News Quality Scoring: an Ensemble of NLP Methods to Assess News Quality

Oday Najad, Tommaso Di Fant, Wageesha Widuranga

Goals of the project

In today's internet-based world news are abundant and often too numerous for anyone to handle. Many articles are either clickbaits, contain no useful information or, in the recent years, are prevalently AI generated.

We hope to build a model that evaluates articles and textual content on the internet in such a way that high quality news, which are well-written and contain lots of information, can be recognized and have more visibility over worse articles.

To achieve this we split the task of evaluating a news article in 6 independent tasks that were researched and implemented separately over the course of the project.

News Evaluation Tasks

01 Writing Quality

02 AI Content Detection

03 Clickbait Detection

04 Information Density

05 Plagiarism Detection

06 Fact checking

Writing quality

Dataset:

Writing quality is subjective and can refer to multiple distinct qualities of a text. To overcome this we train the model on 3 different datasets that have 3 separate tasks. Every dataset has different regression targets:

1. A general grade given to essays (Automatic Essay Scoring)
2. Six scores regarding grammar and syntax (ELLIPSE)
3. A readability score (CLEAR)

Model:

We fine-tuned DistilBert on these datasets mixed together, with a different classification head for each task.

MultiDataset-MultiTask (MDMT) technique: Trains across multiple datasets simultaneously.

Results:

0.05 Mean Square Error, averaged for all tasks and calculated on the validation set

State-of-the-art MSE: 0.03.

AI Content Detection

Dataset:

AI content detection is a wide task, it must work in very different texts, and has to detect all the different AI models' style of writing.
We first trained on a narrow dataset containing human vs AI generated essays (30K essays). Then we trained on a small subset of a more general dataset containing varied long-form text (Artem9k; 1.4M mixed texts).

Model:

We fine-tuned DistilBert on each dataset and produced two different models.
We also compare these two models against an open-source AI content detector.

Results:

The model trained on the small dataset achieved 0.99 F1 on it but very low performance on the unseen bigger dataset.
The model trained on the bigger dataset achieved 0.95 F1 on it and 0.51 F1 on the first dataset.
The open-source model achieved 0.75 F1 on the small dataset and was too slow to even evaluate on the second.

SOTA:

F1-Score: 0.97

Clickbait Detection

Data Tokenization:
uses DistilBertTokenizer ==> converts the titles into input embedding tokens and creates attention masks to attend for relevent different tokens in the sequence.

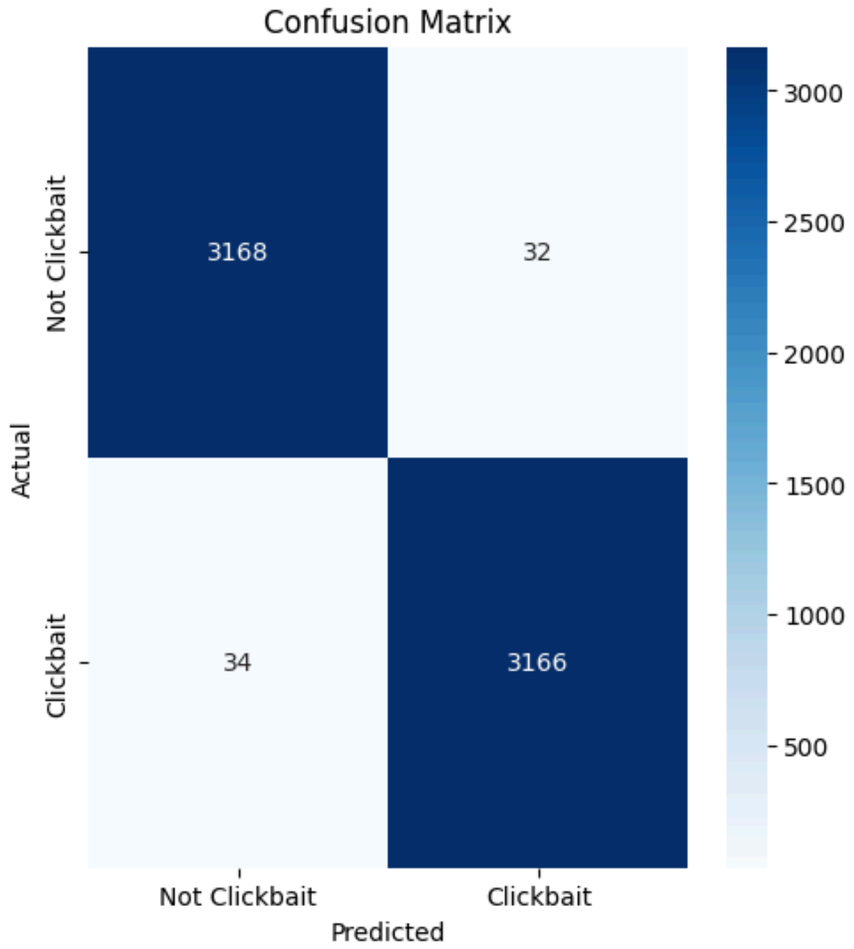
Model Architecture:
We have used the DistilBertForSequenceClassification model which preserves the performance just like the normal Bert model while maintaining computational efficiency. The weights are initialized from the distilbert-base-uncased pre-trained model.

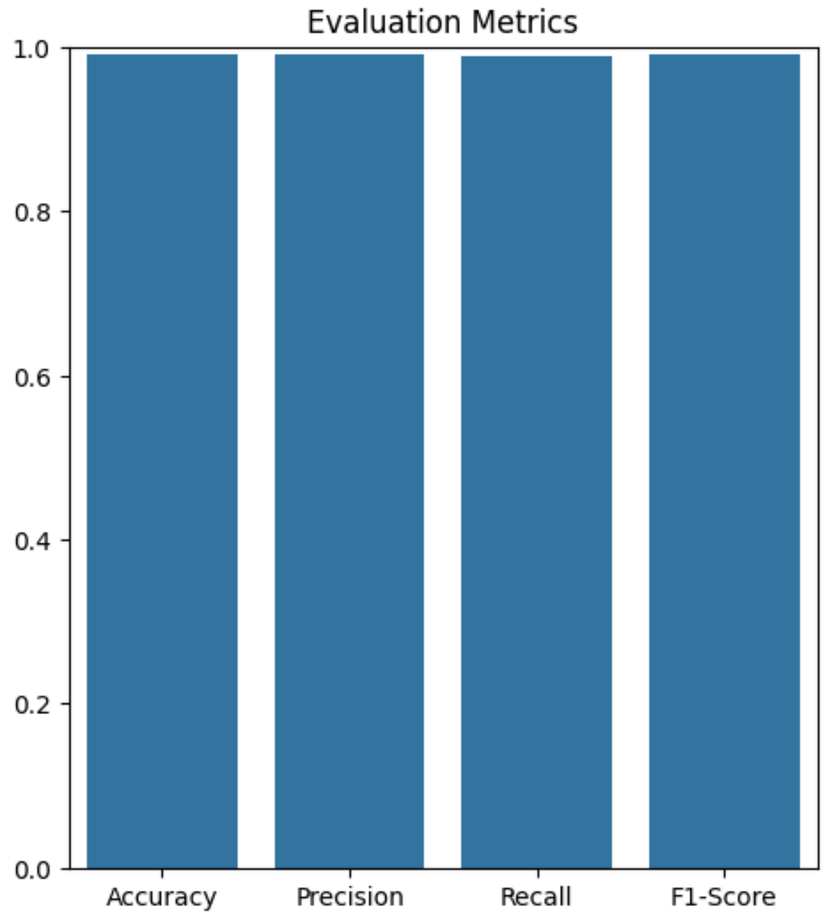
Dataset:
Dataset used is the publicly available clickbait_data dataset.

Training:
used "Trainer" from Hugging Face.

Evaluation Metrics:
Accuracy, Precision, recall and F1-score.

Clickbait Detection





Results:

Accuracy: 0.9897

Precision: 0.9900

Recall: 0.9894

F1-Score: 0.9897

State-of-the-art:

Accuracy: 0.995

F1-Score: 0.993

Information Density

Summarization Using BART:
The summarization model
used in this task is facebook/bart-large-cnn,
a pre-trained version of BART specifically
fine-tuned for summarizing news articles.

Evaluation Using ROUGE:
To evaluate the quality of the generated
summaries, the ROUGE metric is employed.
ROUGE measures the overlap between the
generated summary and the reference summary
by comparing the number of overlapping n-grams,
word sequences, and word pairs.

Dataset:
CNN/DailyMail dataset which contains articles that we have summarized.

ROUGE-1:
measures the overlap of unigrams between the generated
summary and the reference

ROUGE-2:
measures the overlap of bigrams between the
generated summary and the reference

ROUGE-L:
measures the longest common subsequence between
the generated and reference summaries

Information Density

Evaluation:

Accuracy: 0.9897

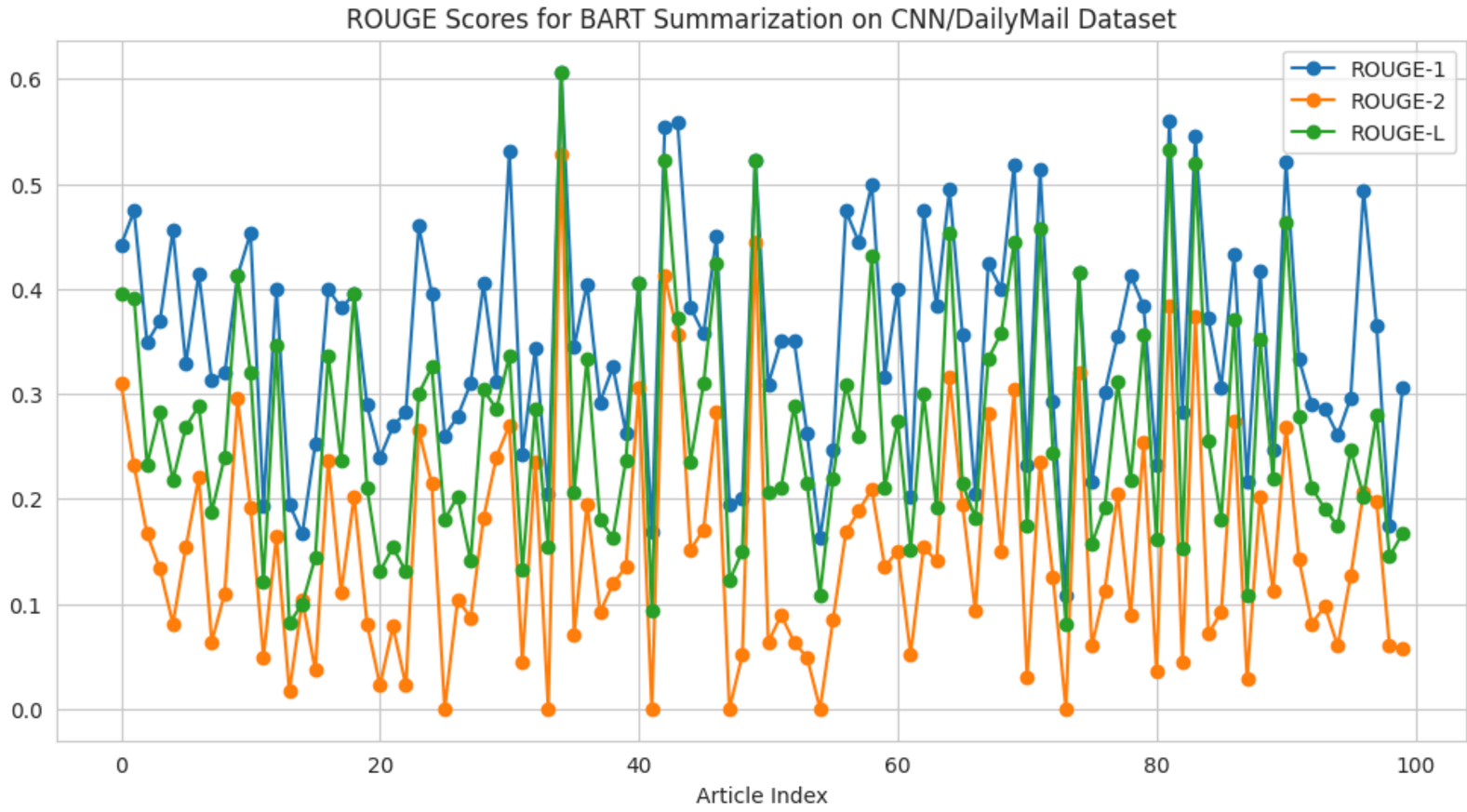
Precision: 0.9900

Recall: 0.9894

F1-score: 0.9897

State-of-the-art:

Accuracy: 0.995



Plagiarism Detection

Methods:

Cosine Similarity - Measures vector-based similarity of text.

Jaccard similarity - Measures the proportion of shared trigrams (intersection) relative to the total number of unique trigrams (union) between two sets.

Methods:

Containment Measure - uses the same set of trigrams, but instead of measuring the ratio between the union set, it measures the ratio in relation to a specific reference set

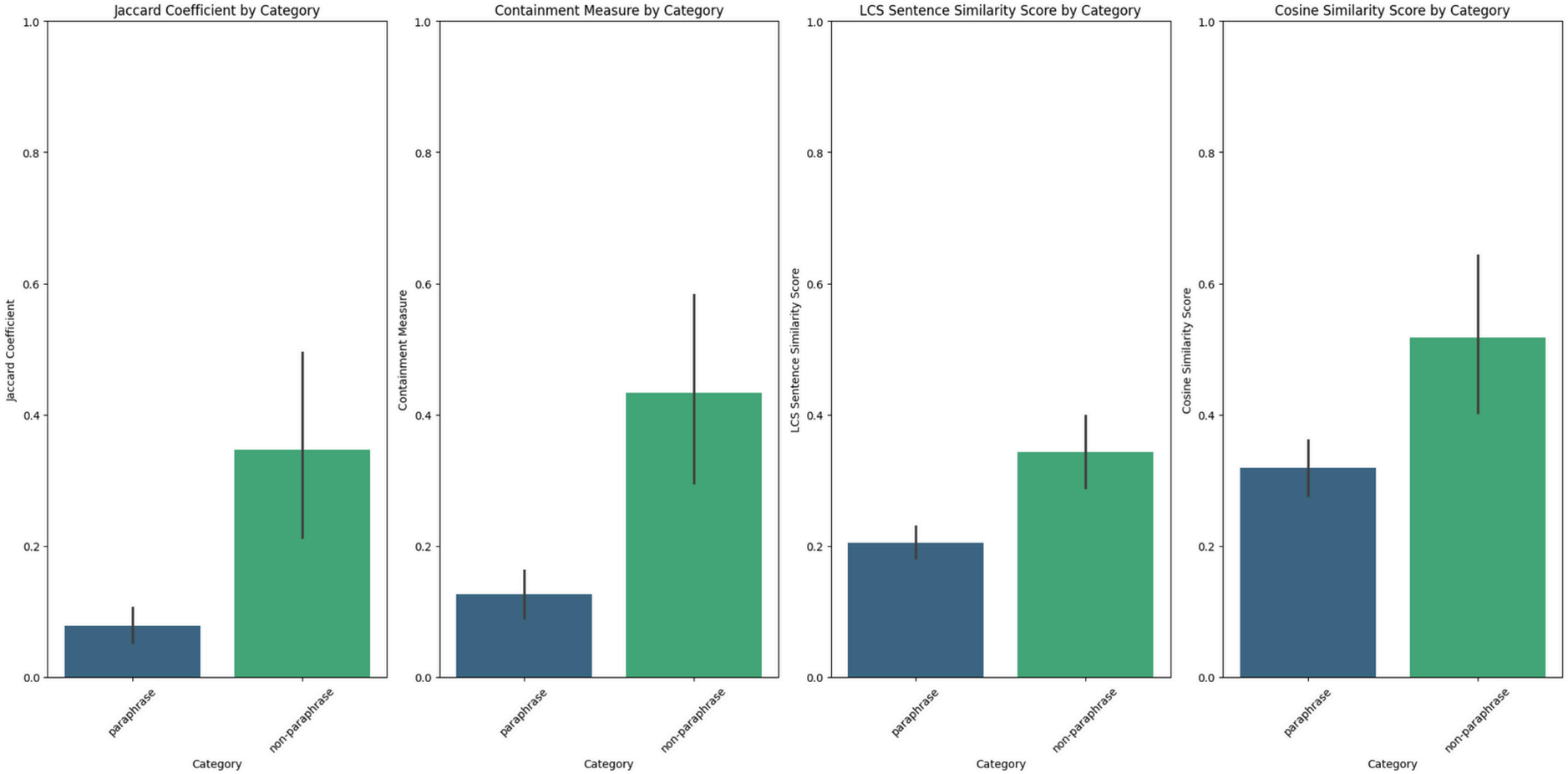
Longest Common Subsequence - This is a technique used in text similarity analysis to identify the longest sequence of words that appear in the same order in both texts, even if they are not consecutive.

Dataset:

Webis-CPC-11 dataset, which is designed for specific tasks, was used. It contains 7,859 candidate paraphrases obtained from Mechanical Turk crowdsourcing

Plagiarism Detection

Results:



Fact checking

Dataset:

FEVER (Fact Extraction and Verification) was created by modifying statements taken from Wikipedia and then verified without the knowledge of the original text

Methods:

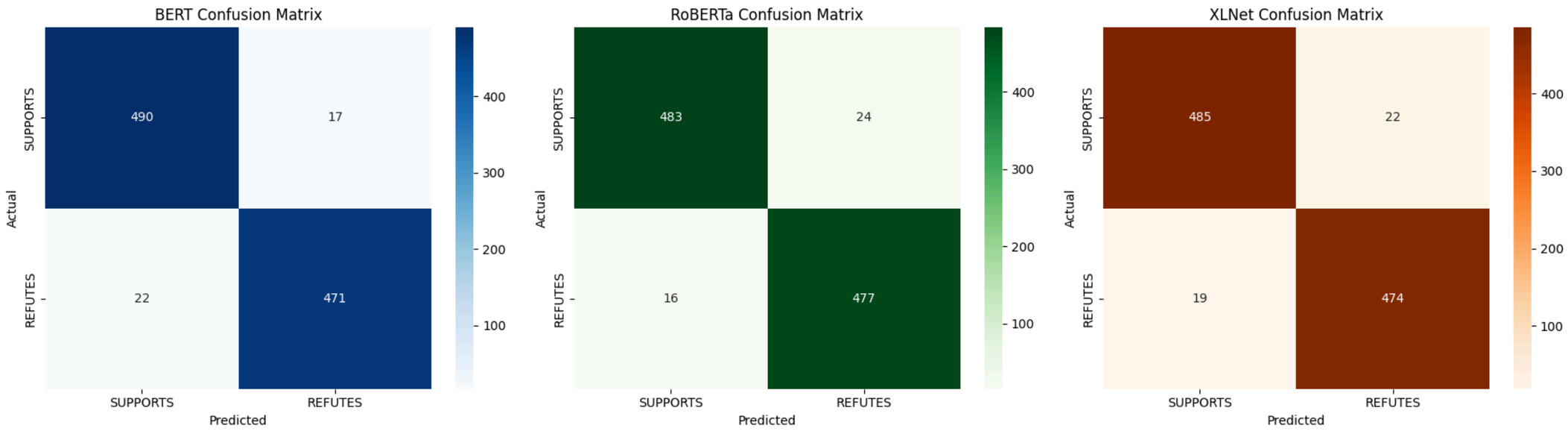
BERT – Model priorities the left and right context of every word in a phrase because it is designed with a bidirectional transformer architecture.

RoBERTa – Improves upon BERT's architecture in several ways. The Next Sentence Prediction (NSP) task is removed, and more data with larger batch sizes and learning rates are used for training.

XLNet – Combines the ideas of the BERT model with a permutation-based training model, where it learns to predict the order of the words in a sentence instead of just masked words.

Fact checking

Results:



Best accuracy: 0.60 achieved with BERT.

SOTA models reach accuracy of 0.75.

Conclusions

The intent of integrating these methods is to come up with a multiple and comprehensive analysis of news content, given that it is more challenging for those who analyze news to get high quality and credible articles in current media.

We think we got desired results in each task with the selected datasets and for best generalization and application more work is required on optimizing and perhaps introducing new approach on diverse data sets.

Implementation Tasks

Name	Task 1	Task 2
Oday Najad	Clickbait Detection	Information Density
Tommaso Di Fant	AI-generated Content Detection	Quality of Writing
Wageesha Widuranga	Plagiarism Detection	Fact Checking

Thank You

Oday Najad, Tommaso Di Fant, Wageesha Widuranga