CSE 635: NLP and Text Mining

Spring 2023

Instructor: Rohini K. Srihari

Group Project Description and Requirements

Overview

The goal of this semester-long project is to provide hands-on experience designing, implementing, evaluating, and demonstrating a complete web mining/text mining/social media mining solution based on a combination of natural language processing (NLP), information retrieval (IR) and machine learning (ML) techniques. You are provided a choice of four topics that broadly fall into the area known as AI for Social Impact. The topics cover generated text detection, fact hallucination detection/prevention, clickbait spoiling, and chatbots. Each project will have a standard dataset and ground truth, enabling quantitative evaluation. Many of these are from past or ongoing challenges and have been attempted by other teams. We encourage you to use any available online tools or platforms to develop your solution. You should strive to produce results in the top 10% of any previously published results on the same dataset. While there is a quantitative evaluation component on a static data set, we are also requiring you to develop a live demo system. This may involve developing a user interface so you can demonstrate the system.

This project will satisfy the MS project requirements specified by the CSE department. While the problem definition and evaluation dataset have been fixed, there is ample room for creativity on your part in further enhancement of the solution and implementation. Be creative, and most importantly, pace yourself properly during the semester. Your project is divided into three phases which are described in more detail later on in this document:

Phase 1: Submission of project proposal and in-person presentation of your proposal. This includes a comprehensive literature review on your selected topic, a necessary step before you begin designing your system!
Phase 2: Interim report describing the evaluation of the baseline system.
Phase 3: Final submission of technical paper, code, and in-class presentation of your end-to-end system.

Project Option 1: Authorship Attribution for Neural Text Generation (Synthetically generated text detection)

Background: In recent years, the task of generating realistic short and long texts has made tremendous advancements. In particular, several recently proposed neural network-based language models have demonstrated their great capabilities to generate texts that are challenging to distinguish from human-written texts with the naked eye. Despite many benefits and utilities of such neural methods, in some applications, determining the "author" of a text in question becomes critically essential (for example, detecting *AIgiarism*).

PROMPT	Putin and Xi are using the coronavirus crisis to extend their control. Across the world, Trump is struggling to keep up
Human	. . . Donald Trump's shock announcement shutting down travel from much of Europe came with the customary back-pedaling common to many of his hastily announced initiatives. At first, Trump said trade and travel would be targeted, in his latest attempt to stem the spread . . .
CTRL	. . . . President Vladimir Putin, meanwhile, says there was no reason why Moscow would spread such diseases. But both leaders have been unable so far to . . . [read more] what do you think about this story? do we need another Ebola scare or two? . . .
GPT	. . . with them in a war that has been going on for decades now - but he's still not strong enough. " " i'm sorry about your family. it must be hard being so far away from home. " she said softly. " yes, we're all very lonely here. there wasn't much time before . . .
GPT2	. . . Putin's paranoia over the contagion of Ebola is a common theme. As his recent visit to Washington shows, he's playing a dangerous game. And here's what Putin wants Putin wants to use Ebola as a pretext for imposing further Russian sanctions on the West. . . .
GROVER	. . . hours after world cup workers hand-picked by Putin to fix problems faced by workers in turkey and Venezuela, a new threat to trump is receding.According to Wikileaks, the Coronavirus outbreak is receding and the Russian leader, Putin,. . .
XLM	. . . with Russia alone ruling over Europe while china under Mao Zedong rules Asia in turn which makes it difficult for him politically but financially as well being president / CEO of America first bank & chairman's nominee at CIA . . .
XLNET	. . . with china or Russia in his efforts at de-destructuring America from the financial crisis that has taken place over the past two years of this century as well a global economic downturn that will continue for many more months into next year . . .
PPLM	. . . the latest developments in Syria, Ukraine, Russia's intervention in the middle east and china's rise to global power. In Moscow, a group of Russian diplomats, who did not want their real names to be reported on the record. . .
FAIR	.. He has ordered his federal agencies to step up efforts to combat it and to warn of new threats. in Washington, Trump's national security adviser, h.r. mcmaster, said on Monday the U.S. government had seen signs of the bug and that countries in the Middle East were on high alert. . . .

Table 1: Snippets of nine texts using the titles of human-written articles as the prompt to neural methods.

Paper: https://aclanthology.org/2020.emnlp-main.673.pdf

Tasks

- **Task 1 (Same Method or Not)** Given two texts T1 and T2, determine if both T1 and T2 are generated by the same NLG method (or human) or not.
- **Task 2 (Human vs. Machine)** Given a text T1, determine if T1 is written by humans or generated by one of k NLG methods.
- **Task 3 (Authorship Attribution)** Given a synthetic text T1, single out one NLG method (among k alternatives) that generated T1.

Dataset: We have data for nine text generators–i.e., one human writer and eight neural machine generators. You are also required to generate articles using the GPT3 and InstructGPT text generators (described in the next section). All the existing eight neural generators require a short prompt to begin their generation and the number of words to generate. These eight generators were chosen because we found they had the best pre-trained models for our task. We used the titles of news articles (written by human journalists) as the prompt and set 500 as the number of words.

Measure	Human	Machine							
		CTRL	GPT	GPT2	GROVER	XLM	XLNET	PPLM	FAIR
# of samples	1,066	1,066	1,066	1,066	1,066	1,066	1,066	1,066	1,066
AVG word count	432.31	530.03	345.03	199	356.76	441.32	452.58	228.89	250.42
SD word count	270.82	73.51	10.79	74.15	114.96	34.67	32.59	64.13	39.94
AVG sentence count	26.87	33.02	32.64	15.68	21.64	3.97	5.02	13.53	17.53
SD sentence count	19.49	21.18	5.55	6.99	9.65	1.71	1.97	4.61	4.88

Table 2: Summary statistics of nine generated texts (one by human and eight by neural methods).

Measure	Human	Machine							AVG
		CTRL	GPT	GPT2	GROVER	XLM	XLNET	PPLM	FAIR	
Flesch Reading Ease	37.97	60.97	68.68	54.49	46.63	46.40	48.94	44.97	51.85	51.21
Flesch-Kincaid Grade	12.79	9.58	8.48	10.27	11.53	11.64	11.28	11.66	10.76	10.89
LIWC-Authentic	25.3	54.28	61.66	15.1	23.76	48.06	80.69	34.27	18.77	40.21
LIWC-Analytic	89.81	51.99	40.93	92.59	89.98	78.61	50.46	73.18	92.89	73.38
LIWC-Article	7.98	1.47	3.18	11.87	8.69	0.59	2.03	2.6	10.05	5.38
Entropy	7.81	8.98	8.01	6.52	7.79	8.99	8.91	7.77	7.41	8.02

Table 3: Linguistic features of nine generated texts.

Dataset link:
https://github.com/AdaUchendu/Authorship-Attribution-for-Neural-Text-Generation/tree/master/data

Dataset Creation

The dataset provided does not contain GPT3 generated data which is more challenging to detect as compared to data generated by other language models. Therefore, you are required to generate GPT3-generated data using the OpenAI API (https://openai.com/api/). The given dataset consists of 1066 prompts; use those prompts to generate articles using the following models and hyperparameters:

GPT3 Model	Max Length	Temperature	Top P
text-curie-001 (GPT3)	500	0.9	1
text-davinci-003 (InstructGPT)	500	0.9	1

Note: OpenAI API is not free. However, if you use your ub email to sign-up, you will receive 18$ worth of credits, which is more than enough to create this dataset. Please divide this task among your teammates, so you don't run out of credits.

Linguistics Features Requirement

We would like you to use several linguistic features in Tasks 1 to 3. Some of the features are(but are not limited to):

- LIWC features.
- Discourse structure of the article.
- Rhetorical Style(RST).
- POS statistics.

Reddit Case Study Requirement

You are also required to carry out a small case study by scraping a small number(>= **500**) of opinionated Reddit articles and generating corresponding GPT3 and InstructGPT texts, thus creating a small dataset. Some of the subreddits of interest are r/changemyview, r/Advice, r/relationships, etc. Please use the Reddit post's title as a prompt for text generation. Also, make sure the Reddit posts you are scraping have a body of considerable length(~500 tokens).

After constructing the dataset, please report the results for Tasks 1 to 3. Also, do a detailed error analysis.

Evaluation Metrics:

- **Task 1:** Macro-averaged F1 is used to evaluate the classification performance.
- **Task 2:** Binary classification performance in Macro-averaged F1 score of "Human vs. Machine" on ten individual test sets.
- **Task 3:** Multi-class classification performance with per-class macro F1 and overall average F1 scores of models.

Bonus points:

Bonus points will be awarded to teams that successfully submit a well-written paper (with the teaching team as co-authors) to the "The 3rd Workshop on Trustworthy NLP", co-hosted with ACL 2023.

Recommended team size: 4 students

Project Option 2: Faithful Benchmark for Information-Seeking Dialogue (Fact Hallucinations Detection and Prevention)

Background: The goal of information-seeking dialogue is to respond to queries with natural language utterances that are grounded on knowledge sources. However, dialogue systems often produce unsupported utterances, a phenomenon known as *hallucination*. To mitigate this behavior, we adopt a data-centric solution and create FAITHDIAL, a new benchmark for hallucination-free dialogues, by editing hallucinated responses in the Wizard of Wikipedia (WOW) benchmark.

Paper: https://arxiv.org/pdf/2204.10757.pdf

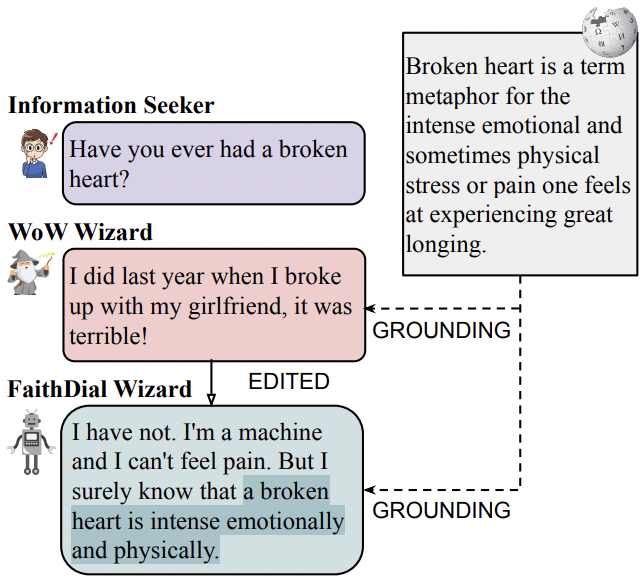


Figure 1: A representative FAITHDIAL annotation: subjective and hallucinated (red) information present in the wizard's utterance of WoW data are edited into utterances faithful to the given knowledge (green). In FAITHDIAL, the wizard assumes the persona of a bot.

Definitions

Definition 2.1 (Faithfulness). *Given an utterance u_n, a dialogue history $\mathcal{H} = (u_1, \ldots, u_{n-1})$, and knowledge $\mathcal{K} = (k_1, \ldots, k_j)$ at turn n, we say that u_n is faithful with respect to $\mathcal{K}$ iff the following condition holds:*

- *$\exists \Gamma_n$ such that $\Gamma_n \vDash u_n$, where $\vDash$ denotes semantic consequence and Γ_n is a non-empty subset of $\mathcal{K}_n$. In other words, there is no interpretation $\mathcal{I}$ such that all members of Γ_n are true and u_n is false.*

Hence, an utterance can optionally be grounded on multiple facts but not none.

Definition 2.2 (INFORMATION SEEKER: A Human). *The information* SEEKER*, a human, aims at learning about a specific topic in a conversational manner. They can express subjective information, bring up a new set of facts independent from the source $\mathcal{K}$, and even open up new sub-topics.*

Definition 2.3 (WIZARD: A Bot). *The Wizard, a bot, aims at conversing in a knowledgeable manner about the* SEEKER*'s unique interests, resorting exclusively to the available knowledge $\mathcal{K}$. They can reply to a direct question or provide information about the general topic of the conversation.*[3]

Dataset

Link: https://mcgill-nlp.github.io/FaithDial/

Dataset	Train	Valid	Test
Turns	36809	6851	7101
Conversations	4094	764	791
Avg. Tokens for WIZARD	20.29	21.76	20.86
Avg. Tokens for SEEKER	17.25	16.65	16.49
Avg. Tokens for KNOWLEDGE	27.10	27.17	27.42
Turns per Conversation	9	9	9

Table 2: Dataset statistics of FAITHDIAL.

(a) FAITHDIAL (b) WOW

Figure 2: Coarse-grained (BEGIN) and fine-grained speech act (VRM) distributions used by wizards in FAITHDIAL and WOW. The inner most circle shows the breakdown of coarse-grained types: Hallucination (red), Entailment (green), Partial Hallucination (yellow), Generic (purple), and Uncooperative (pink). The outer circles show the fine-grained types of each coarse-grained type.

Data Format:

```
{'dialog_idx': 5,
 'response': 'I see. Well, a majority of teller jobs needs ability to
handle the cash and have a diploma from high school.',
```

```
 'original_response': 'i think most bank job require  experience with
handling',
 'history': array(['I have been trying for years to become a bank
teller.'],
       dtype=object),
 'knowledge': 'Most teller jobs require experience with handling cash and
a high school diploma.',
 'BEGIN': array(['Hallucination', 'Entailment'], dtype=object),
 'VRM': array(['Disclosure'], dtype=object)}
```

Tasks

Task 1: Hallucination Critic: Given the conversation history and knowledge, you will have to determine whether the response is hallucinated.
Task 2: BEGIN and VRM Multi-Class Multi-label Classification: Given the conversation history and knowledge, you will have to identify the speech acts (VRM taxonomy such as disclosure, edification, question, acknowledgment, etc.) and the response attribution classes (BEGIN taxonomy) such as hallucination, entailment, etc.
Task 3: Dialogue Generation: Given the conversation history and knowledge, you will have to generate a response that is faithful to the conversation history and knowledge.

Evaluation Metrics:

- **Task 1:** Macro-averaged F1 to evaluate the classification performance.
- **Task 2:** Macro-averaged F1 and confusion matrix to evaluate the classification performance.
- **Task 3:** Critic(percentage of utterances identified as unfaithful, using Task 1 model), BLEU, ROUGE, BERTScore for accessing the generation quality.

Additional Requirements:

- Use additional linguistic features(POS tags, Discourse structure, etc.) as control tokens or any other way to encode those features. Please experiment with multi-encoder pipelines to encode knowledge and conversation history.
- Use contrastive learning loss while training. Come up with more innovative ways to perturb responses(for negative responses) than those given in the paper.
- (Optional) Experiment with curriculum learning to understand if it affects reducing hallucinations.

Caution:

This project has strong baselines with a codebase. If your approach is strikingly similar to the baseline, your project will be heavily penalized and may be reported for AI violation.

Readings:

- Wizard of Wikipedia: Knowledge-Powered Conversational agents https://arxiv.org/pdf/1811.01241.pdf
- BEGIN Benchmark: https://aclanthology.org/2022.tacl-1.62/
- Describing talk: A taxonomy of verbal response modes. (VRM Taxonomy) https://doi.org/10.1002/9781119102991.ch69

Bonus points:

Bonus points will be awarded to teams that successfully submit a well-written paper(with the teaching team as co-authors) to the "The 3rd Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc)", co-hosted with ACL 2023.

Maximum team size: 3 students

Project Option 3: Clickbait Challenge at SemEval 2023 - Clickbait Spoiling

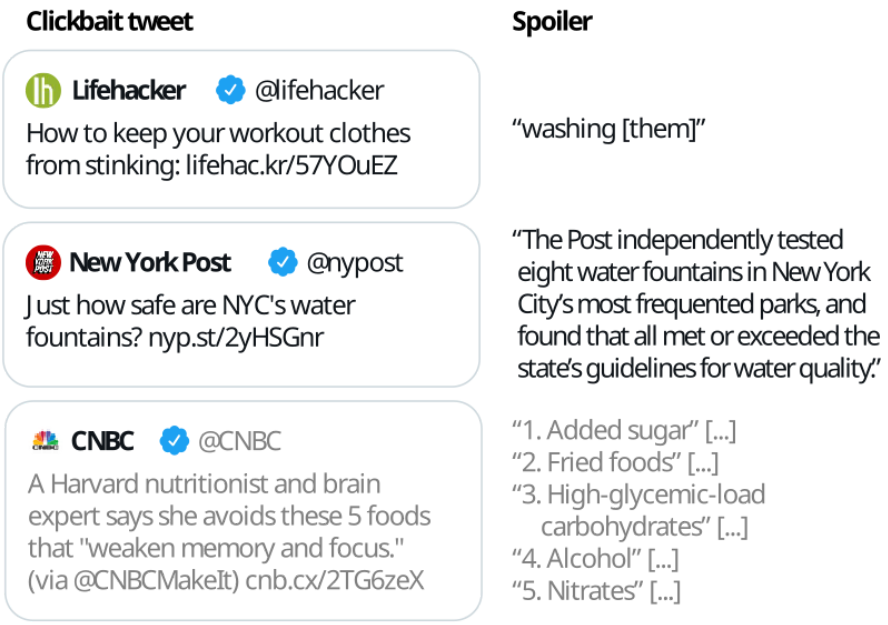
URL: https://pan.webis.de/semeval23/pan23-web/clickbait-challenge.html

