**Classification Assignment**

The aim of this assignment is to take you (nearly) through the full process of text classification from defining a concept to evaluating performance.

**Background**

A key hypothesis in education research is that “time on task” increases student learning. In other words, the more classroom time spent on math, the more students are likely to learn. Of course, teachers and students do not spend 100% of class time learning. Instead, teachers commonly have to spend substantial amounts of time providing students with administrative information and instructions and redirecting student behavior. Your goal is to train a classifier to identify teacher speech which contains predominantly math content.

***The data.*** You have been provided with an excel file which contains every teacher utterance drawn from 112 transcripts of elementary mathematics classrooms. The text column contains the teacher speech. The gold\_standard column contains the (potentially fallible) gold standard label with one of two categories: math instruction or classroom management.

1. Read through the questions below. Then, tell me if you plan to reference ChatGPT while completing the assignment. If you do plan to reference ChatGPT, describe a) how you plan to use it and b) how you will ensure that the final product is your own and that you learn from the activity.
   1. **Yes, we do plan to reference ChatGPT, more specifically to check our understanding of of “math instruction” and “classroom management” and potentially code that is not running properly.**
   2. **In order to ensure that the final product is our own we will not directly use the code given to us by ChatGPT but instead use it as a way to confirm our intuition and as an error handing system.**

**Part 1. Define the Concept and Create a Codebook**

1. Read the two transcripts provided to you. Then, generate reasonable definitions of “math instruction” and “classroom management”. Provide those definition here.
   1. **Math instruction involves direct teaching and activities related to mathematical concepts, principles, and problem-solving. In the transcripts this includes the teacher explaining how to compare numbers using greater than, less than, or equal to signs, demonstrating with base-10 blocks, and students engaging in exercises to understand these concepts.**
   2. **Classroom management involves actions and dialogues aimed at organizing the class, maintaining discipline, and ensuring that students are prepared and attentive for the lesson. In the transcripts this includes the teacher instructing students to put away their belongings, staying silent, organizing groups for activities, and addressing behavioral issues.**
2. Provide an example of an utterance that clearly falls under your definition of “math instruction”. Explain.
   1. **Teacher 46:55**
   2. **and remember Okay, so we need fewer 10s than 85 which numbers in the 10s place?**
      1. **This is clearly math instruction as the teacher is first talking through the math problem followed by her directly asking the students to provide the answer to a math question using base-10 blocks.**
   3. **Teacher 06:21**
   4. **That's 40. How do you know that? These are 10s Finn, how should I fix it? Come on up and fix it. There we go. Now it says 22. Right. So we've got 10, 20, 21 and 22. Nice job. [Inaudible] The number 28. How would I draw that? Ah, okay. Okay, come show me. 2 10s? Okay, good. 2 10s and eight ones. Now we need to compare them. Which number is less than the other number? Nevaeh? Thank you. Tell me 22 or 28?**
3. Provide an example of an utterance that clearly falls under your definition of “classroom management”. Explain.
   1. **Teacher 00:30**
   2. **All right. You know what, go ahead and put your things away now. Put your books away we'll go ahead and get started since everybody's mostly done everybody unpacked shhhh. put your books away. What's this? What's this?**
      1. **This is clearly classroom management because the teacher is organizing the classroom by telling the students to put their things away, gives instructions to organize the class, and is attempting to maintain discipline.**
   3. **Teacher 01:44**
   4. **Hurry up all right, friends. Remember we are having group instruction. Okay, your voice level is zero. No talking. Oops. Okay, raising your hand to speak or ask a question. We are having group instruction. At this time, there's no more getting up out of your seat to go to the restroom. unless it's an emergency. No more drinks of water. Okay. Your eyes are on me hands on top of your desk. We're working on our assignments. And we're putting things where they belong. All right. So we have been working on comparing numbers. Jacob, I need you to pick up your hat love. We have been working on comparing numbers, right? How would we compare two two digit numbers? What do we do? Danielle?**
4. Provide an example utterance that would be hard for a human to classify according to your definitions. Explain.
   1. **Teacher 35:28**
   2. **All right 10 9 8 7 6 5 4 3 2 1 Yes ma'am. Noah you need help? Okay, take all your pages and then put your book away excellent Dane showing me that he's gonna be ready. He's already putting his name on his paper. Get it together Victoria quickly, please. Let's go. Paper should be on top of your desk All right, so friends, now, putting those models in into practice. Okay, putting it into practice. Go ahead and turn your paper over that first paper, turn it over and look at the top. Stop. Here's what we just did. Right? Here's the explanation of it. It says take. Right here, what did they do? They took away 10 less than 33 is 23 Here's my 33 10 less. They circled it and crossed it out. You guys actually took your 10s And would you do? Would you do with that? 10 when it was 10 less? What happened Victoria when he was 10 less? It's they're circling that and crossing it out. They're actually going to take it away. Thank you take it away. So 23 years, 10 less than 33. Here's 33 in the middle. 10 Less is 23. Think 10 More is 4343 is 10 more than 33. Okay, so look at number one. And I'm gonna come around. I'm gonna come around and look at your work. Okay. The number in the middle is what we start with, right? That was our model 70 And we did this one. And we did 41. So we did 70 No, we didn't do we did 37. So 70 and 37. So think 70. What's 10? Less than 70? What's 10 less than 70 60 10? Less than 70 60. What's 10 More than 70? Kayden**
      1. **This is a difficult utterance for a human to classify because the entire beginning of this utterance is clearly classroom management. The teacher is giving instructions on classroom organization and preparing them for an activity. Then, the second part of the text is math instruction.**
5. Based on your definition and the example transcripts, what challenges do you anticipant in training a classifier to identify math instruction?
   1. **A challenging utterance to classify might be when a teacher integrates both classroom management within a math instruction context, such as when organizing students for a math activity or when a teacher has to discipline a student while they are in the middle of providing a long math instruction. These blends directing student behavior (classroom management) with setting up an educational task (math instruction).**
6. What are three additional steps you recommend someone take at this point, before training a classifier, in order to ensure the validity of the concept and of the hand-labelled data?

**Part 2. Split into training, development, and testing.**

Assume that the data provided in the gold\_standard column of teacher\_utterance\_classifications.xlsx contain the labels according to one or more (fallible) human coders who were provided with your codebook.

1. Run the notebook to split the data into training, development, and testing sets (60:20:20). Note that we are performing the split at the transcript level. Why is that important?
   1. **This is important so that we can tune our hyperparameters such as our algorithm of chouse, number of n-grams, lemmas, etc., while also being able to test our model’s performance on a completely held out test set.**

**Part 3. Cleaning**

1. Given your definition and what you observed in the data, what are four pre-processing/feature extraction steps that you would like to test (beyond the tokenization and processing handled by the CountVectorizer defaults)? Explain why each of the five steps may be useful for measuring this specific concept.
   1. **Lemmatization:** 
      1. **Lemmatization is where we reduce words to their base/root form. For example, the words “teaching,” “taught,” and “teaches would all be converted to “teach.” This step is useful because it will help the model generalize better and help reduce the dimensionality of the classifier. With less total words, but same number of total meanings, the model will run faster and more accurately. More specifically to our task in an educational setting, instructional concepts might be expressed in various ways, therefore, lemmatizing will help us better group these forms of instructions.**
   2. **Stop Words Removal:** 
      1. **Removing Stop Words is an important part of most classifiers. In our case though, just removing the “normal” stop words may not be enough. So, we will use the CountVectorizer to remove common English stop words, and add a customized list of words to include classroom-specific but low-information words could be advantageous. For example, frequent but non-informative words like "okay," "right," or "you know" could be removed. This step can help focus the model on more content-specific words that are indicative of math instruction or classroom management.**
   3. **N-Grams:** 
      1. **Beyond single words (unigrams), using bi-grams or tri-grams (sequences of two or three words) can capture more context. For example, phrases like "greater than" or "equal to" might be specifically indicative of math instruction, while phrases like "turn to page" could be more aligned with classroom management. This step can help in understanding the context better than individual words can, which is crucial in distinguishing between content delivery and management instructions. This would have to be done before stop word removal though, otherwise words like “to” might get removed.**
   4. **TF-IDF (Term Frequency-Inverse Document Frequency):** 
      1. **Instead of just counting the frequency of words (as in CountVectorizer), using TF-IDF can help to weigh the words' importance. Terms that occur frequently in a specific document (utterance) but not across documents (other utterances) are considered more important. This is particularly useful in this context because certain terms might be highly relevant to math instruction as mentioned above like specific mathematical terminology or classroom management such as "quiet down," or "listen up" and can help distinguish between the two categories more effectively.**
2. Provide code implementing each of the four pre-processing steps you describe above.

**Part 4. Model Selection**

1. Using your training and development data, I want you to test at least 3 algorithms with at least two hyperparameter options, and two pre-processing options each. In total, you will train 12 models (3\*2\*2). Two of the algorithms need to be algorithms we have covered in class. You will use 5-fold cross-validation to select the hyperparameters for a given algorithm X pre-processing combination (for example, you could use 5-fold cross-validation to select the best the best C value for a lasso regression with lemmatization and then the best C without lemmatization). Describe the combinations that you will test below.
2. Provide the code training the models below. This should include 5-fold cross-validation to select the hyper-parameters.
3. Provide the code assessing the performance of each of the models in the development dataset.
4. Which model had the highest performance? By what criteria? Explain.

**Part 5. Evaluation**

1. Using your testing data, assess the performance of your final classifier. Provide the code below.
2. Describe the performance of your final classifier. In your opinion, is the performance sufficient to use to provide teachers with feedback regarding time spent on task? Explain.
3. What is the approximate number of out-of-class hours you spent completing this assignment?