**Dataless Text Classification**

1. **Semantic Representation**

These approaches typically use Wikipedia to generate semantic representation and the classification is usually composed of three steps: 1) generate semantic representation for both label names and the documents using BOW and ESA representation etc; 2) train a bootstrapped classifier for a feature representation like BOW or ESA; 3) ensemble different machine learning classifiers.

1.1 dataless text classification

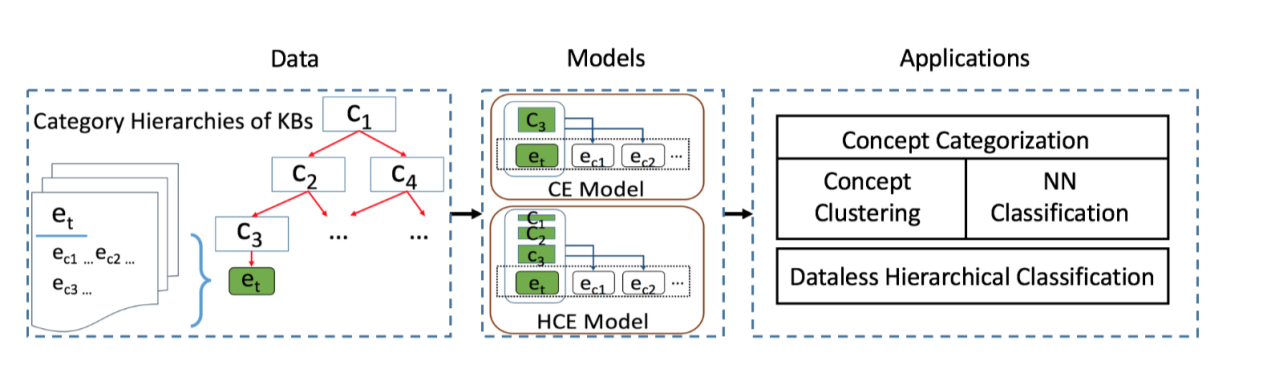
The classification is composed of three steps: 1) generate semantic representation for both label names and the documents using BOW and ESA representation etc; 2) train a bootstrapped classifier for a feature representation like BOW or ESA; 3) induce a new classifier with bootstrapped BOW and ESA classifier by co-training.

1.2 dataless hierarchical text classification

In dataless hierarchical text classification, we have hierarchical labels with 20 low level labels and 6 high level labels. The classification is composed of two steps: a semantic similarity step and a bootstrapping step. 1) In the semantic similarity step, we embed both labels and documents in a semantic space that allows one to compute meaningful semantic similarity between a document and a potential label. While this is a generic step that makes use of external information in the form of the semantic embedding. 2) in the bootstrapping step we adapt to the specific document collection; we use the semantic similarity step to drive a machine learning classifier that iteratively improves the categorization without a need for labeled data. We study dataless classification in the context of two natural hierarchical classification schemes, top-down and bottom-up.

1.3 hierarchical category embedding model (HCE model)

In order to find representations for categories and entities that can capture their semantic relatedness, we use existing hierarchical categories and entities labeled with these categories, and explore two methods: 1) Category Embedding model (CE Model): the model is based on the Skip-gram word embedding model and it replaces the entities in the context with their directly labeled categories to build categories’ context; 2) Hierarchical Category Embedding (HCE Model): it further incorporates all ancestor categories of the context entities to utilize the hierarchical information.



1. **LDA (latent Dirichlet allocation) based methods**

LDA is an unsupervised generative probabilistic model for collections of discrete data such as text documents. In LDA, each document is generated by choosing a distribution over topics and then choosing each word in the document from a topic selected according to the distribution.

2.1 classifyLDA

classifyLDA involves three steps: 1) learn a set of original topics by inferring the unsupervised LDA model over the concerned dataset; 2) manually assign a category label to each topic, and then combine the topics that are associated with a same label into a new single topic; 3) classify test documents using the document-level posterior of those combined topics learned by unsupervised LDA.

2.2 descriptive LDA (DescLDA)

DescLDA incorporates topic modeling. In DescLDA, a describing device (DD), is joined to the standard LDA model to infer descriptive Dirichlet priors (i.e., a topic-word matrix) from a few documents created from descriptive words in category labels/descriptions. These priors can then influence the generation process, making the standard LDA capable of inferring topics for text classification.

The DescLDA-based DLTC method comprises three steps: 1) construct the descriptive documents; 2) induce latent topics with DescLDA; 3) assign category labels to the test documents.

**3. Pseudo-label based Dataless Naive Bayes (PL-DNB)**

PL-DNB is built on semi-supervised naive Bayes. It has an initialization step and a loop step:

1. initialization step

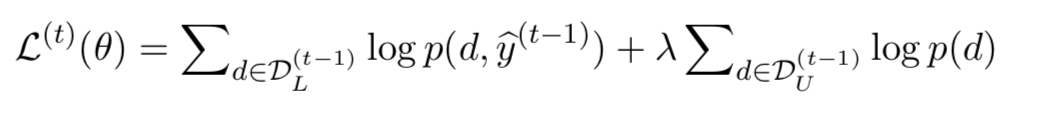
We initialize pseudo-labels ~y(0) for documents using seed word occurrences, leading to a set of documents with pseudo-labels D(0)L and a set of unlabeled documents that contain no seed words D(0)U.

1. loop step

We follow two steps until the maximum iterative number is reached.

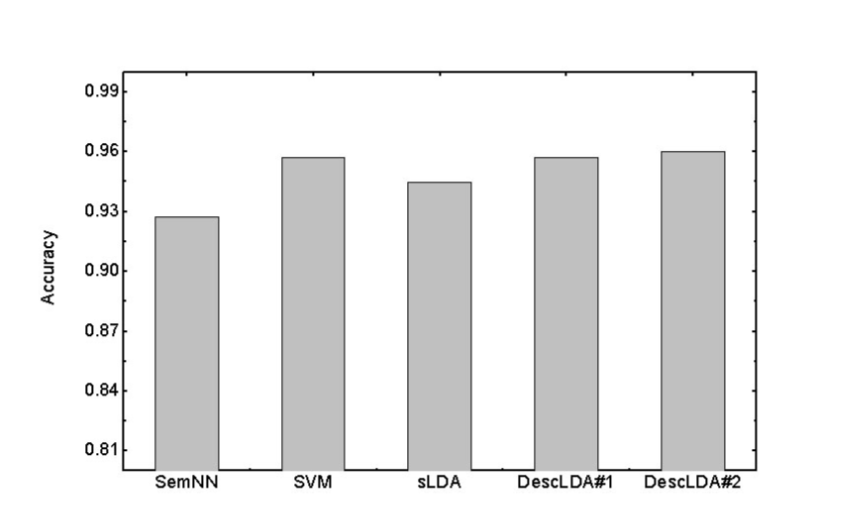
2.1

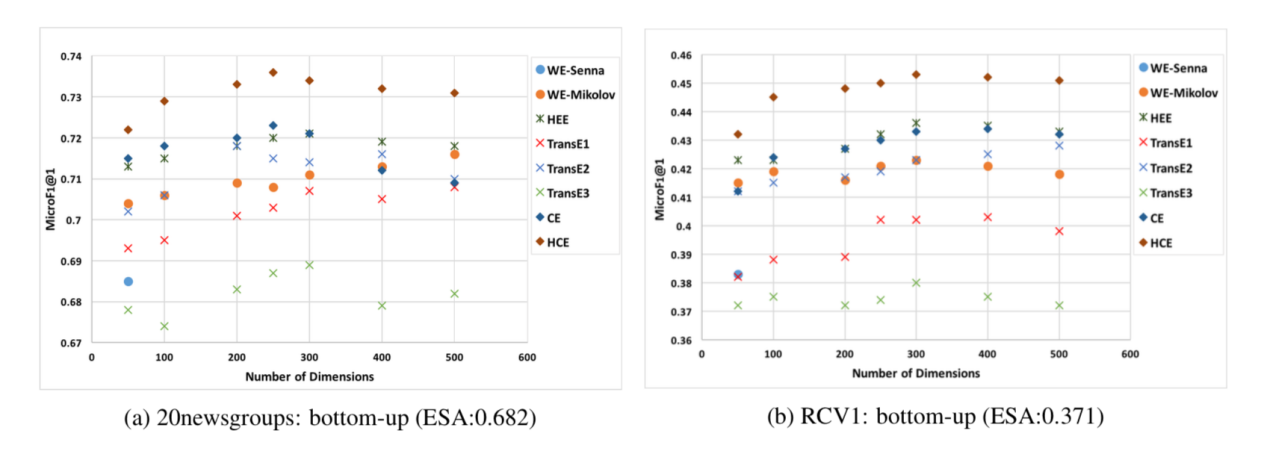
At each iteration t, use the EM algorithm to estimate the naive Bayes parameter θ(t) by maximizing the following semi-supervised objective:

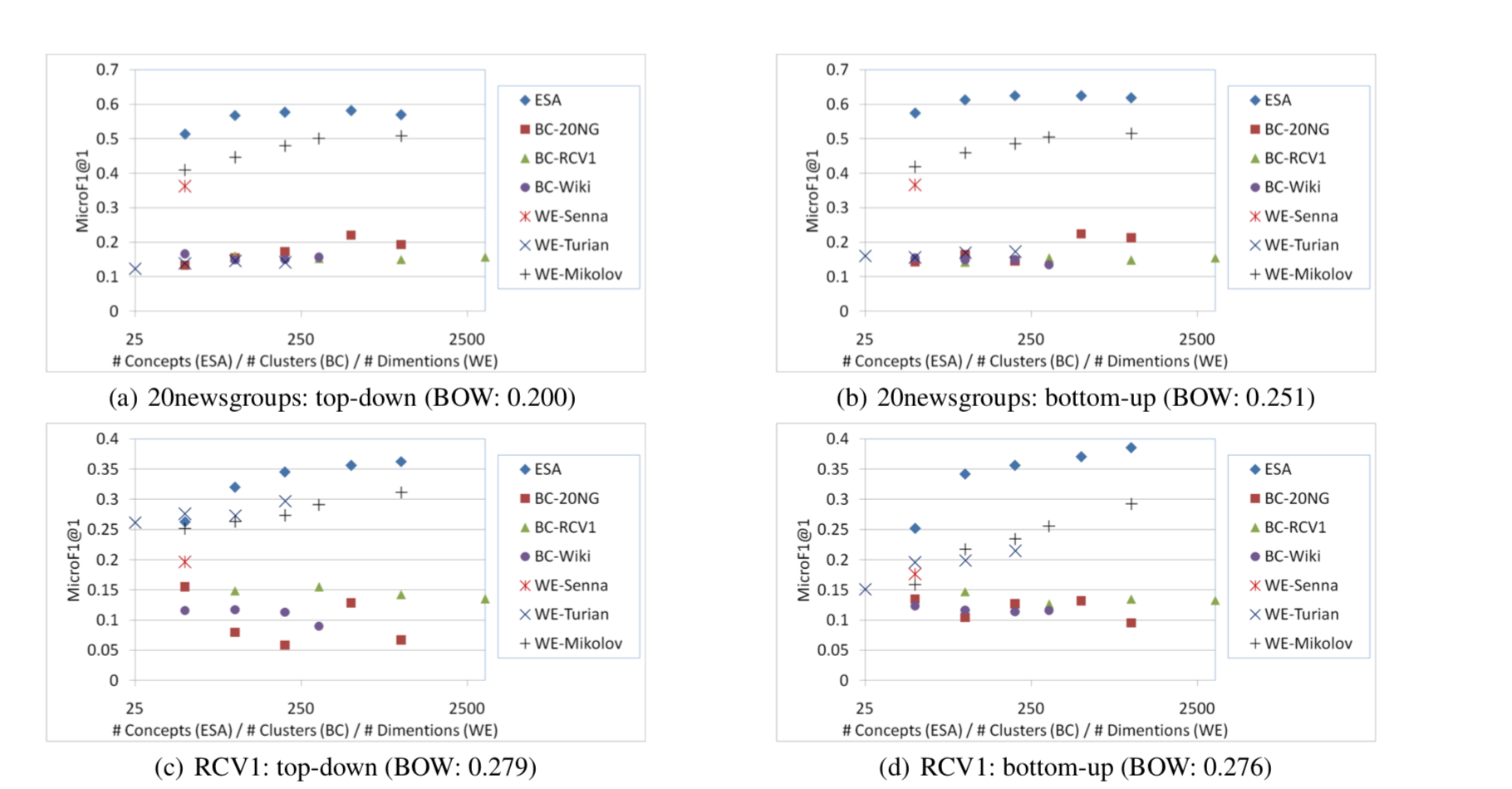


2.2

Given the optimum of θ(t), compute the label posterior distributions for each document. We update pseudo-labels ~y(t) using these label posteriors and seed word occurrences. Only the pseudo-labels with acceptable confidence are left, leading to new document sets of D( t )L and D( t )U.







References:

1. Chang M W, Ratinov L A, Roth D, et al. Importance of Semantic Representation: Dataless Classification[C]//AAAI. 2008, 2: 830-835.
2. Song Y, Roth D. On dataless hierarchical text classification[C]//Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.
3. Li Y, Zheng R, Tian T, et al. Joint embedding of hierarchical categories and entities for concept categorization and dataless classification[J]. arXiv preprint arXiv:1607.07956, 2016.
4. Chen X, Xia Y, Jin P, et al. Dataless text classification with descriptive LDA[C]//Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.
5. Li X, Yang B. A Pseudo Label based Dataless Naive Bayes Algorithm for Text Classification with Seed Words[C]//Proceedings of the 27th International Conference on Computational Linguistics. 2018: 1908-1917.