Synopsis

Clickbait posts link to web pages and advertise their content by arousing curiosity instead of providing informative summaries. Clickbait spoiling aims at generating short texts that satisfy the curiosity induced by a clickbait post.

Task

The following figure illustrates some example inputs and the expected output for clickbait spoiling:



- Task 1 on Spoiler Type Classification: The input is the clickbait post and the linked document. The task is to classify the spoiler type that the clickbait post warrants (either "phrase", "passage", "multi"). For each input, an output like {"uuid": "<UUID>", "spoilerType": "<SPOILER-TYPE>"} has to be generated where <SPOILER-TYPE> is either phrase, passage, or multi.

- Task 2 on Spoiler Generation: The input is the clickbait post and the linked document (and, optional, the spoiler type if your approach uses this field). The task is to generate the spoiler for the clickbait post. For each input, an output like {"uuid": "<UUID>", "spoiler": "<SPOILER>"} has to be generated where <SPOILER> is the spoiler for the clickbait post.

Data [https://zenodo.org/record/6362726#.YsbdSTVBzrk]

The dataset contains the clickbait posts and manually cleaned versions of the linked documents, and extracted spoilers for each clickbait post (the dataset was constructed and published in the corresponding paper). Additionally, the spoilers are categorized into three types: short phrase spoilers, longer passage spoilers, and multiple non-consecutive pieces of text.

The training and validation data is available for download on zenodo. This dataset contains 3,200 posts for training (the file training.jsonl) and 800 posts for validation (the file validation.jsonl). After training and validation, systems are evaluated on 1,000 test posts.

Input Format

The data comes in JSON Lines format (.jsonl) where each line contains a clickbait post and the manually cleaned version of the linked document. For each line, the goal is to classify the spoiler type needed (task 1), and/or to generate the spoiler (task 2).

For each entry in the training and validation dataset, the following fields are available:

- uuid: The uuid of the dataset entry.
- postText: The text of the clickbait post which is to be spoiled.
- targetParagraphs: The main content of the linked web page to classify the spoiler type (task 1) and to generate the spoiler (task 2). Consists of the paragraphs of manually extracted main content.
- targetTitle: The title of the linked web page to classify the spoiler type (task 1) and to generate the spoiler (task 2).
- targetUrl: The URL of the linked web page.
- humanSpoiler: The human generated spoiler (abstractive) for the clickbait post from the linked web page. This field is only available in the training and validation dataset (not during test).
- spoiler: The human extracted spoiler for the clickbait post from the linked web page. This field is only available in the training and validation dataset (not during test).
- spoilerPositions: The position of the human extracted spoiler for the clickbait post from the linked web page. This field is only available in the training and validation dataset (not during test).
- tags: The spoiler type (might be "phrase", "passage", or "multi") that is to be classified in task 1 (spoiler type classification). For task 1, this field is only available in the training and validation dataset (not during test). For task 2, this field is always available and can be used.
- Some fields contain additional metainformation about the entry but are unused: postId, postPlatform, targetDescription, targetKeywords, targetMedia.

The following is a simplified entry in the dataset (line breaks added for readability):


```
{

 "uuid": "0af11f6b-c889-4520-9372-66ba25cb7657",

 "postText": ["Wes Welker Wanted Dinner With Tom Brady, But Patriots QB Had Better Idea"],

 "targetParagraphs": [

  "It’ll be just like old times this weekend for Tom Brady and Wes Welker.",

  "Welker revealed Friday morning on a Miami radio station that he contacted Brady because he’ll be in town for Sunday’s game between the New England Patriots and Miami Dolphins at Gillette Stadium. It seemed like a perfect opportunity for the two to catch up.",
```

"But Brady’s definition of \"catching up\" involves far more than just a meal. In fact, it involves some literal \"catching\" as the Patriots quarterback looks to stay sharp during his four-game Deflategate suspension.",

"\"I hit him up to do dinner Saturday night. He’s like, ‘I’m going to be flying in from Ann Arbor later (after the Michigan-Colorado football game), but how about that morning we go throw?’ \" Welker said on WQAM, per The Boston Globe. \"And I’m just sitting there, I’m like, ‘I was just thinking about dinner, but yeah, sure. I’ll get over there early and we can throw a little bit.’ \"",

"Welker was one of Brady’s favorite targets for six seasons from 2007 to 2012. It’s understandable him and Brady want to meet with both being in the same area. But Brady typically is all business during football season. Welker probably should have known what he was getting into when reaching out to his buddy.",

"\"That’s the only thing we really have planned,\" Welker said of his upcoming workout with Brady. \"It’s just funny. I’m sitting there trying to have dinner. ‘Hey, get your ass up here and let’s go throw.’ I’m like, ‘Aw jeez, man.’ He’s going to have me running like 2-minute drills in his backyard or something.\"",

"Maybe Brady will put a good word in for Welker down in Foxboro if the former Patriots wide receiver impresses him enough."

],

"targetTitle": "Wes Welker Wanted Dinner With Tom Brady, But Patriots QB Had A Better Idea",

"targetUrl": "http://nesn.com/2016/09/wes-welker-wanted-dinner-with-tom-brady-but-patriots-qb-had-better-idea/",

"spoiler": ["how about that morning we go throw?"],

"spoilerPositions": [[[3, 151], [3, 186]]],

"tags": ["passage"]

}

Output Format [validator]

The output format for task 1 and task 2 is identical but other fields are mandatory. Please submit your results in JSON Lines format producing one output line for each input instance.

Each line should have the following format: {"uuid": "<UUID>", "spoilerType": "<SPOILER-TYPE>", "spoiler": "<SPOILER>"}

where:

- <UUID>is the uuid of the input instance.
- <SPOILER-TYPE> is the spoiler type (might be "phrase", "passage", or "multi") to be predicted in task 1. This field is mandatory for task 1 but optional for task 2 (to indicate that your system used some type of spoiler type classification during the spoiler generation).
- <SPOILER> is the generated spoiler to be produced in task 2. This field is mandatory for task 2.

We provide code and example outputs that you can validate your submissions in this GitHub repository.

Evaluation Metrics:

- **Task 1:** Macro-averaged F1, accuracy to evaluate the classification performance.
- **Task 2:** BLEU, METEOR, and BERTScore for accessing the generation quality.

Readings:

Clickbait Spoiling via Question Answering and Passage Retrieval
https://aclanthology.org/2022.acl-long.484/

Bonus points:

Bonus points will be awarded to teams that successfully submit a well-written paper (with the teaching team as co-authors) to the "The 7th Workshop on Online Abuse and Harms", co-hosted with ACL 2023.

Caution:

This project has strong baselines with a codebase. If your approach is strikingly similar to the baseline, your project will be heavily penalized and may be reported for AI violation.

Maximum team size: 3 students

Project Option 4: Topic-Based Empathic Chatbot

Background: The advances in deep learning combined with the availability of large, diverse corpora have led to significant progress in the field of conversational AI, commonly referred to as chatbots. The first phase was targeted at task-oriented applications such as customer service

and relied on rule-based systems. Currently, there is the widespread use of commercial chatbots such as Alexa and Google. At the same time, these systems reflect more advanced machine learning technology, their intended use is primarily to inform, entertain and perform simple tasks for users. More recently, chatbots have been configured to be more empathetic and to reflect various personas in an attempt to engage more deeply with users; the term **socialbots** is now more commonly used. There has been an increased interest in using these socialbots for societal applications, including helping people with amnesia or those experiencing isolation. This project requires developing a socialbot that can engage in open-domain chit-chat and topical conversations.

The aim of this project is to create an end-to-end conversational system that can engage in both open-domain chit-chat and topical conversations. The system should input a query text from the user and generate a text response befitting the query and conversation context. You should extend your IR-based chatbot from CSE 4/535 and enhance it to create an end-to-end conversational system comprising IR and neural-based response generators. Ideally, your system should also implement a "planning component" a.k.a dialogue manager (DM) which (i) determines which generator(s) to invoke in a given conversational state, (ii) chooses the best response from a set of generated candidate responses. Your system should be able to:

1. Respond with empathy.
2. Share opinions whenever required.
3. Share facts whenever suitable. Only limited to the 5 topics from CSE 4/535: Politics, Environment, Technology, Healthcare, and Education.

