proposed approaches on WOS and NYT.

Model	WOS		RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HiMatch	86.20	80.53	84.73	64.11	–	–
HGCLR	87.11	81.20	86.49	68.31	78.86	67.96
PAAMHiA-T5 [1]	**90.36**	81.64	87.22	70.02	77.52	65.97
HBGL	87.36	82.00	87.23	**71.07**	80.47	70.19
HPT	87.16	81.93	**87.26**	69.53	80.42	70.42
HPTD-ELECTRA	87.45	81.67	86.30	68.12	80.54	70.66
HPTD-DeBERTaV3	**87.85**	**82.13**	86.25	66.85	**81.45**	**72.40**

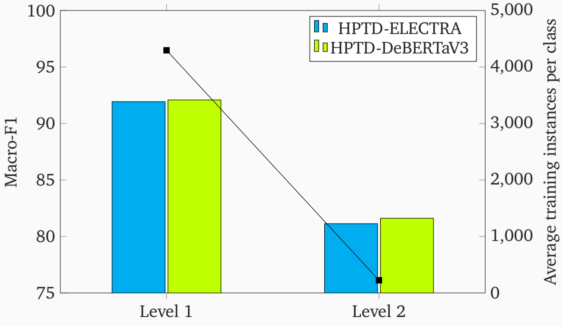
[1] Results obtained using twice the number of model parameters as the other approaches.

Level-wise Results

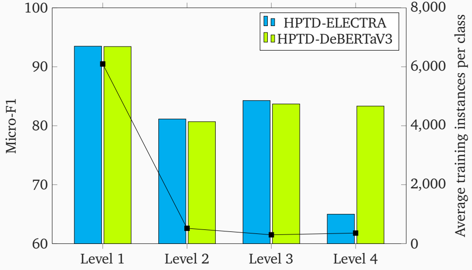
- Classification performance generally decreases for the lower levels of the class hierarchy with fewer average training instances per class.



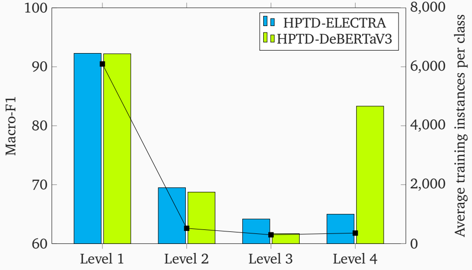
WOS Micro-F1.



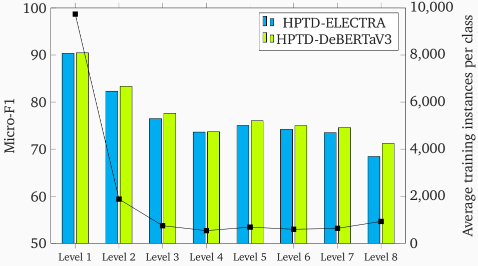
WOS Macro-F1.



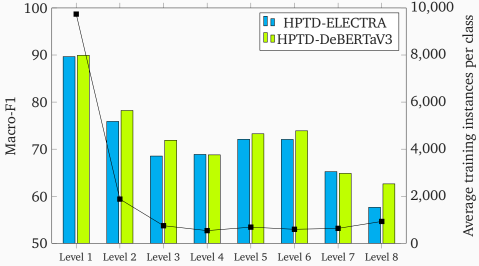
RCV1-V2 Micro-F1.



RCV1-V2 Macro-F1.



NYT Micro-F1.



NYT Macro-F1.

Low-resource Results

- Only use 10% of available training data.
- DeBERTaV3 performs better on the WOS and NYT datasets.
- Macro-F1 scores decrease more than Micro-F1 when using less training data.

Model	WOS		RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HPTD-ELECTRA	82.34 (87.45)	74.75 (81.67)	**80.98** (86.30)	**49.07** (68.12)	75.00 (80.54)	61.28 (70.66)
HPTD-DeBERTaV3	**83.47** (87.85)	**75.74** (82.13)	79.42 (86.25)	45.16 (66.85)	**76.18** (81.45)	**63.38** (72.40)

Part 2 - Language Models with Label-wise Attention Mechanisms

- Background:
  - Label-wise attention mechanisms obtain label-specific document representations from word embeddings.
  - We use two label-wise attention mechanisms to obtain attention weights:
    - Dot Product Attention (DPA):

$$\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{U}_{\mathrm{DPA}}\mathbf{H}^T) \tag{1}$$

    - General Attention (GA):

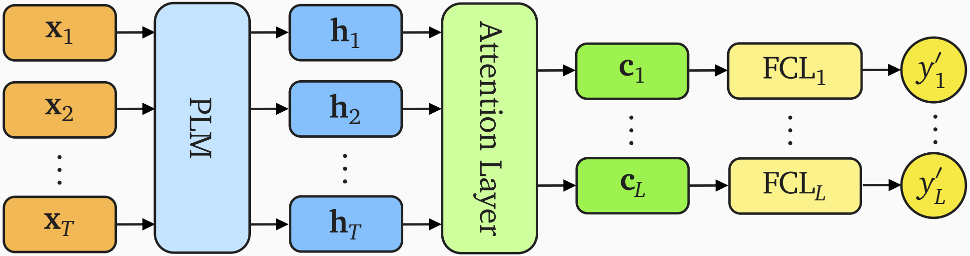
$$\mathbf{Z} = \tanh(\mathbf{Q}_{\mathrm{GA}}\mathbf{H}^T) \tag{2}$$

$$\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{U}_{\mathrm{GA}}\mathbf{Z}) \tag{3}$$

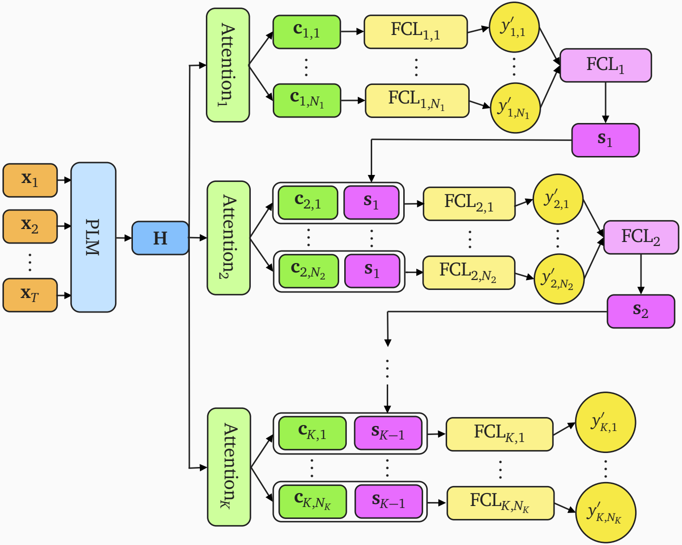
- Objectives:
  - Investigate efficacy of using label-wise attention mechanisms to fine-tune PLMs for HTC tasks.
  - Comparison of different label-wise attention mechanisms.
  - Investigate incorporation of hierarchical class structure into label-wise attention mechanisms.

Model Architecture

- Our approach uses label-wise attention mechanisms to fine-tune PLMs for HTC tasks.



Hierarchical Model Architecture



- Hierarchical label-wise attention (HLA): Separates the label-wise attention mechanisms for each level of the class hierarchy.
- Global hierarchical label-wise attention (GHLA): Extends HLA by concatenating all of the higher-level prediction representations.

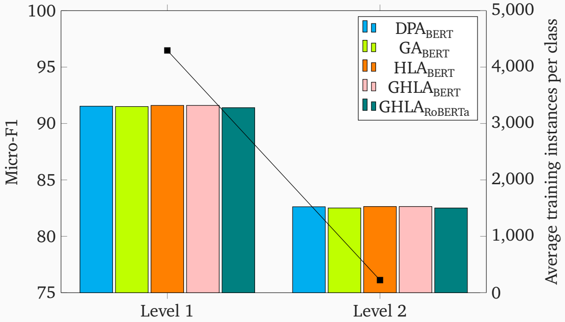
Main Results

- GHLA generally outperforms the other label-wise attention mechanisms.
- Using RoBERTa significantly improves performance on two datasets.
- Using GHLA with RoBERTa outperforms other approaches on RCV1-V2 and NYT.

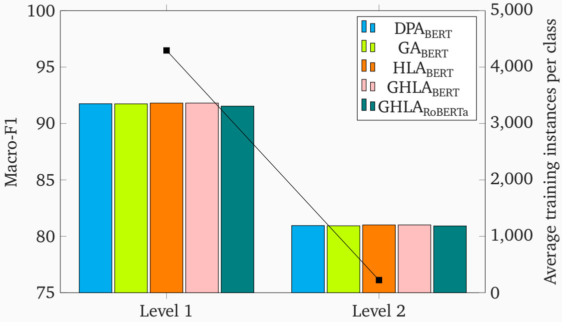
Model	WOS		RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HiMatch	86.20	80.53	84.73	64.11	–	–
HGCLR	87.11	81.20	86.49	68.31	78.86	67.96
PAAMHiA-T5[1]	**90.36**	81.64	87.22	70.02	77.52	65.97
HBGL	**87.36**	**82.00**	87.23	**71.07**	80.47	70.19
HPT	87.16	81.93	87.26	69.53	80.42	70.42
DPA_{BERT}	87.13	81.48	87.07	68.45	79.67	68.27
GA_{BERT}	87.05	81.46	86.88	69.11	80.06	68.56
HLA_{BERT}	87.17	81.55	86.71	68.45	79.60	68.06
$\text{GHLA}_{\text{BERT}}$	87.17	81.55	87.19	68.62	79.67	68.67
$\text{GHLA}_{\text{RoBERTa}}$	87.00	81.44	**87.78**	70.21	**81.41**	**72.27**

[1] Results obtained using twice the number of model parameters as the other approaches.

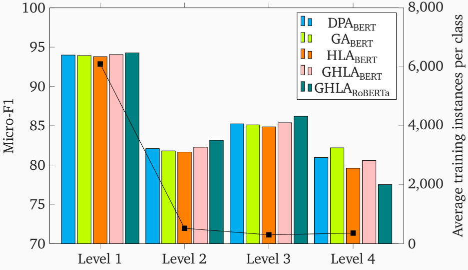
Level-wise Results



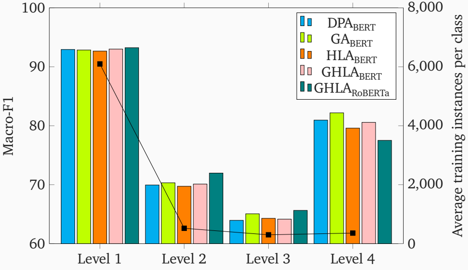
WOS Micro-F1.



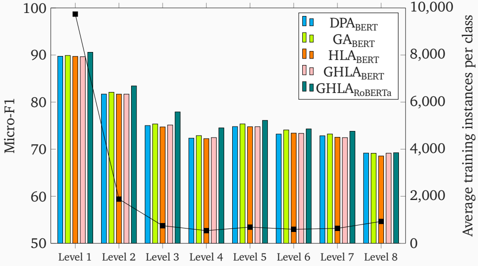
WOS Macro-F1.



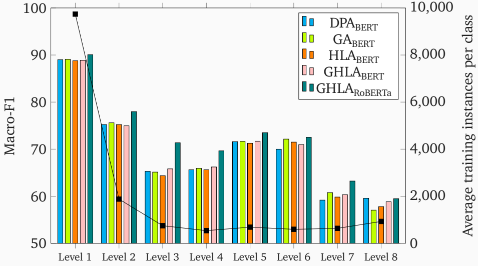
RCV1-V2 Micro-F1.



