



Thai Sentiment Analysis: Accuracy on Public Thai Social-Media and Review Benchmarks

Service version v1.0 · Measured 28 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. The iApp Thai Sentiment Analysis API labels a Thai text as positive, neutral or negative. On the official test split of the Wisersight Sentiment corpus — 2,614 public Thai social-media posts labelled positive, neutral or negative — it returns the correct label for **75.8%** of posts (95% CI 74.1–77.5), with a macro-averaged F1 of 70.2. The errors are mostly between neutral and one of the polar classes: on posts that are positive or negative and answered as either, the service names the right polarity 94.7% of the time. On 1,019 restaurant reviews rated one, two, four or five stars by their own authors, it agrees with the stars on 84.4% (CI 82.2–86.5), mostly by calling a favourable review neutral. When the service reports a confidence of 0.9 or more, which it does for 42.9% of posts, it is right 89.4% of the time. Long texts are the service’s present limit: through the public API a text of more than 964 Thai characters is refused, and should be sent in parts (Section 4.5).

1 Introduction

The Thai Sentiment Analysis API (POST `/v3/store/nlp/sentiment-analysis`) takes one Thai text and returns a label — `pos`, `neu` or `neg` — and a confidence score between 0 and 1. It is used to sort customer comments, social-media mentions and reviews by tone. This paper answers the question a buyer asks first: how often is the label right, on which kinds of text, and where does it go wrong. It uses only public data, so anyone can repeat the measurement.

2 Test Data

Wisersight Sentiment, official test split

2,671 Thai social-media posts and comments about brands, products and services, each labelled by human annotators as positive, neutral, negative, or a question [1]. Released into the public domain (CC0). The 57 posts labelled *question* have no counterpart among the service’s three labels; they are reported separately, leaving 2,614 posts for the three-class figures.

Wongnai restaurant reviews, test split

Thai restaurant reviews, each with the one-to-five-star rating given by its author [2]. A seeded random sample of 1,500 reviews (seed 1337). Stars are not sentiment labels, so we use them only where they are unambiguous: four and five stars as positive, one and two as negative. Three-star reviews are shown but not scored.

The service’s model was trained on public Thai sentiment corpora that include the training splits of both datasets. The test splits are the partitions their publishers hold out for evaluation. Section 4.2 checks that the Wisersight test posts behave as unseen data.

3 Method

Each text was sent unchanged to the production model service, the same instances that answer `api.iapp.co.th`, on 28 October 2026, with at most three requests in flight. To confirm that the public API returns the same

answers, a seeded sample of 300 test posts (seed 2026) was also sent through `api.iapp.co.th`. The labels were identical on all 299 posts answered on both paths; the remaining post, of 1,422 characters, was refused by the public API for its length (Section 4.5).

- **Accuracy:** the share of texts whose returned label equals the reference label. A request that returned an error counts as wrong.
- **Precision, recall, F1** per class, and the **macro-F1**, their unweighted mean, which is not inflated by the large neutral class.
- **95% confidence intervals** by percentile bootstrap over texts, 2,000 resamples (1,000 for macro-F1), seed 1337.

4 Results

4.1 Wisesight test split

Table 1: Three-class results on the 2,614 Wisesight test posts labelled positive, neutral or negative.

Measure	Result
Accuracy	75.8% [74.1, 77.5]
Macro-F1	70.2 [68.3, 72.2]
Accuracy over answered posts only	75.9%
Accuracy over all 2,671 posts, questions counted as wrong	74.2%
Requests that returned an error	4 of 2,671 (0.15%)

Table 2: Per-class precision, recall and F1 (%), Wisesight test split.

Class	Posts	Precision	Recall	F1
Positive	478	66.0	42.3	51.5
Neutral	1,453	75.6	87.3	81.1
Negative	683	81.5	74.7	77.9

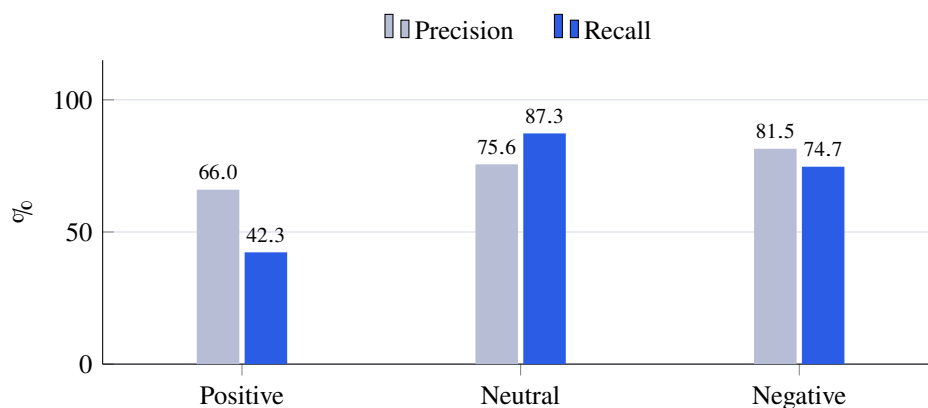


Figure 1: Precision and recall per class on the Wisesight test split.

Table 3: Confusion matrix, Wisesight test split: rows are the reference label, columns the returned label.

Reference	Positive	Neutral	Negative	Error
Positive	202	251	25	0
Neutral	89	1,269	91	4
Negative	15	158	510	0

Polarity reversals are rare: 15 negative posts were labelled positive and 25 positive posts negative, together 1.5% of the test posts. Of the 57 posts labelled *question*, 91.2% were returned as neutral.

4.2 Held-out check

A model that had seen the test posts in training would score on them as it does on its training posts. Table 4 compares the three Wisesight splits; the samples are seeded (seed 1337).

Table 4: Three-class accuracy (%) by Wisesight split.

Split	Posts	Accuracy	95% CI	Macro-F1
Training split (sample)	979	80.7	[78.1, 83.0]	77.3
Validation split (sample)	981	75.1	[72.5, 77.7]	69.9
Test split	2,614	75.8	[74.1, 77.5]	70.2

Accuracy on the training split is 4.9 points higher than on the test split, while the validation and test splits agree. This is the pattern expected when the test split was not used for training.

4.3 Restaurant reviews

Table 5: Wongnai reviews by the author's star rating: share (%) of each returned label.

Stars	Reviews	Positive	Neutral	Negative	Error
1	10	0.0	20.0	80.0	0.0
2	44	4.5	34.1	61.4	0.0
3	481	31.4	64.2	2.7	1.7
4	702	83.5	14.1	0.0	2.4
5	263	90.9	3.8	0.0	5.3

Taking four and five stars as positive and one and two as negative, the service agrees with the author on 84.4% of 1,019 reviews (95% CI 82.2–86.5). It finds 64.8% of the 54 low-rated reviews negative and 85.5% of the 965 high-rated reviews positive; a further 11.3% of high-rated reviews are called neutral. Long restaurant reviews often mix praise with complaints, and the service then tends towards neutral. 39 of the 1,500 reviews (2.6%) returned an error because of their length, and count as wrong above.

4.4 Confidence score

Table 6: Accuracy by returned confidence score, Wisersight test split (answered posts).

Score	Posts	Share (%)	Accuracy (%)
0.90–1.00	1,119	42.9	89.4
0.70–0.90	871	33.4	73.2
0.50–0.70	551	21.1	57.0
below 0.50	69	2.6	42.0

The score is a usable signal. An integrator who needs high precision can act automatically on labels with a score of 0.9 or more and send the rest for review.

4.5 Text length

Two limits apply to long texts.

The public API. The text is sent in the query string of the request, where each Thai character takes nine bytes once encoded. The API accepted an all-Thai text of 964 characters (a request line of 8,306 bytes) and refused one of 965 with HTTP 414. Texts that mix in spaces, digits or Latin letters can be somewhat longer. By this limit, 10 of the 2,671 Wisersight test posts (0.4%) and 159 of the 1,500 sampled restaurant reviews (10.6%) could not be sent in one request.

The model. Sent to the model service directly, texts of up to 1,000 characters were always answered. Above that, the service returned an error (HTTP 500) for 11 of 150 texts between 1,000 and 2,000 characters and for all 32 texts over 2,000; the shortest text that failed had 1,327 characters. The model reads a bounded number of tokens, and a text beyond that is refused rather than shortened.

Table 7: Errors from the model service by text length in characters, over all 6,171 texts sent.

Characters	Texts	Errors	Share (%)
0–500	5,544	0	0.0
501–1,000	445	0	0.0
1,001–1,500	112	2	1.8
1,501–2,000	38	9	23.7
over 2,000	32	32	100.0

Until the service splits long texts itself, send a long text as separate paragraphs of no more than about 900 characters each and combine the labels.

5 Production Performance

Through the public API, from a client in Bangkok, the median end-to-end time per request was 0.33 s and the 95th percentile 0.92 s (299 requests, two in flight). The model itself took a median of 0.53 s per text during the full run, which shared the service with live traffic.

6 Limitations

- **Domain.** Both datasets are consumer text: social media and restaurant reviews. Formal text, such as news or official letters, and domain-specific language were not measured.

- **Three labels.** The service does not detect questions, sarcasm, or mixed sentiment as such; mixed texts tend to be labelled neutral.
- **Training overlap.** The service’s training data includes the training splits of both corpora. Section 4.2 supports, but cannot prove, that the Wiselight test split was not seen.
- **Stars are a proxy.** An author’s star rating is not a sentiment annotation; the review figures measure agreement with the rating.
- **Length.** Every figure is over texts as the datasets give them; the texts that the service refused for their length count as wrong.
- **No competitor comparison.** This paper reports absolute accuracy only.

7 Reproducibility

Both datasets are public. The benchmark script, the scoring script and the raw response for every text are available from iApp Technology on request, sale@iapp.co.th. The public-API check sent 25,726 characters.

References

- [1] A. Suriyawongkul, E. Chuangsuwanich, P. Chormai and C. Polpanumas. *PyThaiNLP/wiselight-sentiment: First release*. Zenodo, 2019. https://huggingface.co/datasets/pythainlp/wiselight_sentiment
- [2] Wongnai. *Wongnai Review Dataset*. 2018. https://huggingface.co/datasets/Wongnai/wongnai_reviews

About iApp Technology. iApp Technology Co., Ltd. is a Thai artificial-intelligence company providing OCR, document understanding, speech, identity verification and large-language-model services for the Thai language.

Web: iapp.co.th · E-mail: sale@iapp.co.th · Tel: +66 2 124 4041