Abstract:

Distinguishing between texts generated by machines and those written by humans has become more difficult due to the swift progress made in natural language generation techniques, especially with big language models. This work uses various big pre-trained language models (LLMs) to handle the challenge of machine-generated text detection. Transparency, trust, and the avoidance of dishonest practices in online interactions and content are the driving forces behind this research. Furthermore, identifying machine-generated text sheds light on the shortcomings of the language models in use today and aids in the fight against the growing threat posed by deep fakes. To do this, a balanced collection of texts produced by machines and humans is gathered from a variety of publically accessible sources. Tokenizing the sentences and eliminating superfluous characters are two preprocessing steps for the gathered data. The dataset is divided into train, validation, and test sets, and labels designating whether the data came from humans or machines are applied. The primary architecture that the authors suggest using is based on the BERT and RoBERTa base models, enhanced with a classification layer to estimate the likelihood of machine-generated text. The models undergo end-to-end training using the labeled data, and their hyperparameters are adjusted to achieve peak performance. The test set is evaluated, and standard categorization metrics are reported. Error analysis is used to find the models' biases, constraints, and failure modes. Potential enhancements such as data augmentation are then considered. By precisely identifying machine-generated text, the suggested method seeks to improve the transparency and reliability of online interactions and information.

Introduction:

Significant progress has been made in the field of natural language processing (NLP) in recent years, especially with the creation of potent artificial intelligence (AI)-based language models. These language models—like OpenAI's GPT-3—have shown remarkable promise in producing writing that is human-like, opening up fascinating new possibilities for use in customer service, virtual assistants, content creation, and other fields. However the rise of machine-generated content has also sparked worries about how it may be abused and the necessity to guarantee the veracity and authenticity of textual data. To address this issue, scientists are investigating several methods for identifying text written by machines. Examining the text's statistical characteristics and linguistic patterns is one such technique. Writing generated by machines frequently possesses specific traits that set it apart from writing written by humans. AI-generated writing can, for instance, be inconsistent, have recurrent phrases, or have an odd word or grammatical structure distribution. Researchers can create models and algorithms that efficiently recognize machine-generated text by utilizing these patterns.

Using other knowledge sources is another strategy. Machine-generated writing frequently produces inaccurate or incomprehensible information because it lacks real-world understanding. The text can be examined for inconsistencies and errors that point to machine-generated content by comparing it to external knowledge bases or fact-checking databases.

Moreover, adversarial methods are being investigated to improve the recognition of text written by machines. In adversarial training, one AI model is trained to produce text by machine while another is trained to distinguish between text that is generated by machines and text that has been written by humans. By going through this iterative process, the models can produce more realistic text and the detection model becomes more proficient at differentiating between the two.

Given the ongoing evolution and improvement of AI models, the field of detection technique development is dynamic. Scholars continuously modify their approaches to stay up to date with the developments in machine learning. In addition, politicians, industry professionals, and researchers must work together to address the moral dilemmas and the dangers posed by the improper use of machine-generated content.

One of the most important areas of natural language processing study is the detection of machine-generated text. Our ability to recognize and classify machine-generated material will help us fight false information, safeguard data integrity, and guarantee the reliability of textual information. As long as this subject continues to progress, we will be able to use AI language models responsibly and ethically, maximizing their advantages while reducing any potential concerns.

Literature Review:

First, the paper talks about a work by Sadasivan et al. (2023) [1], which suggests a way to use lightweight neural paraphrasers like T5 and PEGASUS to undertake paraphrasing assaults to identify machine-generated material. The authors empirically determine the overall variation distance and analyze the overlap between the human and AI text distributions. They also create spoofing assaults with human materials that have been paraphrased and watermarked. The findings emphasize the need for more study in AI text authentication by demonstrating how successful these assaults are against modern detectors.

Alamleh et al. (2023) [2] have conducted another study that aims to differentiate text produced by ChatGPT, a well-known AI language model, from text that is handwritten by humans. The authors assess eleven machine learning models, including deep learning and conventional models, using a dataset of responses from computer science students. For many kinds of prompts, the Random Forest (RF) and Support Vector Machine (SVM) models exhibit excellent accuracy. The acknowledgment of constraints about the size of the dataset and the feature extraction approach highlights the necessity for more extensive and varied datasets as well as improved feature selection techniques.

