Motivation:

As natural language generation techniques like large language models continue to advance rapidly, it has become increasingly difficult to determine whether a given text was produced by a human or machine. The ability to convincingly imitate and generate human-like language at scale opens up opportunities for beneficial applications but also enables harmful deception if not properly attributed or regulated.

Detection of machine-generated text is an important problem with several practical applications. Firstly, it is crucial to ensure transparency and build trust in online content and interactions. Users interacting with AI systems need to be aware of when they are communicating with an algorithm instead of a person. Secondly, detection helps address the rising threat of "deepfakes" - synthetic text, images, audio, and video that aim to mislead or manipulate by disguising the artificial origins. Timely identification of automatically generated propaganda and misinformation can help curb their negative societal impact.

From a research perspective, distinguishing human and machine text provides insights into the subtle differences between natural human language and what current language models are capable of producing. As AI systems become more proficient at language generation, detection performance can act as a benchmark for measuring progress in human-level language understanding. Analyzing what linguistic patterns models struggle with also guides the development of more human-centric models.

Given the potential harms of deceptive text generation at scale and the importance of transparent human-AI interaction, there is a critical need to advance techniques that can determine when language originates from a person versus an algorithm. This work aims to address this important challenge by leveraging large pre-trained language models to accurately detect machine-generated text.

Workflow:

1. Data Collection
   - Collect a balanced dataset of human-written and machine-generated texts from publicly available sources like news articles, social media posts, forums, blogs, etc. Ensure the machine texts are generated using state-of-the-art generative models.
   - The dataset should have enough samples (tens of thousands) to properly train deep learning models. Diversity in data sources also helps generalizability.

2. Data Preprocessing
   - Clean raw texts by removing unnecessary characters, formatting inconsistencies etc.
   - Tokenize the texts into word pieces using a tokenizer like Word2vec, TF*IDF
   - Encode labels (0 for human, 1 for machine) and split the preprocessed data into train, validation, and test sets stratified on the labels.

3. Model Training
   - Choose pre-trained language models like BERT, RoBERTa that have been shown to be effective on text classification tasks.
   - Add a classifier head on top and initialize with pre-trained weights.
   - Train the model end-to-end on the labeled train set while monitoring loss and accuracy on the validation set.
   - Tune hyperparameters like learning rate, batch size, and number of epochs for best validation performance.
   - Create a hybrid model using LSTM, CNN and SVM/SGD in order to detect machine-generated texts.

4. Model Evaluation
   - Evaluate the trained model on the held-out test set.
   - Report standard classification metrics like accuracy, precision, recall, F1-score.

5. Error Analysis
   - Analyze examples where the model predictions differ from true labels.
   - Identify any biases or limitations captured by the current model and data.
   - Try variations like data augmentation, different models, or ensemble approaches to address failure modes.

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.

Dataset Composition:
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.

Proposed Methodology:

Model Architecture:
- We will use 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 in order achieve our goal.

Data Preprocessing:
- Collect a balanced dataset of human and machine-generated texts from SemEval
- Tokenize the texts into word pieces using Tokenizers.
- Encode the labels as 0 for human and 1 for machine-generated text.
- Further split the dataset into training, validation, and test sets stratified on the labels to avoid data leakage.

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.