# Part III: Weakly-Supervised Text Classification

Automated Mining of Structured Knowledge from Text in the Era of Large Language Models
Yunyi Zhang, Ming Zhong, Siru Ouyang, Yizhu Jiao, Sizhe Zhou, Linyi Ding, Jiawei Han
Computer Science, University of Illinois Urbana-Champaign
KDD 2024 Tutorial, Aug 25, 2024

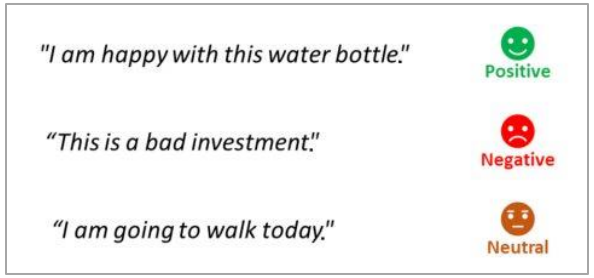
Tutorial Website:

# Outline

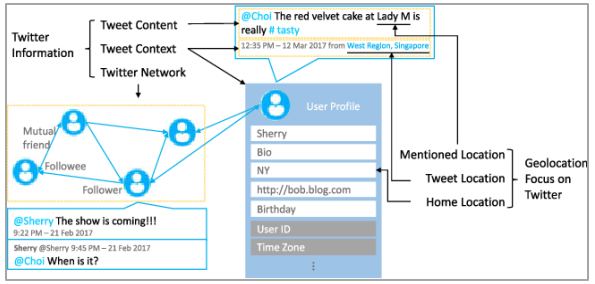
- What is weakly-supervised text classification and why do we care?
- Weakly-supervised flat text classification
  - ConWea [ACL'20], LOTClass [EMNLP'20], ClassKG [EMNLP'21], X-Class [NAACL'21], MEGClass [EMNLP'23], PIEClass [EMNLP'23], CARP [EMNLP'23]
- Weakly-supervised hierarchical text classification
  - WeSHClass [AAAI'19], TaxoClass [NAACL'21], TELEClass [arXiv'24]

# Text Classification

- Given a set of text units (e.g., documents, sentences) and a set of categories, the task is to assign relevant category/categories to each text unit
- Text Classification has a lot of downstream applications

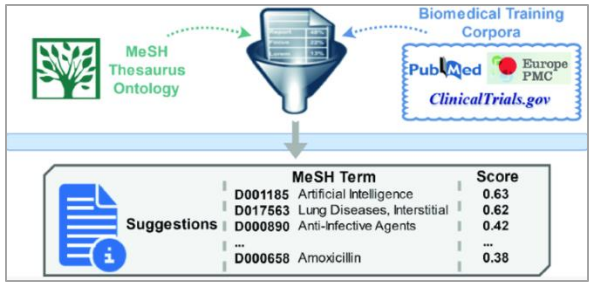



**Sentiment Analysis**

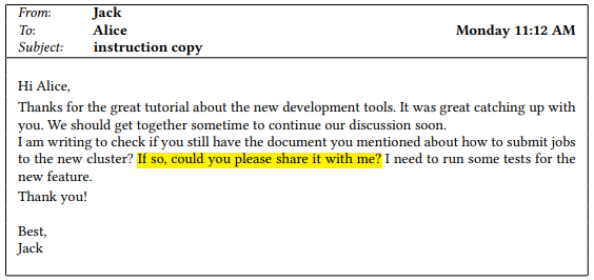



**Location Prediction**


**News Topic Classification**




**Paper Topic Classification**

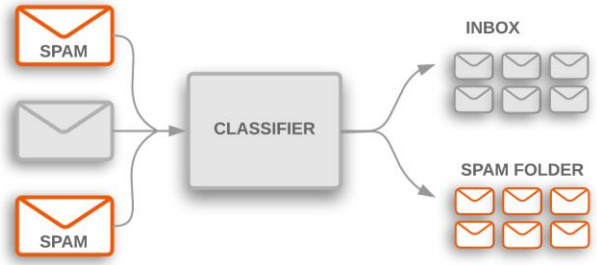



**Email Intent Identification**


**Hate Speech Detection**

# Different Text Classification Settings: Single-Label vs. Multi-Label

- **Single-label**: Each document belongs to one category.
  - E.g., Spam Detection




- **Multi-label**: Each document has multiple relevant labels.
  - E.g., Paper Topic Classification

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

**Abstract**

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5 (7.7 point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

**Related Topics**

Question answering | Language model | Natural language understanding | Named-entity recognition | SemEval | Inference | Winograd Schema Challenge | Sequence labeling

Artificial intelligence | Computer science | Transformer (machine learning model) | View Less

https://academic.microsoft.com/paper/2963341956/

# Different Text Classification Settings: Flat vs. Hierarchical

- **Flat**: All labels are at the same granularity level
  - E.g., Sentiment Analysis of E-Commerce Reviews (1-5 stars)

★★★★★ **It works, it's nice, comfortable, and easy to type on. Not loud (unless you're a key pounder)**

This keyboard works. It's comfortable, sensitive enough for touch typers, very quiet by comparison to other mechanicals (unless, of course, you're a 'key pounder'), and the lit keys are excellent for people like me who tend to prefer to work in a cave-like environment.

https://www.amazon.com/gp/product/B089YFHYYS/

- **Hierarchical**: Labels are organized into a hierarchy representing their parent-child relationship
  - E.g., Paper Topic Classification (the arXiv category taxonomy)

**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:1810.04805 **[cs.CL]**
(or arXiv:1810.04805v2 **[cs.CL]** for this version)

https://arxiv.org/abs/1810.04805

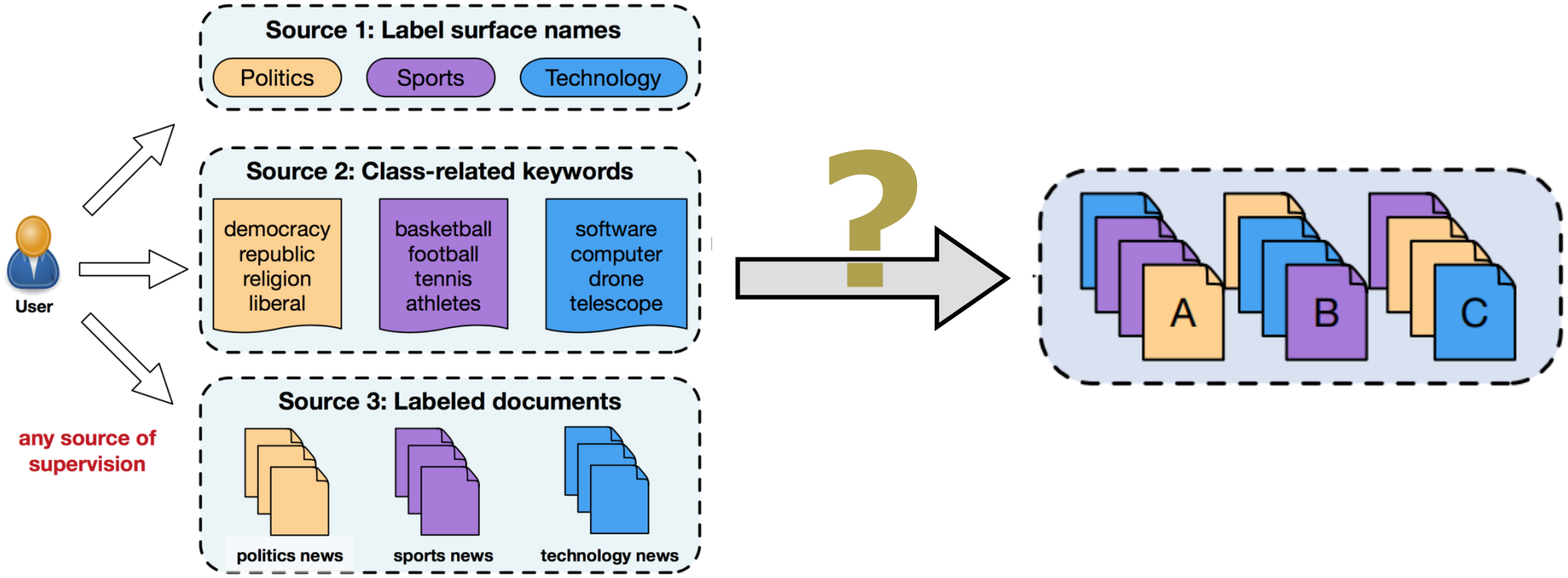
# Weakly-Supervised Text Classification: Motivation

- Supervised text classification models (especially recent deep neural models) rely on a significant number of manually labeled training documents to achieve good performance.
- Collecting such training data is usually expensive and time-consuming. In some domains (e.g., scientific papers), annotations must be acquired from domain experts, which incurs additional cost.