Dataset: You are required to use the following 3 datasets for this project:

1. The CSE 4/535 dataset comprising post-comment pairs from Politics, Environment, Technology, Healthcare, and Education that you already have scraped. You can collect more data around the 5 topics if needed.
2. The Empathetic Dialogues (ED) dataset, which comprises 25k conversations grounded in emotional situations. (https://github.com/facebookresearch/EmpatheticDialogues).
3. The BYU-PCCL chit-chat (CC) dataset that you already used in CSE 4/535.

   Note:

   1. Use the standard training, validation, and testing splits for the ED dataset. The evaluation metrics must be calculated on the test set.

   2. For the CC dataset, create your own training, validation and testing splits by considering the first 80% examples as training, the next 10% as validation, and the remaining 10% as test samples. DO NOT change the ordering of the examples before you create the data splits. Else your results will be incomparable.

   3. You are free to use additional datasets.

Task Definitions: Given a conversation's context C and the current query Q, create a conversational system that yields the most suitable response R: the output of an IR-based model, a neural language model (LM) that learns the conditional distribution p(R|C, Q), or a combination of both. At every turn, the system must generate multiple candidate responses <R_a, R_b, ..., R_n>. A rule-based or neural dialogue manager (DM) should select the most appropriate response. All neural LMs should be trained by minimizing the language modeling loss between the generated response R and the golden response Y.

Evaluation Metrics:

Automatic Metrics

The following metrics should be calculated between the generated response in the test set and the golden response and applies only to the ED and CC datasets. Each metric should be compared against suitable external and internal baselines.

1. Perplexity (https://en.wikipedia.org/wiki/Perplexity)
2. BLEU score (https://en.wikipedia.org/wiki/BLEU)
3. ROUGE score (https://en.wikipedia.org/wiki/ROUGE_(metric))
4. BERT Score (https://github.com/Tiiiger/bert_score)
5. BLEURT Score (https://github.com/google-research/bleurt)

Human Evaluation

We will test your chatbot and rate it on a Likert scale of 1 (low) to 5 (high) on the following metrics:

1. Fluency: Is the generated response natural, fluent and grammatically correct? This metric tests the syntax of the response.
2. Coherence: Is the response appropriate w.r.t the query and context?
3. Empathy: Does the response exude appropriate emotion?
4. Factual: Is the response factually correct (if a fact is expected as a response)

Bonus Points: Bonus points will be awarded to teams that successfully submit a well-written paper (with the teaching team as co-authors) to the "The 5th Workshop on NLP for Conversational AI", co-hosted with ACL 2023.

References:

Proto: https://arxiv.org/pdf/2109.02513.pdf
Huggingface Transformers: https://huggingface.co/docs/transformers/index
TransferTransfo: https://arxiv.org/abs/1901.08149
ACL 2023 Workshops: https://2023.aclweb.org/program/workshops/

Maximum team size: 3 students

What to submit

You should plan on preparing for the following:

1. **Project proposal**: Your proposal must contain the following sections:

- Problem Statement - define the problem you are trying to solve, your objectives.
- Literature Study - background reading on some state-of-the-art results, summarize them.
- Dataset - details on the dataset, and how the dataset is processed and adapted by your system.
- Evaluation - which evaluation metrics are being used?
- Proposed System - high-level architecture of your proposed system followed by a detailed explanation of each component of it.
- Project Plan and Timeline - a clear plan of your project – who does what and the targets for each milestone.

2. In-person presentation of the project plan, and plans for baseline system

3. A midterm report describing the baseline system and initial evaluation results

4. Final in-class presentation

5. Project report in conference paper format

Grading

Milestone 1 (10%): Project Proposal (week of Feb 27th)

- Literature Review
- Project objectives
- Data set, features to be implemented
- Evaluation methodology
- Project plan
- Presentation of project plan

Milestone 2 (15%): Baseline results (week of March 27th)

Milestone 3 (25%): Final Project Presentation (May 10th)

- In class presentation
- Project report (ACL 2023 paper format), code and PPT to be submitted
- All deliverables due by May 12th - Friday

All project related discussion will be conducted through the piazza site for this course.