Pretrained text representations, evolving from context-free word embeddings to contextualized language models, have brought text mining into a new era: By pretraining neural models on large-scale text corpora and then adapting them to task-specific data, generic linguistic features and knowledge can be effectively transferred to the target applications and remarkable performance has been achieved on many text mining tasks. Unfortunately, a formidable challenge exists in such a prominent pretrain-finetune paradigm: Large pretrained language models (PLMs) usually require a massive amount of training data for stable fine-tuning on downstream tasks, while human annotations in abundance can be costly to acquire.

In this tutorial, we introduce recent advances in pretrained text representations, as well as their applications to a wide range of text mining tasks. We focus on \textit{minimally-supervised} approaches that do not require massive human annotations, including (1) self-supervised text embeddings and pretrained language models that serve as the fundamentals for downstream tasks, (2) unsupervised and distantly-supervised methods for fundamental text mining applications, (3) unsupervised and seed-guided methods for topic discovery from massive text corpora and (4) weakly-supervised methods for text classification and advanced text mining tasks.

**TARGET AUDIENCE AND PREREQUISITES**

Researchers and practitioners in the fields of data mining, text mining, natural language processing, information retrieval, database systems, and machine learning. While the audience with a good background in these areas would benefit most from this tutorial, we believe the material to be presented would give both general audience and newcomers an introductory pointer to the current work and important research topics in this field, and inspire them to learn more. Our tutorial is designed as self-contained, so only preliminary knowledge about basic concepts in data mining, text mining, machine learning, and their applications are needed.
- Flat Text Classification [12, 15, 19, 29]
- Hierarchical Text Classification [16, 26]
- Metadata-Aware Text Classification [30–33]

ACKNOWLEDGMENTS

Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

[3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Test Encoders as Discriminators Rather Than Generators. In ICLR.
[6] Xiaotao Gu, Zihan Wang, Zhenyu Bi, Yu Meng, Liyuan Liu, Jiawei Han, and Jingbo Shang. 2021. UCParse: Unsupervised Context-aware Quality Phrase Tagging. In KDD.
[9] Jiaxin Huang, Yu Meng, and Jiawei Han. 2022. Few-Shot Fine-Grained Entity Typing with Automatic Label Interpretation and Instance Generation. In KDD.
[10] Jiaxin Huang, Yuqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. CoRel: Seed-Guided Topical Taxionomy Construction by Concept Learning and Relation Transferring. In KDD.
[13] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative Topic Mining via Category-Name Guided Text Embedding. In WWW.
[14] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance M. Kaplan, and Jiawei Han. 2019. Spherical Text Embedding. In NeurIPS.
[15] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-Supervised Neural Text Classification. In CIKM.
[16] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-Supervised Hierarchical Text Classification. In AAAI.
[17] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and Contrastive Text Sequences for Language Model Pretraining. In NeurIPS.
[18] Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-Supervised Named Entity Recognition with Noise- Robust Learning and Language Model Augmented Self-Training. In EMNLP.
[19] Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text Classification Using Label Names Only: A Language Model Self-Training Approach. In EMNLP.
[20] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. In WWW.
[21] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding. In KDD.
[26] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names. In NAACL-HLT.
[31] Yu Zhang, Shweta Garg, Yu Meng, Xiushi Chen, and Jiawei Han. 2022. Motifclass: Weakly supervised text classification with higher-order metadata information. In WSDM.
[32] Yu Zhang, Yu Meng, Jiaxin Huang, Frank F Xu, Xuan Wang, and Jiawei Han. 2020. Minimally supervised categorization of text with metadata. In SIGIR.
[33] Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. 2022. Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification. In WWW.