- While users cannot afford to label sufficient documents for training a deep neural classifier, they can provide a small amount of seed information:
  - Category names or category-related keywords
  - A small number of labeled documents

# Weakly-Supervised Text Classification: Definition

- Text classification without massive human-annotated training data
  - **Keyword-level weak supervision**: category names or a few relevant keywords
  - **Document-level weak supervision**: a small set of labeled docs

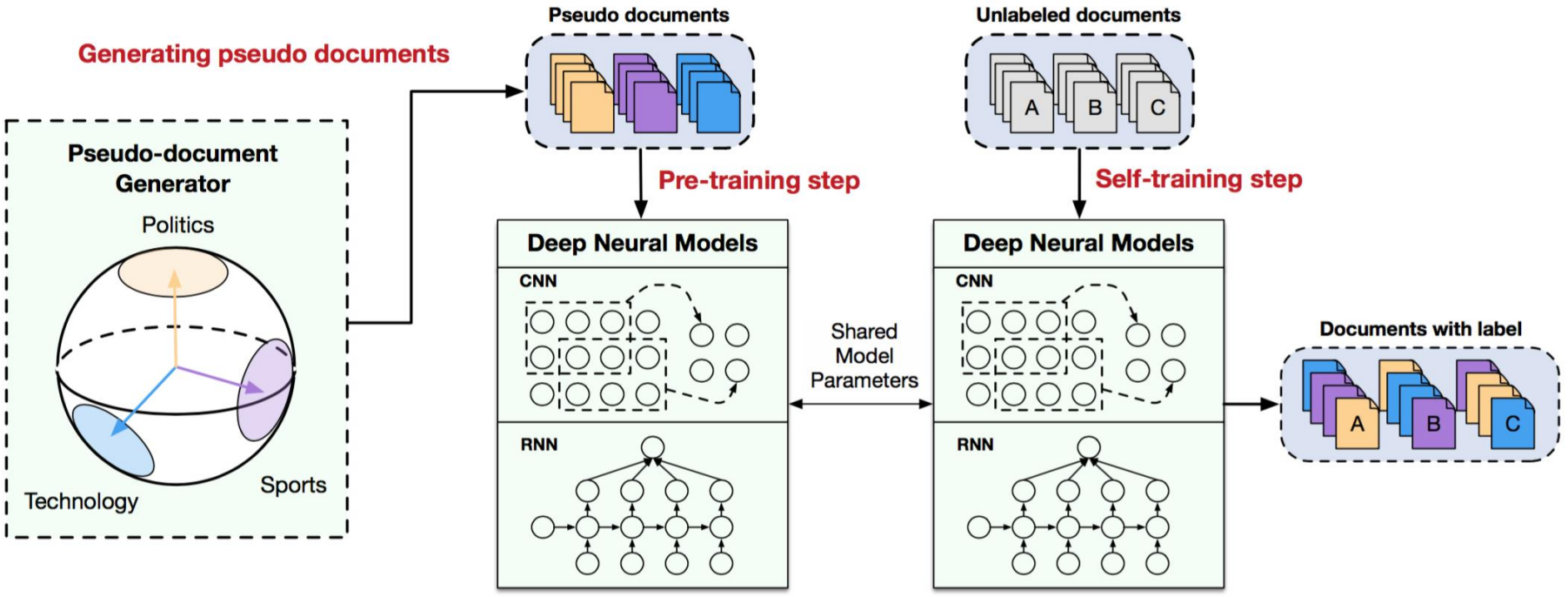


# General Ideas to Perform Weakly-Supervised Text Classification

- Joint representation learning
  - Put words, labels, and documents into the same latent space using embedding learning or pre-trained language models
- Pseudo training data generation
  - Retrieve some unlabeled documents or synthesize some artificial documents using text embeddings or contextualized representations
  - Give them pseudo labels to train a text classifier
- Transfer the knowledge of pre-trained language models to classification tasks

# An Example – WeSTClass

- Embed all words (including label names and keywords) into the same space
- Pseudo document generation: generate pseudo documents from seeds
- Self-training: train deep neural nets (CNN, RNN) with bootstrapping



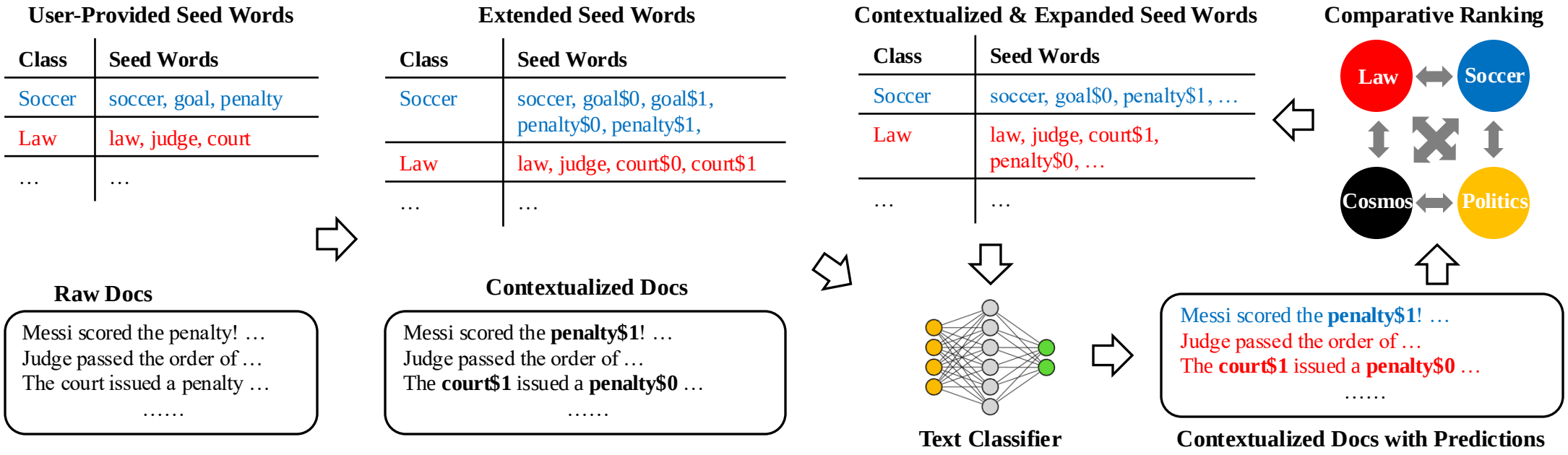

Meng, Y., Shen, J., Zhang, C., & Han, J. "Weakly-supervised neural text classification", CIKM'18.

# Outline

- What is weakly-supervised text classification and why do we care?
- Weakly-supervised flat text classification
  - ConWea [ACL'20], LOTClass [EMNLP'20], ClassKG [EMNLP'21], X-Class [NAACL'21], MEGClass [EMNLP'23], PIEClass [EMNLP'23], CARP [EMNLP'23]
- Weakly-supervised hierarchical text classification
  - WeSHClass [AAAI'19], TaxoClass [NAACL'21], TELEClass [arXiv'24]

# ConWea: Disambiguating User-Provided Keywords

- User-provided seed words may be ambiguous.
  - Messi scored the penalty. ---> Soccer
  - John was issued a death penalty. ---> Law
- Disambiguate the "senses" by clustering the contextualized representations




Mekala, D. & Shang, J. "Contextualized Weak Supervision for Text Classification", ACL'20.

# LOTClass: Find Similar Meaning Words with Label Names

- Find topic words based on label names
  - Overcome the low semantic coverage of label names
- Use language models to predict what words can replace the label names
  - Interchangeable words are likely to have similar meanings

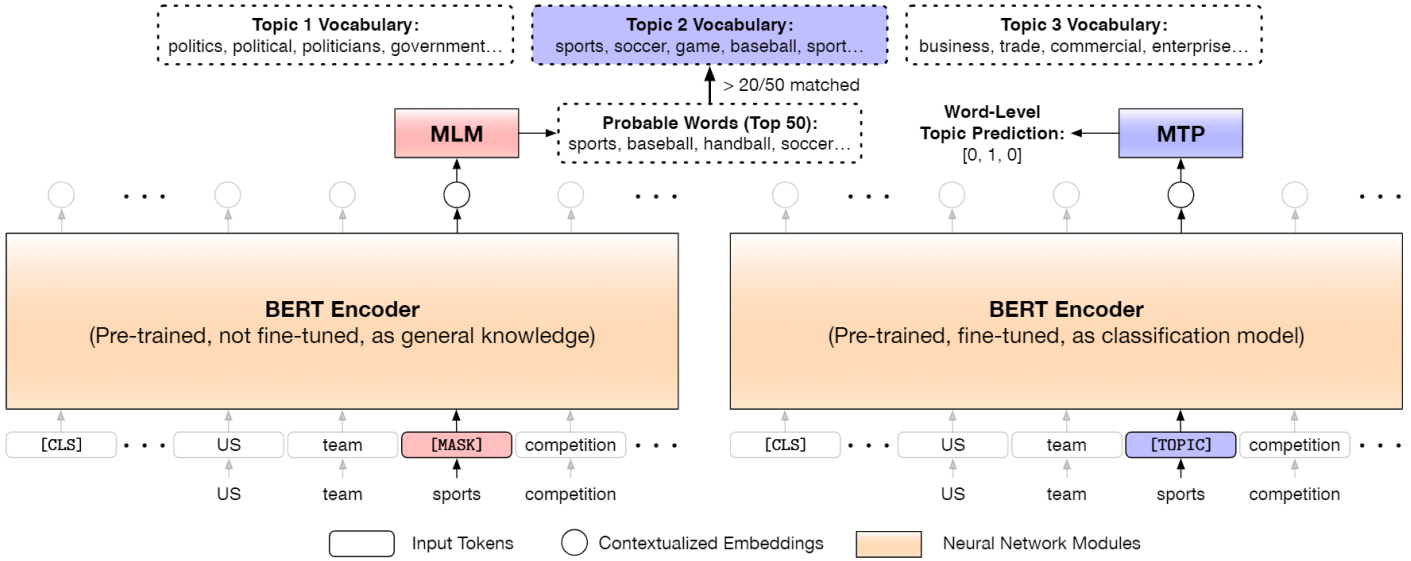
| Sentence | Language Model Prediction |
|---|---|
| The oldest annual US team **sports** competition that includes professionals is not in baseball, or football or basketball or hockey. It's in soccer. | sports, baseball, handball, soccer, basketball, football, tennis, sport, championship, hockey, . . . |
| Samsung's new SPH-V5400 mobile phone **sports** a built-in 1-inch, 1.5-gigabyte hard disk that can store about 15 times more data than conventional handsets, Samsung said. | has, with, features, uses, includes, had, is, contains, featured, have, incorporates, requires, offers, . . . |

Table 1: BERT language model prediction (sorted by probability) for the word to appear at the position of "sports" under different contexts. The two sentences are from *AG News* corpus.

Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., & Han, J. "Text Classification Using Label Names Only: A Language Model Self-Training Approach", EMNLP'20.

# LOTClass: Contextualized Word-Level Topic Prediction

- Context-free matching of topic words is inaccurate
  - "Sports" does not always imply the topic "sports"
- Contextualized topic prediction:
  - Predict a word's implied topic under specific contexts
  - We regard a word as "topic indicative" only when its top replacing words have enough overlap with the topic vocabulary.



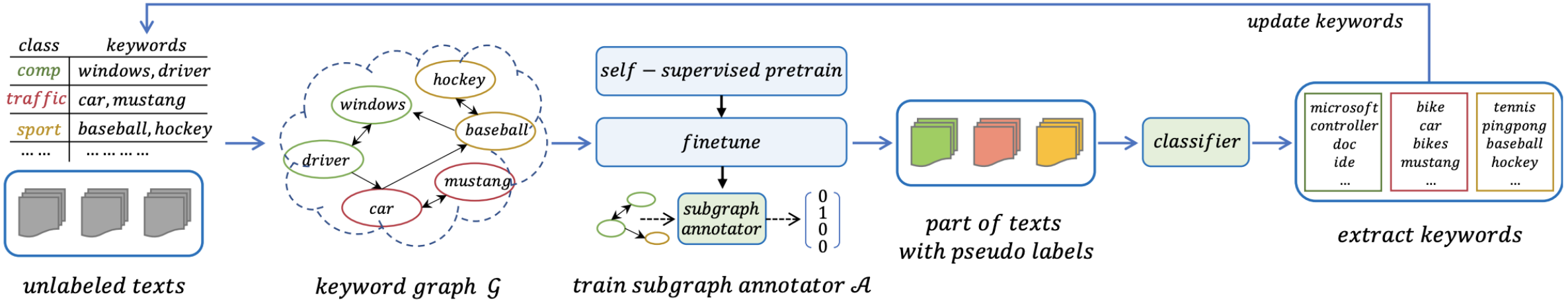
# LOTClass: Experiment Results

- Achieve around 90% accuracy on four benchmark datasets by only using at most 3 words (1 in most cases) per class as the label name
  - Outperforming previous weakly-supervised approaches significantly
  - Comparable to state-of-the-art semi-supervised models

| Supervision Type | Methods | AG News | DBPedia | IMDB | Amazon |
|---|---|---|---|---|---|
| Weakly-Sup. | **Dataless** (Chang et al., 2008) | 0.696 | 0.634 | 0.505 | 0.501 |
| | **WeSTClass** (Meng et al., 2018) | 0.823 | 0.811 | 0.774 | 0.753 |
| | **BERT w. simple match** | 0.752 | 0.722 | 0.677 | 0.654 |
| | **Ours w/o. self train** | 0.822 | 0.850 | 0.844 | 0.781 |
| | **Ours** | **0.864** | **0.889** | **0.894** | **0.906** |
| Semi-Sup. | **UDA** (Xie et al., 2019) | 0.869 | 0.986 | 0.887 | 0.960 |
| Supervised | **char-CNN** (Zhang et al., 2015) | 0.872 | 0.983 | 0.853 | 0.945 |
| | **BERT** (Devlin et al., 2019) | 0.944 | 0.993 | 0.937 | 0.972 |

# ClassKG: Text Classification with Keyword Graph

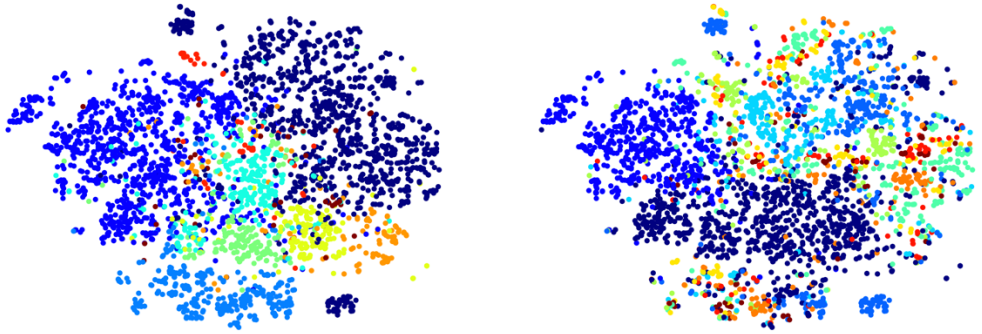
- Existing methods do not consider the correlation among keywords
- ClassKG exploits the correlation among keywords by GNN over a keyword graph, and converts the task to annotating subgraphs
- Iteratively train text classifier and re-extract keywords to update the keyword graph




Zhang, L., Ding, j., Xu, Y., Liu, Y., Zhou, S. "Weakly-supervised Text Classification Based on Keyword Graph", EMNLP'21.

# X-Class: Class-Oriented BERT Representations

- A simple idea for text classification
  - Learn representations for documents
  - Set the number of clusters as the number of classes
  - Hope their clustering results are almost the same as the desired classification
- However, the same corpus could be classified differently



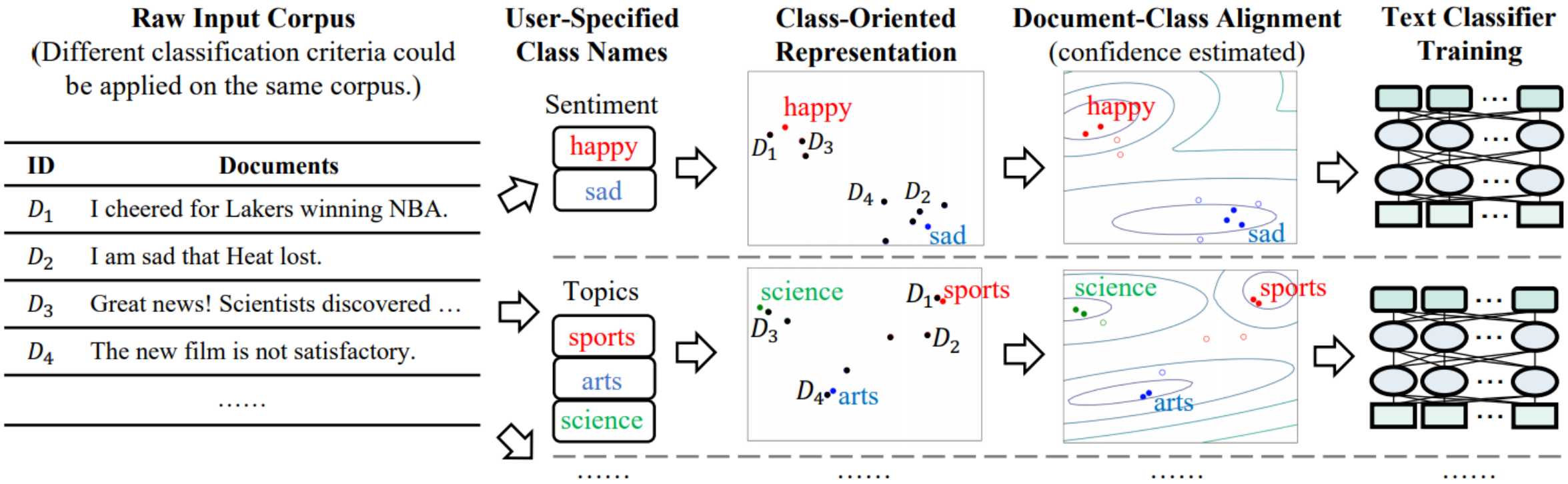
(a) NYT-Topics (b) NYT-Locations

Figure 1: Visualizations of News using Average BERT Representations. Colors denote different classes.

Wang, Z., Mekala, D., & Shang, J. "X-Class: Text Classification with Extremely Weak Supervision", NAACL'21.

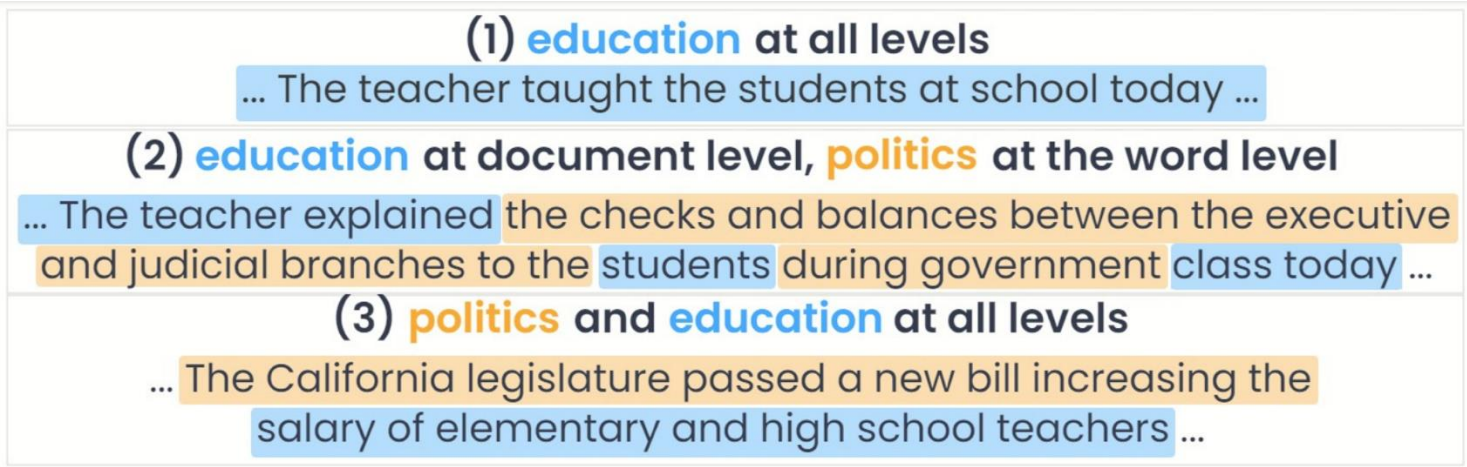
# X-Class: Class-Oriented BERT Representations

- Clustering for classification based on class-oriented representations



# MEGClass: Mutually-Enhancing Text Granularities

- Existing work treats different levels of text granularity (documents, sentences, or words) **independently.**
  - This disregards (1) **inter-granularity class disagreements** (e.g., scientific words in a political document) and (2) the **context** identifiable exclusively through **joint extraction**.

