# ECE 684: Natural Language Processing
# Selecting Optimal Number of Topics for Unsupervised Topic Modeling using Context-Aware Document Embeddings

Mehak Arora

October 28, 2024

# 1 Project Proposal

Topic modeling in Natural Language Processing (NLP) aims to automatically identify latent topics within a collection of documents [1]. Applications of topic modeling extend beyond typical use cases to fields like social media trend analysis, detecting emerging research areas in scientific literature, uncovering hidden patterns in historical archives, and supporting legal document discovery and analysis [2]. Despite the advent of Large Language Models and the subsequent succuss in most NLP tasks, finding hidden topics in text remains an open area of research. Latent Direchlet Allocation (LDA) is a widely used probabilistic modeling technique for topic modeling, that follows the Bag-of-Words assumption, and assigns each word in the corpus a probability distribution over topics [3]. However, the number of topics needs to be decided beforehand and provided as input to the algorithm. BERTopic is another State-of-the-Art topic modeling technique that generates representations for documents as vector embeddings using pre-trained, transformer-based language models, performs dimensionality reduction on these embeddings, and then finds ẗopicsb̈y unsupervised clustering [4]. There are many hyperparameters in the pipeline, changes in which can alter the number of topics found. In this project, **I aim to explore the problem of optimal topic selection.** The shortcoming of the LDA model, as I see it, is the Bag-of-Words assumption, which might not capture semantic information in the sequence of words and the meaning of sentences. Neural topic models (like BERTopic), capture sentence semantics and context much better, but lose the niceness, flexibility, and interpretability of the LDA model in modeling the probability distribution over words. I propose using the BERTopic vector embeddings for optimal LDA topics and model selection. Currently, LDA is performed iteratively with varying number of topics, and the best model is selected based on the one that minimizes perplexity (how well the model fits the data), topic diversity (how diverse each topic is), and topic coherence (how similar documents in the same topic are) [3]. In this project, I will use an additional metric that calculates the similarity in the BERTopic embedding space for each document *after* an LDA run, and use that as an additional metric to quanity how coherent each topic is, as well as the dissimilarity between documents in different topics to quantify how diverse the learned topics are. I will test the performance of this method on data that consists of short documents (a couple of sentences long) about four topics, carefully selected to have as many homonyms as possible. I will then test the performance of the different metrics in selecting the number of topics for the LDA model, and also compare the outcome with the BERTopic model. I will then test out my algorithm on commonly used topic modeling datasets AGNews [5] and 20NEWS [6], which consist of news articles with ground truth on classification.

## References


[1] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," *Information Systems*, vol. 94, p. 101582, 2020.

[2] P. Kherwa and P. Bansal, "Topic modeling: a comprehensive review," *EAI Endorsed transactions on scalable information systems*, vol. 7, no. 24, 2019.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[5] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[6] K. Lang, "Newsweeder: Learning to filter netnews," in *Machine Learning Proceedings 1995*, A. Prieditis and S. Russell, Eds. San Francisco (CA): Morgan Kaufmann, 1995, pp. 331–339. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9781558603776500487