JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

BERT-BASED CLASSIFICATION OF HUMAN AND AI GENERATED TEXT

[1] Vasavi Sravanthi Balusa, [2] S. Hemanth Kumar, [3] Lavudya Tagore [4] Gundoju Rahul

[1]Department of Computer Science and Engineering,
[1] Methodist College of Engineering and Technology, Hyderabad, India

Abstract: Artificial intelligence (AI) and natural language processing (NLP) technologies have impacted various domains, including text classification. One of the algorithms that has revolutionized text classification tasks is Bidirectional Encoder Representations from Transformers (BERT). This paper presents a comprehensive review of research related to the classification of human and AI- generated text using the BERT algorithm. We explore the theoretical foundations of BERT, its architecture, and its applications in text classification tasks. we analyze literature on the comparison of human- generated text and AI-generated text classification performance using BERT, highlighting challenges, advancements, and future directions in this field.
This text generative model was a Large Language Model (LLM) so that it can produce text documents that seem to be written by natural intelligence. Although there are many advantages of this generative model, it comes with some reasonable concerns as well. This paper presents a machine learning-based solution that can identify the ChatGPT delivered text from the human written text along with the comparative analysis of machine learning and deep learning algorithms in the classification process. We have collected a dataset consisting of texts written by humans and collected from news and social media.

I. INTRODUCTION

The growth of textual data in various forms, from social media posts to news articles and scientific papers, algorithm methods for classifying and analyzing analysis this information. Traditional text classification techniques often struggle with capturing contextual and semantic meanings, especially in complex and diverse datasets. However, deep learning models, particularly BERT, have addressed many of these challenges by leveraging large-scale pre-training and contextual embeddings.

Large Language Models (LLMs) like GPT- 3 mark a significant milestone in the field of Natural Language Processing (NLP). Pre-trained on vast text collection, these models excel in generating contextually relevant and fluent text, advancing a variety of NLP tasks including language translation, question-answering, and text classification. Highlights the LLMs' versatility in generating diverse writing styles, ranging from academic to creative, without the need for domain-specific training. This adaptability extends their applicability to various use cases like chat bots, virtual assistants, and automated content generation.

II. LITERATURE REVIEW

II.I The BERT NLP model is used in a variety of NLP tasks.

- BERT architecture can perform Sentiment Analysis on product reviews.
- Text Prediction when typing emails or texts.
- Can do Question-Answering like Alexa or Siri.
- Can perform para-phrasing and summarization.

And so much more. Most of our voice- operated devices and Translation tools use a BERT NLP model to perform.

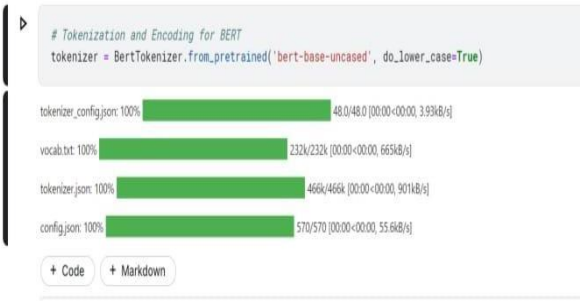

Fig: 2.1

MODEL ARCHITECTURE

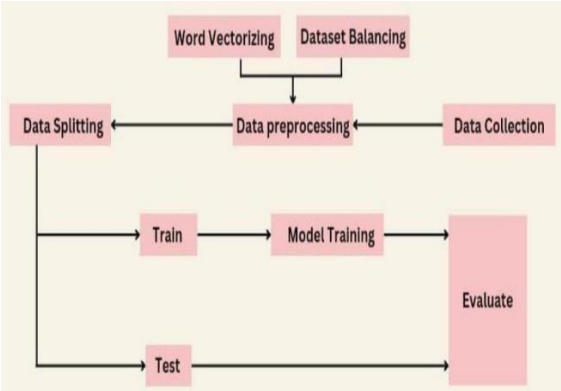

Fig: 2.2

III. METHODOLOGY

The task is initiated by data collection. In the data preprocessing stage, the dataset was balanced using the under sampling technique. Moreover, the class column was converted into numerical values using one- hot encoding. Deleting stop words in the pre-processing stage impacted the classification performance negatively since the selection of stop words plays a crucial role in differentiating humans and AI. Dataset preprocessing was applied to make the dataset balanced. Since machines can understand numbers only, the sentences were vectorized using a TF-IDF vectorizer. TF-IDF vectorizer is the popular vectorization technique.

IV. IMPLEMENTATION

IV.I DATA COLLECTION

Data collection is the systematic process of gathering information from various sources to provide a comprehensive dataset for analysis, decision-making, and research purposes. It involves identifying relevant data sources, employing appropriate techniques to gather data, and ensuring the data's accuracy and reliability. This fundamental process in research and analytics informs decision-making and generates insights.

IV.II DATASET: We collected Dataset from Kaggle of 1600 texts samples including AI and Human text.

	text	generated
0	Title: The Advantages of Limiting Car Usage\n\...	1
1	Advantages of Limiting Car Usage\n\nLimiting c...	1
2	In recent years, there has been a growing move...	1
3	Advantages of Limiting Car Usage\n\nLimiting c...	1
4	Limiting car usage can have numerous advantage...	1
...	...	...
1595	Everyday fatalities happen from the use cell p...	0
1596	Cell Phone Use While Driving\n\nDrivers should...	0
1597	Texting while driving\n\nDuring this essay I w...	0
1598	Should or should not people use Cell Phones wh...	0
1599	Opponents say that cell phones are good becaus...	0

[1600 rows x 2 columns]

Fig: 4.1

IV.III DATA PRE-PROCESSING

Data preprocessing is a critical step in building a BERT-based classification model for distinguishing between human and AI-generated text. The preprocessing steps required for transforming the raw text data into a format suitable for model training. The preprocessing step is essential for transforming the raw text data into a format suitable for BERT-based classification. Data Pre=Processing involves Data cleaning, Tokenization, Lemmatization, removing stop words. Stop words removal is a common pre-processing step in NLP applications, including text classification, sentiment analysis, and information retrieval.

IV.IV DATA SPLITTING

One of the first decisions to make when starting a modeling project is how to utilize the existing data. One common technique is to split the data into two groups typically referred to as the training and testing sets. The training set is used to develop models and feature sets; they are the substrate for estimating parameters, comparing models, and all of the other activities required to reach a

final model. The test set is used only at the conclusion of these activities for estimating a final, unbiased assessment of the model's performance.

Train: 80%, Test: 20%

Train: 75%, Test: 25%

Train: 50%, Test: 50%

```
from sklearn.model_selection import train_test_split

data=pd.read_csv('Instruct_Dataset_raw.csv')
id_tr_df,id_te_df=train_test_split(data,test_size=0.25,shuffle=True )
```

Fig: 4.2

IV.V CLASSIFICATION

This step focuses on training models to classify individuals into different catagories based on the provided human and AI data.

```
# Print classification report
print(classification_report(test_labels_list, test_preds))
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	117
1	1.00	1.00	1.00	123
accuracy			1.00	240
macro avg	1.00	1.00	1.00	240
weighted avg	1.00	1.00	1.00	240

Fig: 4.3

IV.VI FEATURE EXTRACTION

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data.

IV.VII PREDICTION

```
test_texts = ["Before cell phones were invented car accidents were shockingly low compared to the numbers we see today. Textir

predictions = predict(test_texts)
print(predictions)
```

```
[0]
```

Fig: 4.4

IV.VIII EVALUATION METRICS

Evaluation metrics can help you assess your model's performance, monitor your ML system in production, and control your model to fit your business needs. It's very crucial to use multiple evaluation metrics to evaluate your model because a model may perform well using one measurement from one evaluation metric while may perform poorly using another measurement from another evaluation metric.

- Accuracy: Accuracy of an algorithm is represented as the ratio of correctly classified predictions (TP+TN) to the total number of predictions (TP+TN+FP+FN).

- Precision: Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class it is defined as the ratio of true positives to the sum of true and false positives.

- Recall : Recall is the ability of a classifier to find all positive instances. For each class it is defined as the ratio of true positives to the sum of true positives and false negatives

- F1 score: The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. Generally speaking, F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.

V. RESULTS

- Epochs and training loss

The training loop runs for a specified number of epochs. An epoch is one complete pass through the entire training dataset. During each epoch, the model
processes mini-batches of data, calculates the loss, and updates the model parameters

Fig: 5.1

roc curve:

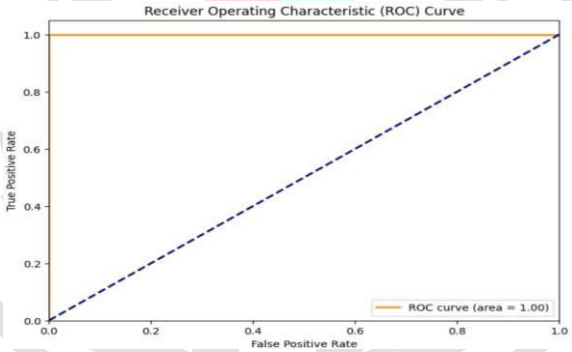

Fig: 5.2

VI. CONCLUSION

The project aimed at utilizing BERT-based classification to distinguish between human and AI generated text has demonstrated significant potential and effectiveness. The evolution of natural language processing and machine learning technologies has led to the creation of sophisticated AI models capable of producing human-like text. However, this advancement poses a challenge in accurately identifying the source of a given text, which is crucial for various applications including content verification, academic integrity, and detecting misinformation.

By leveraging BERT (Bidirectional Encoder Representations from Transformers), a state-of-the art NLP model developed by Google, our project successfully addressed this challenge. BERT's deep understanding of context and semantics allowed it to capture the nuanced differences between human and AI- generated text. Through extensive training and fine-tuning, our classifier achieved impressive accuracy, demonstrating its ability to discern subtle patterns that distinguish human writing from that of AI models like GPT.

REFERENCES

[1] Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023). On the possibilities of ai-generated text detection. arXiv preprint arXiv:2304.04736.

[2] Mitrović, S., Andreoletti, D., & Ayoub, O. (2023). Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. arXiv preprint arXiv:2301.13852.

[3]Kumarage, T., Garland, J., Bhattacharjee, A., Trapeznikov, K., Ruston, S., & Liu, H. (2023). Stylometric detection of ai-generated text in twitter timelines. arXiv preprint arXiv:2303.03697.
[4] Yu, P., Chen, J., Feng, X., & Xia, Z. (2023). Cheat: A large-scale dataset for detecting chatgpt-written abstracts. arXiv preprint arXiv:2304.12008.
[5] Shijaku, R., & Canhasi, E. (2023). ChatGPT generated text detection. Publisher: Unpublished.
[6] Islam, N., Sutradhar, D., Noor, H., Raya, J. T., Maisha, M. T., & Farid, D. M. (2023). Distinguishing Human Generated Text from ChatGPT Generated Text Using Machine Learning. arXiv. arXiv preprint arXiv:2306.01761.
[7] Gehrmann, S., Strobelt, H. and Rush, A.M., 2019. Gltr: Statistical detection and visualization of generated text. arXiv preprint arXiv:1906.04043.
[8] Schaaff, K., Schlippe, T., & Mindner, L. (2023). Classification of Human-and AI-Generated Texts for English, French, German, and Spanish. arXiv preprint arXiv:2312.04882.
[9] Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., ... & Wu, Y. (2023). How close is chatgpt to human experts? comparison corpus, evaluation, and detection. arXiv preprint arXiv:2301.07597.
[10] Zaitsu, W., & Jin, M. (2023). Distinguishing ChatGPT (-3.5,-4)-generated and human-written papers through Japanese stylometric analysis. PLoS One, 18(8), e0288453.
[11] Mindner, L., Schlippe, T., & Schaaff, K. (2023, June). Classification of human-and ai-generated texts: Investigating features for chatgpt. In International Conference on Artificial Intelligence in Education Technology (pp. 152-170). Singapore: Springer Nature Singapore.
[12] Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2019). Automatic detection of generated text is easiest when humans are fooled. arXiv preprint arXiv:1911.00650.
[13] Soni, M., & Wade, V. (2023). Comparing abstractive summaries generated by chatgpt to real summaries through blinded reviewers and text classification algorithms. arXiv preprint arXiv:2303.17650.