Homework III – CS 6474/CS 4803 Social Computing

Grade: Max 150 points; 15% of overall grade (late policy applies)

Due: **April 20, 2026, 11:59pm Eastern Time**

Where to Submit: Canvas

What to hand in: Please complete the required cells in the **HW_III.ipynb file** and ensure that all relevant cell outputs are visible.

Content Warning: This homework involves engagement with material related to opioid use disorder (OUD). Some samples in the dataset may reference drug use, relapse, or treatment experiences, which could be distressing. Please prioritize your well-being while working with the content.

Goal: The goal of this assignment is to gain hands-on experience in **detecting online misinformation** (a key area in Social Computing research) related to a high-stakes public health issue (namely, opioid use disorder). You will examine content (a) written by humans (sourced from Reddit) and (b) generated by a large language model (LLM), in this case, GPT-4 to:

1. **Analyze prevalence differences**: report how frequently a misinformative claim related to OUD is **promoted** and **actively countered (or debunked)** in human-written versus LLM-generated responses.
2. **Analyze framing differences**: examine how human-written and LLM-generated responses differ in their framing, particularly in cases where the misinformative claim is either promoted or countered.

This homework is essentially asking you to do a small-scale replication of the following publication: https://ojs.aaai.org/index.php/ICWSM/article/view/35870 (published at ICWSM 2025). We encourage you to refer to the paper throughout the homework.

The questions in this assignment will test your understanding and application of computational methods for social computing (ranging from traditional transformer-based approaches to more recent LLM-powered techniques), as well as theoretical perspectives on language, such as linguistic variation and narrative framing.

Relevant files: You will primarily work with two items; (1) **HW_III.ipynb** (a Jupyter notebook containing detailed instructions for the tasks to be completed and the questions to be answered) and (2) **resources.zip** (a folder containing the dataset and other relevant files)

Part 1: Data Exploration

Q1: Manifestations of Online Misinformation

A common piece of misinformation about evidence-based treatments for opioid use disorder is that "medications for opioid use disorder (MOUD) simply replace one addiction with another." Refer to this article to read more about it: https://www.shadac.org/news/what-are-moud-medications-opioid-use-disorder-review

Consider the first 10 data points within the "dataset.csv" file (present within the resources folder) to answer the following questions:

Note: The dataset may use the following terms or phrases to refer to medications for opioid use disorder: MOUD, MAT, medications for opioid use disorder, medication-assisted treatment, methadone, suboxone, subs, etc.

Q1 (a): [10 points] Identify way(s) in which the misinformative claim that "medications for opioid use disorder simply replace one addiction with another" manifests within the first 10 human-written responses. Provide a concrete example from the dataset and a brief description of how this manifestation reinforces misinformation (e.g., through anecdotal generalization or false authority).

Q1 (b): [10 points] Identify way(s) in which the misinformative claim that "medications for opioid use disorder simply replace one addiction with another" manifests within the first 10 LLM-generated responses. Provide a concrete example from the dataset and a brief description of how this manifestation reinforces misinformation.

Part 2: Analyze Prevalence Differences

Now that you have gained a basic understanding of the dataset, your task is to use computational methods to analyze and compare how frequently the misinformative claim is (a) promoted and (b) countered across the two types of responses in the dataset (i.e., human-written and LLM-generated responses).

To reiterate, the misinformative claim we are investigating in this homework is "medications for opioid use disorder simply replace one addiction with another."

To complete this part of the assignment, you will develop LLM-based classifiers (using a few-shot learning approach) to detect responses that promote or counter this misinformative claim.

As discussed in class on March 2, few-shot learning is a technique in which we provide a small number of task demonstrations (typically 3-5 labeled examples) within the prompt, in natural language, to guide the LLM on how to perform the task at hand (e.g. classification). In this case, the demonstrations will consist of examples of responses that promote or counter the misinformative claim. Refer to the following article to learn more: https://www.promptingguide.ai/techniques/fewshot.

Steps to complete:

1. LLM setup
2. Develop a classifier to detect responses that promote the misinformative claim
3. Develop a classifier to detect responses that counter the misinformative claim
4. Use the classifiers to report how frequently the misinformative claim is promoted and countered in human-written versus LLM-generated responses.

On completing the above steps, answer the following question:

Q2: [20 points] **Report the following prevalence statistics.**

1. Proportion of human-written responses that promote the misinformative claim:
2. Proportion of human-written responses that counter the misinformative claim:
3. Proportion of LLM-generated responses that promote the misinformative claim:
4. Proportion of LLM-generated responses that counter the misinformative claim:

Based on your findings from the above coding exercise, answer the following questions:

Q3 (a): [10 points] Based on your results, why do you think we observed differences in the prevalence of the misinformative claim between human-written and LLM-generated responses?

Q3 (b): [12 points] Do you think a large language model-based automated approach is suitable for detecting online misinformation? Why or why not?

Q3 (c): [13 points] What potential implications do these findings on prevalence statistics have for online communities and online support-seeking?

Part 3: Analyze Framing Differences

Finally, in this last part of the assignment, you will use computational methods to surface how human-written responses differ from LLM-generated ones in terms of linguistic framing and stylistic patterns. You will further examine how these two sources diverge linguistically when they either promote or counter the misinformative claim.

To investigate these differences, you will use a lexicon-based approach. Specifically, you will work with a psycholinguistic lexicon called LIWC (Linguistic Inquiry and Word Count), which is widely used in social computing research to study stylistic, narrative, and framing properties of text.

Begin by exploring the GitHub repository "liwc-python" (https://github.com/chbrown/liwc-python), which provides a Python implementation of LIWC.

Refer to the README.md file to understand how to:

1. 1. Install the LIWC Python package
2. 2. Load and parse a LIWC lexicon file (.dic). The "resources.zip" folder contains the ".dic" file.
3. 3. Tokenize text and map words to LIWC categories
4. 4. Compute LIWC category counts for a given piece of text

Q4: [25 points] Examine how linguistic framing differs between human-written and LLM-generated responses using LIWC category counts.

1. **Compute normalized LIWC counts**: For each response, compute the normalized LIWC category counts, defined as the proportion of whitespace-separated words in the response that belong to each LIWC category. Perform this computation separately for human-written responses and LLM-generated responses.
2. **Visualize the distributions**: For the following five LIWC categories -- [posemo, negemo, insight, i, you] -- create density distribution plots showing the distribution of normalized category counts across responses. Plot distributions for human-written responses and LLM-generated responses within the same figure to facilitate comparison.
3. **Report summary statistics**: For the five LIWC categories, report the mean normalized value for: (1) human-written responses and (2) LLM-generated responses

Q5: [20 points] Only consider human-written responses. In Part 2, we identified human-written responses that promote or counter the misinformative claim.

In this step, you will examine how linguistic framing differs between human-written responses that promote versus counter the misinformative claim.

1. **Compute normalized LIWC counts**: Compute the normalized LIWC category counts, defined as the proportion of whitespace-separated words in the response that belong to each LIWC category. Perform this computation separately for human-written responses that promote the claim and human-written responses that counter the claim.
2. **Visualize the distributions**: For the following five LIWC categories -- [posemo, negemo, insight, i, you] -- create density distribution plots showing the distribution of normalized category counts across responses. Plot distributions for human-written responses that promote the claim and those that counter the claim within the same figure to facilitate comparison.
3. **Report summary statistics**: For the five LIWC categories, report the mean normalized value for: (1) human-written responses that promote the claim and (2) human-written responses that counter the claim

Q6: [20 points] Only consider LLM-generated responses. In Part 2, we identified LLM-generated responses that promote or counter the misinformative claim.

In this step, you will examine how linguistic framing differs between LLM-generated responses that promote versus counter the misinformative claim using LIWC.

1. **Compute normalized LIWC counts**: Compute the normalized LIWC category counts, defined as the proportion of whitespace-separated words in the response that belong to each LIWC category. Perform this computation separately for LLM-generated responses that promote the claim and LLM-generated responses that counter the claim.
2. **Visualize the distributions**: For the following five LIWC categories -- [posemo, negemo, insight, i, you] -- create density distribution plots showing the distribution of normalized category counts across responses. Plot distributions for LLM-generated responses that promote the claim and those that counter the claim within the same figure to facilitate comparison.
3. **Report summary statistics**: For the five LIWC categories, report the mean normalized value for: (1) LLM-generated responses that promote the claim (2) LLM-generated responses that counter the claim

Q7: [10 points] Answer the following questions based on the above computational analysis:

Q7 (a): How do human-written responses differ linguistically from LLM-generated responses? What factors might explain these differences?

Q7 (b): How do human-written responses (sourced from Reddit) that promote the misinformative claim differ linguistically from those that counter it? What are some potential implications of these differences?

Q7 (c): How do LLM-generated responses that promote the misinformative claim differ linguistically from those that counter it? What are some potential implications of these differences?

Bonus Question (Optional): Analyze Framing Differences Using an Alternative Tool

In this bonus task ([30 points]), instead of LIWC, use another computational tool or method (e.g., the social dimension classifiers used in the original publication: https://ojs.aaai.org/index.php/ICWSM/article/view/35870) to analyze framing differences.

1. **Justify your choice:** Explain why you selected this tool or method and what specific aspect of linguistic framing it captures.
2. **Compute framing differences:** Using your chosen approach, compare human-written vs. LLM-generated responses to identify differences in linguistic framing.
3. **Visualize the results:** Create density distribution plots to illustrate how the two groups differ across the selected linguistic framing dimension(s).
4. **Report summary statistics:** For each dimension or feature, report the mean score for human-written and LLM-generated responses.
5. Provide an interpretation of the findings.