

# CS 2731 Introduction to Natural Language Processing

## Session 15: Project proposal presentations

---

October 18, 2023



University of  
Pittsburgh

School of Computing and Information

# Course logistics

- Project proposal grades and comments were released this morning
- HW2 grades and feedback by the end of the week
- HW3 will be released by Friday (you will get more than a week to do it)
- Syllabus schedule updates will be coming soon
- **Bring a laptop** to class on Monday. It will be an “LLM lab day”
  - Probably do some classification with BERT
  - Announcement will be made on Canvas

# Schedule

1. Connor and Marcelo
2. Jacob and RJ
3. Haoyu, Yuxuan, Qichang
4. Robby and Birju
5. Yixiao, Dhanush, Ahana
6. Norah, Modhumonty, Gina
7. Yuhang and Lingwei
8. Jiyuan, Ming, Qikun
9. Aziz, Bhiman, Atharva
10. Ben, Max, Tom
11. Vincent, Lokesh, Shuhao, Shijia

# Instructions

- Plan for **4 min presentations, 2 min questions**
- Cover at least these key points
  - Project motivation (what is the value of this work?)
  - Super briefly, what 1-2 other related papers have done
  - What data you are planning to use
  - What approach/methods will you be taking
  - Evaluation of your approach (or dataset, if it's a dataset contribution)
- Have each member of the group talk at least a little
- Put your slides in this presentation after your project name slide by **class session, 2:30pm on Wed Oct 18**

# 1. Connor and Marcelo

---

# Unboxing Carcassonne: Learning the Rules Interactively

Connor Sweeney, Marcelo d'Almeida

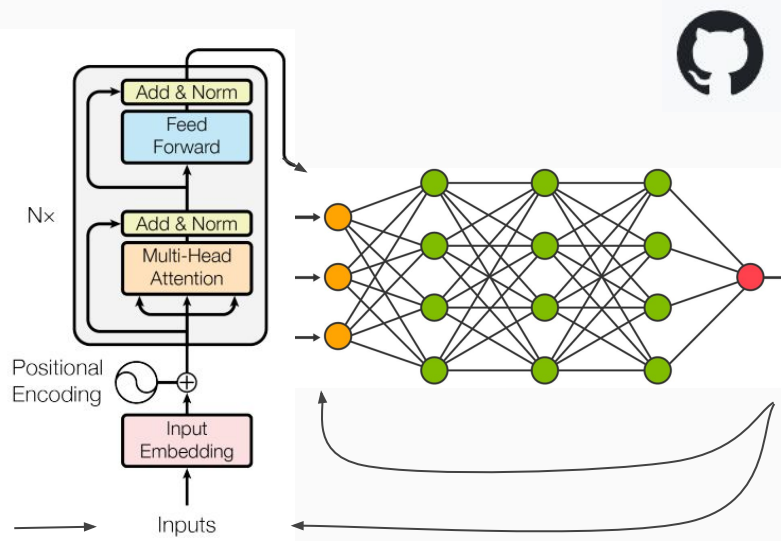
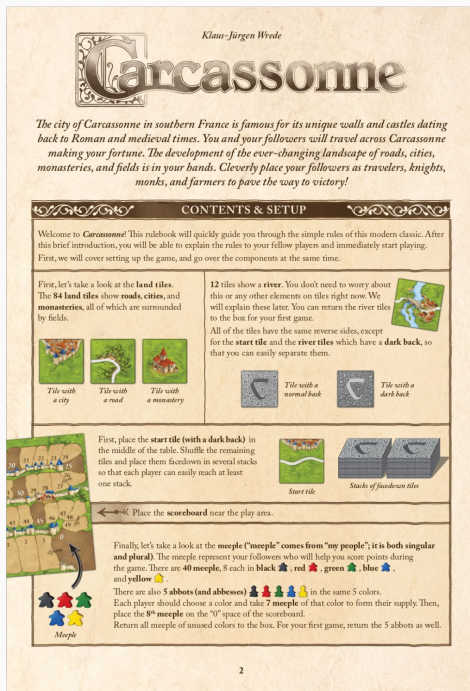
University of Pittsburgh

# Unboxing Carcassonne: Learning the Rules Interactively

Game Manual

Pre-Trained Language Model  
+ RL Agent

Environment

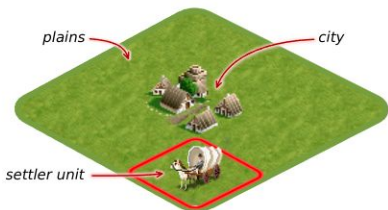


wingedsheep / carcassonne

- ← ["Carcassonne Rulebook v3", [zmangames.com/en/products/carcassonne](https://zmangames.com/en/products/carcassonne)]
- ↑ [Transformer block from "Attention is All you Need", Vaswani et al. 2017. Feedforward network adapted from [vitalflux.com/sklearn-neural-network-regression-example-mlpregressor/](https://vitalflux.com/sklearn-neural-network-regression-example-mlpregressor/)]
- ↖ [[github.com/wingedsheep/carcassonne](https://github.com/wingedsheep/carcassonne)]

# Unboxing Carcassonne: Learning the Rules Interactively

(Branavan et al. 2011)



Relevant text: "Use settlers to irrigate land near your city"

Predicted action words: "irrigate", "settler"

Predicted state words: "land", "near", "city"

Settlers unit, candidate action 1: **irrigate**

Features:

action = **irrigate** and action-word = "irrigate"

action = **irrigate** and state-word = "land"

action = **irrigate** and terrain = plains

action = **irrigate** and unit-type = settler

state-word = "city" and near-city = true

Settlers unit, candidate action 2: **build-city**

Features:

action = **build-city** and action-word = "irrigate"

action = **build-city** and state-word = "land"

action = **build-city** and terrain = plains

action = **build-city** and unit-type = settler

state-word = "city" and near-city = true

(Narasimhan et al. 2017)



is an enemy who chases you



is a stationary collectible



is a randomly moving enemy



is a stationary immovable wall

← ["Learning to win by reading manuals in a monte-carlo framework", S. R. K. Branavan et al. 2011]

↑ ["Grounding language for transfer in deep reinforcement learning", Narasimhan et al. 2017]

↖ ["Grounding language to entities and dynamics for generalization in reinforcement learning", Wang et al. 2021]

(Wang et al. 2021)