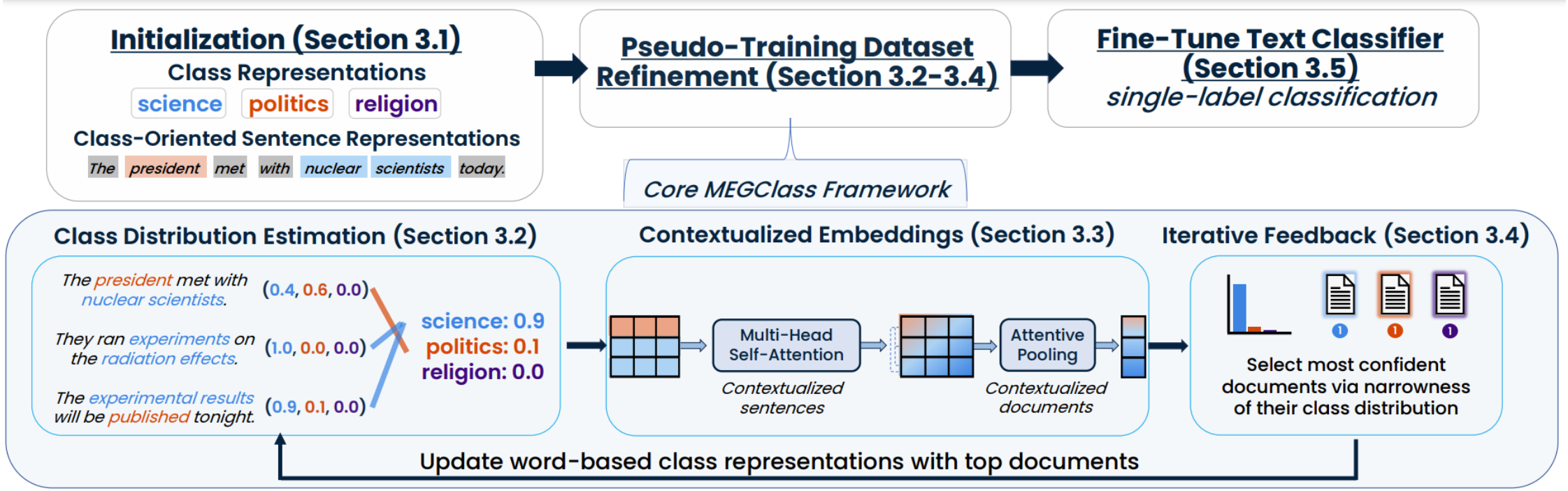



These are the three document types featured within a corpus. While existing methods can only distinguish between (1) and (3), MEGClass *addresses all three types* as well as *minimizes (3)* from the constructed pseudo-training dataset.

Kargupta, P., Komarlu, T., Yoon, S., Wang, X., Han, J. "MEGClass: Extremely Weakly Supervised Text Classification via Mutually-Enhancing Text Granularities", EMNLP'23.

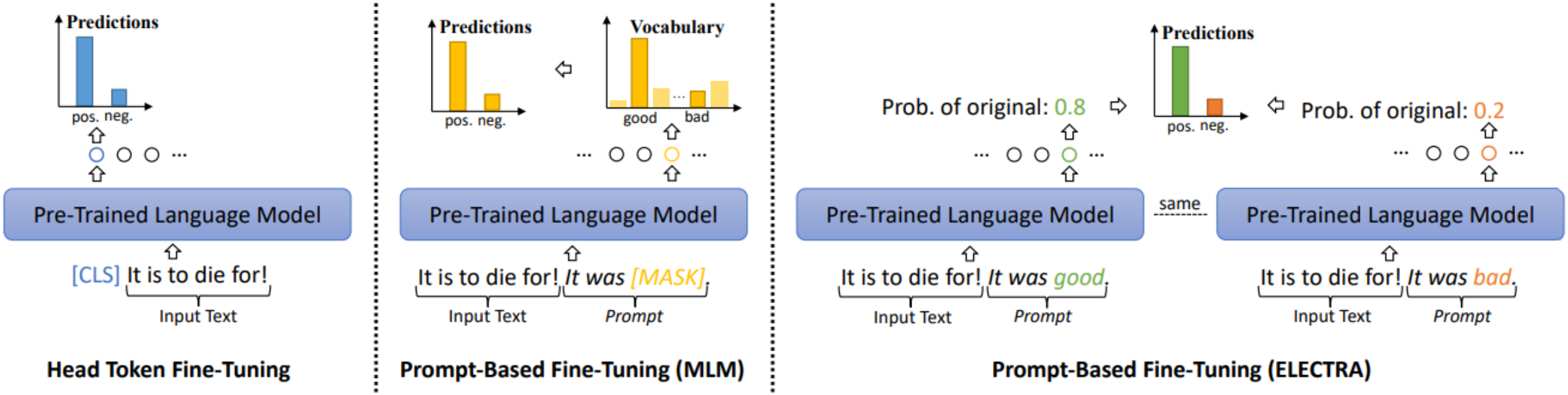
# MEGClass: Mutually-Enhancing Text Granularities

- **Output:** A contextualized document representation that captures the most discriminative class indicators
  - Jointly consider a text's most class-indicative **words**, **sentences**, and **document**-level context.
  - By preserving the heterogeneity of potential classes, MEGClass selects the most **informative class-indicative documents** as *iterative feedback* → enhances the initial word-based class representations.



# PIEClass: Prompt-based Fine-tuning for Text Classification

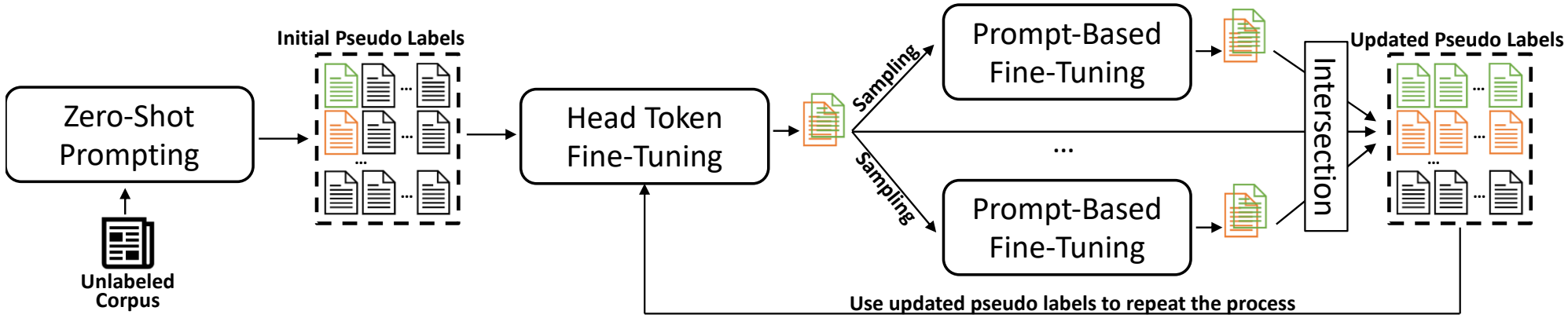
- **Head token fine-tuning** randomly initializes a linear classification head and directly predicts class distribution using the [CLS] token, which needs a substantial amount of training data.
- **Prompt-based fine-tuning for MLM-based PLM** converts the document into the masked token prediction problem by reusing the pre-trained MLM head.
- **Prompt-based fine-tuning for ELECTRA-style PLM** converts documents into the replaced token detection problem by reusing the pre-trained discriminative head.




Zhang, Y., Jiang, M., Meng, Y., Zhang, Y., & Han, J. "PIEClass: Weakly-Supervised Text Classification with Prompting and Noise-Robust Iterative Ensemble Training", EMNLP'23.

# PIEClass: Integrating Head Token & Prompt-based Fine-tuning

- Why do we need prompts to get pseudo training data?
  - Simple keyword matching may induce errors. We use prompts to contextualize the documents.
  - E.g., "*die*" is a negative word, but a food review "It is to *die* for!" implies a strong positive sentiment.
- **Noise-Robust Training with Iterative Ensemble**
  - Uses most confident predictions to improve and expand pseudo labels iteratively
  - Difference with semi-supervised self-training: potential noise in the initial pseudo labels
    - Noise-robustness: use two PLM fine-tuning methods as two views of data with model ensemble




**(1) Zero-Shot Prompting for Pseudo Label Acquisition**

**(2) Noise-Robust Training with Iterative Ensemble**

# PIEClass: Experiment Results

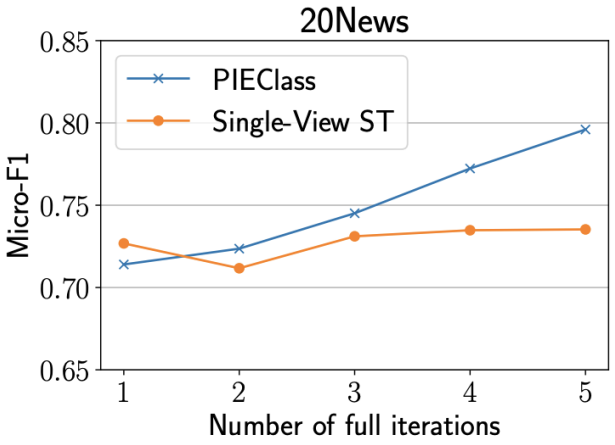
- PIEClass is on par with the fully supervised text classifier on sentiment analysis datasets (i.e., Yelp and IMDB).

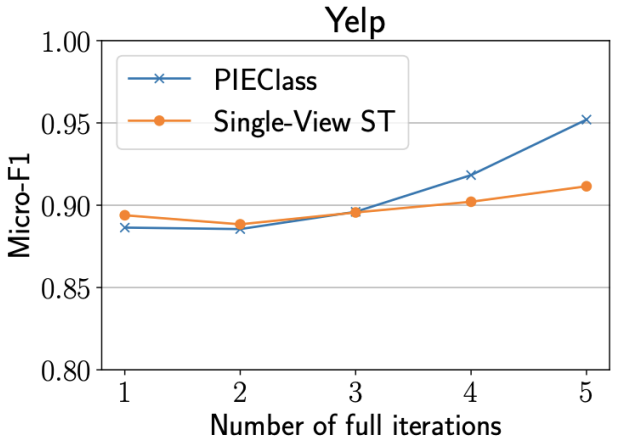
| Methods | AGNews | 20News | NYT-Topics | NYT-Fine | Yelp | IMDB | Amazon |
|---|---|---|---|---|---|---|---|
| **WeSTClass** | 0.823/0.821 | 0.713/0.699 | 0.683/0.570 | 0.739/0.651 | 0.816/0.816 | 0.774/- | 0.753/- |
| **ConWea** | 0.746/0.742 | 0.757/0.733 | 0.817/0.715 | 0.762/0.698 | 0.714/0.712 | -/- | -/- |
| **LOTClass** | 0.869/0.868 | 0.738/0.725 | 0.671/0.436 | 0.150/0.202 | 0.878/0.877 | 0.865/- | 0.916/- |
| **XClass** | 0.857/0.857 | 0.786/0.778 | 0.790/0.686 | 0.857/0.674 | 0.900/0.900 | -/- | -/- |
| **ClassKG**† | 0.881/0.881 | 0.811/**0.820** | 0.721/0.658 | 0.889/0.705 | 0.918/0.918 | 0.888/0.888 | 0.926/- |
| **RoBERTa (0-shot)** | 0.581/0.529 | 0.507/0.445‡ | 0.544/0.382 | -/-‡ | 0.812/0.808 | 0.784/0.780 | 0.788/0.783 |
| **ELECTRA (0-shot)** | 0.810/0.806 | 0.558/0.529 | 0.739/0.613 | 0.765/0.619 | 0.820/0.820 | 0.803/0.802 | 0.802/0.801 |
| **PIEClass** | | | | | | | |
| **ELECTRA+BERT** | 0.884/0.884 | 0.789/0.791 | 0.807/0.710 | 0.898/0.732 | 0.919/0.919 | 0.905/0.905 | 0.858/0.858 |
| **RoBERTa+RoBERTa** | **0.895/0.895** | 0.755/0.760‡ | 0.760/0.694 | -/-‡ | 0.920/0.920 | 0.906/0.906 | 0.912/0.912 |
| **ELECTRA+ELECTRA** | 0.884/0.884 | **0.816**/0.817 | **0.832/0.763** | **0.910/0.776** | **0.957/0.957** | **0.931/0.931** | **0.937/0.937** |
| **Fully Supervised** | 0.940/0.940 | 0.965/0.964 | 0.943/0.899 | 0.980/0.966 | 0.957/0.957 | 0.945/- | 0.972/- |

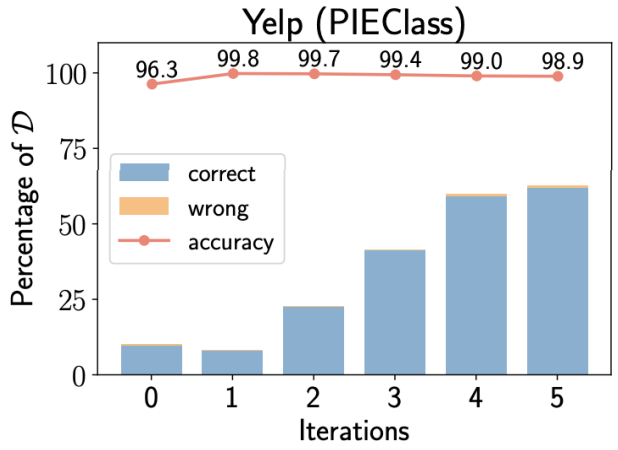
- Why PIEClass can achieve similar performance to the fully supervised method?
  - Annotation errors can affect the fully supervised model if used as training data

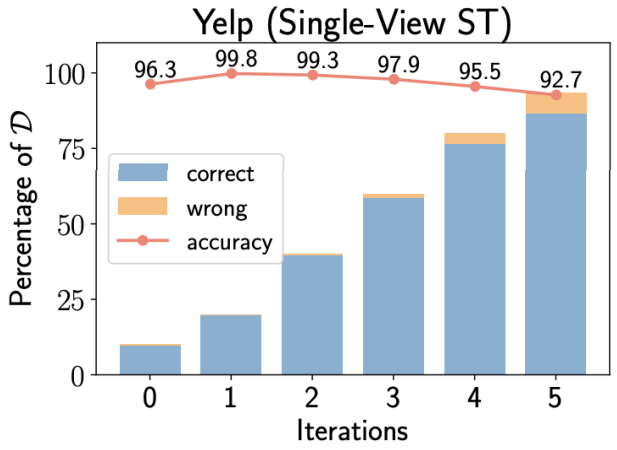
# PIEClass: Study of the Iterative Process

- We show the performance and pseudo label quality of PIEClass and a standard self-training approach for iterations 1 to 5
- **Observation**:
  - Pseudo label quality is more crucial when there are more pseudo labels
  - PIEClass does not blindly expand pseudo labels and can keep the quality of pseudo labels during the expansion process






23

# Zero-Shot Text Classification with LLMs

- Naively prompting LLMs with in-context learning significantly underperform fine-tuned models for text classification
  - Require reasoning abilities to resolve complex linguistic phenomenon
  - The number of ICL examples is limited
- **CARP:** Clue And Reasoning Prompting
  - Step 1: collect local fact evidence as clues such as keywords
  - Step 2: prompting LLMs to go beyond superficial keywords to mine deeper perspectives
  - Step 3: LLM makes the decision based on the reasoning process

Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G. "Text Classification via Large Language Models", EMNLP'23.

# CARP: Clue And Reasoning Prompting

This is an overall sentiment classifier for movie reviews.
First, list CLUES (i.e., keywords, phrases, contextual information, semantic relations, semantic meaning, tones, references) that support the sentiment determination of input..
Second, deduce the diagnostic REASONING process from premises (i.e., clues, input) that supports the INPUT sentiment determination (Limit the number of words to 130).
Third, based on clues, reasoning and input, determine the overall SENTIMENT of INPUT as Positive or Negative.

INPUT: press the delete key

CLUES: delete key
REASONING: The phrase "delete key" implies an action of removing something, which could be interpreted as a negative sentiment.
SENTIMENT: Negative ✓

(a) CARP Zero-shot

This is an overall sentiment classifier for movie reviews.
First, present CLUES (i.e., keywords, phrases, contextual information, semantic relations, semantic meaning, tones, references) that support the sentiment determination of input.
Second, deduce a diagnostic REASONING process from premises (i.e., clues, input) that supports the sentiment determination (Limit the number of words to 130).
Third, determine the overall SENTIMENT of INPUT as Positive or Negative considering CLUES, the REASONING process and the INPUT.

INPUT: noyce films it more as a shocking history lesson than as drama.
CLUES: - Positive: "more," "shocking," "history lesson," "drama." - Negative: None.
REASONING: The input contains positive words such as "more," "shocking," and "drama," which suggest that the sentiment is positive. Additionally, the phrase "history lesson" implies that the film is educational and informative, which is generally seen as a positive thing. Therefore, the sentiment of the input is positive.
SENTIMENT: Positive

INPUT: press the delete key
CLUES: - Positive: None. - Negative: "delete key."
REASONING: The input contains the phrase "delete key," which implies that something is being removed or erased. This is generally seen as a negative action, suggesting that the sentiment of the input is negative.
SENTIMENT: Negative ✓

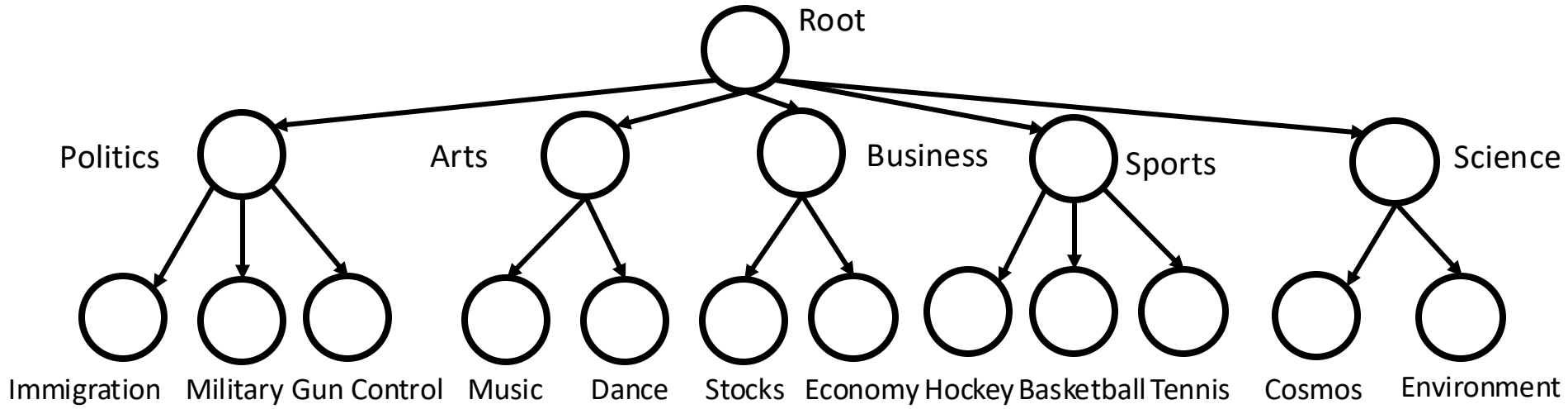
(b) CARP Few-shot (k=1)

# Outline

- What is weakly-supervised text classification and why do we care?
- Weakly-supervised flat text classification
  - ConWea [ACL'20], LOTClass [EMNLP'20], ClassKG [EMNLP'21], X-Class [NAACL'21], MEGClass [EMNLP'23], PIEClass [EMNLP'23], CARP [EMNLP'23]
- Weakly-supervised hierarchical text classification
  - WeSHClass [AAAI'19], TaxoClass [NAACL'21], TELEClass [arXiv'24]

# WeSHClass: Weakly-Supervised Hierarchical Text Classification

- The hierarchy has a **tree** structure. Each document is associated with **one path** starting from the root node. (E.g., the main subject of each arXiv paper.)

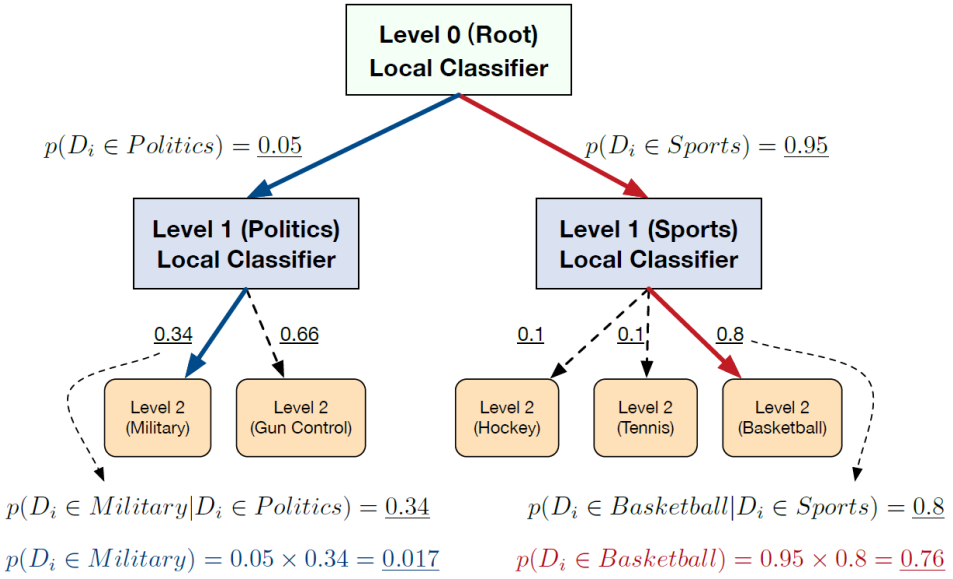



- Keyword-level weak supervision: The name of each node in the taxonomy, or a few keywords for each leaf category
- Document-level weak supervision: A few labeled documents for each leaf category

Meng, Y., Shen, J., Zhang, C., & Han, J. "Weakly-Supervised Hierarchical Text Classification", AAAI'19.

# Hierarchical Classification Model

- Local Classifier Per Node
  - Essentially a flat classification task
  - Follow WeSTClass
- Global Classifier Per Level
  - At each level $k$ in the class taxonomy, construct a global classifier by ensembling all local classifiers from root to level $k$

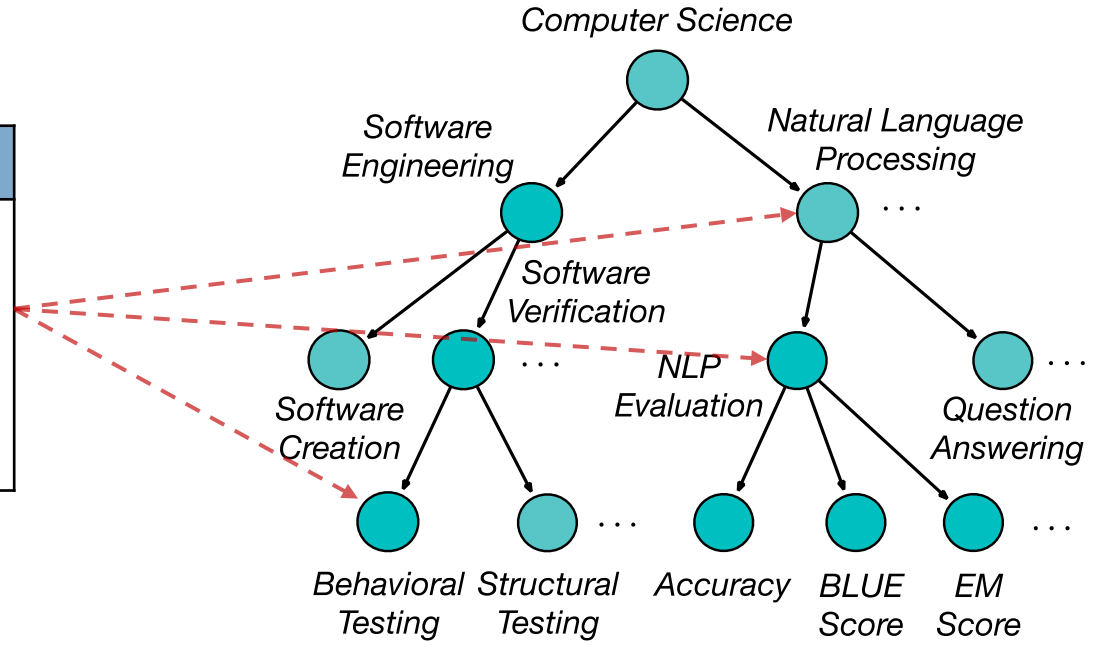



| Methods | NYT | | | | arXiv | | | | Yelp Review | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | KEYWORDS | | DOCS | | KEYWORDS | | DOCS | | KEYWORDS | | DOCS | |
| | Macro | Micro | Macro Avg. (Std.) | Micro Avg. (Std.) | Macro | Micro | Macro Avg. (Std.) | Micro Avg. (Std.) | Macro | Micro | Macro Avg. (Std.) | Micro Avg. (Std.) |
| Hier-Dataless | 0.593 | 0.811 | - | - | 0.374 | 0.594 | - | - | 0.284 | 0.312 | - | - |
| Hier-SVM | - | - | 0.142 (0.016) | 0.469 (**0.012**) | - | - | 0.049 (**0.001**) | 0.443 (**0.006**) | - | - | 0.220 (0.082) | 0.310 (0.113) |
| CNN | - | - | 0.165 (0.027) | 0.329 (0.097) | - | - | 0.124 (0.014) | 0.456 (0.023) | - | - | 0.306 (0.028) | 0.372 (0.028) |
| WeSTClass | 0.386 | 0.772 | 0.479 (0.027) | 0.728 (0.036) | 0.412 | 0.642 | 0.264 (0.016) | 0.547 (0.009) | 0.348 | 0.389 | 0.345 (0.027) | 0.388 (0.033) |
| No-global | 0.618 | 0.843 | 0.520 (0.065) | 0.768 (0.100) | 0.442 | 0.673 | 0.264 (0.020) | 0.581 (0.017) | 0.391 | 0.424 | 0.369 (0.022) | 0.403 (0.016) |
| No-vMF | 0.628 | 0.862 | 0.527 (0.031) | 0.825 (0.032) | 0.406 | 0.665 | 0.255 (0.015) | 0.564 (0.012) | 0.410 | 0.457 | 0.372 (0.029) | 0.407 (0.015) |
| No-self-train | 0.550 | 0.787 | 0.491 (0.036) | 0.769 (0.039) | 0.395 | 0.635 | 0.234 (0.013) | 0.535 (0.010) | 0.362 | 0.408 | 0.348 (0.030) | 0.382 (0.022) |
| Our method | **0.632** | **0.874** | **0.532** (**0.015**) | **0.827** (**0.012**) | **0.452** | **0.692** | **0.279** (0.010) | **0.585** (0.009) | **0.423** | **0.461** | **0.375** (**0.021**) | **0.410** (**0.014**) |

# Weakly-supervised Multi-label Hierarchical Text Classification

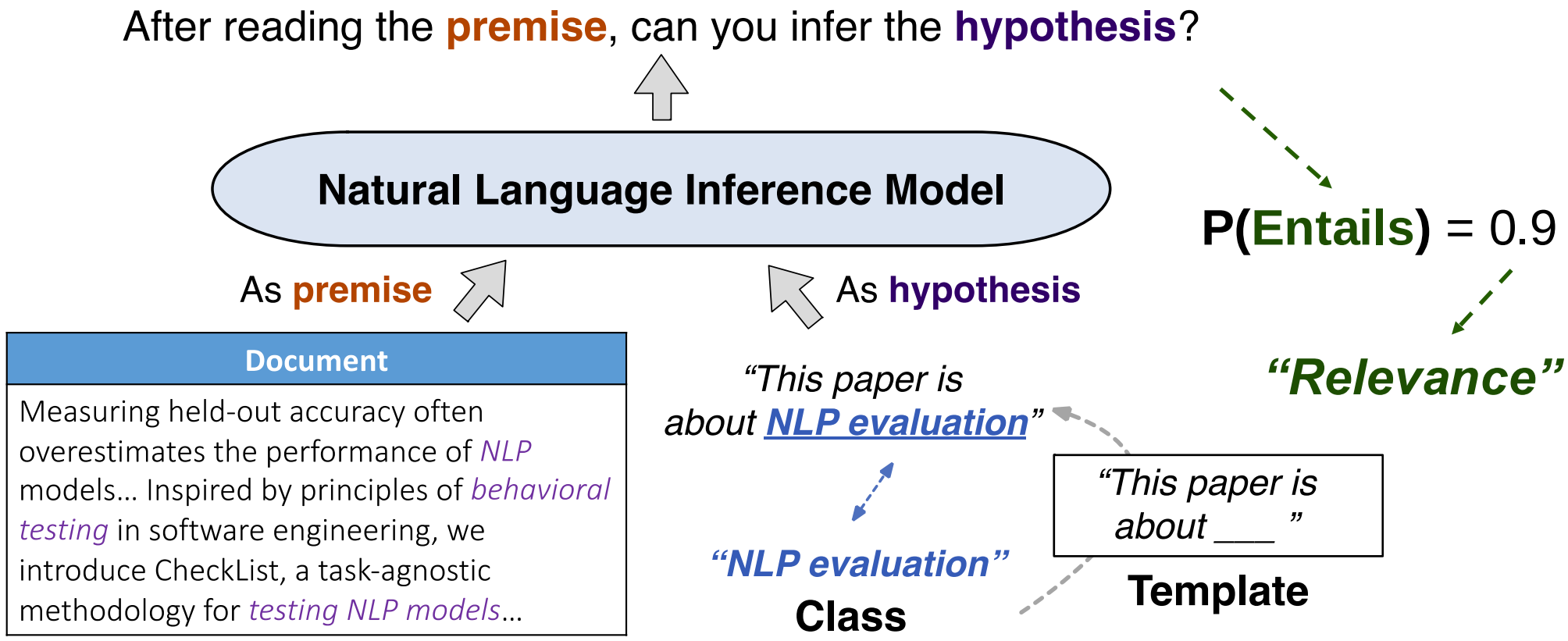
- The taxonomy is a directed acyclic graph (DAG)
- Each document can have **multiple** categories distributed on different paths
- Category names can be phrases and may not appear in the corpus

| Document |
| --- |
| Measuring held-out accuracy often overestimates the performance of ***NLP*** models… Inspired by principles of ***behavioral testing*** in software engineering, we introduce CheckList, a task-agnostic methodology for ***testing NLP models***… |



