Annotation-Informed Modeling I

Training Models on Annotated Data

Jin Zhao

Brandeis University

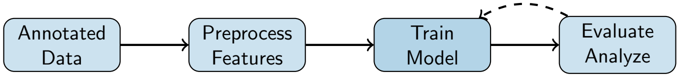
April 13, 2025

Today's Agenda

1. Welcome back from Passover break
2. Training models on annotated data
3. Handling annotation uncertainty
4. Multi-annotator learning
5. Soft labels vs. hard labels
6. Comparing human vs. LLM annotations

Project: Gold standard dataset due
Assignment: HW 4 assigned

The Modeling Pipeline



Today's focus: How annotation quality affects each stage

Standard Training Approach

Using adjudicated gold standard:

1. Take gold label for each instance
2. Split into train/dev/test
3. Train model to predict labels
4. Evaluate on test set

Assumes:

- Single correct label exists
- Gold standard is reliable
- All instances equally informative

But: What about annotation uncertainty?

Handling Annotation Uncertainty

Disagreement carries information

Options:

1. **Ignore uncertainty:** Use gold labels only
2. **Filter uncertain:** Remove low-agreement items
3. **Weight by agreement:** Confident examples matter more
4. **Soft labels:** Train on label distributions
5. **Multi-task:** Model uncertainty explicitly

Key insight: Uncertain examples may be legitimately ambiguous

Soft Labels

Instead of single label, use distribution

Hard label:

- "This is POSITIVE" (one-hot: [1, 0, 0])

Soft label:

- "60% POSITIVE, 30% NEUTRAL, 10% NEGATIVE"
- Label distribution: [0.6, 0.3, 0.1]

From annotator votes:

- 3 annotators: 2 say Positive, 1 says Neutral
- Soft label: [0.67, 0.33, 0]

Training with Soft Labels

Standard cross-entropy loss:

$$L = -\sum_c y_c \log(p_c)$$

Where y is one-hot (hard label)

With soft labels:

$$L = -\sum_c \hat{y}_c \log(p_c)$$

Where $\hat{y}$ is the label distribution

Effect:

- Model learns nuance in uncertain cases
- Less confident predictions for ambiguous items
- Can improve generalization

Multi-Annotator Learning

Don't aggregate – model all annotators

Approaches:

1. **Data augmentation:** Treat each annotation as separate training example
2. **Multi-task learning:** Predict each annotator's label
3. **Annotator modeling:** Learn annotator-specific biases
4. **Ensemble:** Train separate models, combine predictions

Benefits:

- Captures systematic disagreement
- Models perspective diversity
- Better for subjective tasks

Data Augmentation Approach

Simple multi-annotator learning:

Original:

- Text: “Not bad at all”
- Annotations: [Pos, Pos, Neu]

Augmented training data:

- (“Not bad at all”, Positive)
- (“Not bad at all”, Positive)
- (“Not bad at all”, Neutral)

Effect: Model sees all perspectives during training

Human vs. LLM Annotations

Comparing annotation sources

Experiment design:

1. Annotate same data with humans AND LLM
2. Train separate models on each
3. Evaluate both on human-annotated test set
4. Compare performance

Key finding (various studies):

- LLM-trained models often comparable for simple tasks
- Human annotations better for subjective/complex tasks
- Combined often best

When LLM Annotations Work

For model training:

Good scenarios:

- Large training set needed
- Task is objective with clear criteria
- Domain is well-represented in LLM training
- Small quality drop is acceptable

Poor scenarios:

- High-stakes application
- Subjective judgment required
- Novel domain
- Training benchmark models

Practical Considerations

Model selection:

- Start simple (Logistic Regression, SVM)
- Establish baseline before complex models
- Consider data size (transformers need more)

Hyperparameter tuning:

- Use dev set, not test set
- Report best dev configuration
- Don't overfit to dev set

Reproducibility:

- Set random seeds
- Document all preprocessing
- Release code and data

Baseline Models

Always establish baselines:

Simple baselines:

- Most frequent class (majority baseline)
- Random prediction
- Simple rules

Standard ML baselines:

- Bag of Words + Logistic Regression
- TF-IDF + SVM
- FastText

Then consider:

- BERT/RoBERTa fine-tuning
- Domain-specific models

Semester Project: Modeling

For your project:

1. Train at least one model on your gold standard
2. Report baseline performance
3. Analyze errors
4. Connect to annotation quality

Questions to address:

- Does model performance match IAA?
- Which categories are harder?
- What errors does the model make?
- How would you improve annotation?

Next Class: Modeling and Evaluation II

Lecture 22 (Apr 15): Annotation-Informed Modeling II

Topics:

- Evaluation metrics for annotated tasks
- Precision, Recall, F1 for various task types
- Error analysis using annotations
- LLM-as-judge evaluation
- Building evaluation benchmarks

Key Takeaways

1. **Standard training** uses gold labels, ignoring uncertainty
2. **Soft labels** capture annotation distributions
3. **Multi-annotator learning** models all perspectives
4. **LLM annotations** can work for training simple task models
5. **Always establish baselines** before complex models
6. **Model performance** is bounded by annotation quality

Questions?

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu