Article

Not peer-reviewed version

AI vs. Human: Decoding Text Authenticity with Transformers

Daniela Gifu * and Covaci Silviu-Vasile

Posted Date: 25 July 2024

doi: 10.20944/preprints202407.2014.v1

Keywords: large language models; natural language processing; content creation; text authenticity

Article

AI vs. Human: Decoding Text Authenticity with Transformers

Daniela Gifu [1] and Silviu-Vasile Covaci [2]

[1] Institute of Computer Science, Romanian Academy—Iași Branch, Codrescu 2, 700481, Romania; daniela.gifu@iit.academiaromana-is.ro

[2] George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Târgu Mureș, Gheorghe Marinescu 38, 540142, Romania; covaci.silviu-vasile.23@stud.umfst.ro

* Correspondence: drdanielagifu2020@gmail.com; Tel.: +40-742050673

Abstract: In an era where the proliferation of large language models blurs the lines between human and machine-generated content, discerning text authenticity is paramount. This study investigates transformer-based language models—BERT, RoBERTa, and DistilBERT—in distinguishing human-written from machine-generated text. By leveraging a comprehensive corpus, including human-written text from sources such as Wikipedia, WikiHow, various news articles in different languages, and texts generated by OpenAI's GPT-2, we conduct rigorous comparative experiments. Our findings highlight the superior effectiveness of ensemble learning models over single classifiers in this critical task. This research underscores the versatility and efficacy of transformer-based methodologies for a wide range of natural language processing applications, significantly advancing text authenticity detection systems. The results demonstrate competitive performance, with the transformer-based method achieving an F-score score of 0.83 with RoBERTa-large (monolingual) and 0.70 with DistilBERT-base-uncased (multilingual).

Keywords: large language models; natural language processing; content creation; text authenticity

1. Introduction

The proliferation of large language models (LLMs), notably those developed by OpenAI, has blurred the lines between human and machine-generated content, raising significant concerns regarding text authenticity [1,2]. In an era where misinformation dissemination is a pressing issue in every domain [3–5], distinguishing between human-written and machine-generated text is paramount to mitigate risks associated with deceptive content [6,7]. While previous efforts have focused on identifying text generated by specific LLMs or domain-specific models (e.g., ChatGPT), our study aims to tackle the broader task of distinguishing human-written from machine-generated text [8].

Transformer models, such as Bidirectional Encoder Representations from Transformers (BERT) [9], have emerged as powerful tools in Natural Language Processing (NLP) [10,11], demonstrating remarkable capabilities in various tasks, including Text Generation (TG) [12]. Due to its ability to learn contextual representations of words and phrases, Generative Pre-trained Transformer 3 (GPT-3) has demonstrated significant performance across numerous NLP tasks [13]. This model uses the self-attention mechanism, which allows it to assign different weights to each word in the context of a sentence, capturing complex relationships between words and their meanings.

Additionally, transformer models have revolutionized other fields of Artificial Intelligence (AI) and Machine Learning (ML), such as time series analysis, by incorporating self-attention mechanism into specific data [14]. However, the increasing fluency of these models raises questions about the ability to discern effectively between human and machine-generated text [15]. As transformer models continue to advance, they have become the standard for building large-scale self-supervised learning systems [9,13].

 doi:10.20944/preprints202407.2014.v1

We propose an approach centered on transformer models, combined with Bidirectional Long-Short Term Memory (BiLSTM), to predict the source of text and offer a solution to the challenge of discerning text authenticity. This research not only contributes to advancing text authenticity detection systems, but also underscores the versatility and efficacy of transformer-based methodologies in NLP applications.

The main research question addressed in this paper is: How efficient are transformers in building a classifier that can accurately detect human-written text from machine-generated text? We aim to provide insights into this question through our experimental analyses and methodology evaluation.

The structure of this paper is as follows: Section 2 discusses the role of transformers in addressing the problem of text generation, particularly in distinguishing between human-written and machine-generated text. Section 3 outlines the dataset and the method based on transformers. Section 4 delves into the usability and efficiency of the proposed method through a series of tests, followed by concluding remarks in the last section.

Current Survey Mission

This paper examines existing techniques for classifying texts as either human-written or machine-generated, explores their limitations, and proposes several models that leverage contextual understanding to enhance the accuracy and reliability of classification systems.

The main contributions of our research are as follows:

- **Development of Resources**: We contribute to the development of resources for less-resourced languages such as English, Romanian, and Hungarian.
- **Extensive Datasets**: We developed extensive datasets for English, Romanian, and Hungarian, containing both human-authored and machine-generated texts, using several large language models (LLMs).
- **Implementation of Classification Models**: We implemented classification models based on different architectures, including Transformer-based models (such as BERT-base, RoBERTa-base, RoBERTa-large, DistilBERT-base-uncased, XLM-RoBERTa-base, BERT-base-multilingual-cased, and DistilBERT-base-multilingual-cased) and classic machine learning (ML) models, designed to automatically classify texts in several languages.

We release the datasets as open-source resources (available at Papers with Code, AI Crowd, Mendeley Data, and GitHub, accessed on 22 July 2024), along with the codebase (available at GitHub, accessed on 22 July 2024).

