Model Card for **HAN-FakeNews-Detector**

(Hierarchical Attention Network for Binary Fake-News Classification)

Akziz Amin, Lema Santiago, Raj Anshu, Ziehm Leander-Arun

Model version: `v0.1` Compiled: June 6, 2025

1 Model Details

– **Developed by** students at the Deggendorf Institute of Technology, June 2025, v0.1.

– **Vocabulary size at fit time**: 50 000 tokens; minimum word frequency = 2.

– **Architecture:** Hierarchical Attention Network comprising:

  – *Input*: documents truncated to 20 sentences × 50 tokens and mapped to 200-d trainable embeddings.
  – *Word encoder*: one bidirectional GRU layer (50 hidden units each direction, dropout 0.1) with 100-d additive attention.
  – *Sentence encoder*: one bidirectional GRU layer (50 hidden units each direction, dropout 0.1) with 100-d attention.
  – *Classifier*: 100-d document vector → fully connected layer → two logits.

2 Intended Use

Primary intended uses

Assist fact-checkers, journalists and platform moderators by flagging news articles likely to be fabricated or misleading.

Primary intended users

Media-monitoring organisations, newsroom editors, social-media trust & safety teams, researchers studying misinformation.

Out-of-scope uses

– Fully autonomous removal or down-ranking of content without human review.

– Real-time moderation of live audio/video streams.

– Legal or policy enforcement decisions without corroborating evidence.

3 Factors

– **Language**: English-only; performance may drop on other languages or heavy dialect.

– **Domain**: politics, health, celebrity gossip etc. have different linguistic cues.

– **Article length**: short posts <100 words often truncated to one or two sentences.

– **Publication date**: model frozen on 2018 data; concept drift likely as new rumours emerge.

4 Metrics

– Accuracy, Precision, Recall, F1, AUROC, Average Precision.

– Threshold 0.50 (equal-error point on validation ROC).

– Uncertainty: 95 % CI from 100-round bootstrap on the test split.

Table 1: Aggregate performance (stratified 80/10/10 split).

Split	Accuracy	Precision	Recall	F1	AUROC
Validation	0.9993	0.9993	0.9993	0.9993	1.0000
Test	0.9984	0.9984	0.9984	0.9984	0.9999

Robustness checks.
5-fold CV mean ± SD: Acc. = 0.9988 ± 0.0004;
bootstrap 95 % CI: Acc. = 0.9985 [0.9971, 0.9996], AUROC = 0.9999 [0.9998, 1.0000].

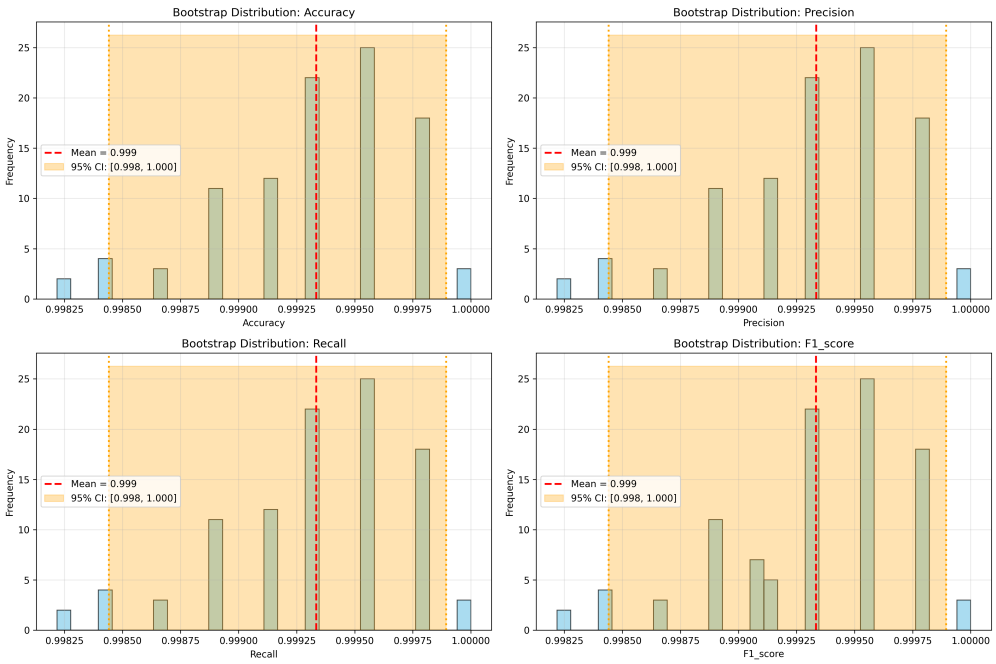

Figure 1: Bootstrap distribution of key metrics.

5 Training Data

– **Source**: Kaggle `Fake.csv` and `True.csv`.

– **Size**: 45 898 articles (fake 23 481, true 22 417).

– **Splits (stratified)**: Train 35 918, Validation 4 490, Test 4 490.

– **Pre-processing**: lower-casing, punctuation kept, NLTK tokenisation; truncation to 20×50; words <2 occurrences mapped to `<UNK>`.

6 Evaluation Data

Same schema as training set.
Held-out test split never used for hyper-parameter tuning.
Additional robustness: 5-fold cross-validation and 100-round bootstrap.

7 Ethical Considerations

– Possible false-positives on satire or opinion pieces.

– Label bias in the original dataset propagates to predictions.

– Drift: new misinformation styles may reduce recall over time.

– Mitigations: periodic re-training, human-in-the-loop review, external bias audits.

8 Caveats & Recommendations

– Not optimised for micro-posts (<20 tokens).

– Retrain every 3–6 months to counter concept drift.

– For multilingual deployment start with multilingual embeddings and fine-tune.