**Vocabulary Terms Collection**

**Identifying Vocabulary Terms**

Identifying vocabulary terms to represent a domain accurately requires expertise in the domain. To eliminate the need for human intervention, a deep learning technique is used to identify relevant terms from an initial collection of predetermined words. The initial set of predefined words is derived from research articles that have used meta-learning to recommend the feature selection algorithm.

Initial set of predefined words:

predefined\_words = ["learning task","classification", "regression","clustering","meta knowledge", "surrogate models", "meta model","base learner", "meta knowledge database", "meta features", "task properties", "meta learning", "metalearning", "advisory method"]

After identifying the terms from Scopus (11426 tokens) and Google Scholar (8 articles, 5000 tokens), the predefined token list is updated to include additional relevant words.

Updated predefined words:

predefined\_words = ["learning task","classification", "regression","clustering","meta knowledge", "surrogate models", "meta model","base learner", "meta knowledge database", "meta features", "task properties", "meta learning", "metalearning",  "advisory method", "domain knowledge", "deep learning model", "deep meta learning", "deep learning method", "classification model", "candidate classification", "base learning", "recommendation model", "meta learning task", "meta learning database", "meta based learning", "feature learning", "ensemble classification model"]

**Data Collection**

The data to identify vocabulary terms are primarily extracted from research articles related to meta-learning. Therefore, two sources have been identified for term extraction:

1. Google Scholar:

- Search phrases: 'meta learning survey' or 'meta-learning survey'

- Initial set: 50 articles downloaded (5 articles excluded as they were not available for download).

2. Scopus database:

- Search phrases: meta-learning AND survey OR meta AND learning AND survey OR meta AND learning OR meta-learning

- Total articles listed: 7117.

**Data Preparation**

**Text Preprocessing:**

- Special characters and digits were removed from the text.

- Lemmatization was applied to standardize words to their base forms (Note: In subsequent rounds, lemmatization was avoided as it generated too many unnecessary words).

- Stop words were removed to focus on important information.

Without performing any text preprocessing, over 34000 tokens were generated per article. After text preprocessing, approximately 5000 tokens were generated per article using a similarity threshold of 0.95 (Google scholar).

**Tokenization and Embeddings:**

- N-grams were generated and joined, and duplicate words were removed.

- Embeddings for the text were calculated using the BERT model.

- Predefined word embeddings were generated using the BERT model.

**Similarity Identification:**

- Similarity between predefined words and words from the document was identified using cosine similarity.

- A threshold of 0.97 was set to identify similar words.

A similarity threshold of 0.95 generated 5,000+ words, so a threshold of 0.97 was used, which generated approximately 700 tokens per research article. The Scopus abstracts, author keywords, and index keywords generated approximately 3,500 keywords for every 100 rows. Hence, the dataset (2,300 rows of Scopus) was merged to remove duplicate tokens (11,400+ tokens).

The remaining Google Scholar articles (37 articles) generated 17,667 tokens (after removing duplicates), and the Scopus database (row number 2301-7117) generated 41775 tokens.

**Identification of Words for Meta-Learning**

All identified similar tokens were manually inspected. This list was sorted, and the final words were used to build vocabulary terms for meta-learning.