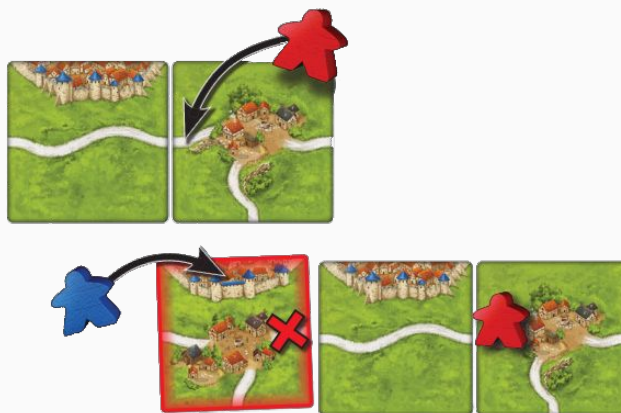
GAME 1 MANUAL

1. at a particular locale, there exists a motionless mongrel that is a formidable adversary.
2. the top-secret paperwork is in the crook's possession, and he's heading closer and closer to where you are.
3. the crucial target is held by the wizard and the wizard is fleeing from you.
4. the mugger rushing away is the opposition posing a serious threat.
5. the thing that is not able to move is the mage who possesses the enemy that is deadly.
6. *the vital goal is found with the canine, but it is running away from you.*



# Unboxing Carcassonne: Learning the Rules Interactively

## Find Valid Actions



Exploration

x

Exploitation

## Evaluate

Language-Informed Agent

x

Language-Unaware Agent  
Scores

Accuracy for Valid Actions

Evaluation of Attention

↑ [Tiles and Meeple from Carcassonne Manual; other images (dice and cards) from the internet]

Questions?

## 2. Jacob and RJ

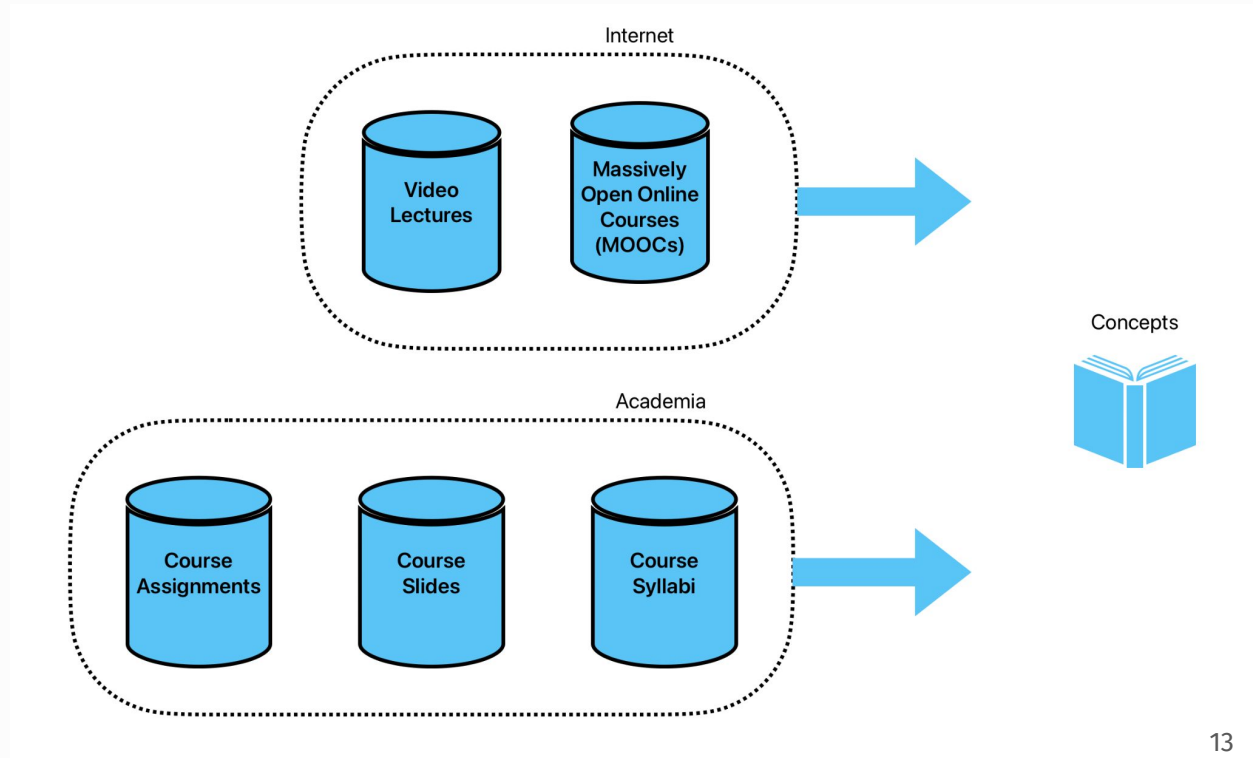
---

# CONCEPT EXTRACTION FROM COURSE MATERIAL

JACOB HOFFMAN AND RAJA KRISHNASWAMY

# MOTIVATION

- Course material data collections available for automatically extracting concepts



# MOTIVATION

- Concept extraction upon course material may:
  - Expedite the learning process for students
  - Help students better understand the main points of the material
  - Liberate instructors from the tedious process of human labeling

- DS-MOCE: Three-stage BIO Labeling Model
  - Lu, M., Wang, Y., Yu, J., Du, Y., Hou, L., & Li, J. (2023). *Distantly supervised course concept extraction in moocs with academic discipline*. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. <https://doi.org/10.18653/v1/2023.acl-long.729>
- MOOC Dataset From Coursera Video Lectures
  - Albahr, A., Che, D., & Albahar, M. (2021). *A novel cluster-based approach for keyphrase extraction from MOOC video lectures*. *Knowledge and Information Systems*, 63(7), 1663–1686. <https://doi.org/10.1007/s10115-021-01568-2>
- Automatic Concept Extraction Using Book Indexes (Less Data)
  - Boughoula, A., San, A., & Zhai, C. (2020). *Leveraging book indexes for automatic extraction of concepts in moocs*. *Proceedings of the Seventh ACM Conference on Learning @ Scale*. <https://doi.org/10.1145/3386527.3406749>

# DATASET - OVERVIEW

1. Create a manually BIO labeled, small-sized dataset using a subset of existing course material (slides and syllabi)
2. Create a dictionary of concepts based on a reputable source
3. Extend the dataset into a full-sized dataset by labeling more documents using a distantly supervised learning model and the dictionary of concepts
4. Split into a training set, a dev set, and a test set

