Investigating Problem-solving approaches of students in MOOCs using natural language processing (NLP)

Dr. Erik Hemberg

ANYSCALE LEARNING FOR ALL

MIT CSAIL Computer Science & Artificial Intelligence Laboratory

Introduction

MOOCs (Massive Open Online Courses): An online course aimed at unlimited participation and open access across the world without prerequisites

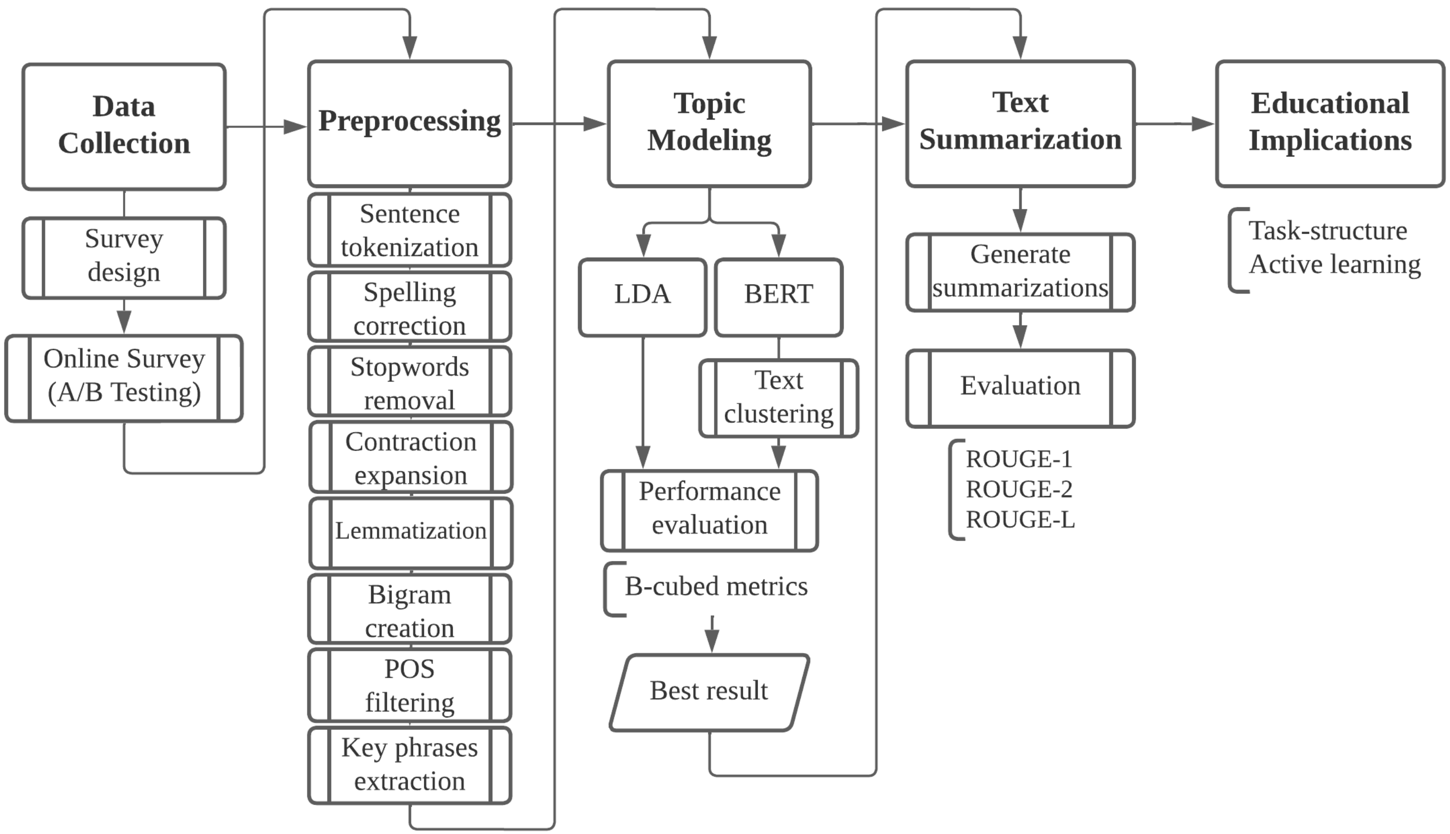
- MOOCs requires different learning engagement and behaviors compared to traditional classroom learning
- MOOCs have limited interaction between students–instructors and the size of class make it hard to monitor and understand the learning behaviors of students
- Gaps in MOOC studies:
    1) Previous research focused on the use of event log or clickstream data
    2) Learners' behavioral analysis or academic performance prediction; limited research on the students' problem-solving approaches in MOOCs

Contributions:

1) Suggesting a research method that can analyze the MOOCs users' problem-solving approaches at scale
2) Proposing an NLP pipeline that applies advanced preprocessing methods, contextual language model, and text summarization model to enhance the accuracy and readability of topic modeling's results
3) Discuss how a Large Language Model could be used to extend MOOC research

Method Overview

- User-generated text data (5,121 students) to analyze their behaviors
- Use of NLP to **scans** and **clean** data, **extract** key information, **detect** topics, and **generate** narrative summaries

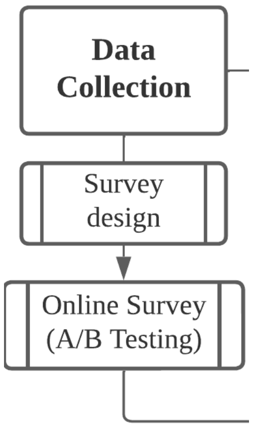


ALFA
ANYSCALE LEARNING FOR ALL

MIT CSAIL
Computer Science & Artificial Intelligence Laboratory

Data Collection

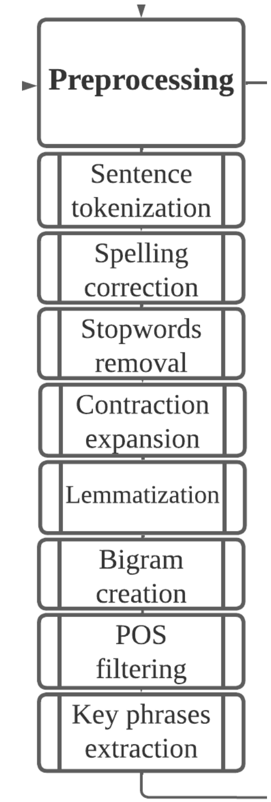
- Data was collected from the MIT CS course, a MOOC hosted by edX (https://www.edx.org/) taught over 9 weeks in Spring 2021
- After an exercise, students were randomly selected using the built-in A/B testing feature to respond to survey asking to describe their problem-solving approaches
- In total, 44,864 responses were collected from 5,121 students
- Selected 7,482 (16.7%) responses to conduct our research

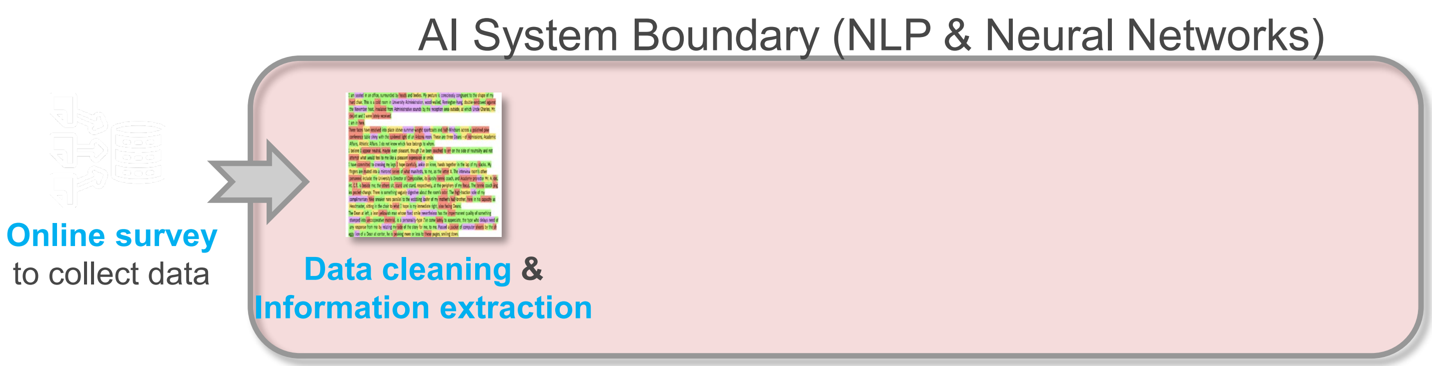


Online survey
to collect data

Text Preprocessing

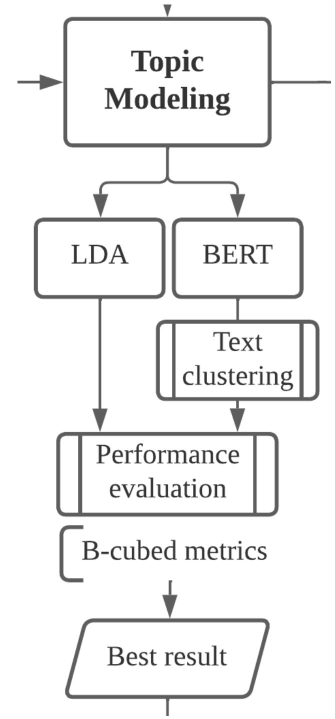
- Employed a chain of 8 preprocessing steps
- In total, 7,482 responses were fed into the preprocessing pipeline. As a result, merely 1,644 responses were remained; approx. 78% of total responses were discarded
- Despite the high loss of data, the performance of topic modeling improved significantly (up to 18 % improvement)

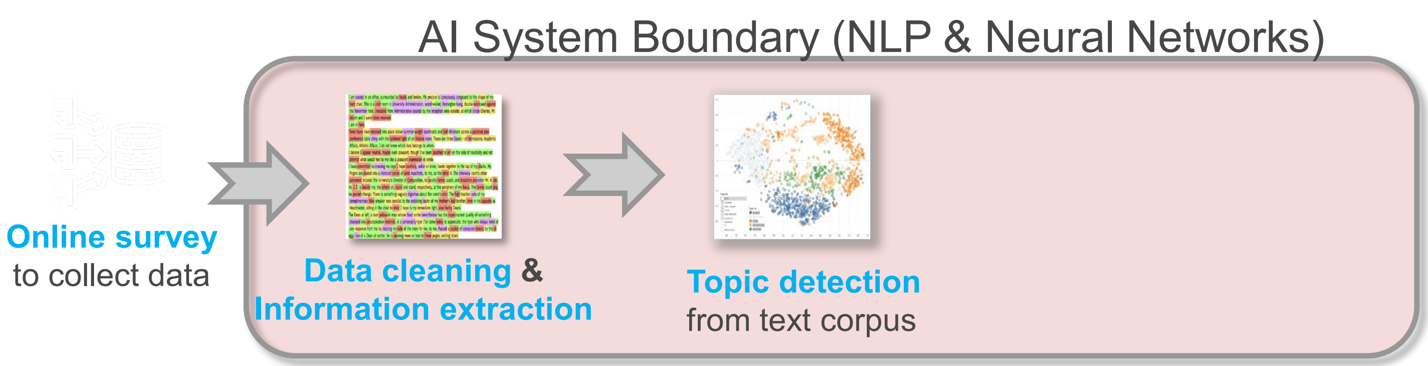




Topic Modeling

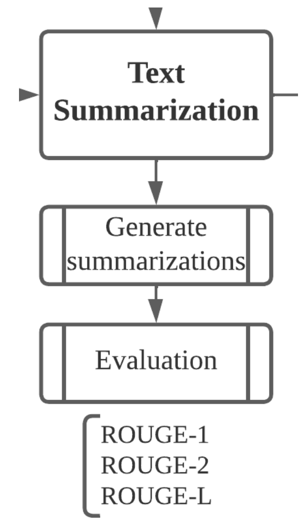
- **Statistic-based Models**: Latent Dirichlet Allocation (LDA) and Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM)
- **Contextual Language Model:** BERT(Bidirectional Encoder Representations from Transformers) + Clustering Algorithms (OPTICS, HAC, DBSCAN)
- B-cubed evaluation metrics to measure the performance
- BERT combined with DBSCAN had the best performance



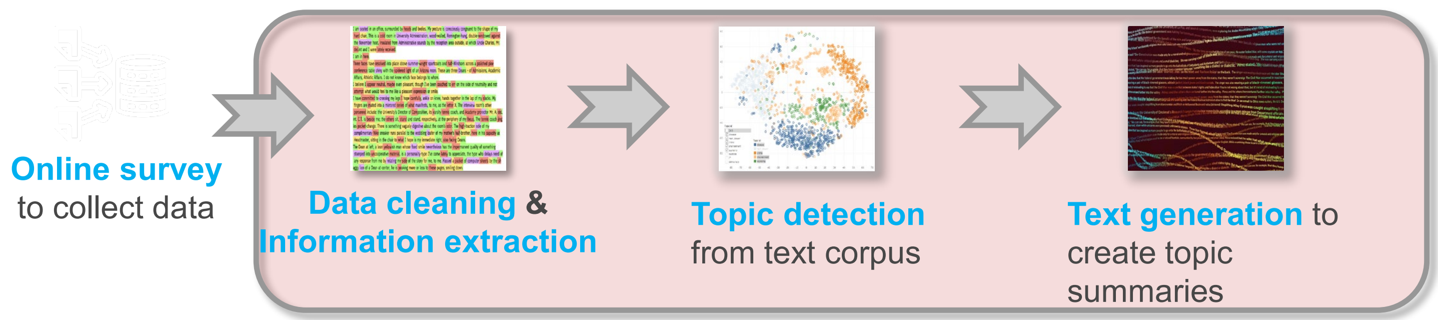


Text Summarization

- Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) to perform the text summarization
- ROUGE metrics and length of generated summaries are used to evaluate the summaries
- Pre-trained model *'pegasus_paraphrase'* is selected to perform the summarization



AI System Boundary (NLP & Neural Networks)



ALFA ANYSCALE LEARNING FOR ALL

MIT CSAIL Computer Science & Artificial Intelligence Laboratory

Educational Implication

a) Task-Structure Framework

Problem-solving methods are first identified, then formulated into subtasks backwards

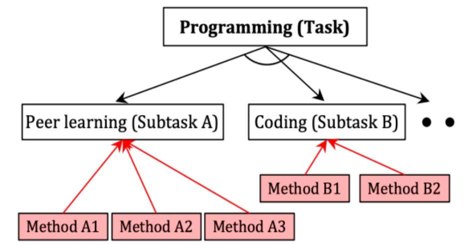
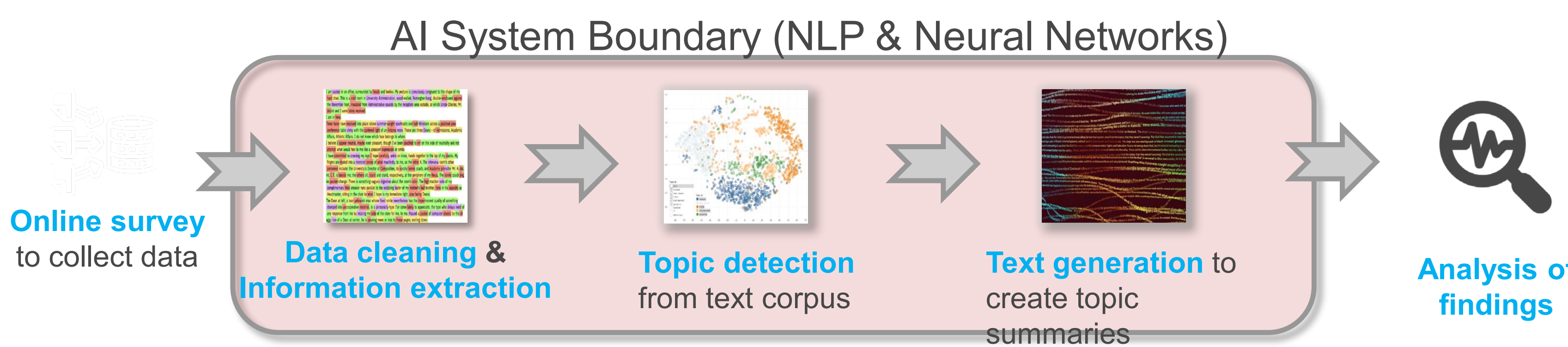


Figure 3-13: An example of bottom-up formulation of a task-structure. Subtasks are formulated based on identified methods

b) Active/Passive Learning Framework

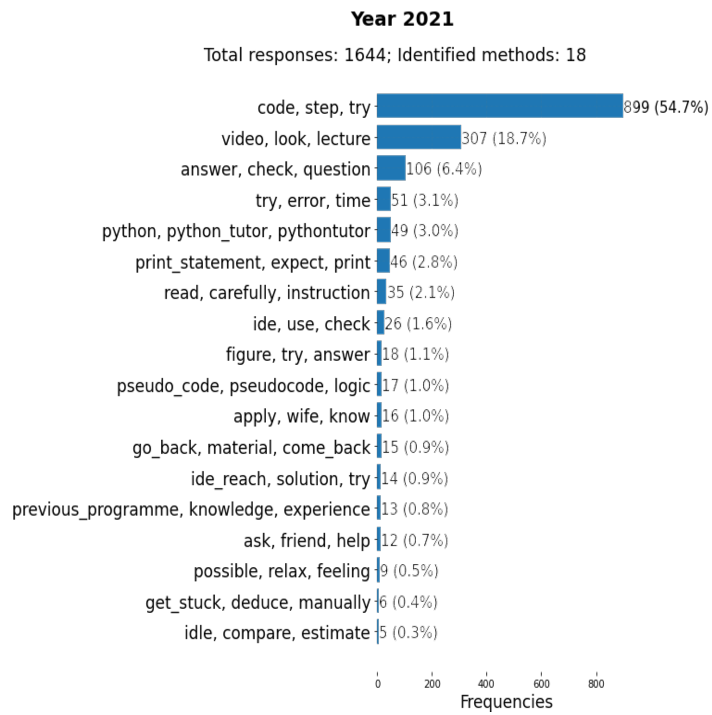
A method of learning in whether students are actively or passively involved in the learning process



ALFA ANYSCALE LEARNING FOR ALL

MIT CSAIL Computer Science & Artificial Intelligence Laboratory

Results



1. **Trial and Error**
2. **View videos**
3. **Check questions with answers**

	Knowledge construction	Social/peer learning	Understanding the problem	Clarifying code logic	Coding and debugging
Active Learning	• Check online resources (ID-17)	• Ask a friend (ID-15) • Use discussion forum (ID-03)	• Write down steps (ID-01)	• Use python tutor (ID-05) • Write pseudo code (ID-10)	• Use print statement (ID-06) • Trial and error (ID-11) • Try on IDE (ID-8, ID-13, ID-18) • Use stack overflow (ID-09)
Passive Learning	• Watch lecture videos (ID-02) • Look lecture material (ID-12) • Use previous knowledge (ID-14)		• Read instructions (ID-07)		• Understand error (ID-04)

ANYSCALE LEARNING FOR ALL

Computer Science & Artificial Intelligence Laboratory

Discussion

Opportunities

a) **Enhancement through transformer model:** The transformer-based topic modeling model significantly outperforms the statistical models. BERT sentence-embedding combined with DBSCAN clustering algorithm performed the best, achieving 78% accuracy compared to 16% by LDA and 24% by GSDMM

b) **Improved Readability:** Integrating the text summarization model into the topic modeling pipeline provided a solution to improving the limited readability of keyword-based results of topic modeling

Challenges

a) **Improved Accuracy vs. Data Loss:** Preprocessing plays a critical role in improving the quality of topic modeling. However, we also had to suffer approx. 78% of data being discarded during the process. This can play as a constraint for small data with a high level of noise.

b) **Capturing multiple approaches:** The shortcoming of this approach is that it can only capture one topic per data entry (single sentence). This approach cannot detect how many methods are used by a single student.

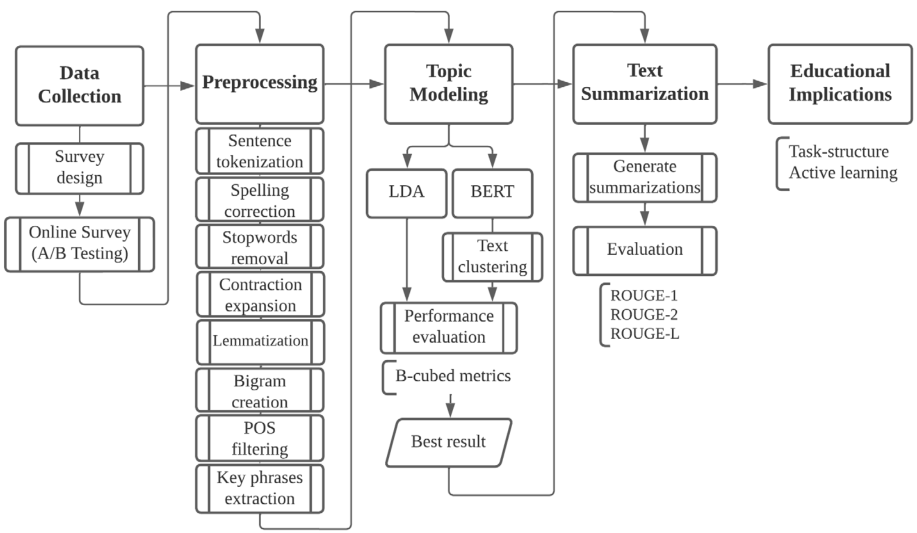
c) **Improving the output quality:** There are rooms for improving the clustering and summarization quality. LLMs can be an option to explore in the future.

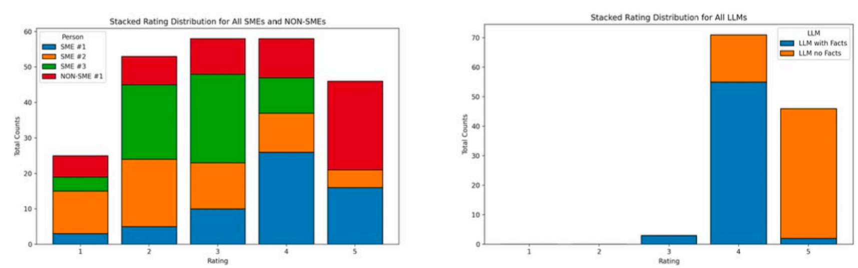
Discussion of applying LLMs

- Assessing Large Language Model's knowledge of threat behavior in MITRE ATT&CK
  - Truthfullness and educational quality?
- Prompts that are engineered with context provide the best answer accuracy with GPT-3.5.
- GPT-3.5 can be used to create multiple choice questions regarding cyber threat knowledge contained in MITRE ATT&CK.

Create multiple choice questions related to the cyber mitigation strategy below. Make a variety of quality questions, with answers, solutions, and explanations. Each question should be made that it tests the ability to apply and/or classify the knowledge tested. Each question should be expert-level (difficulty).
Content to make questions about:
Mitigation: Network Segmentation

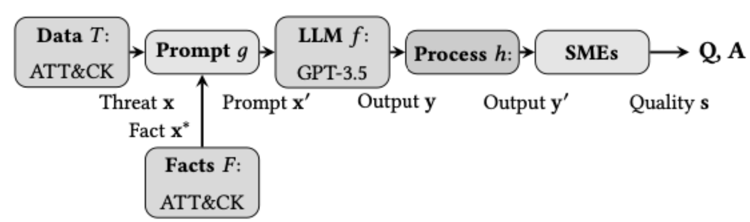
Question: Which technique would be most useful for implementing network segmentation to control lateral movement in a network?
A. Application whitelisting
B. Data encryption
C. VLAN partitioning
D. Antivirus installation
Solution: C. VLAN partitioning
Explanation: VLAN partitioning is a technique directly associated with network segmentation. It creates isolated networks within a larger network, which can help to control lateral movement (the technique hackers use to move through a network once they have gained access) by limiting the ability to communicate between segments.





(a) SME ratings.

(b) LLM ratings.



ALFA ANYSCALE LEARNING FOR ALL

CSAIL Computer Science & Artificial Intelligence Laboratory