NLP Project handout

**Team Members:**

* Rotem Kashani 209073352
* Gal Ein Dor 209070671
* David Koplev 208870279

**Project Description, Data, and Basic Analysis:**

**Project Overview:** The project aims to create a smart system capable of answering questions or providing relevant information from written text. Two methods are utilized:

1. **Seq2Seq LSTM Model:** This model is trained on the AG News Classification Dataset to address queries across various topics.
2. **GPT-2 Fine-tuning:** OpenAI's ChatGPT model undergoes fine-tuning on the same dataset to enhance accuracy in question answering.

**Data Description:**

1. **Dataset:** AG News Classification Dataset
2. **Topics:** News articles are categorized into four topics: World, Sports, Business, and Sci/Tech.
3. **Attributes:** Each article consists of a Title and Description.
4. **Preprocessing:** Data preprocessing involves standard techniques such as tokenization, lemmatization, and removal of English stop-words and special characters.

**Basic Analysis:**

The dataset consists of news articles categorized into four topics.

Preprocessing includes cleaning the text data by lowercasing the text, removing the stop-words, removing special characters, and lemmatizing the text words.

Data is split into train and test sets for model training and evaluation.

**Model Description, Metric, and Results:**

**Seq2Seq LSTM Model:**

**Architecture:** The Seq2Seq LSTM model consists of two main components: an encoder and a decoder.

1.Encoder: Utilizes a Long Short-Term Memory (LSTM) layer to process the input sequence (Title and Description) and encode it into a fixed-size context vector.

2.Decoder: Also employs an LSTM layer, which takes the context vector from the encoder as input and generates the output sequence (predicted Description) step by step.

**Training Procedure:** The model is trained using a sequence-to-sequence learning paradigm, where the input sequence (Title) is mapped to the target sequence (Description). Cross-entropy loss is employed as the optimization objective to minimize the discrepancy between the predicted and actual sequences. The model parameters are updated using the Adam optimizer, which adapts the learning rate for each parameter individually. During training, teacher forcing is utilized, where the decoder is fed the true target sequence at each time step, aiding in faster convergence.

**Results:** The model is trained for multiple epochs, with the loss decreasing gradually over time. After training, the model's performance is evaluated on a separate test set to assess its ability to accurately predict descriptions for given titles.

**GPT-2 Fine-tuning:**

**Architecture:** GPT-2 model is OpenAI's GPT (Generative Pre-trained Transformer) architecture,it is a transformer-based model consisting of a multi-layer stack of self-attention mechanisms, enabling it to generate coherent and contextually relevant text. Fine-tuning involves loading a pre-trained GPT-2 model and adapting its parameters to a specific task, in this case, question answering.

**Training Procedure:** The pre-trained GPT-2 model is fine-tuned on the AG News Classification Dataset using a language modeling objective. During fine-tuning, the model learns to generate responses (answers) based on the context provided (Title and Description of news articles). The parameters of the GPT-2 model are fine-tuned using backpropagation with an appropriate optimizer (e.g AdamW) and learning rate scheduler.

**Results:** After fine-tuning, the performance of the GPT-2 model is evaluated on a test set to measure its accuracy in providing relevant answers to questions posed about news articles. The fine-tuned model's responses are compared against ground truth answers to assess its effectiveness in question answering tasks.

**Comparison and Evaluation:** Both models are evaluated based on their accuracy in answering questions and providing relevant information from written text. The Seq2Seq LSTM model and GPT-2 fine-tuned model undergo separate training and evaluation processes, with their respective architectures and training procedures optimized for their specific tasks. After training and fine-tuning, the models' performances are compared to determine which one achieves higher accuracy and effectiveness in the question answering task.

**Metric Chosen:**

Accuracy: Measured for both Seq2Seq LSTM and GPT-2 models to evaluate their performance in question answering.

**Future Directions:**

1. **Model Improvement:** Experiment with different architectures and hyperparameters for the Seq2Seq LSTM model. Fine-tune larger language models like GPT-3 for improved accuracy in ChatGPT.

2. **Data Augmentation:** Augment the dataset with additional news articles to increase model generalization. Explore methods like back-translation for generating diverse training samples.

3. **Attention Mechanisms:** Implement attention mechanisms in the Seq2Seq LSTM model to improve context understanding. Experiment with attention-based approaches in ChatGPT for better relevance in responses.

4. **Ensemble Methods:** Combine predictions from multiple models for enhanced accuracy and robustness. Implement ensemble techniques like stacking or blending to leverage the strengths of individual models.

5. **Deployment and Integration:** Integrate the models into a user-friendly interface for easy access and interaction. Deploy the system for real-world applications such as news recommendation or question answering platforms.

By implementing these future directions, the smart system can achieve higher accuracy and efficiency in answering questions and providing relevant information from written text.

**Conclusion**

In conclusion, the project encompasses the development and evaluation of Seq2Seq LSTM and GPT-2 models for question answering and information retrieval tasks. Through rigorous training, evaluation, and future improvements outlined, the system aims to provide accurate and efficient responses to user queries, thereby enhancing user experience and utility.