2. Transformers for Human and Machine-Generated

AI technology is increasingly capable of generating text that is difficult to distinguish from human-written content. Initially, traditional machine learning techniques such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), hybrid methods (CNN-LSTM), Support Vector Machines (SVM), and Decision Trees (DTs) were employed. Recently, however, generative models based on transformers, such as BERT [9], RoBERTa [16], DistilBERT [17], have become prevalent.

Transformers are pretrained models designed to accomplish specific tasks. They can be used as they are or can be fine-tuned with additional custom layers to meet specific application needs. BERT and its successors are utilized in various NLP generative tasks, including Machine Translation (MT) [12,18], Question Answering (QA) [19,20], Text Summarization (TS) [21,22], and Text Classification (TC) [23,24]. Transformer models [25,26] are particularly well-suited for text generation tasks requiring contextually rich and coherent text, outperforming traditional neural networks such as CNN, Bi-LSTM, and hybrid CNN-BiLSTM models [27].

Subsequent studies, employed in this study, and advancements have built on the BERT framework, described below.

2.1. BERT

Bidirectional Encoder Representations from Transformers (BERT), introduced by Google AI in 2018 [9], was a groundbreaking advancement in NLP. Unlike unidirectional models, BERT employs a bidirectional approach, considering the context from both the left and right sides of the sequence it aims to understand. Its fine-tuning flexibility allows developers to create high-performance systems by adding an additional output layer on top of the pretrained model.

BERT's architecture is based on the original Transformer, featuring multiple layers, larger feed-forward networks, and more attention heads. It was pretrained on a large corpus using Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) tasks (Figure 1). The input embeddings in BERT are the sum of token embeddings, segmentation embeddings, and position embeddings.

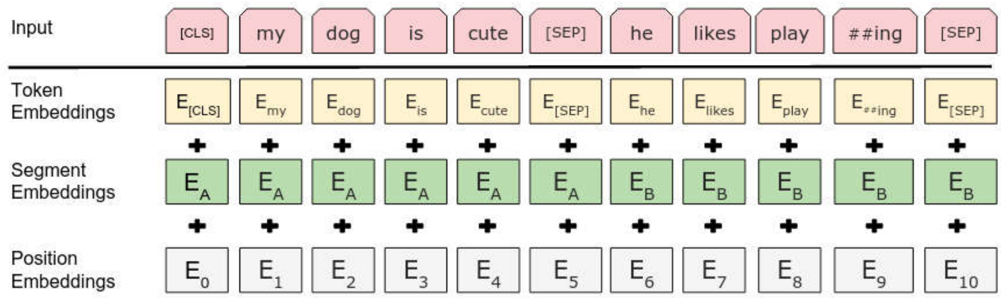

Figure 1. BERT input representation.

The Masked Language Model (MLM) becomes useful for predicting the original vocabulary ID of a masked word based solely on its context, as it can randomly mask some input tokens during training. The next sentence prediction (NSP) task involves training the model to learn the relationships between sentences in a document or text corpus. During training, BERT receives pairs of sentences as input and learns to predict whether they are consecutive in the document.

Subsequent studies and advancements that build on the BERT framework are described below.

2.2. RoBERTa

RoBERTa [21] is a reimplementation of BERT with several modifications to key hyperparameters and embedding techniques [28]. One of the significant improvements in RoBERTa is the use of dynamic masking, which enhances the robustness of semantic text representation. RoBERTa's training involves larger batch sizes and more steps compared to BERT, contributing to its enhanced performance.

A key feature of RoBERTa is the prevention of input sentences from crossing document boundaries, which is crucial for improving contextual understanding. While BERT uses a batch size of 256 sequences and trains for 1 million steps, RoBERTa uses a batch size of 2,000 and trains for 500,000 steps, better adapting to the dynamic masking concept [24,29].

Another important feature of RoBERTa is the use of Byte Pair Encoding (BPE) for tokenization. BPE tokenizes text into subwords extracted from the training corpus based on advanced statistical analysis. By using bytes instead of Unicode characters as the base for subword units, RoBERTa achieves a universal encoding scheme that is more efficient [30,31].

In addition to these features, RoBERTa was trained for a longer period using a larger dataset, including data from sources like Common Crawl, WebText, and other large-scale corpora, which further improves its performance and applicability across various NLP tasks [16,32]. This extensive training on a more diverse dataset contributes to its superior performance in a range of NLP applications, such as text classification, sentiment analysis, and question answering.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.2014.v1

2.3. DistilBERT

DistilBERT, developed by Hugging Face, is a smaller and faster version of BERT that retains 97% of BERT's language understanding capabilities while being 60% faster [17]. This efficiency is achieved by reducing the number of transformer layers and removing certain components, such as token-type embeddings and the pooler used for next sentence prediction.

The architecture of DistilBERT is similar to BERT's, but with notable differences. Specifically, DistilBERT contains only 6 transformer layers, compared to the 12 layers in the BERT base model. Additionally, DistilBERT omits token-type embeddings and the pooler, which are present in BERT. To enhance the quality of sequence representation, DistilBERT introduces a distillation token at the input [17].

DistilBERT's optimized performance makes it an appealing choice for a wide range of applications, offering impressive results with reduced computational requirements [17,33]. This makes it particularly useful in scenarios where computational resources are limited but high performance is still needed.

