

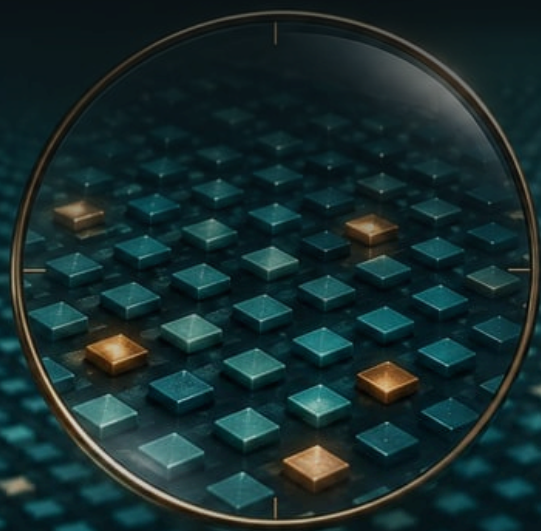
FIELD NOTES N°008 • BENCHMARK

Jev vs trained classifiers

I tested TypeSafe's zero-shot model against fine-tuned and classical ML on three public benchmarks, and looked closely at where each one makes mistakes.

Emilio Gagliardi

KEY FINDINGS • 10 SLIDES



THE TEST

One zero-shot model against models that learn from labels

Jev classifies with no task training. The trained models saw the same labelled rows, scored on the same held-out test rows.

INTENT

CLINC150

150 intents plus out-of-scope requests

SENTIMENT

Airline tweets

Negative, neutral or positive

BANK SUBSCRIPTION

41,188 customer records

Subscribed or not, from a table of mixed fields

THE ANSWER

Jev never beat the strongest trained model. It tied one.

INTENT

DeBERTa ahead by 0.049

macro-F1, 95% interval 0.036 to 0.067

SENTIMENT

Indistinguishable

Difference +0.002 (−0.018 to +0.022), and Jev needed no labels

BANK SUBSCRIPTION

Trained models ahead by 0.109 to 0.126

macro-F1 across logistic regression, random forest and boosting

WHERE IT HOLDS UP

Text, where the label names carry the task

98.5–98.7%

of sentiment messages got the same label across repeat runs

93.9–95.1%

agreement between tuned DeBERTa seeds on the same task

INTENT

Jev refused unsupported requests 20.9 points more often than the trained comparator.

Repeatable labels simplify monitoring. Zero labels meant a first version with no labelling project.

WHERE IT DOES NOT

On a customer table, Jev found one rule

1,998 /
2,000

of Jev's answers matched
the rule "last campaign
succeeded" (`poutcome =
success`)

~500

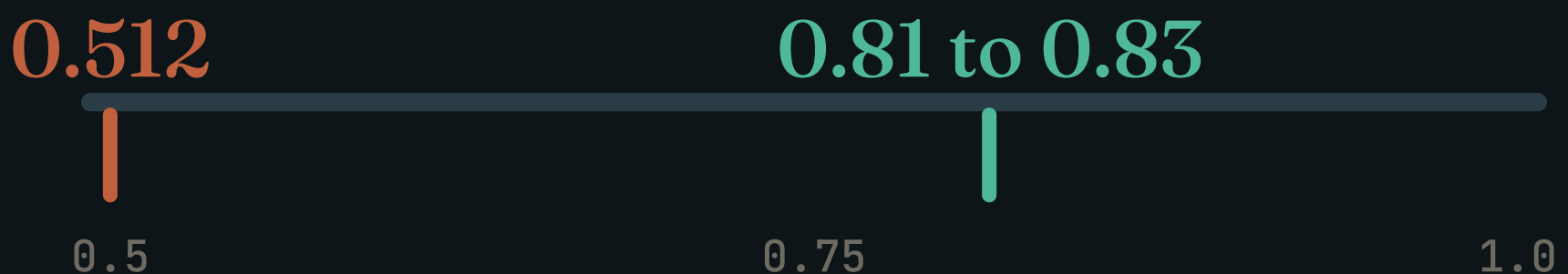
labelled rows were enough
for gradient boosting to
pass Jev

Test the simplest rule first. A macro-F1 of 0.605
looks like skill until a one-line rule scores 0.608.

ERRORS

Can Jev's confidence find its own mistakes?

Error-detection AUC: 0.5 is a coin flip, 1.0 separates wrong from right perfectly.



ON TEXT

A wrong answer gets lower confidence than a right one about 81% to 83% of the time. That makes an error budget workable.

ON BANK DATA

0.512 is chance. Confidence does not say which answers to trust.

AN ERROR BUDGET, COUNTED

1,000 airline messages at a 5% budget

Set the budget first: of the answers acted on without a person, at most 5 in 100 may be wrong. The cut-off was fixed on 2,142 other messages, then applied once.

WHERE THE 1,000 GO



648 handled alone · 352 to a person

INSIDE THE 352



about 127 wrong · 225 right

4.6%

wrong among the 648
handled alone, inside the
5% budget

81%

of all of Jev's mistakes
landed in the review lane

THE SAME DIAL ON BANK DATA

No cut-off meets the budget, so everything goes to a person

WHERE 2,000 RECORDS GO



none handled alone · all 2,000 to a person

INSIDE THE REVIEW LANE



217 wrong · 1,783 right

The automated slice stayed about 10% wrong at every coverage level. When confidence carries no error signal, triage by confidence sends reviewers a random sample.

A staged policy, not a model choice

- 1 **Start zero-shot where the labels explain the task.** No labelling project, and repeatable labels.
- 2 **Write the error budget per error type, with its cost,** before choosing a threshold.
- 3 **Freeze the threshold on validation data,** and give the review lane an owner.
- 4 **Turn every correction into a label** and re-measure when the label pool doubles.

A zero-shot model that matches your baseline has solved a labelling problem, not an accuracy problem.

The full essay has the method, every table and an interactive error-budget chart you can adjust.

emilio-gagliardi.ca/notes/008-jev-vs-trained-classifiers-zero-shot-benchmark.html