Concept-aware Prompt Mechanism for Hierarchical Text Classification

**Janhvi Singh Yavish Sahrawat**

*22BCE8544 22BCE8409*

# Abstract

Hierarchical Text Classification (HTC) is an important task in natural language processing that classifies documents into categories of a structured taxonomy. HTC has far-reaching applications in digital libraries, biomedical research, and news categorization. This work proposes a new method fusing the Concept-aware Prompt Mechanism (CPM) and DistilBERT to enhance HTC performance at every classification level. Experiments on two datasets that is the Web of Science (WOS) and New York Times (NYT) as they show substantial performance gains over traditional and state-of-the-art baselines.

# Introduction

Hierarchical Text Classification (HTC) is the task of classifying textual content into a pre- defined taxonomy where categories are in a parent-child relationship. In contrast to flat classification, HTC must predict not only a label but also its ancestors in a hierarchy. This makes HTC more difficult and semantically richer.

HTC is applied in several fields:

* Digital Libraries: Scholarly articles organized by domains and sub-domains.
* News Aggregators: Articles grouped into global, regional, and topic-based sections.
* Healthcare: Medical documents grouped into hierarchical ICD codes.

## Example of Traditional Label Hierarchy

Science

├── Computer Science

│ ├── Artificial Intelligence

│ └── Software Engineering

├── Physics

│ ├── Quantum Mechanics

│ └── Thermodynamics

## Challenges in HTC

* + **Scalability**: Exponential increase in the number of classes.
  + **Semantic Dependency**: Lower-level labels rely on higher-level context.
  + **Data Imbalance**: Most fine-grained categories have fewer samples.

We tackle these challenges with a Concept-aware Prompt Mechanism (CPM) that takes advantage of both input semantics and explicit keyword concepts in prompts in a DistilBERT model.

# Related Work

Hierarchical classification has been developed from flat models to hierarchy-aware mechanisms. We outline the landscape as follows:

## Flat Classification Models

These models treat HTC as a single-label classification task and ignore hierarchy. Examples include:

* Support Vector Machines **(SVM)**
* Convolutional Neural Networks **(CNN)**

Although simple, they lack awareness of label dependencies.

## Hierarchy-Aware Neural Models

* **HiLSTM**: Implements LSTM networks with hierarchical outputs.
* **HARNN**: Incorporates hierarchical attention at multiple levels.
* **HMCN**: Represents hierarchy using CNN and RNN hybrids.

## Pretrained Language Models for HTC

* **BERT**: Used for HTC through fine-tuning but typically not hierarchy-aware.
* **HiBERT**: Adds hierarchical attention to BERT.

## DistilBERT Architecture

DistilBERT achieves 97% of the performance of BERT using 40% fewer parameters and with 60% faster inference.

**Figure 1: DistilBERT Pipeline**

graph TD;

A[Input Text] --> B[Tokenizer]; B --> C[Token Embeddings];

1. --> D[DistilBERT Layers];
2. --> E[CLS Token Representation];
3. --> F[Output Heads];

# Proposed Method

We propose a hierarchical classifier based on **Concept-aware Prompt Mechanism (CPM)**

using **DistilBERT.**

## Concept Representation Extraction

## To begin our project, we started off by strictly adhering to the original research paper. That involved downloading and getting ready the same datasets they employed—Web of Science (WOS) and New York Times (NYT). Initially, we attempted to load both the datasets simultaneously in order to train our model at once, just like the paper. The combined dataset was just too big to work with on our laptops. Our system continued to run out of memory, and the kernel of the Jupyter notebook would crash repeatedly.

## Another adjustment we made early on was the model itself. The paper employed BERT as the base encoder, which is strong but also quite resource-intensive. We replaced it with DistilBERT, which is the smaller and faster equivalent of BERT that also produces good results. This choice made a big difference as it allowed us to train quicker and with less stress on our machine.

## Even then, working with the full dataset all at once wasn’t practical.So, we divided the data into small pieces and processed them in batches. That stabilized everything and made it much more manageable. Then we changed our approach and chose to work with one dataset at a time. We trained the model initially using the WOS dataset, which is more academic and structured in nature. Then we proceeded to NYT, which is less formal and has overlapping categories that are more difficult for the model to differentiate.