3. Materials and Methods

In our survey, the primary focus was on classifying texts as either human-written or machine-generated using an existing corpus. To differentiate between human and machine-generated text, a few AI Text Classifiers have been developed.

3.1. Dataset

The corpus for this study consists of multiple datasets with comparable text lengths, including both machine-generated and human-written content. Experiments were conducted iteratively across all datasets to provide a comprehensive overview.

To ensure that the detector generalizes well across various domains and writing styles, the human dataset includes texts from diverse domains, specifically:

- **M4 dataset** (https://paperswithcode.com/datasets, accessed on 22 July 2024): Contains human-written text from sources such as Wikipedia, Wiki-How [34], Reddit (ELI5), arXiv, and PeerRead [35] for Chinese, as well as news articles for Urdu, RuATD [36] for Russian, and Indonesian news articles. Machine-generated text is sourced from multilingual LLMs such as ChatGPT, textdavinci-003, LLaMa [37], FlanT5 [38], Cohere, Dolly-v2, and BLOOMz [39];
- **AI Crowd FakeNews Dataset** (https://www.aicrowd.com/challenges/kiit-ai-mini-blitz/problems/fake-news-detection, accessed on 22 July 2024): Contains texts from various news articles and texts generated by OpenAI's GPT-2. The dataset was published by AI Crowd as part of the KIIT AI (mini)Blitz Challenge;
- **Indonesian Hoax News Detection Dataset** (INDONESIAN HOAX NEWS DETECTION DATASET—Mendeley Data, accessed on 22 July) [40]: Contains valid and hoax news articles in Indonesian. It has a simple structure, with CSV files consisting of 2 columns: text and label;
- **TURNBACKHOAX Dataset** (https://github.com/jibranfawaid/turnbackhoax-dataset/tree/main?tab=readme-ov-file#turnbackhoax-dataset, accessed on 22 July 2024): Contains valid and hoax news articles in Indonesian. It has a simple structure, with a CSV file consisting of 3 columns: label, headline, body.

Tables 1 and 2 present the dataset collections.

Table 1. Data Sources in M4 Dataset.

Source	Lang. [1]	Only Human	Source-generated data						
			Human	Davinci003	ChatGPT	Cohere	Dolly-v2	BLOOM	Total
Wikipedia	EN	6.458.670	3.000	3.000	2.995	2.336	2.702	3000	17033
Reddit ELIS	EN	558.669	3.000	3.000	3.000	3.000	3.000	3.000	18.000
WikiHow	EN	31.102	3.000	3.000	3.000	3.000	3.000	3.000	18.000
PeerRead	EN	5.798	5.798	2.344	2.344	2.344	2.344	2.344	17.518

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.2014.v1

arXiv abstract	EN	2.219.423	3.000	3.000	3.000	3.000	3.000	3.000	18.000
Baike/Web OA	ZH	113.313	3.000	3.000	3.000	-	-	-	9.000
RuATD	RU	75.291	3.000	3.000	3.000	-	-	-	9.000
Urdu-news	UR	107.881	3.000	-	3.000	-	-	-	9.000
id_newspapers_2018	ID	499.164	3.000	-	3.000	-	-	-	6.000
Arabic-Wikipedia	AR	1.209.042	3.000	-	3.000	-	-	-	6.000
True & Fake News	BG	94.000	3.000	3.000	3.000	-	-	-	9.000
Total			35.798	23.344	32.339	13.680	14.046	14.344	133.551

[1] Here are the abbreviations provided for the ISO 639-1 language codes: English—EN; Chinese—ZH; Russian—RU; Urdu—UR; Indonesian—ID; Arabic—AR; Bulgarian—BG.

Table 2. Dataset statistics.

Language approach	#Training records	#Testing records
M4—Monolingual	119.757	5.000
AICrowd—Monolingual	232.003	38.666
M4—Multilingual	172.417	4.000
Indonesian Hoax News Detection—Multilingual	600	250
TURNBACKHOAX Dataset—Multilingual	800	316

The M4 input data is organized as JavaScript Object Notation (JSON) records in files with the extension JSON Lines (JSONL).

The structure of each record is very straightforward and intuitive.

Figure 2 and 3 present the structure of the datasets (monolingual/multilingual) for training and development testing.

text	label	model	source	id
This lightening spray calls for dried chamomi...	0	human	wikihow	69365
We consider the problem of mass transport cl...	0	human	arxiv	113114
\nThe paper "Learning to Skim Text" presents a...	1	cohere	peerread	47161
Tesla's solar shingles and Powerwall are innov...	1	chatGPT	reddit	38209
Have you accidentally dropped your precious iP...	1	chatGPT	wikihow	1368
...	...	...	...	...
Josephine "Joyce" Luther Kennard (born May 6, ...	0	human	wikipedia	76820
The fractional Brownian motion with index $\...	0	human	arxiv	110268
Strictly subadditive, subadditive and weakly...	0	human	arxiv	103694
How to Get Into UPenn: Excelling Academically ...	1	chatGPT	wikihow	860
Well, from what little information I have gath...	1	chatGPT	reddit	15795

a)

text	label	model	source	id
Forza Motorsport is a popular racing game that...	1	chatGPT	wikihow	0
Buying Virtual Console games for your Nintendo...	1	chatGPT	wikihow	1
Windows NT 4.0 was a popular operating system ...	1	chatGPT	wikihow	2
How to Make Perfume\n\nPerfume is a great way ...	1	chatGPT	wikihow	3
How to Convert Song Lyrics to a Song'\n\nConve...	1	chatGPT	wikihow	4
...	...	...	...	...
During the Cold War, the United States was po...	1	cohere	reddit	172412
The "continuity thesis" is the idea that ther...	1	cohere	reddit	172413
In the early Middle Ages, the pagan Norse wer...	1	cohere	reddit	172414
There are many similarities between the langu...	1	cohere	reddit	172415
News of Christopher Columbus' voyage to the N...	1	cohere	reddit	172416

b)

