WSC Sports - NLP Team - Home Assignment
Omer Nagar, 302824875

Task

Given:

- (Transcription, binary-label) pairs dataset
- List of params

Build a System that takes as an input a transcript and returns an action (in case an action exists and is valid).

EDA:

Goals:

The purpose of this section is to conduct Exploratory Data Analysis (EDA) to assess:

- **Data Quality**: Look for missing values, duplicate records, and data entry errors. Assess class distribution for imbalance.
- **Data Sufficiency**: Ensure the dataset has enough variety and volume to train a model effectively.
- **Model Complexity Considerations**: Analyze the complexity and patterns in the data to estimate the model size and architecture that might be required.

I performed the following analysis:

1. Number of valid/invalid instances.
2. Label Distribution.
3. Distribution of the transcription length.
4. Word frequency and statistics.
5. Parameters frequency.
6. Parameter and label mutual distribution.

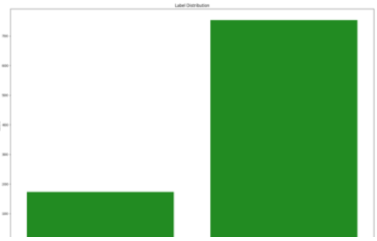
Results:

1. **Number of valid/invalid instances.**

   The data contains several instance with multiple or no parameters in the transcription. Overall, there are:

   Valid instances: 927
   Invalid instances 191

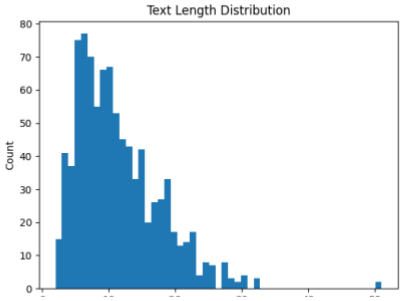
2. **Label Distribution**



The data is unbalanced and there are significantly more positive instances.

Overall:
Positive: 754
Negative: 173

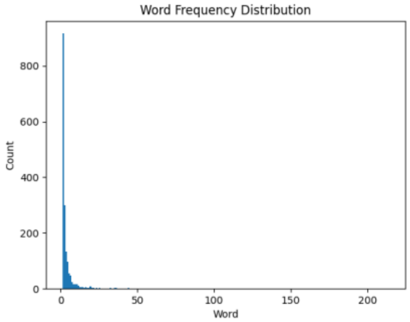
3. **Distribution of the transcription length**.



Most of the transcriptions are short.

Length mean: 17.64
Length Std: 10.514

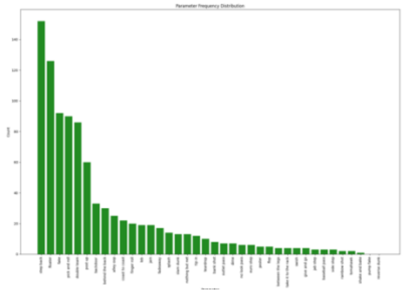
4. **Word frequency and statistics**.



The dataset exhibits a limited variety of words, with many words lacking sufficient representation

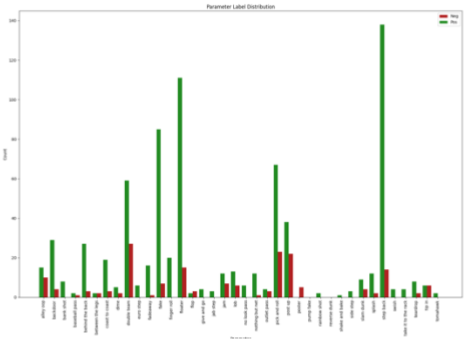
Unique words: 1713
Words frequency mean: 3.7
Words frequency median: 1
Words frequency Std: 10.2

5. **Parameters frequency**.



The Parameters distribution has a tail of values that are underrepresented. There are even two parameters (“reverse dunk”, “pump fake”) that does not appear at all.

6. **Parameter and label mutual distribution**.



several parameters have even higher label imbalance. Moreover, there are missing labels for some of the parameters.

conclusions:

- **Limited data:** data contains under 1K examples and many of the words are underrepresented. I would consider using external basketball related text data to pretrain a model.
- **Label distribution:** there is a need to address the unbalanced labels via sampling/weighted loss.
- **Parameter/label distribution:** I would consider to over-sample underrepresented <parameter, label> pairs.
- **Missing parameters:** some parameters have extremely low appearance, and some does not appear at all. One way to handle it may be to use gen AI to create synthetic training examples.
- **Model size:** the data contains 1K training instances, where the mean length is ~17. The task is to predict a binary label which is not to complex. Therefore, when choosing a model, it should be relatively small one and regularization should be applied during training phase to prevent overfitting.

Methods

Train-Test Split

Due to the significant imbalance in the dataset regarding action-phrases, it's crucial to ensure that every phrase is represented in the test set. To achieve this, I grouped the data based on the <action-phrase, label> pairings and then split each subgroup appropriately. For those instances where a group contains only a single example, I allocated it directly to the test set.

Baseline Model Architecture

According to the above EDA, I experimented with the following architecture:

1. **BERT backbone (freeze-weights)**
   Pretrained BERT model (bert-base-uncased):
   12-layer, 768-hidden, 12-heads, 110M parameters.
2. **MLP**
   2 layers of fully connected with Relu activation, followed by sigmoid
   a. Liner 768-> 256
   b. Relu
   c. Liner 256-> 1
   d. Sigmoid

The baseline does not attend the unbalanced labels and action-phrases.

Training

For the training I used BCE loss, Adam optimizer with learning rate of 2e-5 and weight decay of 1e-2. I used 32 batch size and did a short training of 20 epochs.

Improvements

I think there are several improvements that can be applied in order to cope with the small dataset and the imbalance in action-phrases and labels:

1. **Label weights. (Implemented)**
   Give weights according to 0/1 distribution.
2. **Smart Sampling**
   Sample batches according to <action-phrase, label> distribution
3. **Fine-Tune BERT**
   Leverage additional, unlabeled basketball-related data, ideally from game commentaries, to refine the backbone model. This fine-tuning process aims to enhance the quality of the embeddings.
4. **Use [Mask] token. (Implemented)**
   Hide the action-phrase from the model and replace it with a [MASK] token. This way, the model will have to rely on the surrounding text. This may help in two ways: it can mitigate the action-phrase imbalance if several phrases can replace the [MASK]. Secondly, the model will rely more on the context, which may provide valuable information about the label (whether it took place on the court or not).

Results

	Train-AUC	**Train-Recall**	**Train-Precision**	**Val-AUC**	**Val-Recall**	**Val-Precision**
Baseline	0.7683	0.8844	0.8754	0.7034	0.7048	0.8540
Weighted loss	0.8097	0.7857	0.9277	0.6844	0.6205	0.8512
[Mask]	0.7229	0.9065	0.8752	0.6628	0.7108	0.8676
Weighted loss & [Mask]	0.8348	0.7993	0.9325	0.7052	0.7993	0.8923

Git

htps://github.com/nagar-omer/wsc-interview