SI 630 Homework 3:
Data Annotation and Large Language Models

See Canvas for deadlines!!!

Last updated: March 18 (version 1.1)

1 Introduction

The majority of NLP models require labeled data to learn some task, e.g., classification. That labeled data is produced by humans who sometimes disagree. While some tasks are more objective, such as labeling the part of speech of a word, other tasks are more subjective and the annotator's lived experience and background may influence how they label data. Homework 3 will introduce you to one such subjective task in three ways: (1) labeling data, (2) training models, and (3) analyzing model disagreements.

The task in question is trying to infer the values that an author signals in their writing. A recent SemEval shared task Kiesel et al. [2023] developed one such dataset for this task when examining how authors make arguments for or against some premise. Other related work has tried to examine the moral values of text using Moral Foundations Theory [MFT; Graham et al., 2013]. Models that recognize values are potentially useful in helping analyze arguments for/against topics, analyze how values influence the trajectory of conversations, and examine differences in value priorities across groups (e.g., differences in regions).

Homework 3 is a *three-part* homework focused on teaching you (1) see what is involved in data annotation, (2) build classifiers based on large language models, and (3) understand how model performance relates to annotator variance—and whose judgments the model ends up aligning with. You'll go through the full Machine Learning pipeline from initial unlabeled data to training classifiers. This data curation process is often overlooked as a simple step, but as a practitioner in the field, if you have to annotate your own data (or hire people to label for you), you'll need these skills to ensure your eventual classifier performs well.

This homework has the following learning goals:

1. Gain familiarity with the process of annotation
2. Learn how to compute rates of inter-annotation agreement
3. Learn how to use the Huggingface `Trainer` infrastructure for training classifiers
4. Improve your skills at reading documentation and examples for advanced NLP methods
5. Learn how to use the Great Lakes cluster and access its GPUs
6. Understand the challenges and risks of model alignment by assessing model disagreements

This homework is multi-part because we will use the annotations produced by all students to train and analyze your classifiers. Roughly, the first week will be spent annotating. Once the annotation has finished, we will release the larger set of annotations and teams will develop their classifiers. To get you started, we have released the dataset from the SemEval task which will be in a similar format to what you will receive when we finish annotating. The expectation is that the classifier training part is fairly easy (we're not asking for a fancy model); this part of the assignment is just to get you familiar with how to go through the motions and run deep learning experiments on Great Lakes.

Part 3 will touch on one important aspect of annotation: representation in the labeling. Nearly all supervised NLP models learn to predict a single label. This requires that all the disagreements in annotation are somehow converted to a single label. However, depending on the distribution of *who* is annotated, some voices may ultimately be left out of the final single label. To get a sense of whether this might happen in our data, as a part of your labeling, at the end, you'll take an optional survey to ask you questions about your demographics and value system. **These questions are entirely optional and you can choose to opt-out of any of them by choosing a decline-to-state option.** Once the annotation process in Part 1 is finished, we'll aggregate data and release both aggregated and disaggregated test data for you to evaluate in Part 3; all of this data is anonymous and nothing about the data can be used to link back to your person. The big goal for sharing this data is to get a sense of whether the model's predictions are more closely aligned with specific groups of annotators (if any).

2 Using Great Lakes

This homework requires that you use a GPU for training the models (you can use a CPU for testing them). You can use your own or you can use the GPUs on Great Lakes for free. We have provided you with a course account for Great Lakes, `si630_w24`, which will give you access for many hours. Your Great Lakes jobs are limited to 4 hours of training, which is more than enough to train your deep learning model for this assignment.

Here is a guide on how to access Great Lakes, set up your environment, and submit your jobs: Great Lakes Tutorial. We strongly encourage learning to use Great Lakes for this assignment as (1) you have free access to substantial computational resources and (2) you would want to use it for the next homework and the final project.

3 Class-wide Technical Report

As you will see in this assignment, annotation is challenging, and annotated data—especially, *quality* annotated data—is incredibly valuable. Based on the number of students in SI 630, we will have produced a new and interesting resource for NLP models: A rich set of thousands of responses labeled for their perceived values *and* potentially demographic information on the persons who annotated.

As a class, I would like us to publish our collective resource on argument persuasiveness as a technical report on ArXiv with you all as authors (opt-in) where we share our annotations to contribute to the research community. You would not need to contribute anything beyond your anonymized annotations to be an author. Writing and analysis contributions would move you up in the author order. If there's interest we could even try to publish the technical report as well, which has been done for past course projects, e.g., Kenyon-Dean et al. [2018].[1] You all are doing amazing and it's important to realize that you have the ability to advance the state of NLP and data science in general with the skills. The technical report would be a way to document your contribution.

