Team #5 - BERT Brigade

Binary Machine-Generated Text Detection

Dhyey Nilesh Doshi; 40244534; dhyeydoshi2512@gmail.com

December 12, 2024

Abstract. This project implements a Naive Bayes classifier and introduces a RoBERTa - based and DistilBERT based model in order to compare them and accurately differentiate between human-written and AI-generated text. By using transfer learning with the RoBERTa-base architecture and fine-tuning on a diverse dataset, it achieves a robust **75.5%** test accuracy in detecting machine-generated content. In case of DistilBERT, it achieves a slightly more **78%** test accuracy. Whereas, using a probabilistic classification approach that is - Naive Bayes achieves an accuracy of **68%**. This solution offers a tool for verifying content and mitigating the risks associated with AI-generated content.

Github - `https://github.com/dhy3y/HumanVsAI-Text-Detection`

Github:

Contents

1 Goal of the project

The goal is to develop a model capable of identifying whether a given text is written by a human or generated by an AI language model. This can serve as an important tool for ensuring trust in any form of digital content. It can help fight misinformation.

There were two subtasks given by "Workshop on GenAI Content Detection"[1] from which I did the first one: "English-only MGT detection"

2 Methodology

The approach can be summed up as:

- Evaluating the effectiveness of Naive Bayes with BoW approach
- Understanding the key linguistic features [1] that distinguish human vs. machine text
- Training a model using transfer leaning (RoBERTa and DistilBERT)
- Analysis of Performance metrics

Dataset Preparation: From the original dataset I picked 10,000 lines to work on due to limiting processing power and that RoBERTa has 125M parameters. Preprocessing was done by removing noise such as: URLs, multiple spaces, special characters, numbers and in some experiments, removing stop words.

The dataset was balanced with 50% human and 50% machine text.

Category	Count
AI-generated	5,000
Human-written	5,000
Total Dataset	10,000

Table 1: Dataset Breakdown [2]

[1] In this context, linguistic features refer to the unique elements of language such as syntax and vocabulary choice.

Naive Bayes Classifier

- Bag of Words representation was used
- Stop words were removed
- Word likelihood calculation was done using Laplace smoothing with alpha = 1
- Log probability scoring applied for predictions
- Identified the most informative words

DistilBERT architecture

- DistilBERT base with 66M parameters.
- 5 epochs, AdamW optimizer, Cross entropy loss and 5e-5 learning rate
- Train/val/test split: 80/10/10 with batch size : 16

RoBERTa architecture

- RoBERTa base model with 125M parameters and sequence classification head
- 10 epochs with early stopping, Adam optimizer with 1e-5 learning rate
- Train/val/test split: 80/10/10 with batch size: 16

Appendix A shows the difference between DistilBERT and RoBERTa.

3 Evaluation

3.1 Results

Results for Naïve Bayes:

Class	Precision	Recall	F1-score
human	0.72	0.58	0.64
machine	0.65	0.78	0.71

Table 2: Performance metrics for **Naïve Bayes**

Most significant words according to Naïve Bayes classifier:

wellstructured, commendable, basically, davidson, ive, interplay, daunting, cmv, wellpresented, middlewich, avenues, obviously, versatility, roberts, sorry, insightful, ccp, terrible, sallekhana, innovative, insights, tldr, biodiversity, im, comprehensive, moth, hamsters, itll, definitely, guys, said

Table 3: Significant words

Results for DistilBERT:

Class	Precision	Recall	F1
Human	0.759	0.820	0.789
Machine	0.804	0.740	0.771

Table 4: Performance metrics for **DistilBERT**

Results for RoBERTa classifier:

Class	Precision	Recall	F1
Human	0.737	0.794	0.764
Machine	0.777	0.716	0.745

Table 5: Performance metrics for **RoBERTa-1** (stopwords removed)

Class	Precision	Recall	F1
Human	0.774	0.868	0.818
Machine	0.850	0.746	0.794

Table 6: Performance metrics for **RoBERTa-2** (including stopwords)

3.2 Analysis

The **Naive Bayes** classifier has achieved 68% accuracy, which is not high and there is room for improvement. As seen in Table 2, The F1-scores for human and machine classes are 0.64 and 0.71 with difference being (7%).

Moreover, the recall for machine-generated text is also higher than that for human, which means, it is comparatively better at correctly identifying machine-generated content from the data available.

Looking at the most significant words in Table 3, it offers a lot of insights:

- Words like "wellstructured," "commendable," "well-presented," and "insightful" suggest a more thoughtful approach to writing which **resembles machine generated text**.
- Personal pronouns like "I've" and "I'm" are often used in **human-written text.** And, words like "terrible", "sorry", and "definitely" may indicate emotional tones which is also an indication of the same.

The **DistilBERT** classifier achieved 78% accuracy on test dataset. As seen in Table 4, The F1-scores for human and machine classes are 0.79 and 0.77 respectively, with difference being (2%).
The **RoBERTa** classifier had a training accuracy of 98% and test accuracy of 75.5% which was higher than Naive Bayes, but lower than DistilBERT. The F1-scores for human and machine classes were 0.764 and 0.745 respectively, with difference being (2%).

On the other hand, If **RoBERTa** was used on a dataset **without removing stopwords**, with similar train accuracy, the test accuracy increased to 80.7%. The F1-score for human and machine increased to 0.818 and 0.794 respectively, which was a 5% increase as compared to the above result.

For DistilBERT and RoBERTa,

- The recall is higher for human written text whereas, the precision is higher for machine generated text. Which means it correctly identifies

a higher proportion of human texts. (this was observed to be opposite to Naive Bayes).

  - The key difference between the two models is due to their difference in complexities. RoBERTa is a deep learning model that captures complex patterns, whereas, Naive Bayes is a simpler, probability-based model that may overestimate the presence of machine text due to its simplistic feature dependencies.

- In text pre-processing, if stop-words are removed, it decreased the overall accuracy in RoBERTa.
- Upon increasing the number of epochs, the performance decreased due to overfitting.
- As seen in Figure 1 and 2 below, AUC (Area Under Curve) for ROC and PR curve was 0.83 and 0.86 for both RoBERTa and DistilBERT. But, in the case of RoBERTa without stopword removal, it was increased to 0.87 and 0.90 respectively. Which makes it the best model.

Appendix B explains what are these two curves and why is it helpful to plot them

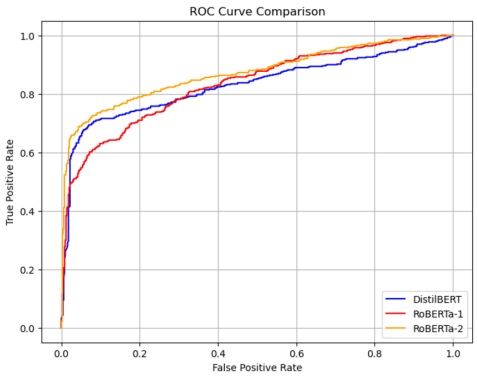

Figure 1: ROC curve

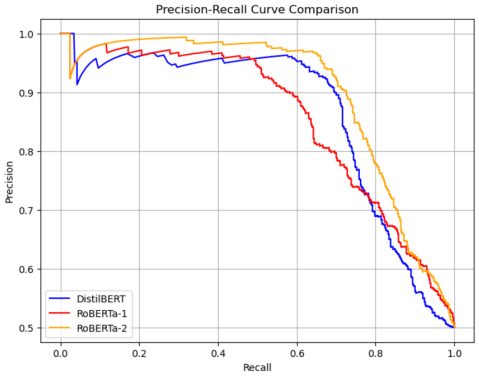

Figure 2: PR curve

Context: Roberta-1 is with stopwords removed. Roberta-2 is without removing stopwords.

4 Role of each team member

As a single-person project, I was responsible for all aspects of this research including data collection, model deployment, testing, poster presentation and writing the report.

5 Limitations

- In Naive Bayes, it is limited to word-level features and no word order/context is used which is the reason of its low performance.
- In RoBERTa, there is a maximum sequence length restriction of 512.
- For RoBERTa, not removing stopwords increased the accuracy. The same was not tested on DistilBERT due to time constraints. As DistilBERT had more accuracy in the original test, **possibly**, it could have achieved more accuracy than RoBERTa when NOT removing stopwords.
- Additionally, there are high computational requirements for running the whole dataset of 600,000 lines.

6 Difference from my original proposal

- Used less dataset than planned. When preparing the proposal, I overestimated my laptop's processing power.
- I had planned to use the Assignment-2 Word2Vec model as well, to compare it with the transformer approach but dropped due to time constraints.
- I was not able to combine multiple models to see if it improves performance.

7 Conclusions

In conclusion, this report provides a comparison of three different approaches to detecting machine-generated text: Naive Bayes, DistilBERT, and RoBERTa. The results demonstrate that transformer-based models significantly outperform traditional probabilistic approaches, with DistilBERT achieving the

highest accuracy of 78%, followed by RoBERTa at 75.5%, and Naive Bayes at 68%.

A key finding was the importance of retaining stopwords in transformer-based models, as proved by RoBERTa's improved performance (80.7% accuracy which became the new highest) when stopwords were included. This suggests that, what we consider the "insignificant words", contain valuable stylistic patterns that help distinguish between human and machine-generated text.

The analysis of model behaviors revealed interesting patterns: while Naive Bayes showed better recall for machine-generated text, both transformer models demonstrated higher recall for human-written content.

8 References

References

[1] Workshop on detecting ai generated content. https://genai-content-detection.gitlab.io/sharedtasks.

[2] Coling 2025 mgt english dataset. https://huggingface.co/datasets/Jinyan1/COLING_2025_MGT_en.

[3] Tung M. Phung. A review of pre-trained language models: from bert, roberta, to electra, deberta, bigbird, and more. dec 2021.

[4] Wikipedia contributors. Receiver operating characteristic. URL https://en.wikipedia.org/wiki/Receiver_operating_characteristic.

[5] Kolena. Pr curve. URL https://docs.kolena.com/metrics/pr-curve/.

A Appendix : DistilBert vs RoBERTa

DistilBERT and RoBERTa are both variants of the BERT (Bidirectional Encoder Representations from Transformers) model.

DistilBERT is smaller and faster:

- It has 40% fewer parameters than BERT

- It is 60% faster than BERT

Whereas, RoBERTa uses the same architecture as BERT but with optimized training [3]

B Appendix : PR and ROC curve

Receiver Operating Characteristic (ROC) Curve is a graph that illustrates the performance of a binary classifier by plotting the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds [4].

The closer the ROC curve is to the top-left corner of the plot, the better the model's performance. ROC curve is the best when you care equally about positive and negative classes, or when the classes are balanced [5].

The **Precision-Recall (PR)** curve is a graph that shows the relationship between precision and recall at various classification thresholds.

The closer the PR curve is to the top-right corner of the plot, the better the model's performance. The PR curve is more useful when dealing with imbalanced datasets [5].

C Appendix: Score comparison

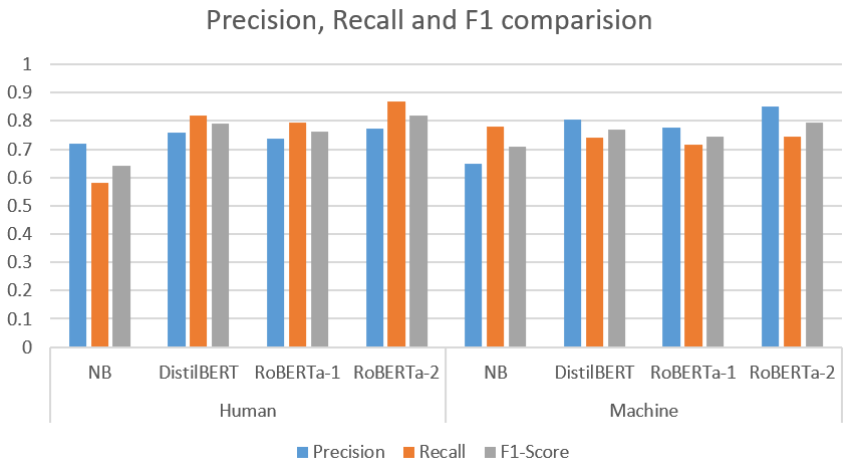

Figure 3: Bar chart comparing Precision Recall and F1-score for all the models tested

Context: Roberta-1 is with stopwords removed. Roberta-2 is without removing stopwords.

D Appendix : Confusion Matrices

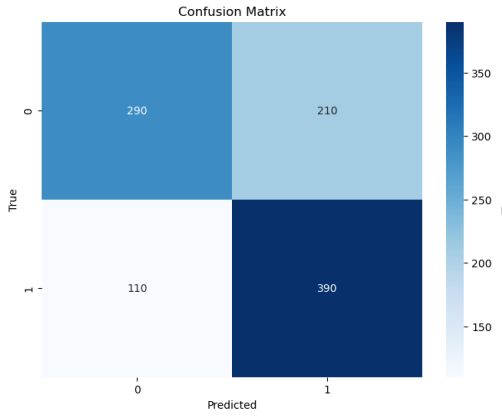

Figure 4: Matrix for NB

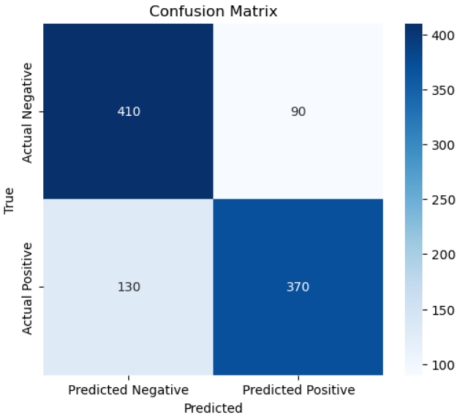

Figure 5: Matrix for DistilBERT

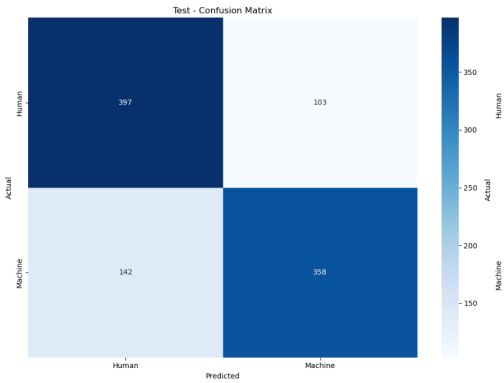

Figure 6: Matrix for RoBERTa-1

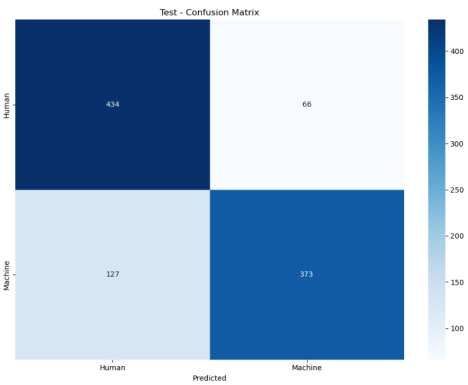

Figure 7: Matrix for RoBERTa-2