# TaxoClass: Document-Class Relevance Calculation

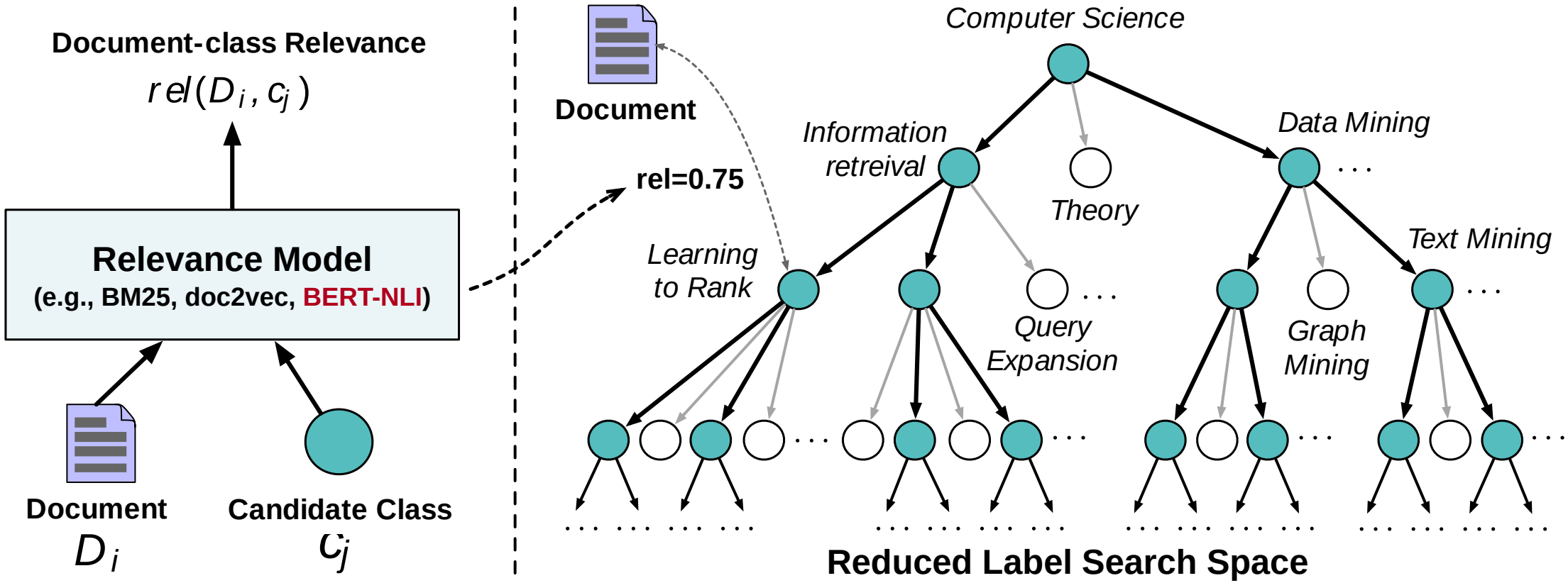
- How to use the knowledge from pre-trained LMs?
- Relevance model: BERT/RoBERTa fine-tuned on the NLI task




Shen, J., Qiu, W., Meng, Y., Shang, J., Ren, X., & Han, J., "TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names", NAACL'21.

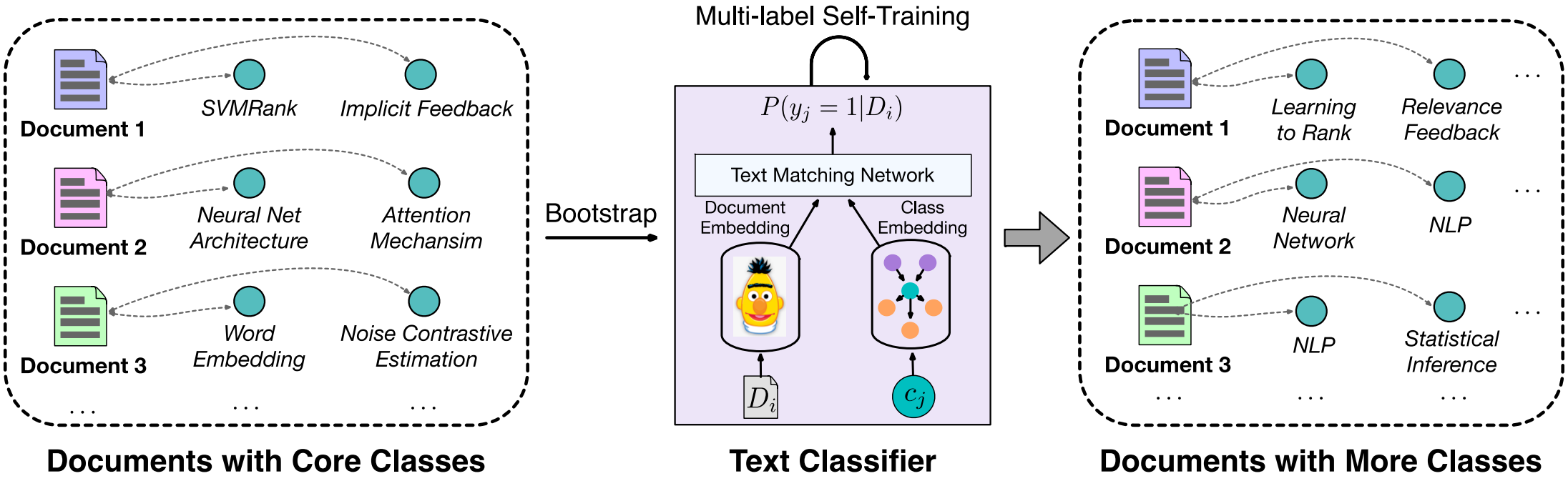
# TaxoClass: Top-Down Exploration

- How to use the taxonomy?
- Shrink the label search space with top-down exploration
  - Use a relevance model to filter out completely irrelevant classes



# TaxoClass: Identify Core Classes and More Classes

- Identify document core classes in reduced label search space
- Generalize from core classes with bootstrapping and self-training



# TaxoClass: Experiment Results

Weakly-supervised multi-class classification method

Semi-supervised methods using 30% of training set

Zero-shot method

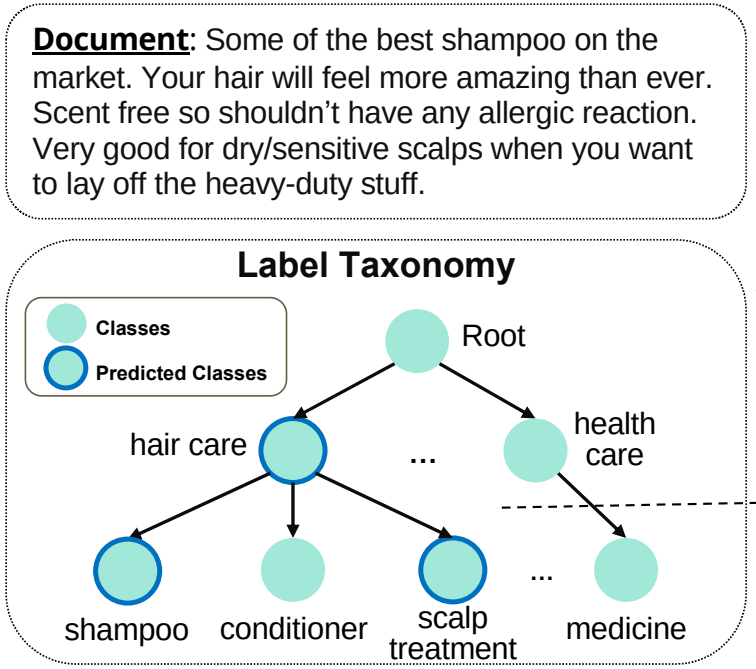
| Methods | Amazon | | DBPedia | |
|---|---|---|---|---|
| | Example-F1 | P@1 | Example-F1 | P@1 |
| **WeSHClass** (Meng et al., AAAI'19) | 0.246 | 0.577 | 0.305 | 0.536 |
| **SS-PCEM** (Xiao et al., WebConf'19) | 0.292 | 0.537 | 0.385 | 0.742 |
| **Semi-BERT** (Devlin et al., NAACL'19) | 0.339 | 0.592 | 0.428 | 0.761 |
| **Hier-0Shot-TC** (Yin et al., EMNLP'19) | 0.474 | 0.714 | 0.677 | 0.787 |
| **TaxoClass** (ours) | **0.593** | **0.812** | **0.816** | **0.894** |

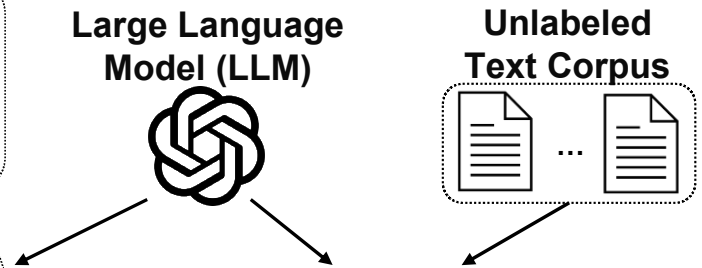
- **vs. WeSHClass**: better model document-class relevance
- **vs. SS-PCEM, Semi-BERT**: better leverage supervision signals from taxonomy
- **vs. Hier-0Shot-TC**: better capture domain-specific information from core classes

$$\textbf{Example-F1} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|true_i \cap pred_i|}{|true_i|+|pred_i|}, \quad \textbf{P@1} = \frac{\#docs\ with\ top{-}1\ pred\ dorrect}{\#total\ docs}$$

# TELEClass: Taxonomy Enrichment and LLM-Enhanced WHTC

- Enrich taxonomy with class-indicative terms mined from the corpus
- Adapt LLMs in the weakly-supervised hierarchical classification task for data pseudo labeling and data augmentation



