The University of Hong Kong

Department of Statistics and Actuarial Science

2022 - 2023 STAT8021 Big Data Analytics Group Project

AI classifiers to distinguish work of ChatGPT from human

Abstract

Being launched within 6 months, ChatGPT has attracted significant attention globally and has already been widely used in various applications. However, "the greater the power, the more dangerous the abuse" accurately describes the threat of mis-usage of ChatGPT. In this project, we built 3 classification models to identify the ChatGPT work from human work. We first collected over 24,000 answer-response pairs from both humans and ChatGPT. After data pre-processing, we studied the characteristics in each group of responses, and created Word2Vec + Fully Connected Neural Network, Word2Vec + Bidirectional LSTM and fine-tuning DistilBERT Transformer. At the end, we evaluated and compared the model performance, then provided explanations of the result, rooms of improvement and possible applications of models.

Group 15

Instructor: Dr. Lequan Yu

1. Introduction

Background: In recent years, large language models (LLM) have been leading the way in AI development. The models are impressive given the ability to handle sophisticated conversations like humans and process information with high accuracy. Among all, the OpenAI developed LLM ChatGPT has gained widespread attention and is considered as one of the most successful applications in this field. Since its launch in November 2022, ChatGPT has achieved unprecedented user growth, attracting one million users in just five days and 100 million users within two months. World-wide users have employed ChatGPT for various tasks, such as writing programming codes, generating ideas, and writing stories etc.

Problem: Despite the promising future offered by ChatGPT, abusing ChatGPT can lead to serious adverse consequences. For instance, ChatGPT enables generation of high quality fake accounts on social media like Facebook and Twitter. In 2023 April, CNBC reported findings of AI generated comments in Amazon products reviews. In education, teachers are worried about students using ChatGPT on their homework and failing to absorb the required knowledge.

Objective: In view of the rapid developments in this field and potential threats, the objective of this project is building a model that can distinguish the work of ChatGPT from human-generated text.

2. Related Work

The rise of large language models has revolutionized natural language processing and enabled a new generation of technology. However, misusing such powerful tools could be worrying and destructive, particularly in the context of propaganda and misinformation generation.

In a research study conducted by Guo and team (Guo, et al, 2023), they compared human evaluations and linguistic analyses of ChatGPT-generated contents with human experts and found that ChatGPT's responses are more focused on the question at hand, whereas humans tend to be more divergent and may shift to other topics. Secondly, ChatGPT provides objective answers, while humans prefer subjective expressions. Lastly, ChatGPT's responses are typically more formal, while humans tend to be more colloquial and use more emotional punctuation and grammar features. They also conducted three different models in detecting contents produced by ChatGPT namely logistic regression model trained on GLTR Test-2 features, a deep classifier for single-text detection, and a deep classifier for QA detection.

In addition to the deep learning model approach, a study by Mitchell et al. (2021) demonstrated AI-generated text tends to occupy negative curvature regions of the model's log probability function. Based on the findings, they built DetectGPT which uses only log probabilities “computed by the model of interest and random perturbations of the passage from another generic pre-trained language mode”. DetectGPT is more discriminative than existing zero-shot methods, notably improving fake news detection when compared to the 20B parameter GPT-NeoX. More relevant academic papers as follows:

3. Dataset

3.1 Data Source

The data source for this project is Huggingface/Hello-SimpleAI/HC3 (https://huggingface.co/datasets/Hello-SimpleAI/HC3). The raw data consist over 24,000 samples and each contains the following 5 columns:

- ID
- question
- human_answers
- chatgpt_answers
- source

Each data represents a combination of question/human-answer/ chatGPT-answer. Human responses are collected from various websites like wiki/reddit, while GPT answers are collected in a clean chat history environment. The first few rows of data are displayed as follows:

id (string)	question (string)	human_answers (sequence)	chatgpt_answers (sequence)	source (string)
"2"	"Why do we still have SD TV channels when H…	["The way it works is that old TV stations got…	["There are a few reasons why we still have SD…	"reddit_eli5"
"3"	"Why has nobody assassinated Kim Jong…	["You ca n't just go around assassinating the…	["It is generally not acceptable or ethical to…	"reddit_eli5"
"4"	"How was airplane technology able to…	["Wanting to kill the shit out of Germans…	["After the Wright Brothers made the first powered…	"reddit_eli5"
"5"	"Why do humans have different colored eye…	["Melanin ! Many of the the first known humans…	["The color of your eyes is determined by the amount an…	"reddit_eli5"
"6"	"Why I can not fabricate a religion…	["Because you 're a minor and your parents…	["The First Amendment to the United States…	"reddit_eli5"

3.2 Data Preprocessing

Data preprocessing is a crucial step in any data science project. In this project, there are two procedures involved in this stage:

1. To avoid data leakage, only human or ChatGPT answers are selected for each question. The selection is random and the total count of each class is the same.
2. Additional column “label” is created to indicate if the answer is generated by ChatGPT or not.
3. Empty responses are removed from our question-reponse data set.

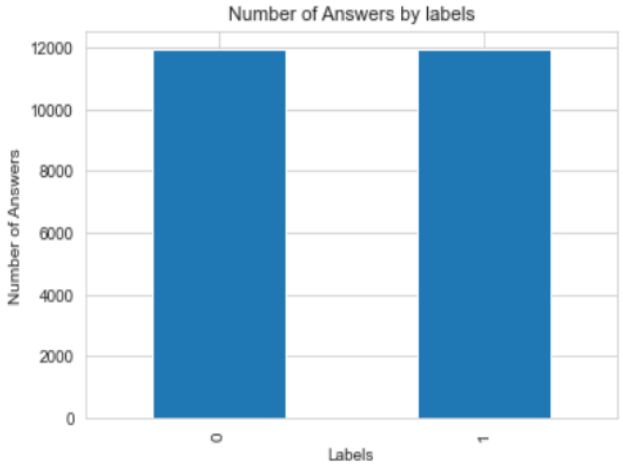
The first few rows of processed data are displayed as follows:

	question	source	labels	answers
0	Can I deduct work equipment I am not required ...	finance	0	Old question, but in the comments of the accep...
1	Can I use losses from sale of stock to offset ...	finance	1	Yes, you can use losses from the sale of stock...
2	How to find the smallest transaction fees and ...	finance	0	The lowest cost way to trade on an exchange is...
3	What to sell when your financial needs change,...	finance	0	You have to understand what risk is and how mu...
4	Unemployment Insurance Through Options	finance	1	Unemployment insurance is a government-run pro...

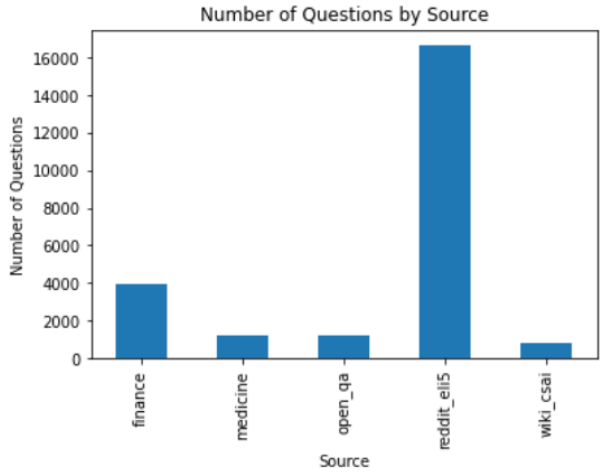
3.3 Exploratory Data Analysis

Before building models, Exploratory Data Analysis is conducted to build understanding of the dataset and identify obvious patterns. Here are the key findings:

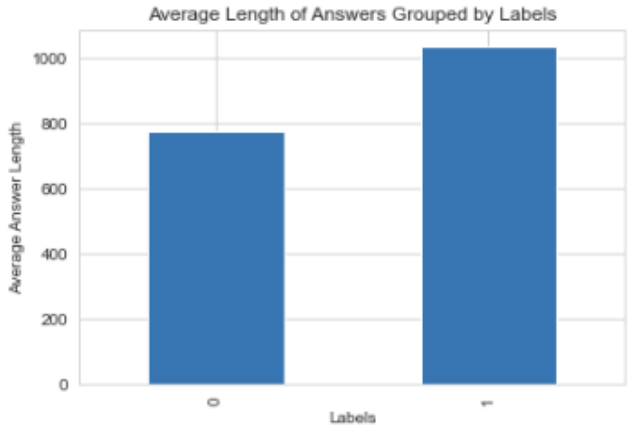
1. Among the 23868 processed records, the sample size is equal between the two groups (answers generated by ChatGPT and human). The balanced dataset can reduce bias in model learning.