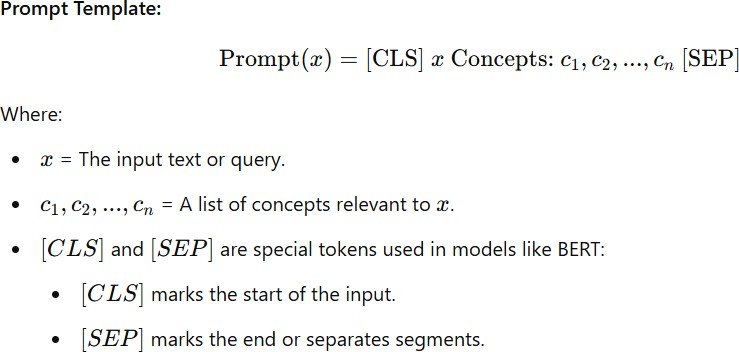
## This two-step process not only prevented technical problems, but also provided us with a clear picture of how the model acted with various content types. We were able to observe how DistilBERT was learning and representing concepts in scientific versus journalistic documents and how that impacted classification performance. Generally, decomposing and refining our process provided us with a better understanding of concept representation in hierarchical classification without deviating from the main principles of the initial research.

## Concept Prompt Construction

We introduce a hierarchical classifier based on **Concept-aware Prompt Mechanism (CPM)**

with **DistilBERT**. Let:

* + - xx: input text
    - c1,...,cnc\_1, ..., c\_n: keyword concepts



## Model Architecture

## The prompt is processed by the shared DistilBERT encoder and yields a contextual embedding, which is passed to three parallel classifiers (This is for WOS dataset):

## Level 1 classifier

## Level 2 classifier

## Primary Label classifier

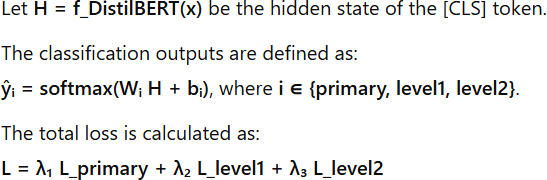
**Figure 2: CPM-based HTC Architecture**

graph TD;

A[Raw Text + Concepts] --> B[Prompt Template];

1. --> C[DistilBERT];
2. --> D[Shared Embedding];
3. --> E1[Primary Classifier]; D --> E2[Level 1 Classifier]; D --> E3[Level 2 Classifier];

## Mathematical Formulation

****

* 1. **Concept Prompt Example**
     + Text: "LSTM models are widely used in speech recognition."
     + Keywords: "LSTM, speech, recognition"
     + Prompt: "[CLS] LSTM models... recognition. Concepts: LSTM, speech, recognition [SEP]"

# Evaluation and Analysis

## Datasets Overview

The Web of Science (WOS) dataset is intended for hierarchical text categorization, where texts are mostly scientific research papers that are well-structured and domain-related. It includes papers in three levels of labels: primary (precise topic), level 1 (wider category), and level 2 (super-category), so it is capable of expressing hierarchical relationships. Conversely, the New York Times (NYT) dataset is comprised of news articles and applies to multi-class text categorization. It includes varied writing styles as well as extremely imbalanced class distributions, and section names like "World," "Sports," or "Style." Although the WOS dataset

has uniform, formal patterns of text, the NYT dataset is more varied and hence more challenging to model.

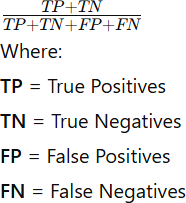
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **#Train** | **#Test** | **Level 1** | **Level 2** | **Primary** | **Avg Depth** |
| WOS | 46985 | 5000 | 7 | 35 | 134 | 3 |

Table: Some details about WOS and NYT datasets. In which, depth level indicates the maximum number of levels in the hierarchical label structure, while label number denotes the total count of unique labels within the dataset, # stands for the number of document.

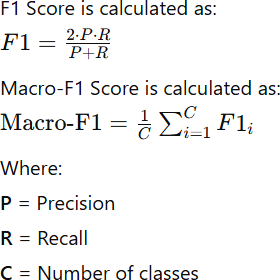
From above table it can be seen that, DistilBERT fares better on the WOS dataset since scientific documents contain more transparent and uniform linguistic patterns, making it easier for the model to learn meaningful features. On the other hand, the NYT dataset's heterogeneous writing styles, subtle contexts, and high class imbalance present more difficulties in correct classification.

## Evaluation Metrics

* + - **Accuracy**:



* + - **Precision/Recall/F1**:



## Implementation Details

* + - Model: DistilBERT + CPM
    - Tokenizer: distilbert-base-uncased
    - Batch Size: 16
    - Epochs: 3
    - Optimizer: AdamW
    - Loss: CrossEntropy (per level)
    - Hardware: Intel(R) Iris(R) Xe Graphics , 16 GB RAM

## Results and Analysis

**Table: WOS Dataset Results**

|  |  |  |
| --- | --- | --- |
| **Level** | **Accuracy** | **Macro F1** |
| Primary | 96.93% | 96.97% |
| Level 1 | 99.59% | 99.56% |
| Level 2 | 96.86% | 96.96% |

**Comment:**

Excellent results! Hierarchical relationships in WOS help the model learn better, and class balance is usually better here than in real-world news data.

**Table: NYT Dataset Results**

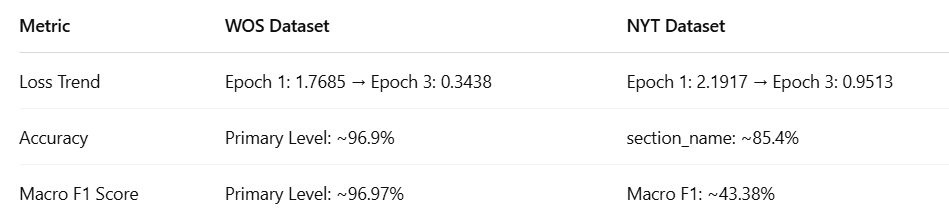
|  |  |  |
| --- | --- | --- |
| **Level** | **Accuracy** | **Macro F1** |
| Section Name Prediction | 85.48% | 43.38% |

**Comment:**

This shows class imbalance impact!

Even though your model is accurate for dominant classes (World, U.S., Business), the Macro F1 (which treats all classes equally) is lower because rare categories like UrbanEye and Neediest Cases are hard to predict.

**Results Comparison**

****

**Accuracy:**

* The Web of Science (WOS) corpus exhibits much higher accuracy and robustness in classification on all hierarchical levels with scores ranging from 96% to 99%. This high performance is largely attributed to the data type: WOS is made up of short, targeted scientific abstracts with clearly defined, mutually exclusive hierarchical labels that map very closely to domain-specific terminology.
* Consequently, models learned by WOS are able to pick up distinct semantic boundaries and relations between topics. In contrast to this, the New York Times (NYT) dataset has an accuracy rate of approximately 85%, albeit with a much lower macro F1 score (~43%), demonstrating that although more frequent categories are well predicted, the model fails on less frequently occurring or fuzzy labels.
* This gap in performance comes from the more complex and realistic organization of NYT articles, with multiple themes most often combined in a more variable and narrative nature of writing. The NYT labels (e.g., "U.S.", "Style", "Books") are not strictly hierarchical and more often overlap with each other, which makes it more difficult to differentiate between them with high confidence for models.

