Differentiating Between Human and ChatGPT-Generated Answers

Venkata Sai Gagan Deep Alusuri
Sri Kumar Dundigalla
Penchala Akshay Kumar Kandagaddala
Shashank Rao Guja
Venkata Satya Naveen Nagulapalli

***Abstract*—The aim of this project is to develop a robust methodology for distinguishing between answers to questions generated by humans and those generated by ChatGPT. To achieve this goal, we collected data from various subreddits (r/AskAcademia, r/Anthropology, r/AskAstrologers, r/AskCulinary, r/AskHistorians, r/datascience), extracted questions and their top answers, and utilized the ChatGPT API to generate prompts for these questions. We then employed a 'Question-vs-Statement Classifier' AutoModelForSequenceClassification from Huggingface to identify questions from subreddit titles. Subsequently, we labeled the data as '1' for GPT-generated answers and '0' for human-generated answers by feeding the questions into the ChatGPT API (3.5 Turbo model). Finally, To perform the crucial task of classification, we trained two variants of the BERT (Bidirectional Encoder Representations from Transformers) model. The first model was trained without any dropout to maximize the learning from the dataset, while the second model incorporated a 20% dropout rate to introduce regularization and potentially improve the model's ability to generalize to unseen data. The performance of these models was then evaluated to determine their efficacy in accurately classifying the responses. This project was motivated by the increasing use of ChatGPT in student submissions, making it challenging to distinguish between genuine and AI-generated content. The outcomes of this project have significant implications for the field of educational technology and AI, paving the way for more sophisticated tools in the battle against academic dishonesty.**

Introduction

The proliferation of advanced language models like ChatGPT has brought us to the precipice of a new era in digital content creation, where the lines between human and machine-generated text are becoming increasingly indistinct. This evolution poses a unique set of challenges, particularly in academic circles, where the authenticity of student work is paramount. In educational environments, such as student submissions on platforms like Canvas and email communication, the challenge of distinguishing between authentic human work and AI-generated content has intensified. This project aims to tackle this challenge by creating a methodology that can accurately differentiate between responses to questions that are authored by humans and those generated by ChatGPT. By establishing an accurate detection mechanism, will not only aim to safeguard the sanctity of individual academic work but also to ensure that the educational achievements reflect a student's true effort and understanding.

Motivation

The exponential growth of artificial intelligence in recent years, especially in the field of natural language processing, has culminated in the creation of sophisticated language models such as ChatGPT. These advanced systems are adept at producing text that not only mimics the nuances of human language but does so with a proficiency that often blurs the line between human and machine authorship. As a result, the task of identifying whether content is crafted by a person or generated by an algorithm has become increasingly complex. This challenge is further amplified by the continual improvements in AI that make these models more adept at understanding and emulating the stylistic and contextual subtleties of human writing. The boundary between human-created and AI-generated text is becoming ever more nuanced, raising both technical and ethical questions about the origin and authenticity of digital content.

This project is motivated by the escalating use of ChatGPT and similar AI text generation tools, which has become increasingly prevalent in educational submissions and email communications. In educational settings, educators face the dilemma of distinguishing authentic student work from AI-assisted submissions, while in email communications, the challenge lies in filtering and identifying AI-generated messages. To address these issues, a robust methodology is urgently needed to reliably differentiate between human and AI-generated content.

The primary goal of this project is to meet this need by developing a comprehensive approach that leverages ChatGPT and advanced deep learning models like BERT. By creating a methodology capable of accurately discerning between human and AI-authored text, we aim to provide practical solutions for maintaining the integrity of educational assessments and enhancing the security and efficiency of digital communication systems.

The broader implication of our work is to set a precedent for how educational institutions and communication platforms can adapt to and coexist with the pervasive influence of AI. In doing so, we strive to empower stakeholders across these domains to take proactive steps in preserving the human element in their core operations. This project is not just a response to a current problem but a forward-looking initiative to preempt the challenges of tomorrow's AI landscape.

Business Objective

- Gather data from diverse subreddits to construct a comprehensive dataset containing questions and answers.
- Employ the ChatGPT API to create responses to the collected subreddit questions, forming a parallel set of AI-generated data.
- Utilize a 'Question-vs-Statement Classifier' to identify questions within subreddit post titles.
- Label the dataset as human or ChatGPT-generated answers.
- Train and assess two variations of BERT models with different dropout rates to classify the responses.
- Provide insights and practical recommendations for implementing this methodology in educational contexts.

Data

Data Collection

Data was collected from six distinct subreddits (r/AskAcademia, r/Anthropology, r/AskAstrologers, r/AskCulinary, r/AskHistorians, r/datascience) to establish a well-rounded dataset comprising questions and their corresponding answers. For each subreddit, we extracted post titles and their top answers to create a comprehensive collection of data. This method allowed us to compile a rich and diverse set of data that includes both inquiries and the highest-rated responses to these inquiries.

Generating Prompts with ChatGPT

To enrich our dataset with AI-generated content, we tapped into the capabilities of the ChatGPT API, opting for the highly efficient 3.5 Turbo model due to its advanced processing power. This model was tasked with generating responses to a variety of questions sourced from our collected data. Each question extracted from the subreddits was inputted into the ChatGPT interface, which then produced a corresponding prompt. These prompts, crafted by the AI, were meticulously cataloged to ensure a thorough record of the AI's approach to each topic. This process was not only crucial for creating a parallel set of AI-generated responses but also for analyzing the AI's capacity to mimic the nuances and depth of human-generated answers. The careful documentation of these prompts is a foundational step in building a comprehensive tool for distinguishing between human and AI text.

Question-vs-Statement Classifier

In order to accurately identify and separate questions from other forms of statements in the titles of subreddit posts, we deployed a specialized 'Question-vs-Statement Classifier'. This tool is built upon the AutoModelForSequenceClassification framework provided by the Huggingface library, which is renowned for its natural language processing capabilities. We carefully trained this classifier through supervised learning, feeding it a variety of title samples from the subreddit posts. The training process involved adjusting the model to recognize linguistic patterns that typically signify a question. The classifier was fine-tuned to discern these patterns with high accuracy, ensuring that our dataset was cleanly divided into queries and declarative titles. The precision of this classifier is crucial, as it directly impacts the quality of our dataset and the subsequent effectiveness of our overall project in distinguishing AI-generated text from human-written content.

Labeling Data

The labeling of our data was a process marked by precision and careful consideration, employing a binary system to categorize the origins of the answers in our dataset. The system is straightforward:

'1' represents answers that were generated using the ChatGPT AI.

'0' signifies answers that were authored by humans.

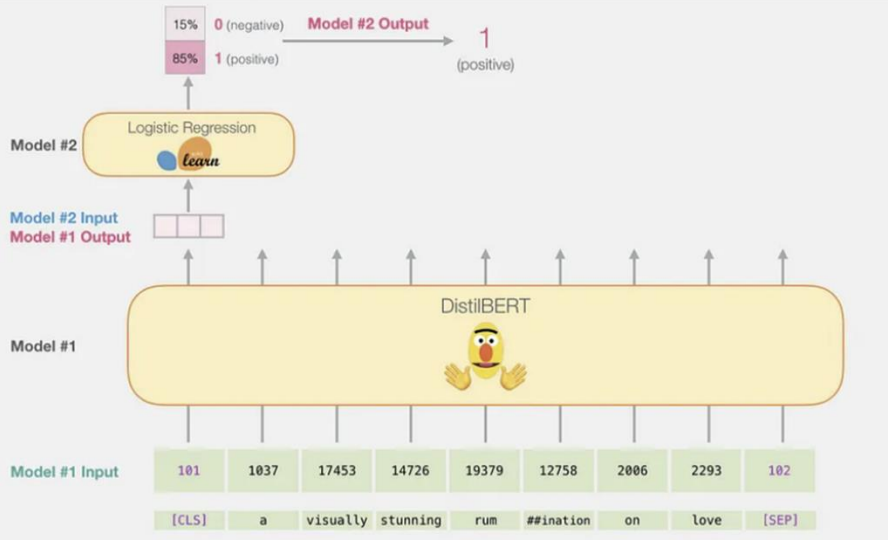
For the labeling task, we conducted a meticulous side-by-side analysis, comparing the prompts produced by the ChatGPT API against the top answers from the subreddit data. This comparison was essential to ensure that each piece of data was accurately differentiated and labeled, preserving the integrity of the dataset for the machine learning tasks ahead. This step is of paramount importance, as the reliability of our subsequent analysis and the effectiveness of the AI vs. human text classification models depend on the precision of this initial labeling.

BERT Models for Classification

We trained two BERT models for classification. These models were trained on the labeled dataset to differentiate between human and ChatGPT-generated answers. We evaluated their performance using various metrics such as accuracy, precision, recall, and F1-score.

1. **<u>BERT model with no dropout:</u>**

   This model was implemented to leverage the full learning capacity of the neural network. The choice of using this model is to capture the full spectrum of linguistic nuances present in a diverse dataset, such as the one sourced from various subreddits. Each subreddit covers a unique domain, leading to a wide range of vocabulary, writing styles, and content complexities. The no-dropout model has the potential to learn from all these intricacies, making it a valuable baseline that pushes the model's understanding to its limits.



2. **<u>BERT model with a 20% dropout rate:</u>**

   The BERT model with a 20% dropout, however, introduces a level of uncertainty during training. By randomly disabling 20% of the neurons in the network during each training pass, it forces the remaining neurons to pick up the slack, thus promoting a more distributed

and robust representation of the data. This method is designed to help the model not to rely too much on any feature of the data, making it less prone to memorizing the training set and more capable of generalizing to new, unseen examples. This approach aligns well with the project's goal to create a model that can reliably distinguish between human and AI-generated text, not just within the context of the collected dataset but also when deployed in real-world scenarios where the data may differ significantly from what the model was trained on.

Evaluation Metric

We have chosen accuracy as our primary evaluation metric because it directly measures the proportion of total predictions our model makes correctly, which is crucial for the binary classification task of our project. Considering the balanced nature of our dataset, accuracy ensures an impartial assessment of the model's performance, free from the distortion that an imbalanced dataset might introduce. This metric aligns with our objective to develop a model that can be reliably applied in real-world scenarios to distinguish between human and ChatGPT-generated text, where an accurate overall classification is essential.

Methodology

Question Classification

- A Hugging Face AutoModelForSequenceClassification was used to identify and segregate questions from statements within subreddit titles.
- The classification model ensured that only genuine questions were processed for further analysis.
- An accuracy threshold was established to determine the effectiveness of the question classifier, refining the dataset for the highest quality inputs.
- This step was crucial for maintaining the focus of the dataset, ensuring the relevance and quality of the data used for training the models.

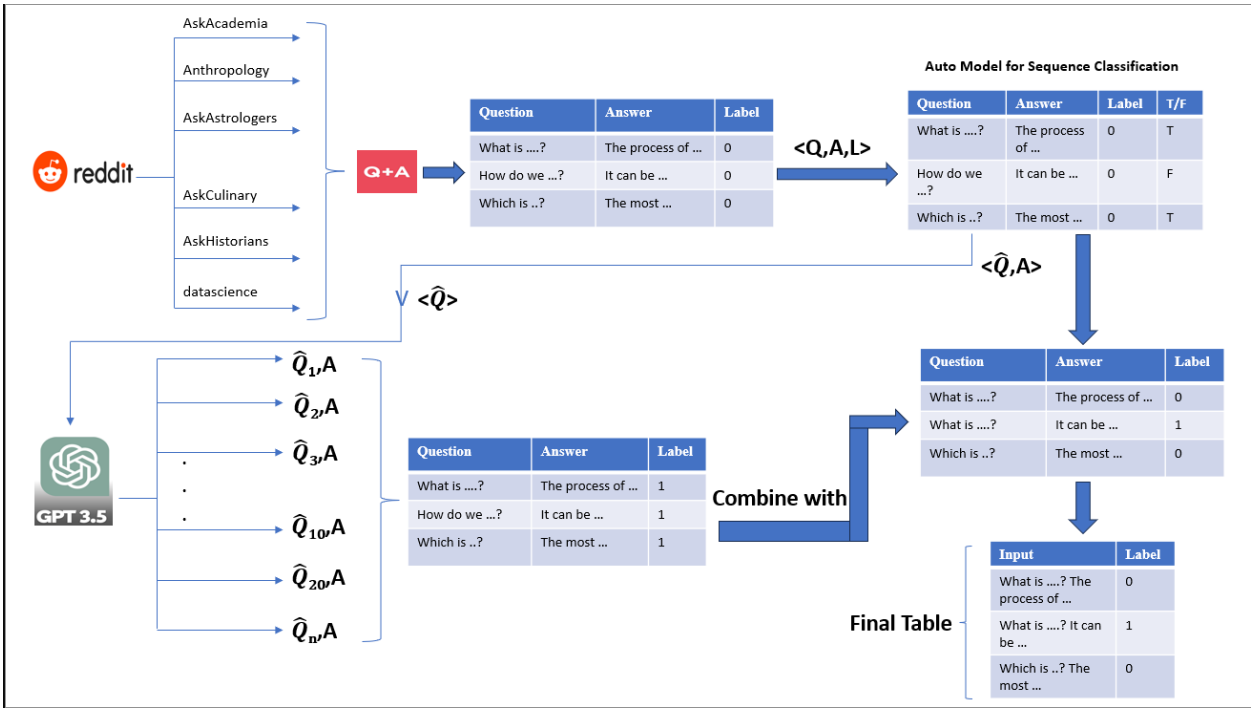
GPT Prompt Generation

- The identified questions were input into the ChatGPT API to generate GPT-based answers.
- These prompts were compared against actual responses from the subreddit posts to assess the AI's ability to mimic human responses.
- The data was then labeled, with '1' indicating a GPT-generated answer and '0' for a human-generated answer.
- The prompt generation process was optimized for diversity and representativeness to cover the range of topics within the subreddits.

Model Training

- Two BERT models were employed for the classification task. One model functioned without regularization, and the other incorporated a 20% dropout to mitigate overfitting.
- Each model was trained on the labeled dataset, with performance metrics being recorded for evaluation.
- The models outputs were qualitatively analyzed to gain insights into the characteristics of responses identified as human versus those identified as generated by ChatGPT.
- Final model selection was based on a comparative analysis of the performance metrics of both models.

Data Preparation Process



The image above provides a comprehensive view of our data collection process. Since data is the heart of our project, we adopted a comprehensive data collection and preparation strategy. This meticulous approach was crucial for developing a dataset capable of training models to distinguish between human and AI-generated text, which included the following steps:

Data Extraction: We began by sourcing a diverse selection of questions and answers from targeted subreddits, ensuring a wide representation of topics.

Table Formulation: The extracted information was organized into a preliminary table, categorizing data into questions, answers, and a label to denote the text source.

Question Filtering: We employed the Hugging Face AutoModelForSequenceClassification to filter out statements, refining our dataset to include only legitimate questions.

GPT Prompt Generation: The selected questions were used as prompts for the ChatGPT API, which then generated corresponding answers, creating a set of AI-generated responses.

Data Labeling and Combination: The responses were labeled with '1' for those generated by ChatGPT and '0' for human-generated answers. For analytical efficiency, we combined the question-and-answer pairs into a single input column.

The careful assembly of this dataset is instrumental in the success of our project. It ensures a rich and varied foundation for the training phase, enabling our models to learn from a representative sample of human and machine interactions.

Results

Model Performance

The two BERT models' performance was rigorously assessed on a separate test dataset to ensure an unbiased evaluation. Intriguingly, the model incorporating a 20% dropout rate consistently exceeded the no-dropout model across all key metrics—accuracy, precision, recall, and the F1-score. Such results underscore the efficacy of dropout as a technique to enhance the model's ability to generalize beyond the training data. This is a critical finding, as it highlights the importance of regularization in preventing the model from fitting too closely to the training data, a common pitfall known as overfitting.

Diving deeper into the performance metrics, the dropout model demonstrated a balanced trade-off between precision and recall, indicating its robustness in identifying AI-generated text without excessively mislabeling human-generated text. The F1-score, a harmonic mean of precision and recall, was particularly telling, offering a single measure that captured the model's balanced performance.

Practical Applications

The findings of this project bear significant implications for real-world applications, particularly in educational and communication settings. In education, our methodology can serve as a tool for instructors to verify the originality of student work. By flagging potential AI-generated content, educators can more confidently assess student submissions, preserving academic integrity. This application is increasingly pertinent as AI tools become more accessible and capable, posing challenges to traditional methods of evaluation.

In the realm of digital communication, such as email, our model provides a framework for distinguishing between messages authored by individuals and those generated by AI. This capability could be instrumental in combating spam, identifying phishing attempts, and generally improving the authenticity of digital correspondence. As AI-generated content becomes more sophisticated, the ability to filter and verify the source of messages will become crucial for security and efficiency.

Moreover, the methodology could be extended to other domains where text authenticity is essential, such as in legal documents, journalistic sources, and online content creation. By adapting our models to specific linguistic features and patterns within these domains, we can contribute to upholding the integrity and trustworthiness of information across various platforms.

In conclusion, the project not only demonstrates the technical feasibility of using BERT models with dropout for text classification tasks but also opens avenues for practical implementations that address pressing challenges in the digital age. As AI continues to advance, tools like ours will become vital in maintaining the veracity of human communication and creation.

Conclusion

In conclusion, our project stands as a testament to the innovative use of machine learning to address the contemporary challenge of discerning human from ChatGPT-generated text. Through a rigorous process of data collection across various subreddits, prompt generation via the advanced ChatGPT API, and the strategic application of a 'Question-vs-Statement Classifier,' we have curated a rich dataset that reflects the complexities of natural language. The models employing 20% dropout and those without it yielded comparable outcomes. However, opting for the model with a 20% dropout rate is advisable as it better addresses the issue of overfitting. This consideration is crucial for the robustness and generalizability of the model, making it the preferred choice for implementation. This model not only exhibits superior performance across standard metrics but also promises significant practical utility. It offers a reliable solution for educational institutions seeking to uphold academic integrity and for digital communication platforms aiming to filter out AI-generated content. Our approach, thus, sets a benchmark for leveraging AI in maintaining the authenticity of text, which is crucial in this era of burgeoning AI-generated content.

Future Work

In future work, we aim to broaden the scope of our dataset, incorporating an even wider array of sources beyond the subreddits already utilized. This expansion would ensure a richer diversity of questions, capturing a broader spectrum of linguistic styles and topics, which can further refine the model's accuracy. We can employ the same approach for a subject by training the model on its previous submissions and questions and use it to detect if the future submissions are generated by GPT. This can also be integrated into emails and other areas of interest where this distinction is necessary. Exploring alternative deep learning architectures, such as GPT-3 or transformer-based models with different configurations, might provide new insights and enhance the model's ability to generalize across varied textual datasets.

Moreover, the implementation of our current methodology into practical applications warrants exploration. Educational institutions could adopt this model to detect AI-assisted cheating, thereby reinforcing academic honesty. Email systems could utilize it to filter out sophisticated spam that mimics human writing, thus maintaining the integrity of digital communication. By adapting our approach to these real-world applications, we could effectively mitigate the growing challenges posed by the sophistication of AI-generated content, ensuring that the digital exchange of information remains authentic and reliable.