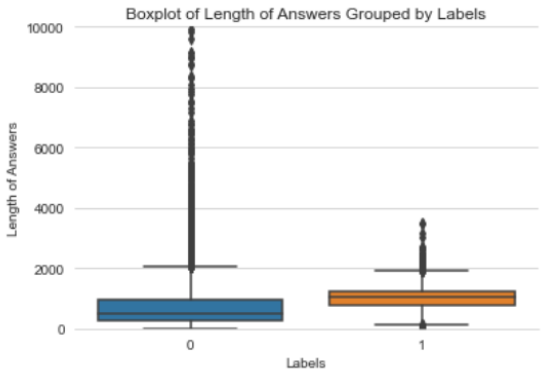


2. The data are sourced from five different categories, with the majority of data coming from Reddit.



3. The average length of ChatGPT answers is longer than human-generated responses, with a lower variance. This finding indicates that ChatGPT generally generates more detailed and elaborate responses compared to humans.





4. Models

4.0 Naive Length Classifier

As demonstrated in the elementary data analysis, there exists a significant correlation between the GPT label and the lengt h of the response. In order to establish a baseline model for comparative purposes, a simple decision stump was develope d.

Decision: *If length of an answer is longer than 118, it would be classified as generated by ChatGPT". Otherwise, it woul d be classified as "generated by human"*

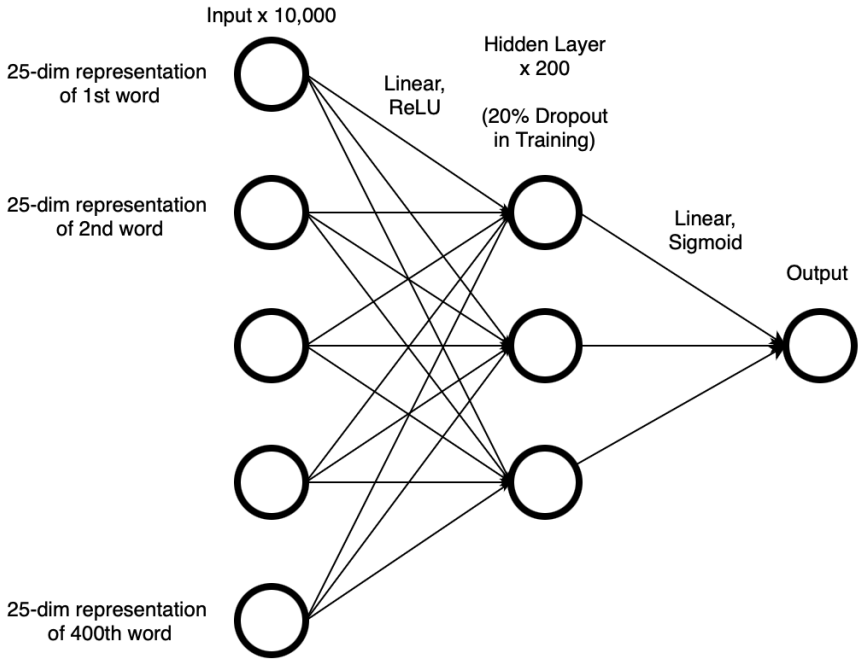
The training accuracy is 73.30%, and testing accuracy is 72.64%. It is a baseline for our model performance.

4.1 Fully Connected Neural Network

The naive classifier above does not consider the contextual meaning of the text. To extract the contextual meaning of the response, a word embedding is required before inputting it into a deep learning model. There are various word embedding algorithms available, including Word2Vec, ELMO, and others. However, the simplest Word2Vec model is adopted as a f oundational milestone to more advanced models, which captures the meaning of a single word but does not account for th e contextualized meaning of the entire text.

The model begins by using the NLTK tokenizer to tokenize the response paragraph into individual words. The Word2Vec model is then applied to obtain a 25-dimensional vector representation of each word. These representations are concatenat ed into a variable-length vector. Input tokens are limited to 400 words to maintain a reasonable size for our neural networ k. Any additional input beyond this limit is truncated, with any short responses being padded with zero values at the end. This process results in a sentence representation vector with a length of 10,000.

To effectively study the prior question and improve our classification, a variant of the model is built. The same Word2Ve c model is used to tokenize the question and limit our question input tokens to 100 words and answer input tokens to 400 words. This process resulted in a sentence representation vector with a length of 12,500. A separate neural network will b e learnt and used to compare with the original.



The fully connected neural network is composed of one hidden layer. First, a linear function is used to map the sentence v ector of either 10,000 or 12,500 dimensions to 200 hidden nodes. Next, a BatchNorm layer is applied to normalize the out put, followed by a ReLU activation layer. To reduce the potential for overfitting, an additional dropout layer is inserted w ith a probability of 20%. After that, a linear function is applied to map from the 200 nodes to a single output node, follow ed by a sigmoid function.

Using hyperparameter tuning techniques, we used the AdamOptimizer with a learning rate of 0.005 and trained the model for 5 epochs. An overfitting problem was observed when training with 10 epochs or more.

4.2 Bidirectional LSTM

To establish a better understanding of the language structure in answer responses, the simple neural network model is repl aced with a more modern Bidirectional Long Short-Term Memory (Bi-LSTM) model. The word embedding details are si milar to those used previously, except that the model is fed word by word instead of using one lengthy vector. The input c onsists of 400 25-dimensional word embeddings for each sentence.

The bidirectional LSTM (Bi-LSTM) is a type of recurrent neural network that includes two separate recurrent layers: one layer processes the input sequence in a forward direction, and another processes it in a backward direction. This allows th e network to capture both past and future context for better performance in binary classification tasks. In the model, there is a single hidden layer with 200 hidden nodes for the forward direction and another 200-nodes hidden layer for the back ward direction. After that, a linear function is applied to the 400-dimensional output of the model, followed by a sigmoid function for binary classification.

Using hyperparameter tuning techniques, AdamOptimizer is applied with a learning rate of 0.005 and trained the model f or 20 epochs. More epochs are used for Bi-LSTM because its convergence rate is much slower compared to the simple ne ural network in the previous model.

4.3 DistilBERT Transformer

The Transformer model has been recognized as the state-of-the-art deep learning model in Natural Language Processing (NLP). Its self-attention mechanism allows it to excel in language understanding. Nonetheless, training Transformers fro m scratch can be challenging due to their large size and massive corpus of text data . As a result, a pre-trained Transform er library is utilized and fine-tuned with the training dataset. A variant of the BERT family, namely DistilBERT is emplo yed because of its smaller computational power demand with most of the original model's capabilities being preserved. Th e DistilBERT model can fit in using a modern customer-level GPU.

Before fitting the data into the DistilBERT model, the first 200 tokens of each answer are selected in order to speed up th e training process. Then, the pretrained "distilbert-base-uncased" transformer is employed from the hugging face and fi tted into the "DistilBertForSequenceClassification" model. AdamWOptimizer is used to be the model optimizer with a learning rate of 0.00005, a weight decay of 0.01 and training epochs of 3. These parameters were chosen based on experi mentation.

Epoch	Training accuracy	Validation accuracy
1	99.75%	99.37%
2	99.56%	98.81%
3	99.57%	98.64%

5. Discussion

5.1 Satisfactory result with over 90% accuracy for all 3 models

With reference to the consolidated table below, all 3 models yield satisfactory results with over 99% accuracy, precision, r ecall and F1 score in the validation stage. In order to test whether our models are well-developed and ready to put into ap plication, we further extracted 40 answers from the internet to be the testing dataset. Obvious differences are observed in t he testing set where the metrics for simple neural network (answers only) are around 92% while LSTM and transformer m odels reach 97% and 98% respectively. It is consistent with our expectation that the latters are more complex than a simpl e neural network with more layers and parameters to capture more complex relationships in the dataset. Moreover, Huggi ngface Transformer models are designed to handle natural language processing (NLP) tasks including text classification w hich is expected to work better.