Parallel to this, Katib et al. (2023) [3] suggest using the Long Short-Term Memory Recurrent Neural Network (TSA-LSTM RNN) model in conjunction with the Tunicate Swarm Algorithm to distinguish between text produced by ChatGPT and text produced by humans. The authors use a variety of feature extraction methods, including countvectorizer, word embedding, and TF-IDF, in conjunction with the LSTM RNN model for detection and classification. The outcomes show that the TSA-LSTM RNN system outperforms other current techniques in distinguishing between text generated by ChatGPT and human language.

Moreover, Desaire et al. (2023) [4] address the difficulty of precisely identifying text written by AI when ChatGPT is explicitly instructed to write in the style of a chemist. The authors provide a technique that trains an XGBoost machine learning model for classification using text features that are taken out of paragraphs. The outcomes demonstrate excellent accuracy in differentiating between paragraphs composed by humans and those created by AI, even in situations when ChatGPT deliberately tries to fool detectors. However, restrictions on the method's applicability to scientific writing and the paucity of published specifics are underlined, indicating the need for more investigation and openness.

A work by Vygon and Mikhaylovskiy (2023) [5], combines triplet loss-based embeddings and kNN classification to target keywords in speech data. Despite not having anything to do with text, this study emphasizes how crucial efficient feature extraction and classification methods are for finding particular patterns and traits in machine-generated data. The outcomes show how successful the suggested strategy is, producing cutting-edge outcomes on benchmark datasets. Nonetheless, the acknowledgment of limitations concerning the size and scope of the dataset indicates the necessity for additional investigation and enlargement of the study.

Dataset:
The Machine-Generated Text Detection Dataset is a comprehensive collection designed for training and evaluating models focused on discerning between human-generated and machine-generated texts. The dataset comprises 119,756 text samples, meticulously curated to ensure diversity and balance in terms of sources, genres, and origins.
The dataset is evenly split between human-generated texts and machine-generated texts, featuring 63,351 samples of the former and 56,406 samples of the latter. This balanced distribution is crucial for robust model development and unbiased evaluation.
Also, the dataset contains another 5000 text samples for validation purpose.

Proposed Methodology:

Model Architecture:
- We will use the BERT and RoBERTa base model as our main architecture. BERT is pre-trained on a large corpus of text using masked language modeling which allows it to learn contextual relations between words. This makes it suitable for our task.
- On top of BERT, we will add a classification layer with sigmoid activation to output a probability of the text being machine-generated or human-generated.
- We will create a hybrid model consisting of LSTM, CNN and SVM/SGD which will be trained on the dataset to achieve our goal.

Data Preprocessing:
During the preprocessing stage of the data, we select a well-balanced dataset from SemEval that includes texts produced by both machines and humans. We capture intricate linguistic structures by breaking texts down into fine-grained word parts using advanced tokenization with the BERTTokenizer. The labels are encoded (0 for human-generated, 1 for machine-generated), and a representative split of the training (95,806 samples) and testing (22,950 samples) sets is ensured by a stratified 80-20 split. To reduce any biases and create a solid and credible dataset for the construction of machine learning models later on, shuffling occurs during the split.

Error Analysis:
- Analyze examples where model predictions differ significantly from true labels.
- Identify failure modes and biases if any based on text content or generator.
- Perform additional experiments like data augmentation to address identified issues.

References:

[1] Sadasivan, V.S. et al. (2023). "Can ai-generated text be reliably detected?" arXiv.org. Available at: https://arxiv.org/abs/2303.11156

[2] Alamleh, H. et al. (2023). "Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning." 2023 Systems and Information Engineering Design Symposium (SIEDS).

[3] Katib, I. et al. (2023). "Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning." Mathematics, 11(15), 3400.

[4] Desaire, H. et al. (2023). "Accurately Detecting AI Text when ChatGPT is Told to Write like a Chemist." arXiv.org. Available at: https://arxiv.org/abs/2304.25683

[5] Vygon, A. and Mikhaylovskiy, A. (2023). "Learning Efficient Representations for Keyword Spotting with Triplet Loss." arXiv.org. Available at: https://arxiv.org/abs/2305.18942