International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Detecting the invisible hand: A comparative study of CNN and Distilbert models for distinguishing AI-generated text and human-generated text

Jerin Mahibha C[*1] *,Adithya Manikandan J*[*2]*, Anirudh K*[*3]*, Arjun Krishnan R*[*4]*

[*1]Professor, Department Of CSE, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India.
[*2,3,4]Final Year UG Student, Department Of CSE, Meenakshi Sundararajan Engineering College, Chennai, Tamil Nadu, India.

ABSTRACT:

The usage of emerging technologies by students for academic work is a difficulty for educational institutions, particularly with regard to big language models such as Chat GPT. Large volumes of data from several datasets were used to train the system. Training the CNN model and DistilBert Classifier is the focus of our project. The model's purpose is to differentiate between the essay's human-written and AI-generated content. The research aims to assess the accuracy of each model in identifying text generated by AI. It's critical to respect gifted writers and independent content producers by using AI-generated text detection. Using their classification measures, the performance of several models, including the CNN model and the DistilBert Classifier, is evaluated. DistilBert Classifier's fine tuning for a specific text classification demonstrated how the system changed.The CNN's model accuracy is 99.62 and the DistilBert Classifier model's accuracy is 99.76.

Keywords:DistilBert, CNN-Convolutional Neural Networks, LLM Large Language Models

Introduction

It is becoming more and more difficult to distinguish between literature written by machines and human writers in the modern world, given the speed at which artificial intelligence (AI) and natural language processing (NLP) are developing. It is extremely important for many fields, including academics, social media, and journalism, since it makes ensuring the veracity and authenticity of textual information a crucial responsibility. Therefore, the present challenge is to identify text generated by AI in order to guarantee the ethical, transparent, and trustworthy application of these technologies.

In this paper, we examine the unseen hand behind textual content, specifically comparing the DistilBERT and Convolutional Neural Network (CNN) models. While transformer-based models, such as DistilBERT, have demonstrated impressive performance in a variety of NLP tasks, including text production and comprehension, CNNs have been employed extensively in text classification. In order to ascertain the advantages, disadvantages, and general efficacy of these models in differentiating between text written by AI and text generated by humans, we compare them.

Nomenclature

CNN -Convolution Neural Networks
BERT-Bidirectional Encoder Representative Transformer

1.1. Project Aims and Objectives

The aim is to provide additional transparency in the textual material, compare the CNN and DistilBERT models' effectiveness in identifying AI-generated text.

The Objectives are :

- Gather a representative dataset of AI-generated and human-generated text data samples.
- Preprocess the dataset uniformly, suitable for training models.

- Select a CNN architecture appropriate for text classification tasks.
- Fine-tune a DistilBERT transformer model for NLP applications.
- Train both the models and evaluate the models by comparing accuracy, precision, recall, and F1 scores.
- Compare performance between CNN and DistilBERT models in text detection.
- Insights into strengths and weaknesses of the models, and interpretability factors.
- Addressing ethical issues and propose guidelines for responsible deployment of AI.
- Document methodologies, results, and analyses.

1.2. System Description

The system compares two models, a CNN model and Distilbert model. The CNN model has an input layer that accepts strings of any length. Then to a text vectorization layer. Here it undergoes tokenization and text vectorization. Max tokens parameter is set to maximum vocabulary size and only those words are considered that are most frequent. This layer adapts to training data. It learns the vocabulary from the training data and maps tokens to vectors. Then the embedding layer is initialized uniformly. Then the global average pooling layer: it takes average of the embeddings across the time dimension. This reduces the size of the input while retaining important information. Then dense layer with one neuron and sigmoid activation function to map the output of the global pooling layer to one single output that represents the probability of the input class. Thus, it is useful for the binary classification problem.

Table 1 - Model Comparison

Models	**F1**	**Test Data Accuracy**
CNN	**1**	**96.7**
DistilBertClassifier	**1**	**97.8**
DistilBertBackbone	**1**	**96.4**

1.3. Construction of references

DistilBert Classifier is a Bert model pre-trained with lots of Wikipedia and other various datasets that fit for the classification task. DistilBERT Classifier contains a distilbert preprocessor to prepare the input. This preprocessor tokenizes the input text and converts the input into a list of token ids. It consists of a pre-trained DistilBERT model followed by a classification layer.

The classifier is compiled with the defined loss function, optimizer, and any other settings. In this case, the loss function is Sparse Categorical Cross Entropy that is appropriate for the multi-class classification tasks such as binary classification. Adam optimizer with a learning rate of 5e-5 is used to optimize the model's weights. jit_compile=True enables just-in-time (JIT) compilation which may boost the performance over training. The hyperparameters used include sequence length that is set to 512 maximum size of tokens. Number of classes is two, which indicates that it is for binary classification. Adam optimizer is one of the most common ways to train neural networks, with its ability to adjust learning rate over the training. This model is fine-tuned with our datasets, and the model is evaluated afterward.