One key drawback of these models is the lengthy training time involved. In this project, these 3 models were trained in va rious computer setup so no direct apple-to-apple comparison can be inferred here. For model 1, simple neural network, an d model 2, LSTM, and model 3 Transfomer, these were trained with GPU.

Model	Dataset	Accuracy	Precision	Recall	F1 score
Model 1 - Word2Vec Embedding + Simple NN	Validation (A)	99.46%	99.46%	99.46%	99.46%
	Testing (A)	92.39%	92.39%	92.40%	92.39%
	Validation (Q+A)	99.54%	99.54%	99.54%	99.54%
	Testing (Q+A)	92.35%	92.39%	92.39%	92.35%
Model 2 - Word2Vec Embedding + Bi-LSTM	Validation (A)	99.42%	99.42%	99.42%	99.42%
	Testing (A)	97.53%	97.56%	97.51%	97.53%
Model 3 - Fine tuning using Huggingface transformer model	Validation (A)	99.57%	99.57%	99.57%	99.57%
	Testing (A)	98.64%	98.66%	98.65%	98.64%

5.2 Questions not important with only answers being sufficient

One interesting finding is that with reference to the table above, in model 1, neural network, the validation and testing acc uracy for the dataset with and without questions are very similar. Initially it' s believed that by including questions in the model it will return more accurate classifications but indeed the model only looks at the answers in the training process. As mentioned in the above section, Guo and team (2023) found that some key differences between ChatGPT and human a nswers are that ChatGPT answers are usually written in an organized manner which are usually longer, more detailed, for mal and objective than human answers. The difference could be caused by "answers" being truncated in this project as a compromise to intensive memory and computation power required for a full model.

5.3 Application of models

Given the ability to distinguish between ChatGPT and human generated write ups, the classification models in this project can be applied in a wide range of settings, such as fake account detection in social media platforms and homework-plagiarism detection by education institutions. In addition, It can be used as a new InstructGPT and be the reinforcement learning target for other LLM, in order to generate more human-like answers.

5.4 Future work

The project uses only GPT 3.5 data while its development is rapid. While GPT-3.5 was released in November 2022, GPT-4 was released in March 2023 less than half a year which is a multimodal language model which can handle inputs of both images and text. (Dilmegani, 2023) Therefore, the development of the classifier is ongoing. With more data on GPT-4 answers the model can be further trained to incorporate the development of answers from interactive AI chatbox. Also, full response, instead of truncated one, shall be used if computation resources allow.

Another area of improvement is that currently it can classify answers between pure human and pure ChatGPT. It does not work when it is a mixture of both. While this project is a first step to the development of these classifiers, more work is needed to highlight which part is from human and from ChatGPT in a mixed answers.

6. Conclusions

The development of ChatGPT has brought about a new era in the application of AI in our lives. However, it is important to consider the negative impacts of such technology. To mitigate these issues, tools are needed to identify whether the materials are generated by AI or human beings. This project applies a simple neural network, LSTM and Huggingface Transformer in building a classification model to distinguish between answers generated by ChatGPT and those generated by humans with satisfactory accuracy. The technology has the potential to be applied in various settings. Nevertheless, it is crucial to continue researching and developing and training these tools to keep pace with the development of the interactive AI chat box and generative language model.

7. Referenece

Annierpalmer. (2023, April 25). People are using A.I. Chatbots to write Amazon Reviews. CNBC. Retrieved May 1, 2023, from https://www.cnbc.com/2023/04/25/amazon-reviews-are-being-written-by-ai-chatbots.html

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv preprint arXiv:2108.13751.

Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023, January 26). Detectgpt: Zero-shot machine-generated text detection using probability curvature. arXiv.org. https://arxiv.org/abs/2301.11305

Open AI (2022). Aligning language models to follow instructions. Retrieved Apr 4, 2023, from https://openai.com/research/instruction-following

C. Dilmegani (2023). GPT4: In-depth Guide in 2023. Retrieved Apr 4, 2023, from https://research.aimultiple.com/gpt4/