Figure 2. Mono- and multilingual Training Dataset.

text	label	model	source	id
Giving gifts should always be enjoyable. Howe...	1	bloomz	wikihow	0
Yveltal (Japanese: ユベルタル) is one of the main a...	1	bloomz	wikihow	1
If you'd rather not annoy others by being rude...	1	bloomz	wikihow	2
If you're interested in visiting gravesite(s) ...	1	bloomz	wikihow	3
The following are some tips for becoming succe...	1	bloomz	wikihow	4
...	...	...	...	...
The paper deals with an interesting applicatio...	0	human	peerread	4995
This manuscript tries to tackle neural network...	0	human	peerread	4996
The paper introduced a regularization scheme t...	0	human	peerread	4997
Inspired by the analysis on the effect of the ...	0	human	peerread	4998
\n- You definitely need to report misclassific...	0	human	peerread	4999

a)

text	label	model	source	id
Giving gifts should always be enjoyable. Howe...	1	bloomz	wikihow	0
Yveltal (Japanese: ユベルタル) is one of the main a...	1	bloomz	wikihow	1
If you'd rather not annoy others by being rude...	1	bloomz	wikihow	2
If you're interested in visiting gravesite(s) ...	1	bloomz	wikihow	3
The following are some tips for becoming succe...	1	bloomz	wikihow	4
...	...	...	...	...
The paper deals with an interesting applicatio...	0	human	peerread	4995
This manuscript tries to tackle neural network...	0	human	peerread	4996
The paper introduced a regularization scheme t...	0	human	peerread	4997
Inspired by the analysis on the effect of the ...	0	human	peerread	4998
\n- You definitely need to report misclassific...	0	human	peerread	4999

b)

Figure 3. Mono- and multilingual Testing Dataset.

There are mainly three major differences if we compare the datasets used for training the models and the dataset that will be used for final evaluation:

 doi:10.20944/preprints202407.2014.v1

- The task formulation is different;
- Human text was upsampled to balance the data;
- New and surprising domains, generators, and languages will appear in the test sets. Real test sets will not include information about generators, domains, and languages.

Nevertheless, the test dataset includes BLOOMZ[1] outputs (for monolingual language) that are not included in the training set. Moreover, the model is prepared for real-world application scenarios.

3.2. System Overview

The architecture (Figure 4) is based on BERT-based transformers (BERT, RoBERTa, DistilBERT) using the HuggingFace library.

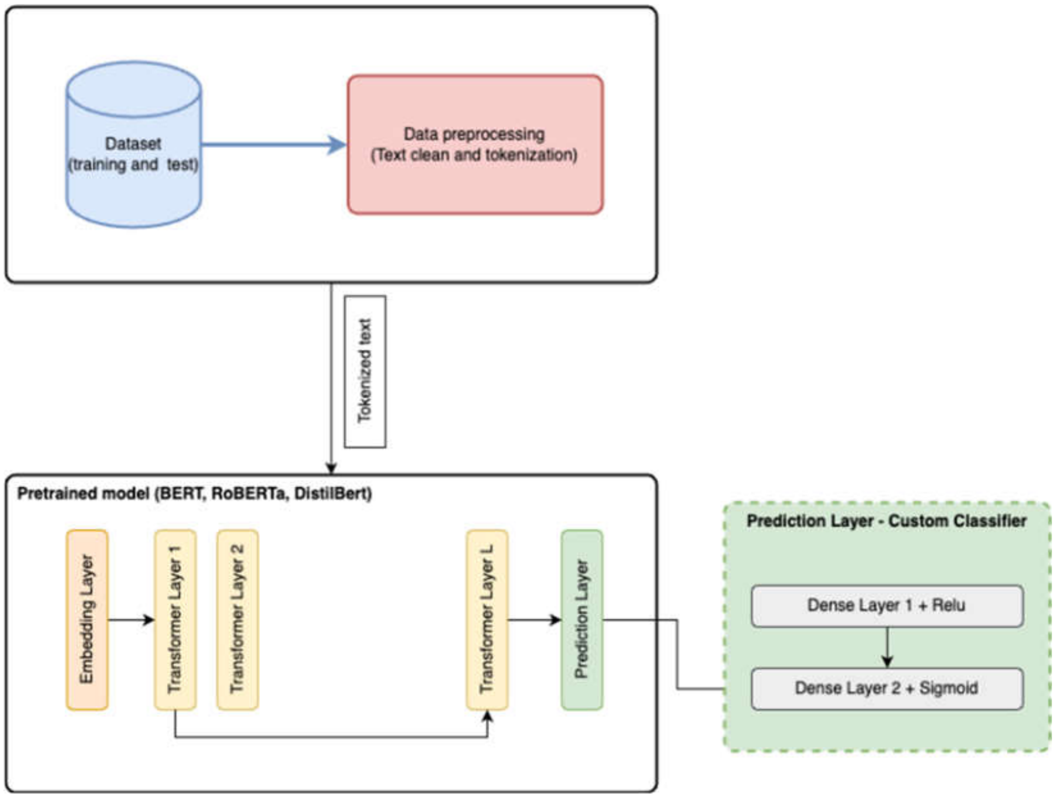

Figure 4. Architecture.