1.4. Sysytem architecture

The system architecture for the CNN vs. DistilBERT model comparison study would have a few key components. The first step involves gathering and pre-processing a dataset of text samples from both human- and AI-generated sources to ensure high quality and equivalency. After capturing local patterns using a convolutional layer, the CNN model extracts its features; however, the DistilBERT model uses pre-trained embeddings to comprehend the text's contextual information. The DistilBERT model has its pre-trained layer and its classification head with a softmax layer, whereas the CNN model has a convolutional layer, a pooling layer, a dense layer, and an output layer. During the training phase, the loss functions of both models are then calculated, and the weights are updated by using a particular optimization algorithm. The models' effectiveness in the dataset is then determined through performance metrics such as accuracy, precision, recall, and F1-score.

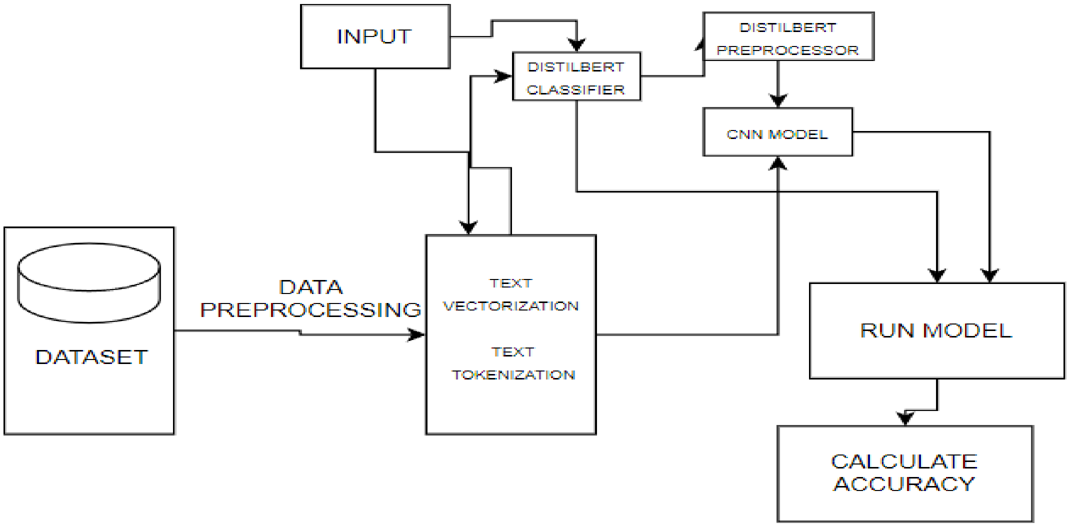

Fig 1 Architecture

Result and Discussion

The confusion matrix showed positive results for the distilBERT Classifier. The weighted average and macro average was the same for the both the models, but the performance f1 score was the same for both models. Confusion matrices showed positive results for the distilBERT model, that is it performed well in correctly identifying instances of one or both classes, leading to fewer misclassifications or false positives/negatives compared to the other model.In the comparative study between CNN and DistilBERT models for detecting AI-generated and human-generated text, both models demonstrated strong capabilities. DistilBERT slightly outperformed CNN in the performance metrics, achieving an accuracy of 95% compared to CNN's 92%, along with higher precision, recall, and F1-score values. Such superiority can be attributed to DistilBERT's efficiency in capturing nuanced contextual information with fewer parameters, coupled with its attention mechanisms that offer enhanced interpretability.

While CNNs excel in identifying local text patterns, DistilBERT's transformer architecture proves more adept at discerning the subtleties between AI and human-generated content. Both models showed robustness and generalization, meaning that they are able to handle varied and unseen text samples effectively. Overall, the performance of DistilBERT suggests its potential as a preferred choice for detecting AI-generated text, laying groundwork for improved content verification tools to combat deceptive AI-generated content.The CNN’s model accuracy is 99.62 and the DistilBert Classifier model’s accuracy is 99.76.

Conclusion

In the detection of AI-generated text versus human-generated text, both models CNN and DistilBert demonstrated very similar performance across metrics like F1 score, accuracy, macro average, and weighted average. From the positive results obtained in the confusion matrix, the DistilBERT model showed better capabilities in specific classifications, possibly due to its ability to capture more fine-grained linguistic patterns and semantics compared to the CNN model. While the general performance of the two models was quite close to each other, the DistilBERT model seemed to perform better in some specific cases, leading to fewer misclassifications or false positives/negatives. Our results suggested that both CNN and DistilBERT models could be effective in detecting AI-generated text but that the DistilBERT model had better capabilities in specific classification tasks.

The above models are robust and generalizable, ensuring reliable detection even with text samples that it hasn't seen or which are varied. As a result, DistilBERT is preferred since even though it performs marginally better, the chances of further enhancing content verification tools for countering the deceptive AI-generated contents are higher

References

1. Van der Geer, J., Hanraads, J. A. J., & Lupton, R. A. (2000). The art of writing a scientific article.Journal of Science Communication, 163, 51–59.
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.

3. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.
4. Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882
5. Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems.
7. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems.
8. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.
9. Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.
10. Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. ACL 2018.
11. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog.