Hierarchical Text Classification using Language Models with Global Label-wise Attention Mechanisms

Jaco du Toit[1,2] and Marcel Dunaiski[1,2]

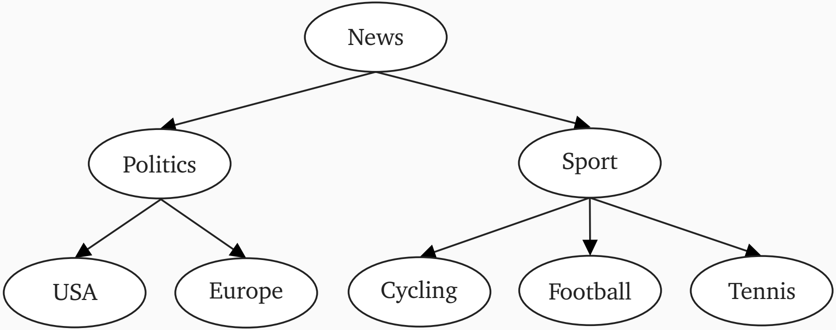
[1]Computer Science Division, Department of Mathematical Science
Stellenbosch University
[2]School for Data Science and Computational Thinking
Stellenbosch University

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

forward together
sonke siya phambili
saam vorentoe

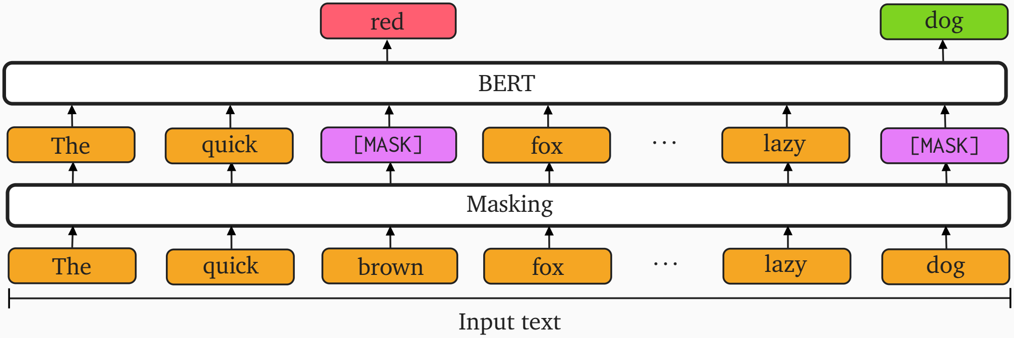
Hierarchical Text Classification

- Objective: Classify text documents into classes from a structured class hierarchy.
- Improves organisation and navigation of large document collections.
- Allows users to select the level of granularity that they prefer.



Transformer-based Language Models

- Trained through self-supervised learning tasks on large amounts of textual data.
- Attention mechanisms obtain contextually and semantically aware word embeddings.
- BERT: Uses the masked language modelling pre-training task.
- RoBERTa: Improved BERT architecture that is trained on more data for longer.



Label-wise Attention Mechanisms

- Label-wise attention mechanisms obtain label-specific document representations of the token representations obtained by the language model.
- Places more weight on the most important features for each class separately.
- We use two label-wise attention mechanisms to obtain attention weights:
  - Dot Product Attention (DPA):

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

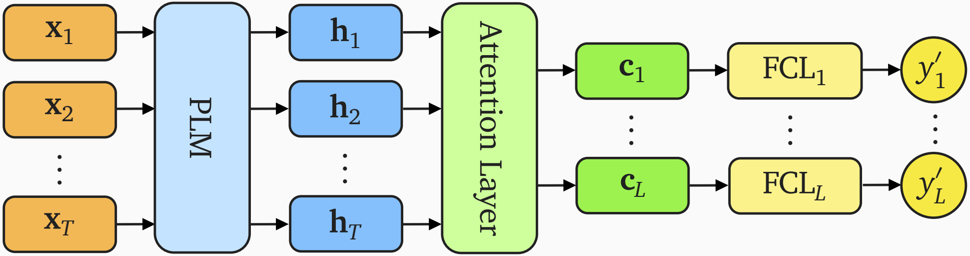
  - General Attention (GA):

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

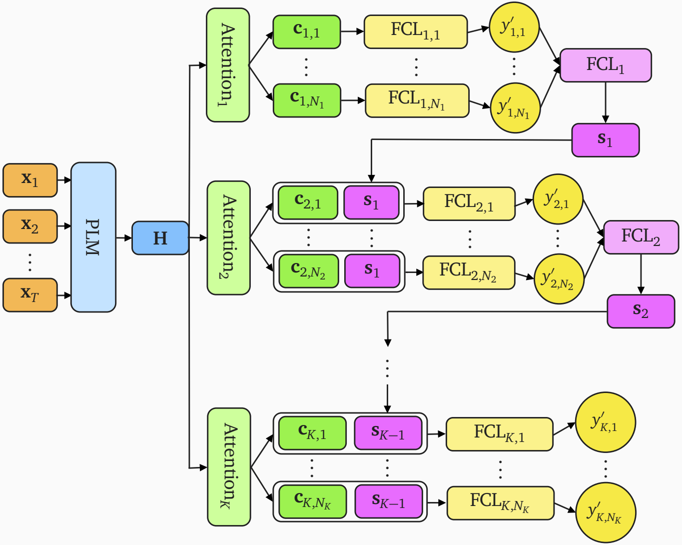
$$\boldsymbol{\alpha} = \mathrm{softmax}(\mathbf{U}_{\mathrm{GA}}\mathbf{Z}) \tag{3}$$

Model Architecture

- Text tokens (orange) are passed to the pre-trained language model which obtains representations for each token (blue).
- The token representations are used by the label-wise attention mechanism to obtain label-specific document representations (green).
- The label-wise document representations are used to obtain the confidence scores for the document belonging to each class (yellow).



Hierarchical Model Architecture



- Hierarchical label-wise attention (HLA): Separates the label-wise attention mechanisms for each level of the class hierarchy.
- Output at a level is used to obtain a prediction representation which is concatenated to the lower-level label-wise representations.
- Global hierarchical label-wise attention (GHLA): Extends DPA by concatenating all of the higher-level predictions to the label-wise document representations.

Experiments

- Perform experiments on three 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

- Evaluation metrics:
  - Micro-F1: Averages performance over all testing instances.
  - Macro-F1: Equally weighs performance for each class.

Main Results

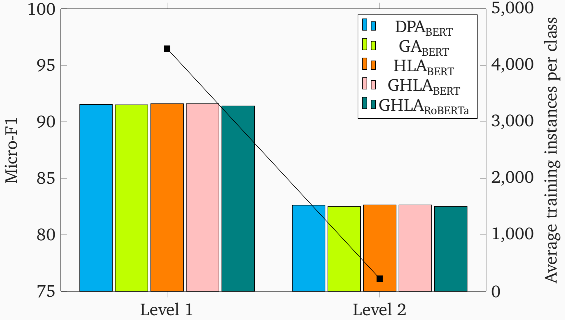
- GHLA generally outperforms the other label-wise attention mechanisms.
- Using RoBERTa significantly improves performance on two datasets.
- Using GHLA with RoBERTa outperforms previously proposed approaches on the RCV1-V2 and NYT datasets.

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**

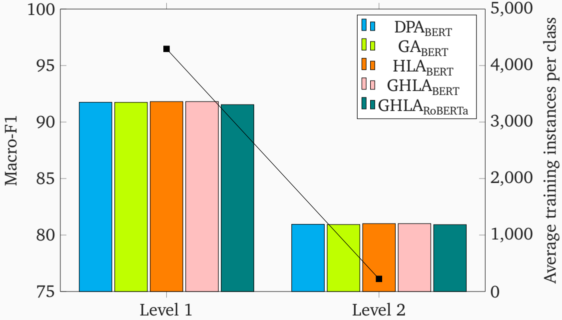
[1] Results obtained using twice the number of model parameters as the other approaches.

Level-wise Results

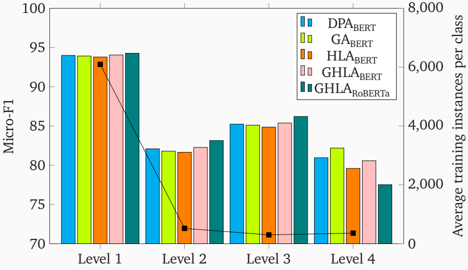
- We evaluate the classification performance at each level of the class hierarchy separately and determine the correlation with the average number of training instances.
- Classification performance generally decreases for the lower levels of the class hierarchy with fewer average training instances per class.



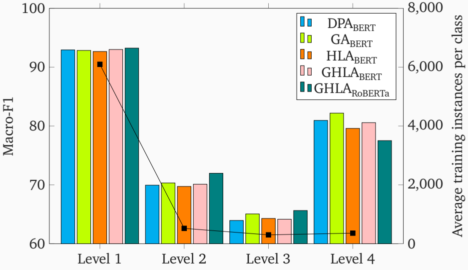
WOS Micro-F1.



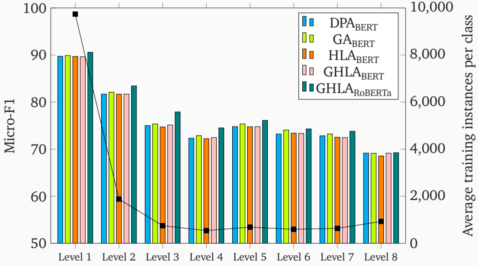
WOS Macro-F1.



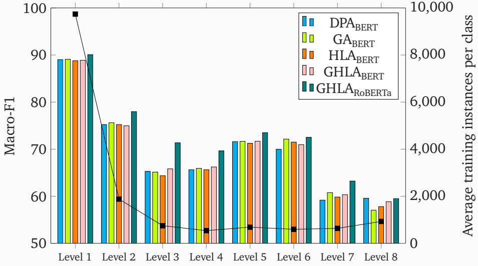
RCV1-V2 Micro-F1.



RCV1-V2 Macro-F1.



NYT Micro-F1.



NYT Macro-F1.

Low-resource Results

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

Conclusion

- Using label-wise attention mechanisms to fine-tune pre-trained language models is an effective approach for hierarchical text classification.
- Our label-wise attention mechanism effectively leverages the natural language understanding capabilities of the language model and the hierarchical class structure to improve classification performance.
- Using RoBERTa as the underlying language model generally improved classification performance over using BERT.
- RoBERTa significantly improved low-resource performance.

Thank you!

Any questions?