

# Inter-Annotator Agreement

## Adjudication & Modeling Introduction

Jin Zhao

Brandeis University

March 30, 2025

# Today's Agenda

- 1 Resolving annotator disagreements
- 2 Adjudication strategies
- 3 Creating gold standard datasets
- 4 LLM-assisted adjudication
- 5 Introduction to modeling with annotated data
- 6 From annotations to trained models

*Note: No class April 1 (Passover), No class April 6-8 (Passover Break)*

**Assignment:** HW 3 due

# Why Resolve Disagreements?

## **Disagreements are inevitable**

### **But ML models need single labels:**

- Supervised learning requires ground truth
- Multiple labels per instance is problematic
- Evaluation needs clear correct answers

### **Creating gold standard:**

- Authoritative, adjudicated labels
- Used for training and evaluation
- Represents “correct” annotation

## Common approaches:

### ① Majority voting:

- Most common label wins
- Simple, scalable
- Requires odd number of annotators

### ② Expert adjudication:

- Expert reviews disagreements
- Higher quality, more expensive

### ③ Discussion:

- Annotators discuss until consensus
- Time-intensive but educational

### ④ Probabilistic:

- Weight by annotator reliability
- More sophisticated

# Majority Voting

## Simple and effective

### Process:

- 1 Collect labels from all annotators
- 2 For each item, select most frequent label
- 3 Handle ties (random, expert review, or skip)

### Advantages:

- Easy to implement
- Scalable to large datasets
- No additional annotation needed

### Disadvantages:

- Ignores annotator quality differences
- Majority can be wrong
- Doesn't improve guidelines

# Expert Adjudication

## For higher quality gold standard

### Process:

- 1 Identify items with disagreement
- 2 Expert reviews each case
- 3 Expert makes final decision
- 4 Optionally: update guidelines based on patterns

### When to use:

- High-stakes tasks
- Creating evaluation benchmarks
- Complex annotation requiring domain expertise

**Cost:** More expensive and slower

## New approach for 2025

### Options:

#### ① LLM as tie-breaker:

- When annotators disagree, ask LLM
- Use as third “vote”

#### ② LLM reasoning:

- Ask LLM to explain which label is correct
- Human reviews LLM reasoning

#### ③ LLM confidence:

- Accept human majority if LLM confidence low
- Review if LLM disagrees with high confidence

**Caution:** LLM shouldn't be sole adjudicator for evaluation data

# Creating Gold Standard Dataset

## Complete workflow:

- 1 Multiple annotators label data
- 2 Calculate IAA to verify quality
- 3 Identify disagreements
- 4 Apply adjudication strategy
- 5 Verify adjudicated labels
- 6 Document process

## Best practices:

- Keep original annotations (for analysis)
- Document adjudication decisions
- Track problematic patterns
- Update guidelines for future annotation

# From Annotations to Models

## The ML pipeline:



**Annotation quality directly affects model quality**

## Classification tasks:

- Input: text features (BoW, embeddings)
- Output: predicted label
- Algorithms: Logistic Regression, SVM, Neural Networks

## Sequence labeling tasks:

- Input: token sequences
- Output: BIO label sequences
- Algorithms: CRF, BiLSTM-CRF, BERT-NER

**Key insight:** Model learns patterns from annotations

## Classification:

- Accuracy, Precision, Recall, F1
- Confusion matrix

## Sequence labeling:

- Token-level accuracy
- Entity-level Precision/Recall/F1
- Exact match vs. partial match

## Important:

- Evaluate on held-out test set
- Report multiple metrics
- Compare to baseline

# Annotation Quality and Model Performance

## The relationship:

- Higher IAA  $\rightarrow$  cleaner training signal
- Cleaner signal  $\rightarrow$  better model
- Better model  $\rightarrow$  higher accuracy

## Upper bound:

- Human agreement is ceiling for model
- If humans agree 80%, model likely  $\leq 80\%$
- (Though models can sometimes beat individuals)

**Rule of thumb:** Expect model  $F1 \approx$  Human IAA

# Common Modeling Pitfalls

## Data issues:

- Training on unadjudicated data
- Test set contamination
- Class imbalance

## Evaluation issues:

- Overfitting to dev set
- Cherry-picking metrics
- Not comparing to baseline

## Process issues:

- Not documenting preprocessing
- Inconsistent tokenization
- Not releasing data/code

## After Passover break (Apr 15):

### Deliverables:

- ① Adjudicated gold standard dataset
- ② Documentation of adjudication process
- ③ Final IAA report
- ④ Updated guidelines (if changed)

### Next steps:

- Train baseline model
- Evaluate on gold standard
- Analyze errors
- Prepare final report

# Next Class: Modeling and Evaluation I

## Lecture 21 (Apr 13): Annotation-Informed Modeling I

*Note: No class April 6-8 (Passover Break)*

### Topics:

- Training models on annotated data
- Handling annotation uncertainty
- Multi-annotator learning
- Soft labels vs. hard labels
- Human vs. LLM annotation comparison

**Project:** Gold standard dataset due

**Assignment:** HW 4 assigned

# Key Takeaways

- 1 **Adjudication** resolves disagreements to create gold standard
- 2 **Majority voting** is simple but may ignore annotator quality
- 3 **Expert adjudication** produces higher quality but costs more
- 4 **LLMs can assist** adjudication but shouldn't be sole judge
- 5 **Model quality** is bounded by annotation quality
- 6 **Document** your adjudication process for reproducibility

## Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

✉ [jinzhao@brandeis.edu](mailto:jinzhao@brandeis.edu)