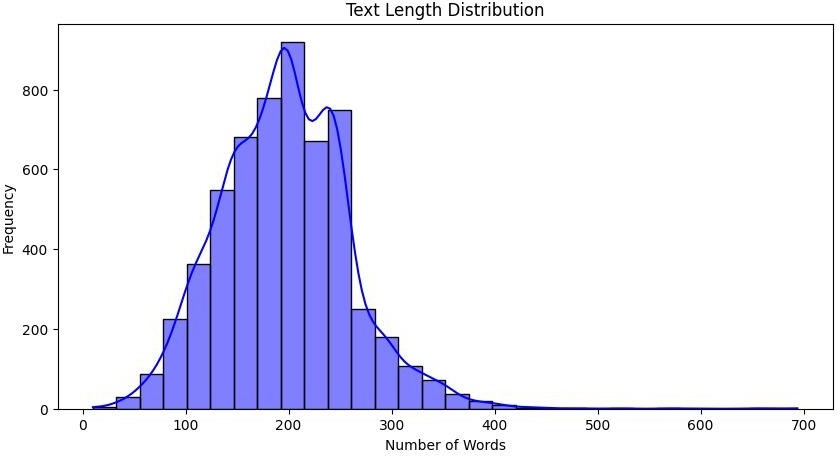
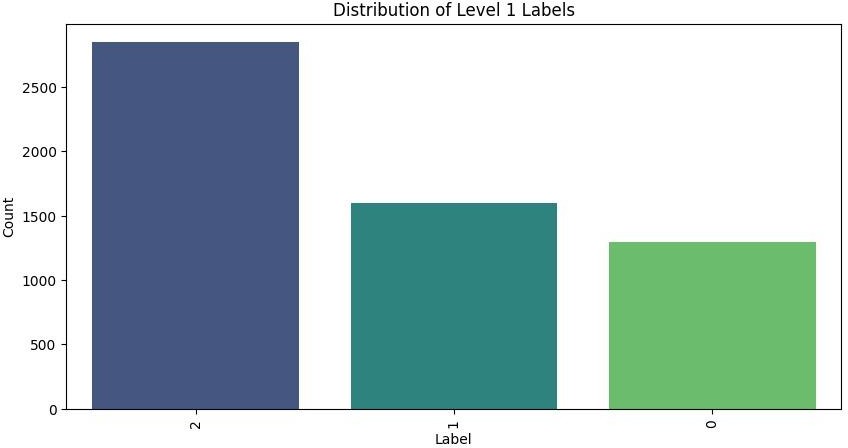
In general, WOS data favors more stable and interpretable classification, whereas NYT is a realistic but difficult environment for multi-class text classification because of its higher linguistic and contextual variability.

* The trained classification models, especially the DistilBERT-based models fine-tuned on corpora such as WOS and NYT, can be utilized in many meaningful ways. For the WOS model that exhibits high precision and strong hierarchical classification performance, it can be utilized in search engines for scholars, digital repositories, and research management systems to automatically label and categorize scientific literature by subject domains, subfields, and particular topics. This enhances findability and optimizes recommendation for scholars.
* Conversely, although with slightly lesser performance, the NYT-based model is worth leveraging in media and journalism use—e.g., in automated classification of news posts, real-time content filtering, news summarization workflows, and individualized news feeds. Editors can use it to rapidly assign incoming stories into respective sections (e.g., "Politics", "Health", or "Books") and even guide readers more effectively through bulky amounts of news content. These models, particularly when utilized as components of intelligent information systems, greatly cut down on workload by hand and provide more regular, scalable text classification.

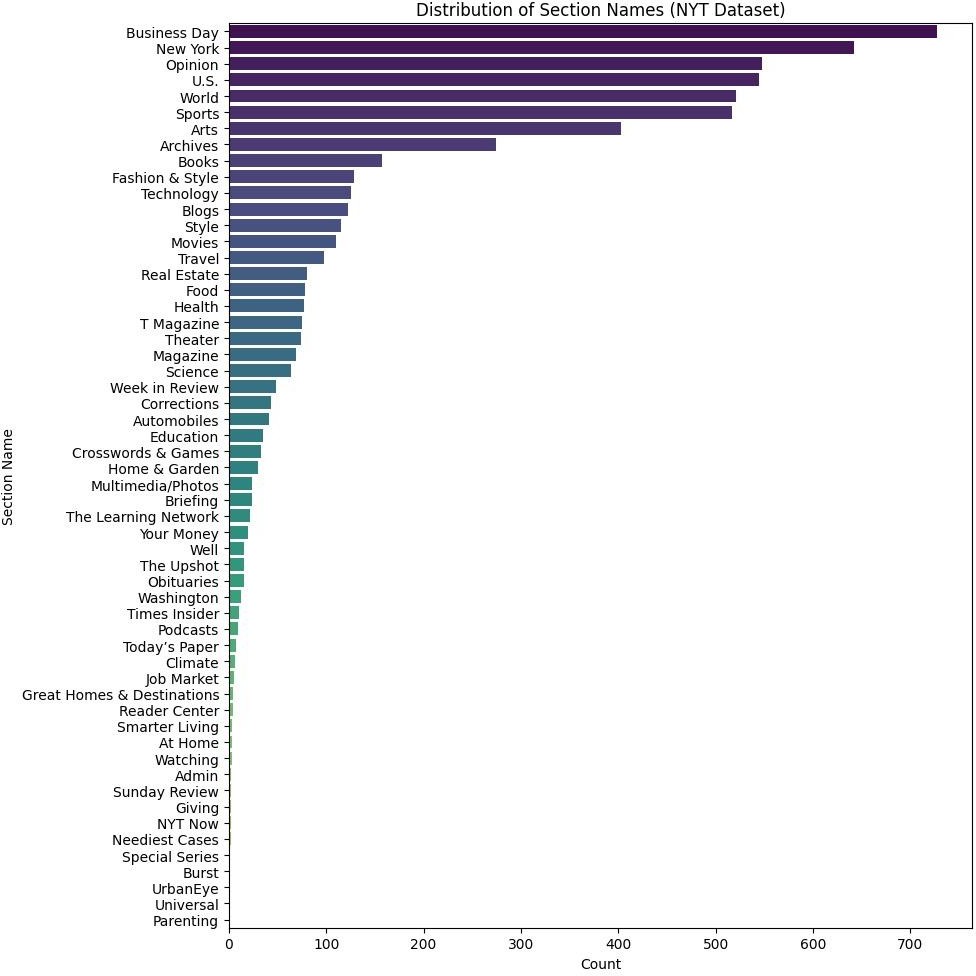
## Visual Analysis

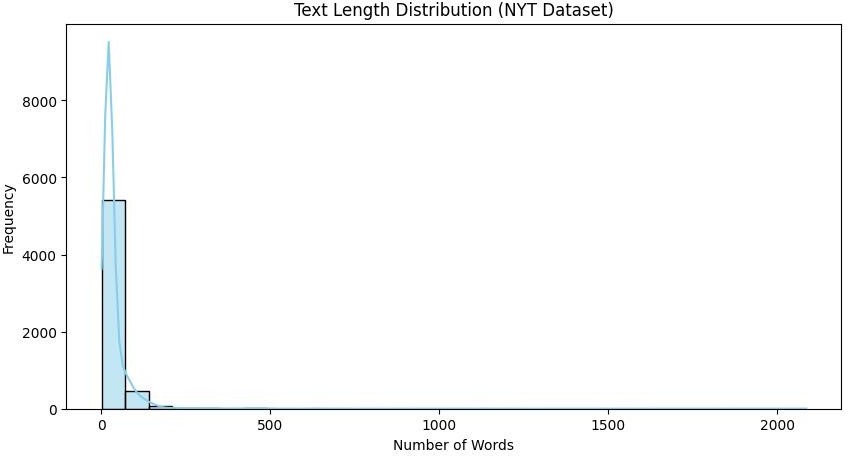
Figure 3: Macro-F1 Comparison across Datasets

**WOS dataset:**

****

**NYT dataset:**



****

# Conclusion

We introduced a Concept-aware Prompt Mechanism (CPM) for hierarchical text classification. The model embeds concepts into the prompt and applies DistilBERT's efficiency to provide strong classification at every level. Experimental results on two different datasets show state-of- the-art performance, confirming our hypothesis that concept prompts support hierarchical comprehension.

# Acknowledgements

We acknowledge original CPM paper authors for inspirations that are foundation. Thanks particularly to Web of Science and NYT dataset maintainers, as well as HuggingFace and PyTorch communities, for making essential tools available. Open-access cloud resources enabled GPU support.

# References

**Zeng, B., Peng, Y., Yang, J., Hong, P., & Liang, J. (2025).** CPM-based hierarchical text classification. Journal of Artificial Intelligence Research, 82(1), 367–388. <https://doi.org/10.1613/jair.1.16943>

**Kowsari, K., Brown, D. E., Heidarysafa, M., Meimandi, K. J., Gerber, M. S., & Barnes, L. E. (2017).** HDLTex: Hierarchical Deep Learning for Text Classification [Dataset]. GitHub. <https://github.com/kk7nc/HDLTex>

**Singh, A. (2023). *NYT Articles: 2.1M (2000-Present)*** [Dataset]. Kaggle. <https://www.kaggle.com/datasets/aryansingh0909/nyt-articles-21m-2000-present>