RCV1-V2 Macro-F1.



NYT Micro-F1.



NYT Macro-F1.

Low-resource Results

- HLA, DPA, and GHLA perform best on WOS, RCV1-V2, and NYT respectively.
- Using RoBERTa significantly improves performance across the three datasets.

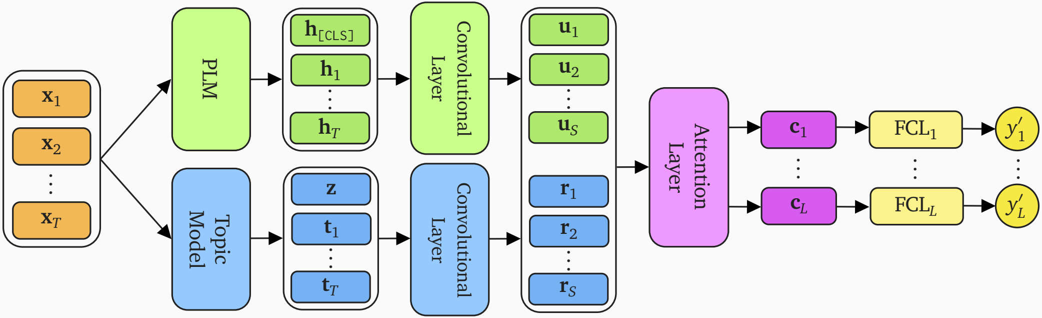
Model	WOS		RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
DPA_{BERT}	79.41 (87.13)	67.51 (81.48)	82.81 (87.07)	52.33 (68.45)	72.29 (79.67)	49.29 (68.27)
GA_{BERT}	79.45 (87.05)	67.50 (81.46)	82.79 (86.88)	49.32 (69.11)	72.54 (80.06)	48.75 (68.56)
HLA_{BERT}	79.53 (87.17)	67.73 (81.55)	82.74 (86.71)	51.87 (68.45)	72.28 (79.60)	45.84 (68.06)
$\text{GHLA}_{\text{BERT}}$	78.39 (87.17)	67.03 (81.55)	82.51 (87.19)	50.74 (68.62)	72.42 (79.67)	50.04 (68.67)
$\text{GHLA}_{\text{RoBERTa}}$	**79.76** (87.00)	**68.98** (81.44)	**84.45** (87.78)	**55.24** (70.21)	**75.70** (81.41)	**57.85** (72.27)

Part 3 - Combining Language and Topic Models

- Background:
  - Topic models extract abstract topics from a corpus of documents.
  - Previous approaches have shown that combining the features extracted from topic models with language model features improves text classification performance.
- Objectives:
  - Investigate if the combination of these features improves performance on HTC tasks.
  - Compare the use of these feature extraction approaches to previously proposed approaches which fine-tune PLMs.

Model Architecture

- Our approach uses topic and language models to extract features which passed to a convolutional neural network (CNN) with label-wise attention and classification layers.



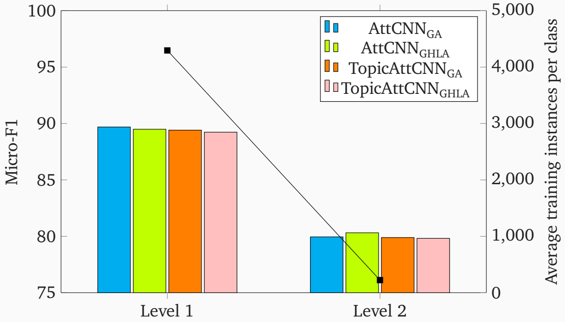
Main Results

- Compare feature combination (TopAttCNN) to only language model features (AttCNN).
- Using features extracted from the topic model generally decreases performance.
- This approach performs significantly worse than previously proposed approaches.

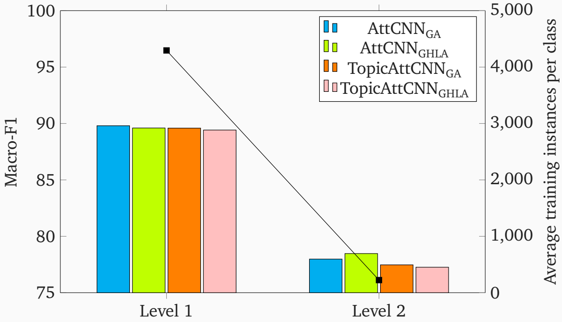
Model	WOS		RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HiMatch	86.20	80.53	84.73	64.11	–	–
HGCLR	87.11	81.20	86.49	68.31	78.86	67.96
PAAMHiA-T5[1]	**90.36**	81.64	87.22	70.02	77.52	65.97
HBGL	**87.36**	**82.00**	87.23	**71.07**	**80.47**	70.19
HPT	87.16	81.93	**87.26**	69.53	80.42	**70.42**
$\text{AttCNN}_{\text{GA}}$	84.93	78.57	84.67	62.48	77.07	64.08
$\text{TopAttCNN}_{\text{GA}}$	84.76	78.07	**84.72**	62.33	76.88	64.18
$\text{AttCNN}_{\text{GHLA}}$	**85.00**	**79.02**	84.54	**63.11**	76.94	**64.57**
$\text{TopAttCNN}_{\text{GHLA}}$	84.64	77.86	84.51	60.32	**77.08**	64.35

[1]Results obtained using twice the number of model parameters as the other approaches.

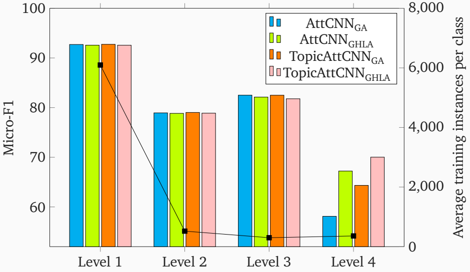
Level-wise Results



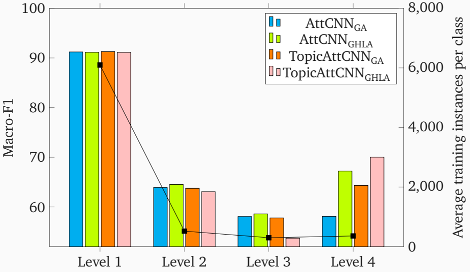
WOS Micro-F1.



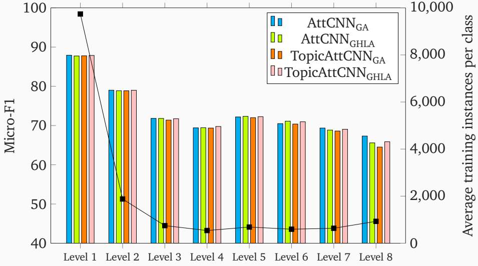
WOS Macro-F1.



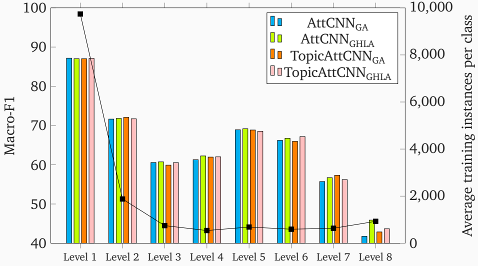
RCV1-V2 Micro-F1.



RCV1-V2 Macro-F1.



NYT Micro-F1.



NYT Macro-F1.

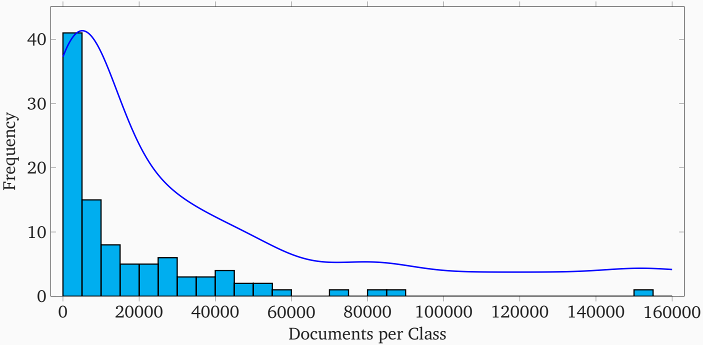
Low-resource Results

- $\text{AttCNN}_{\text{GHLA}}$ and $\text{TopAttCNN}_{\text{GA}}$ approaches perform the best on the WOS and NYT datasets respectively.

Model	WOS		RCV1-V2		NYT	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
$\text{AttCNN}_{\text{GA}}$	75.30 (84.93)	61.76 (78.57)	78.20 (84.67)	46.75 (62.48)	70.80 (77.07)	50.58 (64.08)
$\text{AttCNN}_{\text{GHLA}}$	**75.49** (85.00)	**63.89** (79.02)	78.32 (84.54)	**47.44** (63.11)	71.04 (76.94)	50.88 (64.57)
$\text{TopAttCNN}_{\text{GA}}$	72.37 (84.76)	57.52 (78.07)	**78.69** (84.72)	46.71 (62.33)	**71.18** (76.88)	**51.34** (64.18)
$\text{TopAttCNN}_{\text{GHLA}}$	73.47 (84.64)	59.99 (77.86)	78.35 (84.51)	45.87 (60.32)	70.58 (77.08)	50.96 (64.35)

Part 4 - Introducing Three New Benchmark Datasets

- Motivation:
  - Only RCV1-V2 is accompanied with a detailed creation methodology.
  - Current benchmark datasets are imbalanced.
- Objectives:
  - Create three new datasets in the domain of research publications.
  - Evaluate best-performing approaches to provide baseline for future research.



Journal-based classification schema

- The Journal Topics (JT) classification schema assigns categories to each journal and classifies a publication based on the journal it is published in.
- Journal-based classifications have been shown to be unreliable and inaccurate.

Publication	JT_{L1} (6)	JT_{L2} (52)
"Can Creditor Bail-in Trigger Contagion? The Experience of an Emerging Market..."	Social Sciences	Business Economics
"Dissecting the genre of Nigerian music with machine learning models. Music Information..."	Natural sciences	Information, computer & communication technologies
"The complementarity of a diverse range of deep learning features extracted from video content for video recommendation. Following the popularisation of media streaming, a number of video streaming services are..."	Engineering Natural sciences	Electrical & electronic engineering Engineering sciences (other) Information, computer & communication technologies

Citation-based classification schema

- The Citation Topics (CT) classification schema clusters publications based on citation relationships such that clusters form distinct classifications.
- Does not allow publications to belong to multiple research fields.

Publication	CT_{L1} (10)	CT_{L2} (326)	CT_{L3} (2457)
"Can Creditor Bail-in Trigger Contagion? The Experience of an Emerging..."	Social Sciences	Economics	Economic Growth
"Dissecting the genre of Nigerian music with machine learning models. Music Information..."	Electrical Engineering, Electronics & Computer Science	Knowledge Engineering & Representation	Statistical Tests
"The complementarity of a diverse range of deep learning features extracted from video content for video..."	Electrical Engineering, Electronics & Computer Science	Knowledge Engineering & Representation	Collaborative Filtering

Journal Topics Filtered classification schema

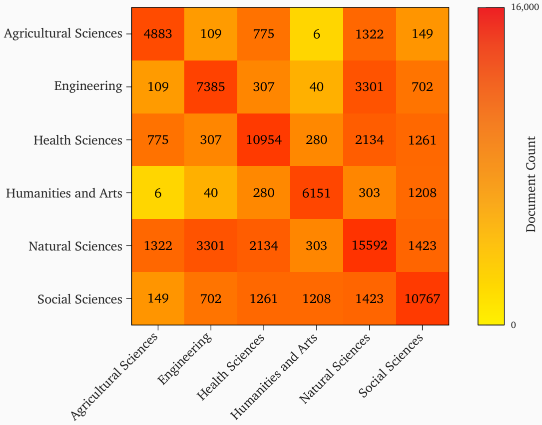
- We proposed the Journal Topics Filtered (JTF) schema that combines the journal- and citation-based classification schemas to create a new categorisation which leverages their respective advantages.
- We used the co-occurrence counts of the JT_{L2} and CT_{L2} classes to map each CT_{L2} class to one or more JT_{L2} classes.
- We created new class assignments which are formed by removing assignments and categories that do not form clear mappings between the two classification schemas.
- The aim of this approach is to increase the probability that an individual document is correctly classified.
- Our proposed approach also allows documents to belong to multiple classes.

Dataset Creation

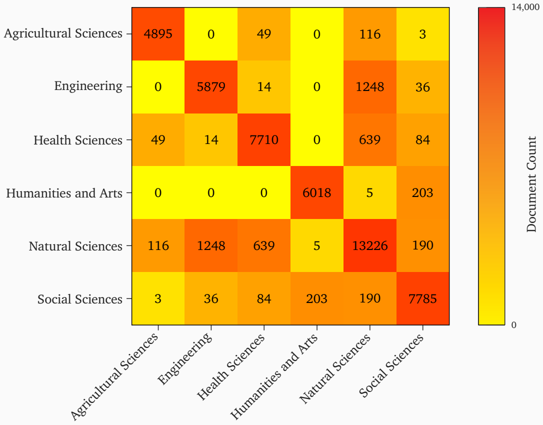
- Randomly sampled 5000 papers from Web Of Science for each of the CT_{L2} classes.
- WOS_{JT}: Randomly sampled 1000 documents for each JT_{L2} class.
- WOS_{CT}: Randomly sampled 200 documents for each CT_{L2} class.
- WOS_{JTF}: Randomly sampled 1000 documents for each JTF_{L2} class.

Dataset	Levels	$Classes_{L1}$	$Classes_{L2}$	Avg. Classes	Train	Dev	Test
WOS_{JT}	2	6	52	2.93	30,356	6,505	6,505
WOS_{CT}	2	10	326	2.00	45,640	9,780	9,780
WOS_{JTF}	2	6	46	2.25	30,048	6,439	6,439

First-level co-occurrence counts



$\mathrm{WOS}_{\mathrm{JT}}$.



$\mathrm{WOS}_{\mathrm{JTF}}$.

Classification Results

- We evaluated our best performing approaches on the three newly created datasets.
- $\text{GHLA}_{\text{RoBERTa}}$ and HPTD-DeBERTaV3 generally outperform the other approaches.
- Performance on WOS_{JTF} is significantly better than the other two datasets.

Model	WOS_{JTF}		WOS_{JT}		WOS_{CT}	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
HPT	84.97	82.13	67.62	61.71	73.25	**61.87**
HPTD-ELECTRA	84.75	81.70	67.19	60.91	71.39	58.41
HPTD-DeBERTaV3	85.68	**82.93**	68.35	62.19	**73.45**	61.27
$\text{GHLA}_{\text{RoBERTa}}$	**85.72**	82.92	**68.38**	**62.38**	73.34	61.29

Conclusion

- We proposed three new hierarchical text classification approaches which use the natural language understanding capabilities of pre-trained language models.
- We showed that the Hierarchy-aware Prompt Tuning for Discriminative PLMs (HPTD) approach effectively leverages the pre-trained knowledge of the language model.
- We showed that the global hierarchical label-wise attention mechanism (GHLA) uses the hierarchical class structure information to improve classification performance.
- We showed that using the features extracted from topic models does not always improve classification performance.
- We developed three new benchmark datasets in the domain of research publications.

Thank you!

Any questions?