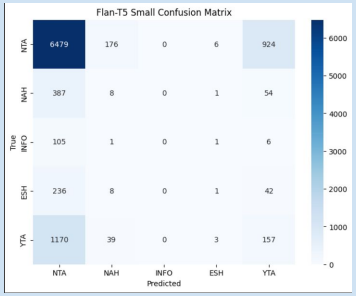
Reddit AITA Evaluation Strategy

Classifications

- Statistical Report

	precision	recall	f1-score	support
NTA	0.77	0.85	0.81	7585
NAH	0.03	0.02	0.02	450
INFO	0.00	0.00	0.00	113
ESH	0.08	0.00	0.01	287
YTA	0.13	0.11	0.12	1369
accuracy			0.68	9804
macro avg	0.20	0.20	0.19	9804
weighted avg	0.62	0.68	0.65	9804

- Confusion Matrix



- Matthews Correlation Coefficient (MCC)

Justifications

- ROGUE (1, 2, L, LSum)

Metric: rouge

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is a set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation.

Note that ROUGE is case insensitive, meaning that upper case letters are treated the same way as lower case letters.

This metrics is a wrapper around Google Research reimplementation of ROUGE: https://github.com/google-research/google-research/tree/master/rouge

- COMET

Metric: comet

Crosslingual Optimized Metric for Evaluation of Translation (COMET) is an open-source framework used to train Machine Translation metrics that achieve high levels of correlation with different types of human judgments (HTER, DA's or MQM). With the release of the framework the authors also released fully trained models that were used to compete in the WMT20 Metrics Shared Task achieving SOTA in that years competition.

- BLEURT

BLEURT: Learning Robust Metrics for Text Generation

Thibault Sellam, Dipanjan Das, Ankur P. Parikh

Text generation has made significant advances in the last few years. Yet, evaluation metrics have lagged behind, as the most popular choices (e.g., BLEU and ROUGE) may correlate poorly with human judgments. We propose BLEURT, a learned evaluation metric based on BERT that can model human judgments with a few thousand possibly biased training examples. A key aspect of our approach is a novel pre-training scheme that uses millions of synthetic examples to help the model generalize. BLEURT provides state-of-the-art results on the last three years of the WMT Metrics shared task and the WebNLG Competition dataset. In contrast to a vanilla BERT-based approach, it yields superior results even when the training data is scarce and out-of-distribution.