The model was pretrained [9,41–43] on large generic datasets and fine-tuned for specific tasks like text classification, named entity recognition, and sentiment analysis [44].

As a baseline, we chose the RoBERTa-base pretrained model, fine-tuned with a sequence classification/regression head on top.

The model was trained and evaluated on the same dataset mentioned before. Table 2 contains the baselines' hyperparameters. Additionally, we used Cross-Entropy loss as the loss function, as we are dealing with a binary classification task. The model is built in such a way that via the Sigmoid function at the end, it should output a probability of 0 (= no AI-generated text) and 1 (= AI-generated text). As an optimizer, AdamW, an improved version of Adaptive Moment Estimation (Adam), is significant in training deep learning (DL) models. The learning rate value was set to 2e-5.

The average results for the baseline monolingual setup across three runs for the RoBERTa-base pretrain-dataset are 0.74, and respectively 0.72 for multilanguage, based on the xlm-roberta-base pretrain-dataset.

---

[1] BLOOMZ, a variant of BLOOM model, supports 46 human languages. Hugging Face reports that the 7 billion-parameter BLOOMZ runs three times faster on the Intel Habana Gaudi2 compared to the A100-80G.

 doi:10.20944/preprints202407.2014.v1

This model has 2 layers, 768 hidden units, 12 heads, and 125 million parameters.

Table 3. Baseline model: Hyperparameter Optimization.

Hyperparameter	Values
Learning rate	2e-5
Batch Size	16
Epochs	3
Weight decay	0,01

Fine-tuned models. The main objective of the experiment was to obtain a fine-tuned model that could outperform the baseline model.

Various variations of models/approaches were trained, and ultimately, we decided to combine Hugging Face's Transformers library with PyTorch and Scikit-Learn libraries.

Additionally, a custom classifier class was applied on top of pretrained models to identify the correct label for our texts. The classifier consists of 2 dense layers: the first layer with 768 neurons (for "base" versions) / 1024 neurons (for "large" versions), and the second layer with 32 neurons (for "base" versions) / 8 neurons (for "large" versions).

Since we have a binary classification task, we use one neuron for the output layer and the sigmoid function (which returns values between 0 and 1) as the activation function for our neural network. The number of neurons in the first layer actually represents the number of neurons in the output layer of the pretrained models (768 for "base" versions, and 1024 for large models).

For classifications task based on neural networks, we used Activation Function (Rectified Linear Unit—ReLU Sigmoid for hidden layers) loss function, and AdamW, as optimizer. The learning rate value was set to 1e-5.

Table 4. Fine-tuned model: Hyperparameter Optimization.

Hyperparameter	Values
Learning rate	1e-5
Batch Size	8
Epochs	5

As pretrained models, we tested BERT-base, RoBERTa-base, RoBERTa-large, as well as DistilBERT-base-uncased for the monolingual setup, and XLM-RoBERTa-base, BERT-base-multilingual-cased, DistilBERT-base-multilingual-cased for the multilingual setup, models provided by the Transformers library.

For monolingual experiments, as expected, RoBERTa-large provided the best results with an accuracy of 0.83, but the training process took approximately 10 hours.

Using the DistilBERT-base-multilingual-cased model for monolingual experiments also yielded promising results, with less power consumption, within approximately 3 hours. Thus, it can be considered a very good alternative to RoBERTa or BERT. It is important to note that we need to use different pretrained models for each subtask (monolingual and multilingual), as there are separate models optimized for multilingual tasks.

In order to reduce training time, GPUs were used for model training and inference. All experiments were conducted on a Mac Studio machine, as detailed in the results section.

3.3. Experiments

The experimental setup involved preprocessing the dataset, feature engineering, and modeling using different transformer architectures.

- **Preprocessing**

We created a custom PyTorch DataSet class for loading data and performing basic preprocessing steps:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.2014.v1

(1) Text Cleanup: Removing HTML tags, special characters such as # and @, punctuation, and multiple spaces.

(2) Basic preprocessing: Tokenization

- **Feature Engineering**

For this survey, Bag of Words (BoW) and Word to Vectors (word2vec) models were used.

- **Modelling**

For pretrained, transformers like BERT-base, RoBERTa-base, RoBERTa-large, DistilBERT-base-uncased/XLM-RoBERTa-base, BERT-base-multilingual-cased, DistilBERT-base-multilingual-cased) combined with a custom classifier consisting of 3 layers with varying numbers of neurons responded promising.

To adjust the learning rate for different parameters, Adaptive Moment Estimation (ADAM) optimizer was chosen.