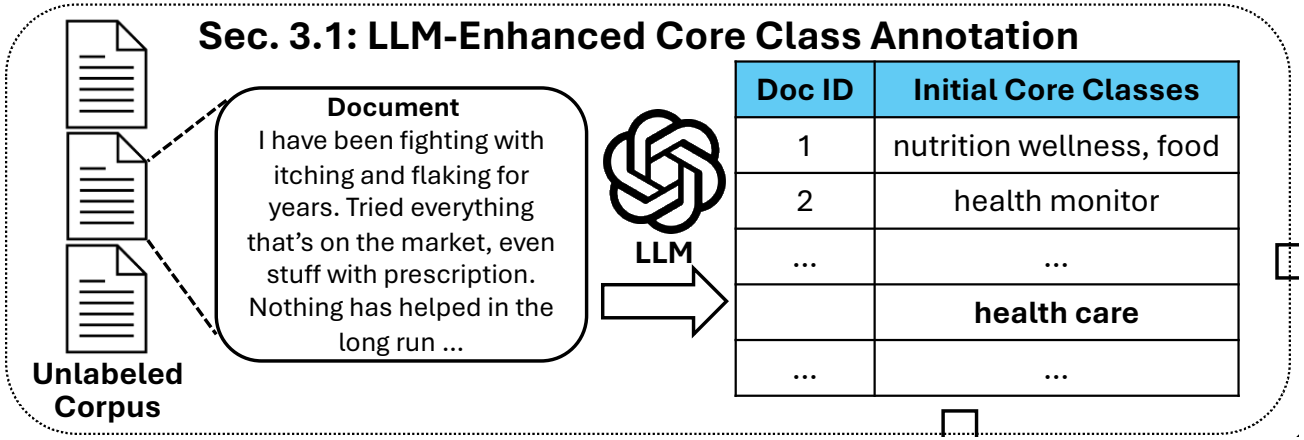

**Taxonomy Enrichment**

| Class | Enriched Key Terms |
|---|---|
| **hair care** | hair color, curly, dryer |
| **scalp treatment** | dandruff, Hask Placenta |
| **shampoo** | flakes, itching, clean |
| **conditioner** | moisture, soft hair |

Zhang, Y., Yang, R., Xu, X., Li, R., Xiao, J., Shen, J., Han, J., "TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision", arXiv'24

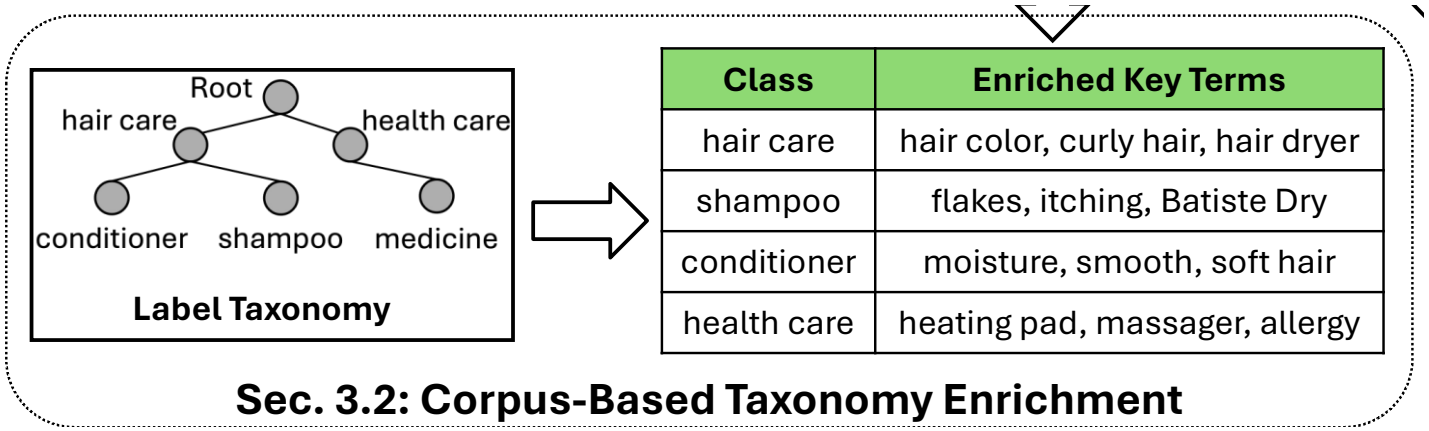
# TELEClass: LLM-Enhanced Core Class Annotation

- Enhance the core class annotation process of TaxoClass with LLMs
- Use top-down tree search with semantic similarity to reduce the candidates core class for LLM annotation
  - First use LLMs to enrich each class with keywords for better understanding
- Apply LLMs to select core classes from the candidates



# TELEClass: Corpus-based Taxonomy Enrichment

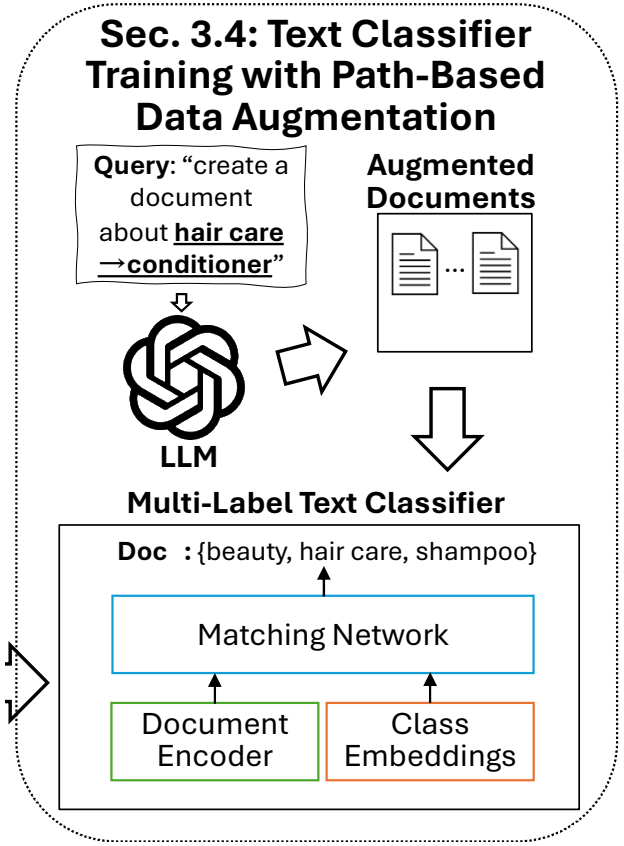
- First step gives roughly classified documents for each node
- For each set of siblings, we find a set of key terms for each node that are (1) frequent for this node, (2) infrequent for other sibling nodes, and (3) semantically similar with the class name
- These corpus-based features are used to refine the LLM-based core classes



# TELEClass: Path-based Data Augmentation

- The core class annotation can miss some infrequent classes which are common in the hierarchical label space
- Apply LLMs to further generate pseudo documents
  - Generate document for **each path from root to leaf**
  - For example, a path "hair care -> shampoo"

  distinguishes hair shampoo with carpet shampoo
- Finally, train classifier with **both** refine core classes and generated documents



# TELEClass: Performance

- Strong performance comparing with TaxoClass and zero-shot LLM

| Supervision Type | Methods | Amazon-531 | | | | DBPedia-298 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Example-F1 | P@1 | P@3 | MRR | Example-F1 | P@1 | P@3 | MRR |
| Zero-Shot | Hier-0Shot-TC† | 0.4742 | 0.7144 | 0.4610 | — | 0.6765 | 0.7871 | 0.6765 | — |
| | ChatGPT | 0.5164 | 0.6807 | 0.4752 | — | 0.4816 | 0.5328 | 0.4547 | — |
| Weakly-Supervised | Hier-doc2vec† | 0.3157 | 0.5805 | 0.3115 | — | 0.1443 | 0.2635 | 0.1443 | — |
| | WeSHClass† | 0.2458 | 0.5773 | 0.2517 | — | 0.3047 | 0.5359 | 0.3048 | — |
| | TaxoClass-NoST† | 0.5431 | 0.7918 | 0.5414 | 0.5911 | 0.7712 | 0.8621 | 0.7712 | 0.8221 |
| | TaxoClass† | 0.5934 | 0.8120 | 0.5894 | 0.6332 | 0.8156 | 0.8942 | 0.8156 | 0.8762 |
| | TELEClass | **0.6483** | **0.8505** | **0.6421** | **0.6865** | **0.8633** | **0.9351** | **0.8633** | **0.8864** |
| Fully-Supervised | | 0.8843 | 0.9524 | 0.8758 | 0.9085 | 0.9786 | 0.9945 | 0.9786 | 0.9826 |

- Once trained, only needs very low inference cost
  - GPT-4 needs ~$2,500 for inference on DBPedia

| Methods | Amazon-531 | | | DBPedia-298 | | |
|---|---|---|---|---|---|---|
| | F1 | P@1 | P@3 | F1 | P@1 | P@3 |
| ChatGPT | 0.516 | 0.681 | 0.475 | 0.482 | 0.533 | 0.455 |
| ChatGPT-L | 0.662 | 0.857 | 0.644 | 0.665 | 0.830 | 0.649 |
| GPT-4‡ | 0.699 | 0.822 | 0.689 | 0.605 | 0.652 | 0.592 |
| TELEClass | 0.648 | 0.851 | 0.642 | 0.863 | 0.935 | 0.863 |

# References


- Kargupta, P., Komarlu, T., Yoon, S., Wang, X., Han, J. "MEGClass: Extremely Weakly Supervised Text Classification via Mutually-Enhancing Text Granularities", EMNLP'23.
- Mekala, D. & Shang, J. "Contextualized Weak Supervision for Text Classification", ACL'20.
- Meng, Y., Shen, J., Zhang, C., & Han, J. "Weakly-supervised neural text classification", CIKM'18.
- Meng, Y., Shen, J., Zhang, C., & Han, J. "Weakly-Supervised Hierarchical Text Classification", AAAI'19.
- Meng, Y., Zhang, Y., Huang, J., Xiong, C., Ji, H., Zhang, C., & Han, J. "Text Classification Using Label Names Only: A Language Model Self-Training Approach", EMNLP'20
- Wang, Z., Mekala, D., & Shang, J. "X-Class: Text Classification with Extremely Weak Supervision", NAACL'21
- Zhang, Y., Jiang, M., Meng, Y., Zhang, Y., & Han, J. "PIEClass: Weakly-Supervised Text Classification with Prompting and Noise-Robust Iterative Ensemble Training", EMNLP'23
- Shen, J., Qiu, W., Meng, Y., Shang, J., Ren, X., & Han, J., "TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names", NAACL'21
- Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G. "Text Classification via Large Language Models", EMNLP'23.
- Zhang, L., Ding, j., Xu, Y., Liu, Y., Zhou, S. "Weakly-supervised Text Classification Based on Keyword Graph", EMNLP'21.
- Zhang, Y., Yang, R., Xu, X., Li, R., Xiao, J., Shen, J., Han, J., "TELEClass: Taxonomy Enrichment and LLM-Enhanced Hierarchical Text Classification with Minimal Supervision", arXiv'24

# Q&A

Tutorial Website: