FALL TERM 2025

CSE 594: HUMAN AI INTERACTION AND SYSTEMS

ASSIGNMENT 2

SUBMITTED BY: VANSHIKA SINGLA

UM UNIQUE NAME: VANSHIII

UM ID: 54433790

AI-Assisted Task Boundary Tone Detection

Abstract

This project studies how humans and AI can jointly detect tone boundaries in short chat messages. The task asks participants to judge whether a message sounds Consensual Flirt or Too Forward. The AI model provides its guess, a confidence score, and a short rationale. Humans confirm or correct the label. The goal is to evaluate human-AI complementarity: the balance between AI speed and human social sensitivity and to evaluate how AI assistance changes accuracy and decision speed compared to humans alone and AI alone.

Task Description

Each trial shows one short message with the AI's label, confidence, and rationale (for example "hedge or uncertainty phrasing"). Participants choose the final label: Consensual Flirt or Too Forward, or pick Not sure for edge cases. Trials are brief so a random 10-minute subset can be used later in Assignment 3.

This is a setting that benefits from AI assistance but should not be fully delegated. The model recognizes obvious pressure cues such as prove it or send a pic, but humans handle hedges, teasing, and mixed signals better.

Assignment 3 will use a random subset of these 200 trials, designed to be completed within about 10 minutes.

Dataset

- Study dataset: 200 messages
  - 50 clear consensual
  - 50 clear too forward
  - 50 ambiguous positive leaning
  - 50 ambiguous negative leaning
- Training dataset: 600 separate messages for model training only.
- Columns:
  input_text, ground_truth_label, model_output_label, model_score, human_rating_placeholder, ai_readable_label, ai_rationale
- Messages are short (< 25 words), synthetic, and non-explicit. A Not sure option will be available in Assignment 3.

I designed the task and wrote both datasets myself, creating a balanced and ambiguous sample to explore human-AI complementarity.

All messages are synthetic and non-explicit. The task avoids sensitive or real data and is ethically appropriate for classroom studies.

Model and Implementation

A small model trained from scratch

- TfidfVectorizer on character n-grams (3–5)
- Logistic Regression (max_iter = 300)
- 10 percent of training used as validation to choose threshold that maximizes F1.
- Fixed random seeds and local runtime only (no external API).
- Repro files include requirements.txt, a2_runinfo.txt, and README.md.

Results

Accuracy: 0.8000

F1: 0.8333

Precision: 0.7143

Recall: 1.0000

Chosen threshold: 0.150 val_f1: 1.0000

Per class:

TOO_FORWARD P=1.000 R=0.600 F1=0.750 n=100

CONSENSUAL P=0.714 R=1.000 F1=0.833 n=100

Interpretation: The tuned threshold favours recall on *Consensual* cases. The model catches all consensual items but marks 40 *Too Forward* messages as consensual. This is acceptable because the goal is to ensure no safe messages are wrongly flagged. Humans then review the borderline pushy ones.

Metric choice: Accuracy and F1 summarize performance; precision and recall show safety trade-offs. Per-class scores make the imbalance visible.

Error Analysis

Top buckets from a2_error_analysis.csv

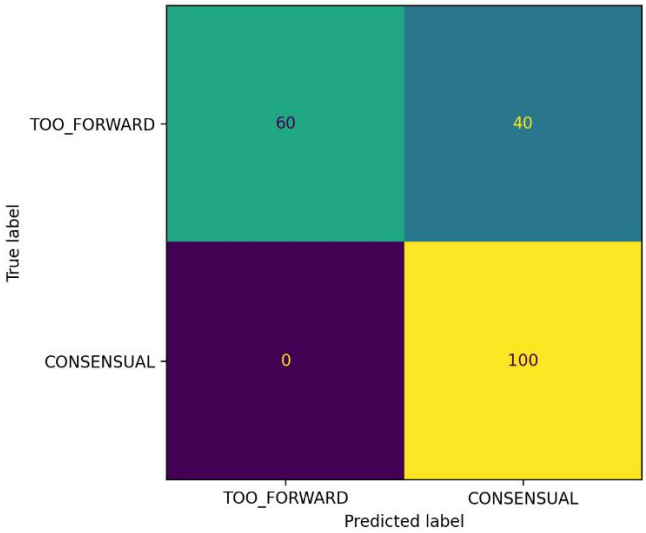
- guilt_push 54.35 %
- hedges_uncertainty 17.39 %
- teasing_push 13.04 %

Summary: Guilt phrases like *you never reply* and *you do not care* sound friendly to the model, which then labels them as consensual. Hedges such as *maybe* or *not sure* blur intent. Teasing lines like *come on* sit on the boundary. These are the situations where human judgment adds the most value.

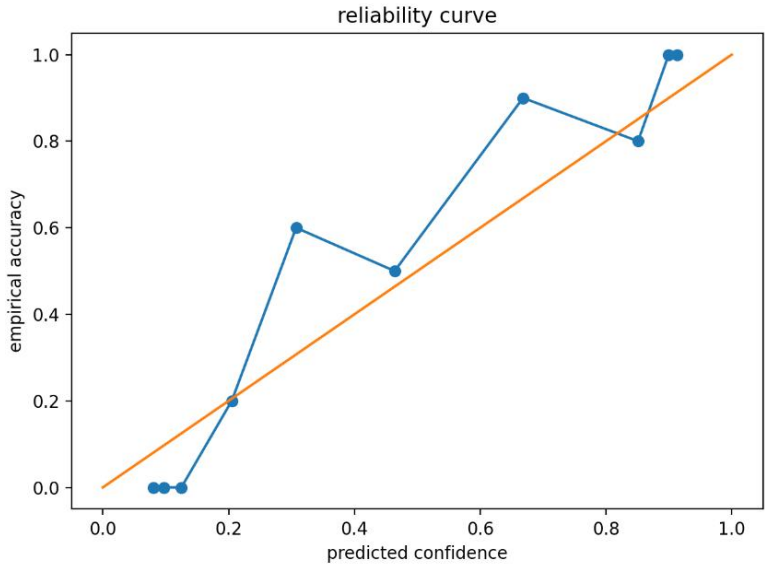
The first dataset was too clean and produced perfect scores. I added ambiguous cases to make errors and complementarity visible.

Figures

- Confusion matrix (a2_confusion_matrix.png): 100 consensual correct, 60 too forward correct, 40 too forward predicted as consensual.



- Calibration curve (a2_calibration.png): mid-range confidence is fairly reliable; extremes are less so. Confidence should guide but not decide.



- Example model output:

	input_text	ground_truth_label	model_output_label	model_score	human_rating_placeholder	ai_readable_label	ai_rationale
2	u know u want to but i will not push	1	1	0.3178629852459108		CONSENSUAL_FLIRT	lexical pattern features
3	this could be fun if u r comfortable with it	1	1	0.8334185152174288		CONSENSUAL_FLIRT	boundary or consent wording
4	is it okay if i call you cute	1	1	0.9095733075728915		CONSENSUAL_FLIRT	boundary or consent wording
5	want to grab coffee this week if you are free	1	1	0.9038298722180597		CONSENSUAL_FLIRT	lexical pattern features
6	come over to my place now no excuses	0	0	0.09358405709682588		TOO_FORWARD	lexical pattern features
7	hey send me a pic right now i do not care if you are busy	0	0	0.07471712061002658		TOO_FORWARD	deadline or time pressure
8	hey would you be comfortable with a dinner date	1	1	0.9018392106901123		CONSENSUAL_FLIRT	boundary or consent wording

Predicted Ranking for Assignment 3

1. Human + AI
2. Human only
3. AI only

Reason: AI gives speed and consistency but overlooks social context. Humans fix guilt and teasing mistakes while staying fast.

Submitted Files

Included files

- a2_report.pdf
- a2_study_with_outputs.csv
- a2_study_dataset.csv
- a2_study_dataset_perfect.csv
- a2_training_dataset.csv
- a2_model.pkl
- train_eval.py
- a2_metrics.txt
- a2_error_analysis.csv
- a2_confusion_matrix.png
- a2_calibration.png
- a2_dataset_summary.txt
- a2_runinfo.txt
- requirements.txt
- README.md

The grader can reproduce results by running one command from the README in a clean environment with Python and the listed packages.