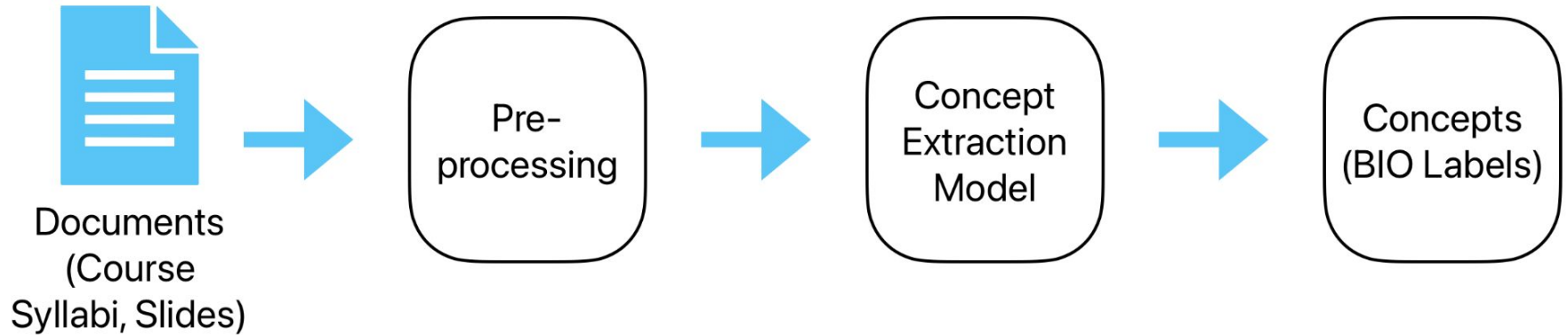


# DATASET - EXAMPLE BIO-LABELED ENTRY

<u>text</u>	the	operating	system	uses	interrupts	to	implement	system	calls
<u>label</u>	O	B	I	O	B	O	O	B	I

- B = Beginning, I = Inside, O = Outside
- Concepts =
  - Operating System
  - Interrupts
  - System Calls

# APPROACH - OVERVIEW

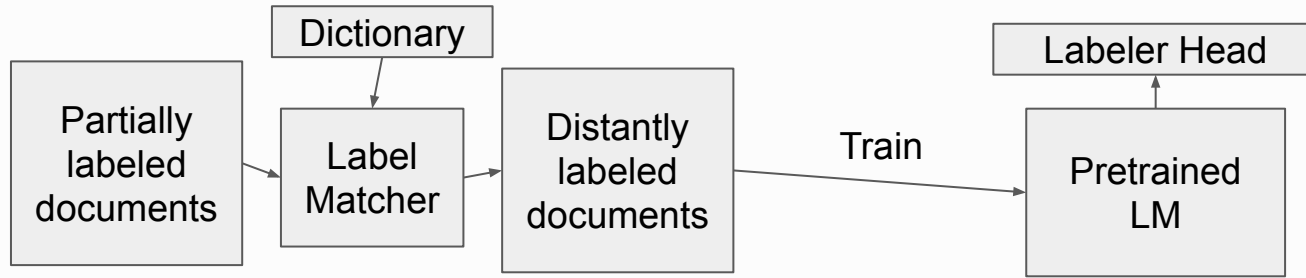


## APPROACH - CONCEPT EXTRACTION MODEL

- To do the concept extraction, we adapt the general approach used by DS-MOCE: self-training of a pre-trained LM with a classifier head on distant labels.

# APPROACH - DISTANT LABELING

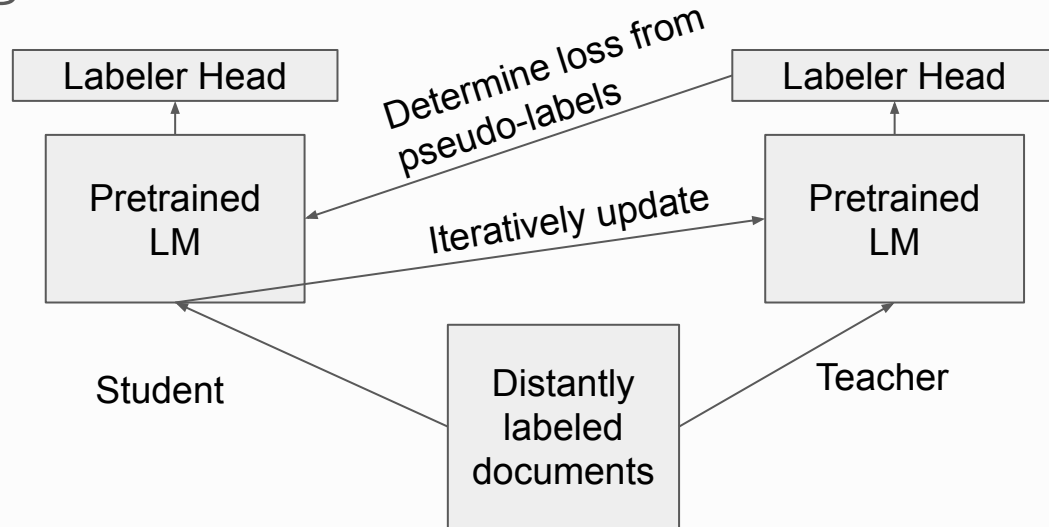
- One way we can solve this problem is to “fake” labels.
  - Use some list of concepts to get labels. Train the model on that, and employ pretrained models like BERT to add additional knowledge.



- This data will be noisy, however, so we will need to refine it to get anything useful from it.

# APPROACH - SELF-LEARNING

- So, we plan to additionally perform self-learning on the data. We plan to accomplish this through teacher-student self learning



- Run the model on the test set and compare it to the labeled concepts to compute:
  - Accuracy
  - Precision
  - Recall
  - F1-score

# QUESTIONS?

### 3. Haoyu, Yuxuan, Qichang

---



# Motivation

- **Resource Efficiency:** Traditional supervised machine learning requires a significant amount of labeled data, which can be expensive and time-consuming to obtain. Active learning, by selectively querying the instances it finds most informative, promises to reduce the number of labels required, leading to more efficient use of annotation resources.
- **Possible Improved Model Performance:** By concentrating on the most ambiguous and challenging instances, active learning can lead to better model generalization. This is particularly beneficial when dealing with complex datasets where misclassifications can have significant consequences.

# Related Work

- The paper “Interactive evaluation of classifiers under limited resources” aims to address the issue of evaluating classifier accuracy under limited resources. The paper proposes interactive algorithms to make the most of the limited number of true labels for evaluating classifier accuracy across the entire dataset.
- In the paper "D-CALM: A Dynamic Clustering-based Active Learning Approach for Mitigating Bias", the author explored the possibility of infusing clustering with active learning to overcome the bias issue of both active learning itself and traditional annotation methods.

# Dataset and Methods

- Dataset:
  - The RCV1 dataset, a benchmark dataset on text categorization.
  - News20 binary dataset, a binary form of the text classification UCI News 20 dataset
  - Stanford Sentiment Treebank (SST)
- Methods:
  - Employ an uncertainty-based active learning strategy to select which text samples require labeling to maximize model performance. We will also use a specific NLP model architecture(BERT) as our baseline and compare our approach against it.

# Evaluation

- We will evaluate the performance of our approach using common NLP performance metrics such as **accuracy, precision, recall, F1 score**, etc. Additionally, we will analyze the **learning curves** and **the size of data used** to assess the effectiveness of active learning.

## 4. Robbie and Birju







---

# Project Overview


- Motivation: How much does the language that political commentators use influence language use within the communities which consume their content?

## Podcast Charts

# Spotify — United States of America — News

1		The New York Times The Daily
2		The Daily Wire The Ben Shapiro Show
3		NPR Up First
4		NPR NPR News Now
5		Strike Force Five
6		The First TV The President's Daily Brief
7		BBC World Service Global News Podcast





### Ben Shapiro

r/benshapiro

Join

Posts Youtube Channel Spotify Soundcloud Custom Flair Request

Create Post

Hot New Top

PINNED BY MODERATORS

1 Posted by u/Linuxthekid **The Mod Who Banned You** 19 hours ago

**Weekly Ben Shapiro Show Discussion thread**

0 Comments Share Save

27 Posted by u/Linuxthekid **The Mod Who Banned You** 7 days ago

**5 Places you can donate to help Israel!** Pinned moderator post

6 Comments Share Save

Posted by u/AnakinSkycocker5726 Facts don't care about your feelings 2 hours ago

**HAMAS BLEW UP A HOSPITAL ON ACCIDENT AND IS GOING TO TRY TO BLAME IT ON ISRAEL: Hundreds reported killed in Gaza hospital explosion amid rocket barrage** Daily Wire

# Paper: Trumping Hate on Twitter?

- Overview
  - The authors sought to determine if Donald Trump's 2016 campaign caused an increase in the amount of hate speech or white nationalist rhetoric used in everyday political discourse.
- Data
  - A set of political Tweets mentioning Donald Trump or Hillary Clinton posted between July 2015 and July 2017.
  - The ADL hate speech database.
  - A set of white nationalist or alt-right subreddits and mainstream political subreddits.
- Methods
  - Dictionary method to count number of tweets with hate speech language.
  - Naive bayes to classify tweet as either white nationalist or mainstream.



# Paper: Language Use & Listener Engagement in Podcasts

- Overview
  - Authors: Spotify Employees
  - Research Goal: How does specific language use in podcasts influence listener engagement?
- Data
  - Spotify Podcast Dataset
    - Transcripts representing the first 10 minutes (most relevant to engagement metrics) of popular podcasts
- Methods
  - “Manual” statistical investigation
    - Determine what language attributes to include in their models
  - Classification Models
    - Input: Podcast Transcript
    - Output: “High Engagement” vs “Low Engagement”
    - Model accuracies were around 70-80%

# Project Specifics

- Data
  - Political Media
    - Spotify Podcast Dataset
    - YouTube podcast transcripts
  - Online Community
    - Subreddit associated with show
    - Tweets mentioning show name or hosts
    - YouTube comments on video
- Method: Snapshot language models
  - N-gram language model
  - Recurrent neural language model
- Evaluation
  - Language model perplexity
    - How confused are our language models after being trained on media and tested on user's comments?
    - For general communities not associated with a single creator, which media most closely mirrors their language?
  - Tracking a single user's language over time

## 5. Yixiao, Dhanush, Ahana

---

# What we want to do and the found related works

We are trying to build a Hate Speech (HS) target classifier, and here are two previous works we found that tried to solve this task:

[Targeted Identity Group Prediction in Hate Speech Corpora.](#) (Sachdeva et al., WOH 2022)

[Robust Hate Speech Detection in Social Media: A Cross-Dataset Empirical Evaluation.](#) (Antypas & Camacho-Collados, WOH 2023).

This first one sets this task as the main goal of the paper, using a RoBERTa based encoder to achieve the accuracies show in the table below, it used MHS dataset to train the model and evaluate it on two other datasets.

HateCheck Corpus		
Identity Group	Accuracy (Chance)	F1 Score
Disability	0.989 (0.869)	0.957
Gender	0.978 (0.739)	0.954
National Origin	0.986 (0.875)	0.941
Race	0.981 (0.871)	0.926
Religion	0.984 (0.869)	0.935
Sexual Orientation	0.993 (0.852)	0.974

Gab Hate Corpus		
Identity Group	Accuracy (Chance)	F1 Score
Disability	0.972 (0.969)	0.237
Gender	0.954 (0.927)	0.636
National Origin	0.868 (0.846)	0.402
Politics	0.788 (0.710)	0.557
Race	0.873 (0.781)	0.622
Religion	0.924 (0.827)	0.773
Sexual Orientation	0.981 (0.954)	0.780

The latter one is focusing on building a Merged Dataset of the current HS datasets and then use the task of target prediction as an example to illustrate that using more data for fine-tuning usually will boost model performance. The table below is the accuracies achieved in this paper.

model	Train	sexism	racism	disability	sexual orientation	religion	other	not-hate	AVG
TimeLMs	All Datasets	<b>72.2</b>	<b>72.9</b>	<b>74.2</b>	<b>76.9</b>	<b>52.6</b>	<b>58.8</b>	<b>90.6</b>	<b>71.6</b>
	HateX	52.1	16.5	0	58.8	31.8	5.8	86.0	35.9
BERTweet	All Datasets	<b>73.1</b>	<b>72.5</b>	<b>74.1</b>	<b>77.6</b>	<b>48.6</b>	<b>59.3</b>	<b>90.9</b>	<b>70.9</b>
	HateX	47.8	6.8	0	43.9	0	0	85.5	26.3
RoBERTa	All Datasets	<b>70.4</b>	<b>72.4</b>	<b>73.9</b>	<b>76.5</b>	<b>47.3</b>	<b>55.5</b>	<b>90.3</b>	<b>69.5</b>
	HateX	50.5	16.3	0	67.9	29.1	7.7	85.5	36.3
BERT	All	<b>68.9</b>	<b>66.3</b>	<b>75.5</b>	<b>69.3</b>	<b>40.3</b>	<b>54.9</b>	<b>93.3</b>	<b>66.9</b>
	HateX	40.4	16.0	0	66.2	15.9	0	85.4	32.0
SVM	All	<b>62.7</b>	<b>67.0</b>	<b>71.5</b>	<b>70.5</b>	4.1	<b>49.0</b>	<b>59.11</b>	<b>81.9</b>
	HateX	20.1	6.0	0	54.9	<b>6.8</b>	0	84.5	24.6
Baseline (most frequent)		0	0	0	0	0	0	84.0	12.0

The latter trained their target classifiers, which include a RoBERTa based model, on both individual datasets and the Merged Dataset that includes the MHS dataset used in the former paper. Below is the summary of the individual datasets collected into the Merged Dataset, three of them have been annotated with targets that include 6 categories.

Dataset	Binary		Multiclass					
	hate	not-hate	racism	sexism	sexual orientation	disability	religion	other
HateE	5303	7364	2474	2829		-		
MHS	2485	5074	735	784	251	21	246	10
DEAP	3727	105	3727			-		
CMS	1237	10861	-	1237		-		
Offense	1142	12547			-			
HateX	2562	5678	757	492	407	30	239	143
LSC	889	1267			-			
MMHS	5392	-	472	764	512	1387	224	2033
HASOC	1237	4348			-			
AYR	393	1246	42	343		-		
AHSD	1363	4088			-			
HTPO	351	2647			-			
HSHP	1498	426	9	1489		-		
<b>Total</b>	<b>27,579</b>	<b>55,651</b>	<b>8,216</b>	<b>7,938</b>	<b>1,170</b>	<b>1,438</b>	<b>709</b>	<b>2,186</b>

As the paper pointed out the F1 score for the disability category is 0 when training the model solely on one individual dataset, HateX, which has a much lower sample size in the disability category compared with the merged dataset. Furthermore, we can see that, for the religion category, the F1 score is the lowest compared with other categories, and it's no surprise that it has the lowest sample size compared with other categories!

## So, what are we going to do based on the former work?

1. As we previously saw, data size matters and matters a lot in building a good classifier, one improvement we could do is simply including more training data for the classifier, here are some of the datasets(with HS targets annotated) we found online that might be used for training the model.

[Social Bias Frames: Reasoning about Social and Power Implications of Language](#) (Sap et al., ACL 2020)

[Learning from the Worst: Dynamically Generated Datasets to Improve Online Hate Detection](#) (Vidgen et al., ACL-IJCNLP 2021)

[Latent Hatred: A Benchmark for Understanding Implicit Hate Speech](#) (EISherief et al., EMNLP 2021)

[ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection](#) (Hartvigsen et al., ACL 2022)

2. The previous work tries to classify the HS into 6 categories, and it's clear that the category(religion) with a much lower sample size has a much lower accuracy being achieved. An question we can ask is what would happen if we include more detailed categories for the HS to be classified into

It's obvious that the number of predicted categories is a big thing to decide on, some dataset annotates the HS with 13 categories(Hartvigsen et al., ACL 2022), some datasets annotate with around 80 categories(Vidgen et al., ACL-IJCNLP 2021) or even around 1000 very detailed categories(Sap et al., ACL 2020). Depending on the datasets we have, we need to choose a certain categorization for building the model, and we might do exploration on this aspect.

3. Depending on the accuracy that the classifier we'll build can achieve, datasets we have at hand and the workload, we might try to use the built classifier to do data analysis on existing datasets to answer some interesting questions like:

- On certain platforms (like twitter), what is the distribution of the hate speech targets? On a per-year scale, as time goes on, how does the hate speech target distribution change?
- Is there a difference between the target distributions for implicit and explicit hate speech?
- How correct will the model be if we use it for identifying the hate speech target for implicit hate speech compared with explicit hate speech?

## Evaluation Metrics

There's no golden benchmark for training and testing our models due to the fact that HS target prediction is a rather small and specific research area, and essentially, we are not really trying to build "new" model, but rather making minor changes to the current existing models being already widely studied in HS research to see if they can perform well for target prediction, in this light, benchmark might be too serious for us to evaluate the model. What we can do instead is to compare our model's performance with the models of the two previously mentioned papers on the same test sets.

The specific metrics we'll use will be F1 score, macro F1 and weighted F1, which are also used in the two previously mentioned papers(so that we can compare ours with them), due to the highly imbalanced nature of HS dataset.

It is important to note that we cannot use the average score(no matter it is macro F1 or weighted F1) to represent the model's practical value, every target group is living being, any form of average would be inherently biased towards all the groups being studied, the more accurate model comparison should be based on F1 score for each individual category. Because we're still not sure about the target categorization method we're going to use and might try to predict more detailed target categories, the evaluation metrics used here might be even less quantitatively and practically meaningful. Therefore, In addition to using the previously mentioned metrics for model comparisons, we will calculate PR AUC as threshold agnostic metric for complementing F1 score for each category, and accuracy over chance to ensure genuine model learning.

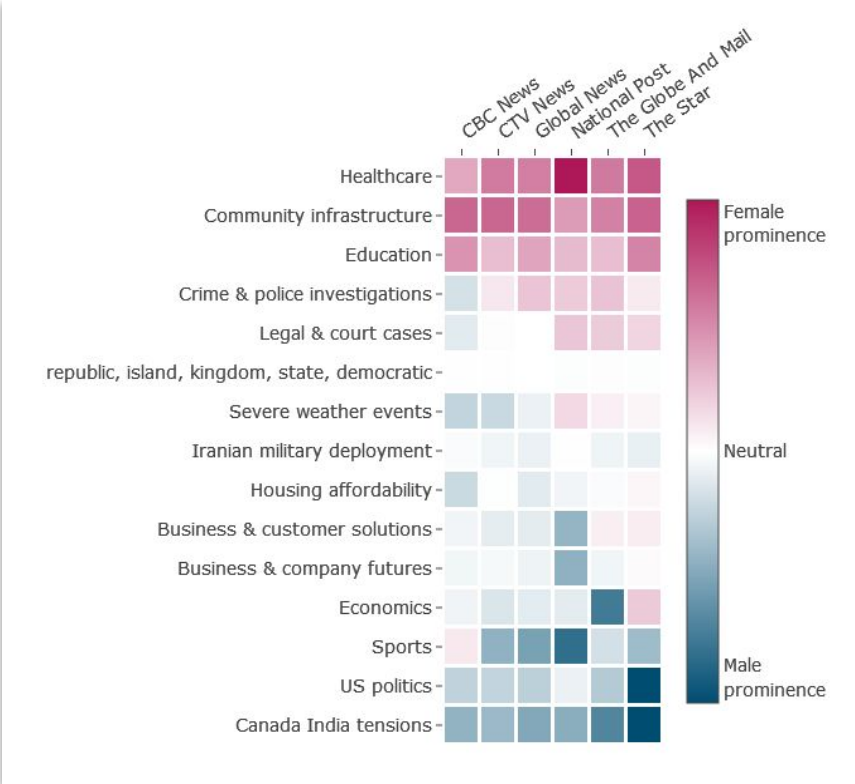


## 6. Norah, Modhumonty, Gina

---

# Topic Modelling for Gender Bias

Explore topic models and examine gender bias in regional news text by topic.



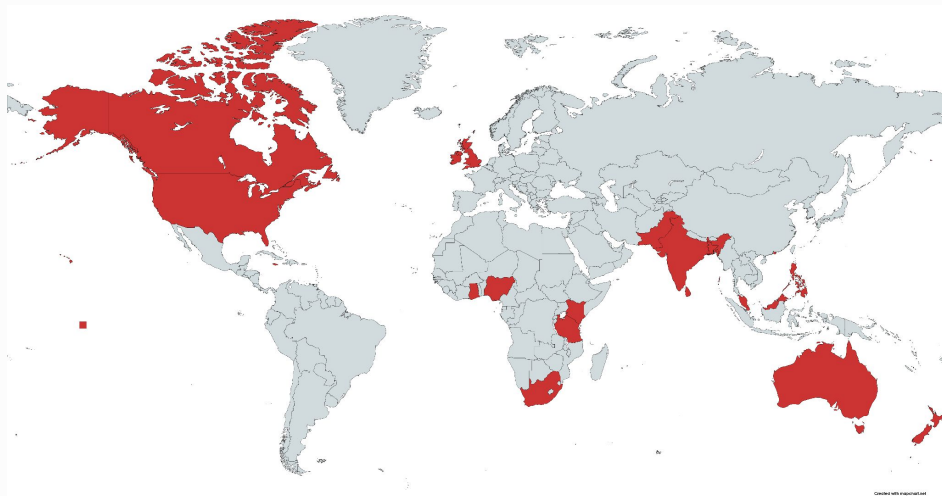
# Related Work

- **Gender Bias in the News: A Scalable Topic Modelling and Visualization Framework<sup>1</sup>**
  - Canadian news
  - Uses times referenced and times quoted as metric
- **Does Gender Matter in the News? Detecting and Examining Gender Bias in News Articles<sup>2</sup>**
  - Analyzing headlines and abstracts
  - Methodology for discovering implicit and explicit gender biases in news

1: <https://www.frontiersin.org/articles/10.3389/frai.2021.664737>

2: <https://dl.acm.org/doi/fullHtml/10.1145/3442442.3452325>

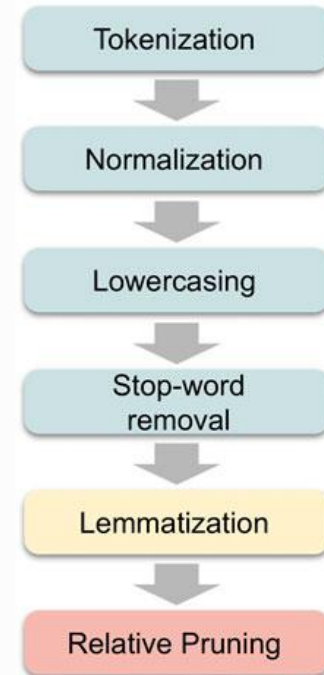
- **News On the Web (NOW) Corpus<sup>1</sup>**
  - Over 18 billion words of data from online news sources
  - English texts from 20 countries spanning the globe
  - Dates ranging from 2010 to present
  - Subset of this data will be used for the project



Countries included in NOW News Corpus: AU, BD, CA, GB, GH, HK, IE, IN, JM, KE, LK, MY, NG, NZ, PH, PK, SG, TZ, US, ZA

# Method - Experimentation

- Data pre-processing
- Testing multiple topic models for performance and quality of topics
  - Latent Dirichlet Allocation (LDA)
  - Structural Topic Model (STM)
  - BERTopic
  - Top2Vec



# Method - Gender Bias Analysis

## Method - Gender Bias Analysis

- Best/favorite model from original experimentation
- Segment data into regions
- Pull topics from regions
- Analyze data for gender bias by regional topic

# Evaluation

## Topic Evaluation

- Quantitative: Perplexity, Coherence, Topic Diversity Score
- Qualitative: Topic Visualization

## Gender Bias Evaluation

- Adhering to gender binary for now
- Representation of gender
- Stereotyping
- Sources or Expert Opinions

## 7. Yuhang and Lingwei

---



# Motivation

- **Speech acts**, which are the actions that speaker intend with utterances (actions like asking questions or making requests), playing a crucial role in understanding the intentions of a speech.
  - **Locutionary Act:** "It's hot in here" has a locutionary act that conveys the information that the speaker believes it's warm in the current environment.
  - **Illocutionary Act:** It includes various speech act types such as assertions, questions, requests, commands, promises, and so on.
  - **Perlocutionary Act:** It relates to how the listener interprets and responds to the speaker's words. It may be the compliance with the request.
- **Emojis** may help automated systems determine speech acts.
- Build a **new dataset** and evaluate it using a simple classifier with interpretable features and see if emojis are informative.

# Related Work

- Tweet Acts: A Speech Act Classifier for Twitter

Logistic regression model with manually labeled tweets.

They explored speech act recognition on Twitter by treating it as a multi-class classification problem

# Datasets and Methods

- Collect data with emojis from social platform like Reddit (older Reddit data) or Twitter dataset (no longer access for them)
  - **annotate** the dataset with emojis
  - **annotate** the dataset without emojis (removing emojis from sentences)
  - will assign **labels** according to the **annotation guide of speech acts**
- Build classifiers using machine learning algorithms (e.g., LSTM, BERT)  
training them separately on datasets with and without emojis to see if including emojis resulting in better performance

# Evaluation

- compare the performance on **different datasets**  
(training separately on datasets with and without emojis)
- use metrics such as accuracy, precision, recall, and F1 score to **assess the emojis' effectiveness.**



## 8. Jiyuan, Ming, Qikun

---

# Motivation

- Lyrics are the soulful bridge that connects the rhythm to human emotions.
- Sentiment analysis could be a powerful tool that can unravel the profound layers of human emotions articulated in song lyrics.
- The existing lyric emotion classifiers have limited performance. We desire a better one.

# Related Papers

- Patra et al. [1] reported that they get accuracy of 51.56% for music mood classification by using clips.
- Çano et al. [2] reported that they get accuracy of 74.25% for music lyrics classification.
- Xia et al. [3] reported a s-SVM-based function and got accuracy of 73.2% for music classification.

[1] Braja Gopal Patra, Dipankar Das, Sivaji Bandyopadhyay. Automatic Music Mood Classification of Hindi Songs. 2013

[2] Erion Çano, M. Morisio. MoodyLyrics: A Sentiment Annotated Lyrics Dataset. 2017

[3] XIA, Y., WANG, L., & WONG, K.-F. (2008). Sentiment vector space model for lyric-based song sentiment classification. International Journal of Computer Processing of Languages, 21(04), 309–330.



# Data

- NJU\_MusicMood V1.0: contains 777 songs with lyrics, and each song comes with one label, representing its sentiment: angry, happy, relaxed, or sad.
- A back-up dataset, containing 1000 songs with emotion labels created by crowdsourcing.

# Methodology

- A hybrid approach, combining different deep learning models:
  - RNN
  - BERT
  - LSTM
  - Lexicon-based sentiment analysis
- Other algorithm or method:
  - SVM

# Evaluation

- Split our dataset into:
  - Training set
  - Validation set
  - Testing set
- Performance metrics will include
  - Accuracy
  - Confusion Matrix
  - Precision and Recall
  - F1-score

## 9. Aziz, Bhiman, Atharva

---

# **Fairness Analysis of Human/AI-Generated Summaries of Student Reflections**

Bhiman Kumar Baghel

Abdulaziz Alotaibi

Atharva Vichare

# Motivation

- Education is Important
- Student Reflections help maintain the quality of education
- #Reflection >>>> #Professor, Hard to analyze -> So Summarize

## Reflection Summarization Research:

### Luo and Litman (2015)

- Introduced "student coverage" concept emphasizing topics frequently mentioned by students.
- Proposed a student coverage-assisted phrase-based summarization algorithm.

### Luo et al. (2016)

- Enhanced the previous model by assessing phrases for their informativeness and student alignment.

**Indicating some bias towards Majority**

# Problem Statement

- How can we ensure that automatic summarization algorithms are unbiased, especially when summarizing student reflections?
- Do these algorithms fairly represent both the majority and minority topics within reflections?
- Do the demographics and gender of students influence their outputs?
- It is essential to analyze and ensure fairness in these systems to detect and address any biases.

# Related Work

## Research on Fairness in Summarization/Scoring Algorithms:

**Litman et al. (2021)**

- Evaluated fairness in Automated Essay Scoring (AES).
- Identified biases across gender, race, and socioeconomic status.

**Huang et al. (2023)**

- Investigated bias in opinion summarization, not/ill considering opinion diversity.

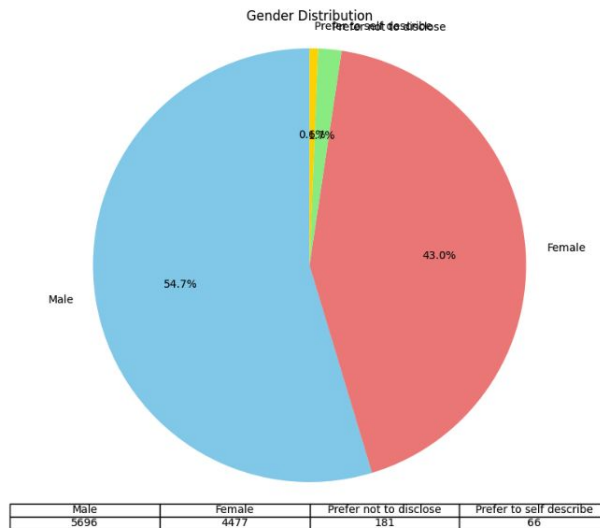
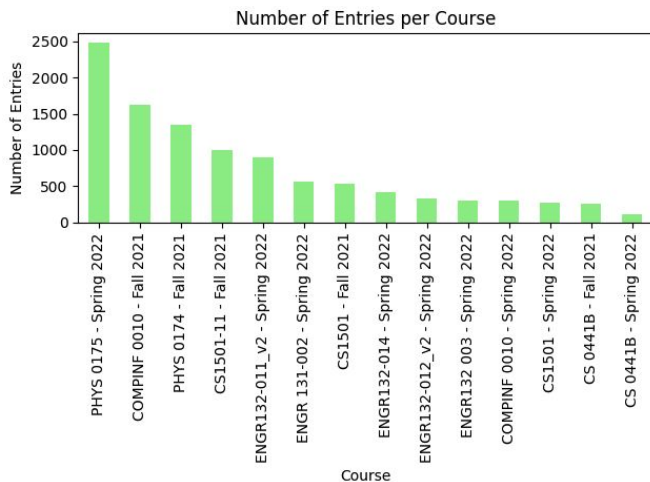
**Dash et al. (2019)**

- Highlighted that some summarization algorithms present socially prominent user groups in a manner divergent from their original data representation.



# Dataset - CourseMirror (Fan et al., 2015)

- 10k Student Reflections, each answering three questions:
  - 1) Describe what you found most interesting in today's class
  - 2) Describe what was confusing or needed more detail
  - 3) Describe what you learn about how you learn.
- Includes abstractive, extractive, and phrase-based summaries of these reflections, produced by both humans and AI



# Research Plan

1. Demographic statistical analysis
2. Exploratory text/correlations with demographic information
3. Comparison between AI and human summarizations with specificity score
4. Build a classifier to see how well it can predict demographics from the text



# Q&A

## 10. Ben, Max, Tom

---

# Motivation

- Creating patient notes can be a time consuming tasks for physicians and nurses
- As a summary of other data (lab, physio, diagnoses, etc.) collected throughout the patient's stay in the hospital, discharge notes can be particularly challenging
- These notes require significant amounts of domain specific knowledge, making it difficult for various tasks, especially generation
- Being able to generate patient notes can be helpful in downstream tasks like detecting medication / diagnoses outliers

# Related Work

- [1, 2] are language models fine-tuned on clinical data from the MIMIC-III database but do not address the model capability of text generation and has not been evaluated on such tasks.
- [3] is a model finetuned on the MIMIC-III dataset. This model focuses on generating discharge notes to be used as artificial data for training other models. Their model expects partially written discharge notes as an input, requires EHR features to be formulated in text, does not take temporal information into account, uses only a small set of hand picked EHR features, and only generates alternative discharge notes.

[1] Huang, K et al. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission <https://arxiv.org/abs/1904.05342>

[2] Alsentzer, E et al. 2019. Publicly Available Clinical BERT Embeddings <https://arxiv.org/abs/1904.03323>

[3] Amin-Nejad, A et al. 2020. Exploring Transformer Text Generation for Medical Dataset Augmentation <https://aclanthology.org/2020.lrec-1.578.pdf>

# Methodology

- Goal: generate clinical (discharge) notes using only structured time series data
- EHR features: physiological signals, medications, procedures, lab results
- Two encoders (TS and text) and one decoder (text)
- Training
  - text encoder/decoder: finetune pretrained language models (e.g. BERT, GPT, LLaMA)
  - data: MIMIC-III note events data in
  - unsupervised: mask out words, context prediction, NLI
  - EHR encoder: match the representations of the text encoder
- Inference and Evaluation
  - input: EHR time series features only
  - encode with the EHR encoder
  - decode into discharge summaries using the text decoder

# Dataset and Evaluation

- MIMIC-III dataset: demographic
  - EHR records, diagnoses, notes
  - widely used in researches on clinical data
- Same evaluation metrics for text generation as [3]
  - negative perplexity (neg. PPL)
  - BLEU score
  - ROUGE-2 and ROUGE-L
- Adapt the learned encoders and hidden representations to downstream tasks
  - mortality prediction or discharge prediction
  - evaluation strategy from representation learning
  - evaluates the quality of features learned

[3] Amin-Nejad, A et al. 2020. Exploring Transformer Text Generation for Medical Dataset Augmentation <https://aclanthology.org/2020.lrec-1.578.pdf>



# 11. Vincent, Lokesh, Shuhao, Shijia

---

# Motivation and Goal

## **Motivations:**

- Summarization ability is important in data collection and analysis. It shows the efficiency of data collection and information evaluation.
- The category of movie is diverse and topics are plenty.
- People love movies. This model can save time for searching a movie.

## **Goal:**

An automatic text summarization machine based on movie subtitles

**Example:** An ex-hitman comes out of retirement to track down the gangsters who killed his dog and stole his car.

# Related Work

- A Combined Extractive With Abstractive Model for Summarization [1]  
First the top-k important sentences are extracted by using Encoder model and use Beam search to rewrite syntactic blocks and extracted sentences. Trained on News datasets.
- Extractive Summarization Considering Discourse and Coreference Relations based on Heterogeneous Graphs [2]  
Focuses on extractive summarization with heterogeneous graph based model that incorporates both discourse and coreference relations.
- Movie Summarization based on Indonesian Subtitles with Restricted Boltzmann Machine [3]

[1] W. Liu, Y. Gao, J. Li and Y. Yang, "A Combined Extractive With Abstractive Model for Summarization," in IEEE Access, vol. 9, pp. 43970-43980, 2021, doi: 10.1109/ACCESS.2021.3066484.

[2] Huang & Kurohashi, EACL 2021 <https://aclanthology.org/2021.eacl-main.265>

[3] S. I. G. Situmeang, R. K. Lubis, F. J. N. Siregar and B. J. D. C. Panjaitan, "Movie Summarization based on Indonesian Subtitles with Restricted Boltzmann Machine,"

# Dataset

## **CMU Movie Summary Corpus:**

A dataset contains movie basic information and summary.

For 42,306 movies.

## **IMDB OpenSubtitles Corpus:**

A dataset contains movie subtitles of multiple languages.

Large database of movie and TV subtitles and includes a total of 1689 bitexts spanning 2.6 billion sentences across 60 languages

[1] Learning latent personas of film characters.

[2] OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles (<https://aclanthology.org/L16-1147>) (Lison & Tiedemann, LREC 2016)

# Project Methodology

## 1. Collecting and Merging the data

Match summaries and subtitles from existing open datasets

Collect more data if needed - wikipedia for summaries, opensubtitles for subtitles

## 2. Training

### a. Extractive Model Fine Tuning:

(Extracting sentences from subtitles)

Model: BERTSUM/T5 small

Input Data: Subtitles

Output: Top k sentences

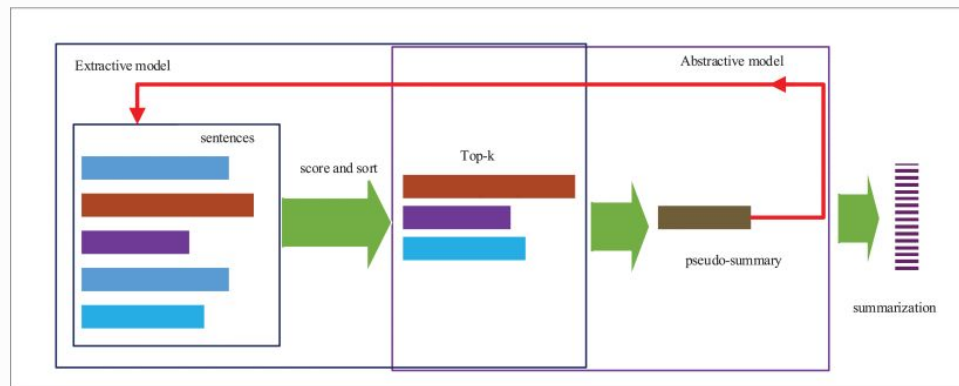
### b. Abstractive Model Fine Tuning:

(Abstracting summary for the movie)

Model: Model T5/FLAN-T5

Input Data: Output of Extractive model

Output: Summary of the Movie



# Evaluation

- Extractive Summarization: Comparing performance with LexRank, LSA
- Abstractive Summarization:

Evaluation of the abstractive model will be done using: F1 scores, ROUGE  
Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scoring algorithm

- ROUGE 1 - overlap of unigrams
- ROUGE 2 - overlap of bigrams
- ROUGE L - Longest common subseq.

$$R = \frac{W_{matched}}{M},$$
$$P = \frac{W_{matched}}{N},$$
$$F1 = \frac{2 \times P \times R}{P + R}$$