4 Part 1: Annotating Data

Part 1 will have you annotating data to infer the values an author might have based on a piece of their writing. Specifically, you'll be examining answers to general questions made on Reddit and assess whether the author of the answer is likely to agree with specific kinds of value-based judgments. This is a highly subjective task, yet people's answers are not random. Based on your own lived experience and interactions, you will likely make inferences about what another person values based on what they write.

For the value system, we'll be using Moral Foundations Theory [MFT; Graham et al., 2013]. MFT is a common way of assessing a person's general values with respect to specific dimensions. We have simplified the overall questions you'll ask but for this task, you'll be rating an author's agreement on *nine* questions. We'll use these questions later to get a sense of the author's estimated values. The annotation task has more instructions but the general theme is to use your first instinct when rating and to not spend lots of time reading and trying to hypothesize about what another might have meant. There are no right or wrong answers here so use your best judgments.

Given the technical issues, you'll be asked to annotate 100 total questions. We *strongly* recommend spreading the annotation out over time, e.g., do 20 per day for five days or do 20 in different sittings throughout the day. This will make the task easier and more manageable. Doing only 1 or 2 at a time will be slower though since you won't stay as familiar with the rating questions so we encourage small blocks—you'll be amazed at how many you can get through in a 20-minute burst!

At the end of annotation, you'll be asked to fill in a series of questions about your own values and demographics. These questions should be filled in for completeness, but you can answer "decline to answer" for any of them you don't feel comfortable. The hope is that enough folks will opt-in when answering these questions so that we can look for what patterns emerge in how annotators' different values/backgrounds might influence the ratings. All data shared is fully anonymized. For the value-based questions, we'll construct an aggregate score based on your answers and share only that (no answers to individual questions).

■ **Problem 1.** (25 points) Annotate 100 instances using the annotation system at http://curry.si.umich.edu:8080/. When creating a new user login, do not use your uniqname password (you can just use any old regular one that is at least 6 characters).[2]

■ **Problem 2.** (5 points) In a paragraph, describe your thoughts on the annotation process. As a few prompts, consider what did you learn as a result of annotating? Was the task easy or hard as you went through it? If you had to do a "next iteration" on the task in your future job, how

[1] `https://aclanthology.org/N18-1171/`

[2] The passwords on this system are salted and hashed (i.e., not plaintext) but it's best to be safe and keep your umich password secret.

would you update the task? You are welcome to write anything in your reflection about how it has changed your perception of annotation.

5 Part 2: Recognizing Persuasive Arguments

Part 2 will have you developing classifiers using the very powerful Huggingface library.[3] Modern NLP has been driven by advancements in Pretrained Large Language Models (LLMs) which, like we discussed in Week 3, learn to predict the word sequences. Substantial amounts of research have shown that once the models learn to "recognize language," the parameters in these models (the weights in the neural network) can quickly be adapted to accomplish many NLP tasks. We'll revisit the task from Homework 1: classification.

Most of the work on LLMs has focused on BERT or RoBERTa, which are large masked language models. These models have been shown to generalize well to a variety of new tasks. We'll use RoBERTa for our homework, which should be feasible to train for our task.

This part of the assignment will have you using Great Lakes cluster, which has provided free GPUs for us to use.[4] You will need to use Great Lakes for this assignment to be able to train the deep learning models. We have made it an explicit learning goal as well, so that you are comfortable with the system for Homework 4 and using the cluster for your course projects. We have provided a guide to using Great Lakes for you to get started.

■ **Problem 3.** If you don't have an account on Great Lakes, request one now at `https://arc.umich.edu/login-request`

For training your LLM-based classifier, we'll use the `Trainer` class[5] from `huggingface` which wraps much of the core training loop into a single function call and support many (many) extensions and optimizations like using adaptive learning rates or stopping your model if it converges to avoid overtraining. The class's documentation has many options, which can be a it overwhelming at first. We have released a starter notebook for Part 2 to get you started working with the SemEval data.

■ **Problem 4.** (15 points; code) Develop your code in the Part 2 notebook and ultiamte use `huggingface`'s `Trainer` class to train a multilabel classifier to predict the values in an argument.

■ **Problem 5.** (5 points; PDF) When you train the final multilabel model, use Weights & Biases to record performance on the development set during training. Show the plots in your PDF and in a few sentences, describe which version of the model (i.e., the parameters at a specific time point) you would choose and why.

■ **Problem 6.** (10 points; code + kaggle) Use your trained `huggingface` model to predict the values score on the test dataset we released from Part 1 and submit it to Kaggle. **NOTE: This**

[3] `https://huggingface.co/`

[4] `https://arc.umich.edu/greatlakes/user-guide/`

[5] `https://huggingface.co/docs/transformers/main_classes/trainer`

dataset is not yet out. Your grade will depend, in part, on how well your model performs. **You only get one submission to Kaggle, so make it count!**

6 Annotation Evaluation

NOTE: This part is not possible until we have Part 1 finished

Once you have the code for training the model ready, we can also put it to use to examine how annotator backgrounds might influence the model performance. Specifically, you'll analyze how different (anonymized) annotators' backgrounds might influence performance and how models align more or less with different groups of annotators.

Using the held-out development data we created in Part 1, let's examine the data in more detail to see how annotators agreed (or disagreed) on the labels.

■ **Problem 7.** (10 points; PDF) Calculate the inter-annotator agreement (IAA) on the disaggregated training and development data using Krippendorf's α for both the nominal and interval setting. Report both numbers. In a few sentences describe what you see and what you think the difference (if any) means for how much annotators agreed.

Our value-inference annotation task is subjective and an annotator's backgrounds or values *might* influence their choices. We'll look at identities through two kinds of assessments. The first will look at IAA within specific subgroups. Identity is complex and while it can be useful to analyze subgroups, please be aware that our analysis is only aimed at capturing high-level trends and likely omits details that we could observe with a larger sample size or with a richer description of identity (e.g., examining intersectional identities).

For looking at subgroups that have categorical answers, we'll examine differences within groups of annotators who gave the same answer to *one* of the demographic questions, e.g., folks who selected a particular gender identity. Due to sparsity, we'll only work with groups that have at least 3 people in them; this avoids drawing non-representative conclusions from a small sample size.

For scalar questions like politics and the MFT scores, we'll group the data as follows. For each MFT dimension, divide the users into thirds (terciles) according to their scores for that dimension (i.e., sort by score, then take the top third, middle third, bottom third). You can then compare the ratings for the ”high”, ”middle”, and ”low” groups. For politics, group annotators into three groups: those selecting any of the “conservative” labels, those with the “moderate” rating, and those selecting any selecting a “liberal” label.

■ **Problem 8.** (15 points; PDF) For each of the groups, calculate Krippendorf's α using an interval scale. Use a barplot and label each bar according to the group. In a few sentences describe what you see and what you think the difference (if any) means for how much different subgroups of annotators agreed. What do you think this result means for the data quality?

In our second analysis of annotator subgroups, let's look at whether the classifier you trained is biased towards predicting the mean ratings for one group or another.

■ **Problem 9.** (15 points; PDF) First, use your model to get the scores on the development data. Second, using the disaggregated development data we provide and the different demographic subgroups from the earlier problem, for each subgroup, create a separate development dataset that is the average of the scores for just the individuals in that group. Third, calculate the difference between the model's predictions and each subgroup's predictions. Fourth, plot the mean score reference and 95% confidence intervals for each group. Fifth, in a few sentences, describe what you see and what you think the difference (if any) means for model performance. Do you think the classifier performance is representative of all groups? If so, say why; if not, suggest what you think would be needed to fix it.

7 What to submit

See Canvas for the official deadlines. This is a multipart assignment and you'll submit materials for this assignment to different places:

- Complete all your annotations on the Potato tool
- Submit your notebooks for Part 2 and Part 3
- Submit a PDF of your plots and answers for Parts 1, 2, and 3
- Submit the test set predictions for the Part 1 test set on Kaggle (NB: you only get one upload!)

8 Academic Honesty

Unless otherwise specified in an assignment all submitted work must be your own, original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the University's policies on Academic and Professional Integrity may result in serious penalties, which might range from failing an assignment, to failing a course, to being expelled from the program. Violations of academic and professional integrity will be reported to Student Affairs. Consequences impacting assignment or course grades are determined by the faculty instructor; additional sanctions may be imposed.

Copying code from any existing Trainer or huggingface code implementation is considered grounds for violation of Academic Integrity and will receive a zero. Code generated through automated or AI-assisted tools, such as Co-Pilot will also receive a zero.

References

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pages 55–130. Elsevier, 2013.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1171. URL `https://aclanthology.org/N18-1171`.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.313. URL `https://aclanthology.org/2023.semeval-1.313`.