

MLPC – Team Toothpaste



- Aral Cincim, Sandro Müller, Linus Madlener, Kevin Eberl

Data Split

Best split to avoid data leakage?

Data Splitting

- 80% for training to generalize
- 10% for validation, reducing overfitting
- 10% for testing, test unbiased performance

Data Leakage

- When similar frames are in training & test set
- Potential solution: split by files first
- Another issue: Annotation too similar for certain classes and text embeddings

Evaluation

Which evaluation criterion?

Balanced Accuracy Score

- Imbalanced data: classes not equally distributed -> no favoring majority class
- *Balanced Accuracy* = $\frac{1}{2} \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right)$

Worst vs. best performance

- Baseline classifier predicting majority class -> 75% (drums)
- TODO: Score of best classifier