- **Prediction**

Since this model returns probabilities between 0 and 1, we use a 50% threshold for target classification. Predictions are stored using the given test dataset, which includes test IDs and sample targets, and a prediction file is generated based on the model's predictions. For evaluation of both subtasks, we employ sklearn.metrics, calculating Accuracy (Acc), Precision (P), Recall (R), and F-score (also known as the F1 score or F-measure).

For the multilingual subtask, we use different pretrained models, selecting custom pretrained models optimized for multilanguage tasks.

4. Results

The experiments were conducted on a Mac Studio machine. In terms of performance, the Mac Studio is equipped with Apple's M1 Max Chip, featuring a 10-core CPU, an integrated 24-core GPU, and a maximum memory bandwidth of 400GB/s, according to the official specifications.

The number of epochs was set to 3 for all experiments conducted (refer to Tables 5 and 6).

Table 5. Performance metrics for monolingual subtask.

Model	Acc (%)	P (%)	R (%)	F-score (%)	Model runtime (min.)
Baseline	74				
RoBERTa-large	83	84	83	**83**	607
RoBERTa-base	81	83	81	81	166
BERT-base	71	74	71	70	162
DistilBERT-base-uncased	68	73	68	66	77

Table 6. Performance metrics for multilingual subtask.

Model	Acc (%)	P (%)	R (%)	F-score (%)	Model runtime (min.)
Baseline	69				
XML-RoBERTa-base	68	70	68	68	522
BERT-base-cased	63	68	64	61	415
DistilBERT-base-uncased	70	71	71	**70**	203

The results table presents a comparison of metrics for the tested models, including accuracy, precision, recall, and F-score, along with the time required for training and evaluation of each model.

5. Discussion

The experiments conducted in this study focused on evaluating the performance of various pretrained models from the BERT family, which were fine-tuned with a custom classifier to detect AI-generated text.

- For **monolingual models**, the results revealed notable insights. Specifically, the RoBERTa-large model, in conjunction with a custom classification layer, demonstrated the highest performance levels among all tested models. This performance exceeded baseline results observed in competitions such as SemEval-2024 Task 8. Despite its slightly lower accuracy, DistilBERT showcased efficient resource utilization. Additionally, the RoBERTa-base model exhibited performance closely comparable to that of RoBERTa-large while boasting significantly faster training times. Particularly noteworthy was the performance of a hybrid model combining a pre-trained model with DistilBERT alongside a custom classifier. Despite a marginally lower accuracy of 0.68, this model exhibited com-mendable precision at 0.73, underscoring its resource efficiency and satisfactory performance.
- In the case of **multilingual models**, the results indicated lower accuracy levels and longer training times due to the larger dataset. Interestingly, the DistilBERT model surpassed its teacher, BERT, in this subtask, achieving an accuracy of 0.70 compared to the baseline accuracy of 0.68. This outcome suggests the necessity for distinct approaches when addressing monolingual and multilingual tasks.

Our findings underscore the inherent challenges in distinguishing between hu-man-written and machine-generated text. While transformer models exhibit promise in this domain, further research is imperative to enhance model robustness and ad-dress the limitations observed in real-world scenarios.

By leveraging alternative approaches and refining feature engineering techniques, future investigations can contribute significantly to the advancement of AI-generated text detection systems.

6. Conclusions

This study provides valuable insights into the effectiveness of transformer models in identifying AI-generated text. Moving forward, research efforts should explore alternative approaches, such as A Lite BERT for Self-supervised Learning of Language Representations (ALBERT) [45], and incorporate advanced feature engineering techniques to improve model performance and robustness. Addressing these aspects can significantly contribute to ongoing efforts to improve AI-generated text detection systems.

This study offers a detailed evaluation of pretrained models BERT, RoBERTa, and DistilBERT for detecting AI-generated texts. Despite achieving good results, distinguishing machine-generated from human-written text remains challenging, especially with unseen data during training. Future research should explore other methods like ALBERT [36], more advanced feature engineering, and the combination of machine learning techniques to enhance model robustness. By addressing these limitations and incorporating the suggested improvements, this study can significantly contribute to ongoing efforts to improve AI-generated text detection systems.

Author Contributions: Conceptualization, D.G. and S.-V.C.; methodology, D.G. and S.-V.C.; software, S.-V.C.; validation, S.-V.C.; formal analysis, S.-V.C.; investigation, D.G.; resources, D.G.; data curation, D.G.; writing—original draft preparation, D.G.; writing—review and editing, D.G.; visualization, S.-V.C.; supervision, D.G.; project administration, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: As open-sources, the following datasets are released (https://paperswithcode.com/datasets;https://www.aicrowd.com/challenges/kiit-ai-mini-blitz/problems/fake-news-detection; INDONESIAN HOAX NEWS DETECTION DATASET—Mendeley Data; https://github.com/jibranfawaid/turnbackhoax-dataset/tree/main?tab=readme-ov-file#turnbackhoax-dataset, accessed on 22 July 2024), and the codebase (SemEval2024Task8/subtaskA/detector.py at main · SilviuCovaci/SemEval2024Task8 GitHub, accessed on 22 July 2024). Additional information is available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 July 2024 doi:10.20944/preprints202407.2014.v1

