Team ID: 5
Team Name: BERT Brigade

Binary Machine-Generated Text Detection

UNIVERSITÉ Concordia UNIVERSITY

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.

Goal of the Project

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

Methodology

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

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

For Naive Bayes Classifier

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

For 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

For RoBERTa architecture [1]

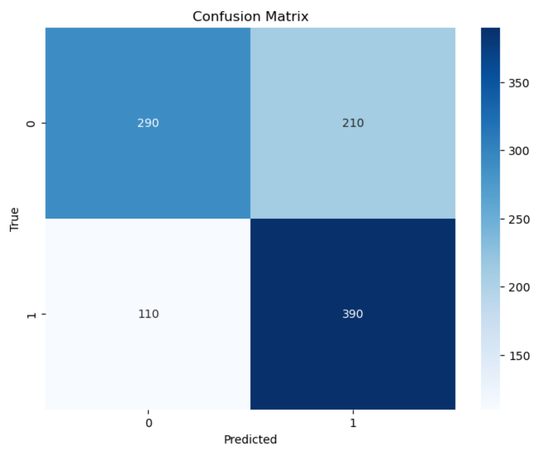
- RoBERTa base model with 125M parameters and sequence classification head
- Max sequence length: 512 tokens
- Adam optimizer with 0.00005 learning rate
- Train/val/test split: 80/10/10 with batch size : 16
- 10 epochs with early stopping

Results

Results for Naive Bayes:

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

Table 1. Performance metrics for NB



Result for **most significant words**: well-structured, commendable, basically, davidson, i've, interplay, daunting, cmv, well-presented, middlewich, avenues, obviously, versatility, roberts, sorry, insightful, ccp, terrible, sallekhana, innovative, insights, tl;dr, biodiversity, i'm, comprehensive, moth, hamsters, it'll, definitely, guys, said

Results for RoBERTa classifier:

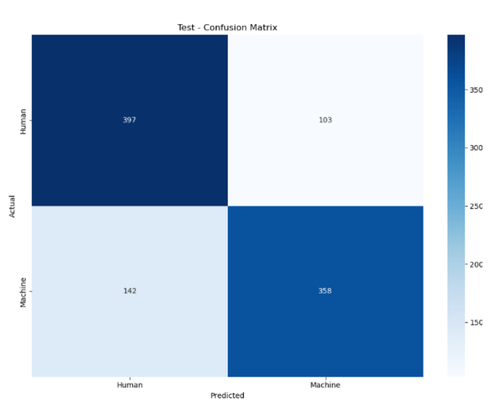
- Training Accuracy: 0.9802
- Test set results :

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

Table 2. Performance metrics for RoBERTa

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

Table 3. Performance metrics for RoBERTa without removing stopwords



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

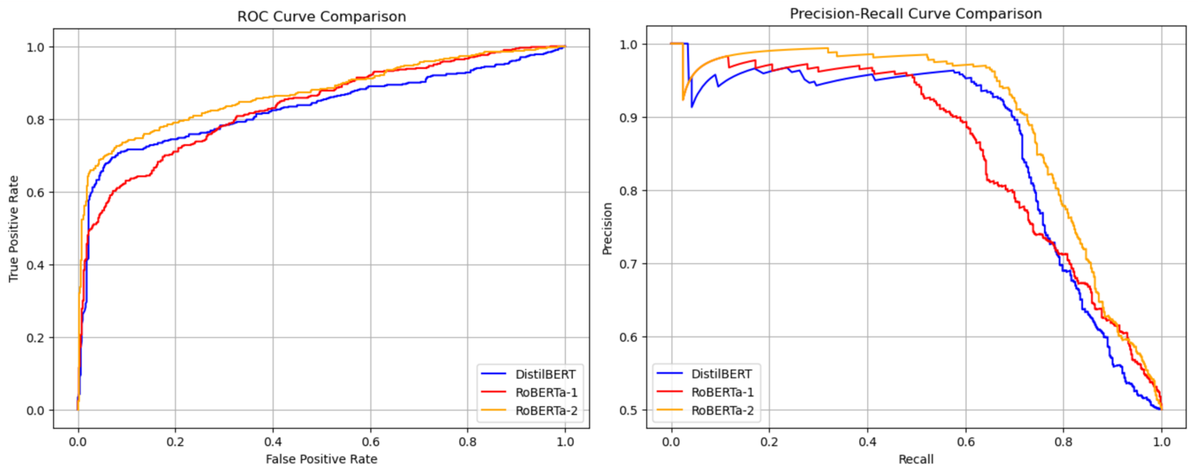


Figure 1. ROC curve Figure 2. PR curve

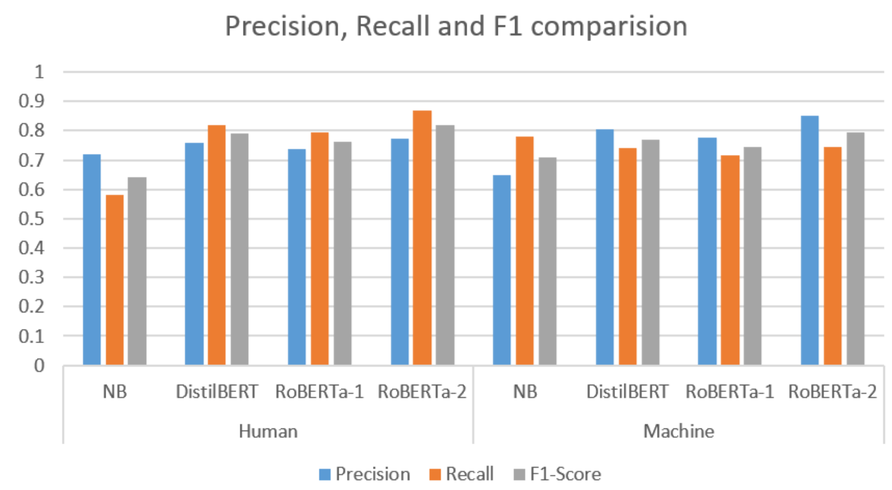


Figure 3. Precision, Recall and F1-score comparison for all the models

Analysis

The Naive Bayes classifier has achieved 68% accuracy, which is not high and there is room for improvement. The recall for machine-generated text is higher than for human, which means it is comparatively better at correctly identifying machine-generated content from the data available.

- The most significant words identified by NB offers 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 and indication of the same.

For RoBERTa classification: It has training accuracy of 98% and test accuracy of 75.5% which was higher than the above model. For Roberta without removing stopwords, with similar train accuracy, the test accuracy increased to 80.7%. Whereas, for DistilBERT, it was 78%.

- 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. (opposite to NB)
- 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.
- AUC for ROC and PR curve was 0.83 and 0.86 for RoBERTa and DistilBERT but in RoBERTa without stopword removal, it was 0.87 and 0.90 respectively.

Limitations and Future Work

- Limitations:
  In NB (Limited to word-level features and no word order/context used). In RoBERTa (Maximum input length restriction, high computational requirements for the whole dataset)
- Future work:
  In NB (Implementing N-gram).In RoBERTa (Finding resources to run the whole dataset, adding multi-language support)

References

[1] D. Munoz, *Classification with RoBERTa and TPUs*, https://www.kaggle.com/code/dimasmunoz/text-classification-with-roberta-and-tpus, Accessed: 2024-11-28, 2020.