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
ADAM	Adaptive Moment Estimation
AI	Artificial Intelligence
BoW	Bag of Words
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long-Short Term Memory
BLOOM	Big Science Large Open-science Open-access Multilingual Language Model
BPE	Byte Pair Encoding
CNN	Convolutional Neural Networks
DTs	Decision Trees
DL	Deep Learning
DistilBERT	Distilled BERT
GPT	Generative Pre-trained Transformer
JSON	JavaScript Object Notation
JSONL	JSON Lines
LLMs	Large Language Models
LSTM	Long Short-Term Memory
MLM	Masked Language Model
ML	Machine Learning
MT	Machine Translation
NLP	Natural Language Processing
NSP	Next Sentence Prediction
P	Precision
QA	Question Answering
R	Recall
ReLU	Rectified Linear Unit
RoBERTa	Robustly Optimized BERT
SVMs	Support Vector Machines
TC	Text Classification
TG	Text Generation
TS	Text Summarization
word2vec	Word to Vectors

References

1. Weidinger, L.; Mellor, J.F.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; et al. Ethical and social risks of harm from Language Models. *arXiv* 2021, arXiv:2112.04359.
2. Solaiman, I.; Brundage, M.; Clark, J.; Askell, A.; Herbert-Voss, A.; Wu, J.; Radford, A.; Krueger, G.; Kim, J.W.; Kreps, S.; et al. Release Strategies and the Social Impacts of Language Models. *arXiv* 2019, arXiv:1908.09203.
3. Gîfu, D. An Intelligent System for Detecting Fake News. *Procedia Computer Science*, open access journal, edited by Yong Shi, ELSEVIER, Vol. 221, 2023, pp. 1058-1065, DOI: 10.1016/j.procs.2023.08.088.
4. Ermurachi, V.; Gîfu, D. UAIC1860 at SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, (SemEval-2020), Association for Computational Linguistics, Barcelona, Spain, 2020, pp. 1835-1840.
5. Gîfu, D. Utilization of Technologies for Linguistic Processing in an Electoral Context: Method LIWC-2007. *Proceedings of the Communication, Context, Interdisciplinarity Congress*, 19-20 Nov. 2010, Vol. 1, "Petru Maior" University Publishing House, Târgu-Mureș, 2010, pp. 87-98.
6. Ma, Y.; Liu, J.; Yi, F.; Cheng, Q.; Huang, Y.; Lu, W.; Liu, X. AI vs. Human-Differentiation Analysis of Scientific Content Generation. *arXiv* 2023, arXiv:2301.10416.
7. Ouatu, B.; Gîfu, D. Chatbot, the Future of Learning? *Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education*, Springer, 2020, pp. 263-268.

8. Wang, Y.; Mansurov, J.; Ivanov, P.; Su, J.; Shelmanov, A.; Tsvigun, A.; Whitehouse, C.; Afzal, O. M.; Mahmoud, T.; Sasaki, T.; Arnold, T.; Aji, A. F.; Habash, N.; Gurevych, I.; Nakov, P. M4: Multi-Generator, Multi-Domain, and Multi-Lingual Black-Box Machine-Generated Text Detection, May 24. *arXiv* 2023, arxiv: 2305.14902.
9. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Association for Computational Linguistics*, 2019.
10. Manoleasa, T; Sandu, I.; Gîfu, D.; Trandabăț, D. FII UAIC at SemEval-2022 Task 6: iSarcasmEval—Intended Sarcasm Detection in English and Arabic. *Proceedings of the 16th International Workshop on Semantic Evaluation*, (SemEval-2022), Association for Computational Linguistics, Seattle, Washington, US, 2022, pp. 970-977.
11. Alexa, L.; Lorenț, A.; Gîfu, D.; Trandabăț, D. The Dabblers at SemEval-2018 Task 2: Multilingual Emoji Prediction. *Proceedings of the 12th International Workshop on Semantic Evaluation*, (SemEval-2018), Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2018), New Orleans, Louisiana, United States, 2018, pp. 405-409.
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. *Proceedings of Advances in Neural Information Processing Systems*, Vol. 30, 2017.
13. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 2020.
14. Niu, P; Zhou, T.; Wang, X.; Sun, L.; Jin, R. Attention as Robust Representation for Time Series Forecasting. *arXiv* 2024, arXiv: 2402.05370v1.
15. Clark, E.; August, T.; Serrano, S.; Haduong, N.; Gururangan, S.; Smith, N.A. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, 1–6 August 2021.
16. Liu Y.; Lapata, M. Text Summarization with Pretrained Encoders. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China. Association for Computational Linguistics, 2019, pp. 3730–3740.
17. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *arXiv* 2020, arXiv:1910.01108.
18. Tang, G.; Sennrich, R.; Nivre, J. An Analysis of Attention Mechanisms: The Case of Word Sense Disambiguation in Neural Machine Translation. *Proceedings of the Third Conference on Machine Translation: Research Papers*, Belgium, Brussels. Association for Computational Linguistics, 2018, pp. 26–35.
19. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text, 2016, pp. 2383-2392. DOI: 10.18653/v1/D16-1264.
20. Yang, Y.; Cer, D.; Ahmad, A.; Guo, M.; Law, J.; Constant, N.; Abrego, G. H.; Yuan, S.; Tar, C.; Sung, Y.-H.; et al. Multilingual universal sentence encoder for semantic retrieval. *arXiv* 2019, arXiv:1907.04307
21. Liu, Y., Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov: RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* 2019, arXiv:1907.11692
22. Zhang, H.; Cai, J.; Xu, J.; Wang, J. Pretraining-Based Natural Language Generation for Text Summarization. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Hong Kong, China. Association for Computational Linguistics, 2019, pp. 789–797.
23. Sun, C., Qiu, X., Xu, Y., & Huang, X. How to Fine-Tune BERT for Text Classification? *China National Conference on Chinese Computational Linguistics*, Springer, Cham, 2019, pp. 194-206.
24. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A. J.; Hovy, E. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020, pp. 5754-5764. https://arxiv.org/abs/1906.08237
25. Awalina, A.; Krisnabayu, R. Y.; Yudistira, N.; Fawaid, J. Indonesia's Fake News Detection using Transformer Network. *6th International Conference on Sustainable Information Engineering and Technology*, Malang Indonesia: ACM, 2021, pp. 247–251. DOI: 10.1145/3479645.3479666.
26. Azizah, A. F. N.; Cahyono, H. D.; Sihwi, S. W.; Widiarto, W. Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection. *arXiv* 2023, arXiv:2308.04950.
27. Zecong, W.; Jiaxi, C.; Chen, C.; Chenhao, Y. Implementing BERT and Fine-Tuned RobertA to Detect AI Generated News by ChatGPT. *arXiv* 2023, arxiv:2306.07401.
28. Zhang, H.; Shafig, M. O. Survey of Transformers and Towards Ensemble Learning Using Transformers for Natural Language Processing. *Journal of Big Data*, 11, 25, 2024, DOI:10.1186/s40537-023-00842-0.

 doi:10.20944/preprints202407.2014.v1

29. Clark, K.; Luong, M. T.; Le, Q. V.; Tannenbaum, K. ELECTRA: Pre-Training Text Encoders as Discriminators Rather than Generators. *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, 2020. https://arxiv.org/abs/2003.10555
30. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, 2016, pp. 1715-1725. DOI: 10.18653/v1/P16-1162
31. Gage, P. A New Algorithm for Data Compression. *C Users Journal*, 12(2), 1994, pp. 23-38. DOI: 10.5555/150181
32. Raffel, C.; Shinn, C.; Roberts, A. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140), 2020, pp. 1-67. https://arxiv.org/abs/1910.10683
33. Smith, J.; Lee, K.; Kumar, A. Optimizing Transformer Models for Mobile Devices: A Comparative Study. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*, 2021, pp. 234-245. DOI: 10.48550/arXiv.2105.07893
34. Koupaee M.; Wang, W. Y. Wikihow: A Large-Scale Text Summarization Dataset. *arXiv*2018, arXiv:1810.09305, DOI: 10.48550/arXiv.1810.09305
35. Kang, D.; Ammar, W.; Dalvi, B.; van Zuylen, M.; Kohlmeier, S.; Hovy, E.; Schwartz, R. A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long Papers), New Orleans, Louisiana. Association for Computational Linguistics, 2018, pp. 1647–1661.
36. Shamardina, T.; Mikhailov, V.; Cherniavskii, D.; Fenogenova, A.; Saidov, M.; Valeeva, A.; Shavrina, T.; Smurov, I.; Tutubalina, E.; Artemova, E. Findings of the Ruatd Shared Task 2022 on Artificial Text Detection in Russian. *arXiv*2022, arXiv:2206.01583, DOI:10.48550/arXiv.2206.01583.
37. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; Lample, G. Llama: Open and Efficient Foundation Language Models. *arXiv*2023, arXiv:2302.13971.
38. Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V. Y.; Huang, Y.; Dai, A. M.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; Wei, J. Scaling Instruction-Finetuned Language Models. *arXiv*2022, arXiv:2210.11416.
39. Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T. L.; Bari, M. S.; Shen, S.; Yong, Z. X.; Schoelkopf, H.; Tang, X.; Radev, D.; Aji, A. F.; Almubarak, K.; Albanie, S.; Alyafeai, Z.; Webson, A.; Raff, E.; Raffel, C. Crosslingual Generalization Through Multitask Finetuning. *arXiv*2022, arXiv:2211.01786.
40. Faisal, R.; Inggrid; Y.; Rosa, A. A. Indonesian Hoax News Detection Dataset. *Mendeley Data*, V1, 2018. DOI: 10.17632/p3hfgr5j3m.1
41. Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep Contextualized Word Representations. *arXiv*2018, DOI:10.48550/arXiv.1802.05365.
42. McCann, B.; Bradbury, J.; Xiong, C.; Socher, R. Learned in Translation: Contextualized Word Vectors. *Advances in Neural Information Processing Systems* 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), 2017, pp. 6294–6305. Curran Associates, Inc.
43. Howard J.; Ruder, S. Universal Language Model Fine-Tuning for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers, 2018, pp. 328–339, Melbourne, Australia. Association for Computational Linguistics.
44. Wolf, T.: Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; Rush, A. M. Huggingface's Transformers: State-of-the-Art Natural Language Processing. *arXiv*2019, arXiv:1910.03771.
45. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. *arXiv*2020, arXiv: 1909.11942v6.