I am currently working on aspect-based sentiment analysis of restaurant reviews. For example: "The pizzas we ordered were excellent, although I wouldn't say the portions were big enough." If I condition on the food quality of the entity "pizzas", this review would be positive ("pizzas[...] were excellent") but conditioning on the food quantity, the review would be negative ("wouldn't say the portions were big enough"). Which model architecture would you recommend me to use to perform this aspect-based sentiment analysis for the given target terms?

Aspect-based sentiment analysis (ABSA) is a subtask of sentiment analysis that aims to identify sentiment towards specific aspects or attributes of a given entity, such as food quality or service in a restaurant review. To tackle this problem, you can consider the following model architectures:

1. **BERT**: BERT (Bidirectional Encoder Representations from Transformers) is a powerful pre-trained language model that can be fine-tuned for various NLP tasks, including aspect-based sentiment analysis. You can fine-tune a BERT model on your specific dataset by training it to classify sentiment for each target term.
2. **XLNet**: Similar to BERT, XLNet is another transformer-based pre-trained language model that can be fine-tuned for ABSA tasks. It addresses the limitations of BERT's bidirectional context representation by using a permutation-based approach, potentially yielding better results.
3. **Aspect-based CNN (AB-CNN)**: This is a convolutional neural network (CNN) architecture specifically designed for ABSA tasks. The model uses two parallel CNNs to capture both target and context information. You can train this model on your dataset to predict sentiment for target terms.
4. **LSTM with Attention**: Long Short-Term Memory (LSTM) networks are a type of recurrent neural network (RNN) that can capture long-range dependencies in the text. By combining LSTM with an attention mechanism, you can allow the model to focus on relevant parts of the input text for a specific aspect while predicting sentiment.
5. **Hierarchical Attention Networks (HAN)**: HAN is another model that combines LSTM and attention mechanisms. It employs a hierarchical structure, capturing information at both the word and sentence levels. This structure allows it to model both local and global context, which can be particularly helpful for ABSA tasks.

Choose the model that best suits your needs and computational resources. BERT, XLNet, and other transformer-based models are likely to yield the best results but may require more resources for fine-tuning and inference. CNN or LSTM-based approaches may be more computationally efficient but could deliver slightly lower performance.

Regardless of the architecture you choose, remember to fine-tune the model on a labeled dataset that includes the target terms and their corresponding sentiment labels, so it can learn to predict sentiment for those specific aspects.

Given that I use BERT, what should I pass to the model as input?

To use BERT for aspect-based sentiment analysis, you will need to preprocess the input text and target term(s) in a way that BERT understands. Here's how you can prepare the input:

1. **Tokenization**: Tokenize both the review text and the target term(s) using the BERT tokenizer. This will convert the text into subword tokens compatible with BERT's vocabulary.
2. **Special tokens**: Add the special tokens [CLS] and [SEP] to the input. The [CLS] token should be placed at the beginning of the input sequence, while the [SEP] token should be inserted after the review text and after the target term(s). For example, if the review is "The pizzas we ordered were excellent, although I wouldn't say the portions were big enough." and the target term is "pizzas", the input should look like:

[CLS] The pizzas we ordered were excellent, although I wouldn't say the portions were big enough. [SEP] pizzas [SEP]

1. **Segment IDs**: Create segment IDs to differentiate between the review text and the target term. Assign a segment ID of 0 to the tokens in the review text and 1 to the tokens in the target term. In our example:

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1

1. **Position IDs**: Generate position IDs for the input tokens, ranging from 0 to the maximum sequence length minus 1. Position IDs allow the model to understand the position of each token in the input sequence.
2. **Padding**: Pad the input sequence with [PAD] tokens to reach the maximum sequence length used during training. Also, pad the segment IDs and position IDs accordingly with 0s.
3. **Attention mask**: Create an attention mask for the input sequence. The attention mask is a binary vector that indicates whether a token is a [PAD] token (0) or not (1).

Once you've prepared the input sequence, including tokens, segment IDs, position IDs, and attention mask, you can feed them into the BERT model to obtain the sentiment prediction for the given target term. You will need to fine-tune the BERT model on your aspect-based sentiment analysis dataset beforehand to ensure it can make accurate predictions for your specific task.

And what should I pass as model input to Roberta?

RoBERTa is an optimized version of BERT and follows a similar input structure. However, there are some differences in tokenization and special tokens. To prepare input for RoBERTa, follow these steps:

1. **Tokenization**: Tokenize both the review text and the target term(s) using the RoBERTa tokenizer. This will convert the text into subword tokens compatible with RoBERTa's vocabulary.
2. **Special tokens**: Add the special tokens <s> and </s> to the input. The <s> token should be placed at the beginning of the input sequence, while the </s> token should be inserted after the review text and after the target term(s). For example, if the review is "The pizzas we ordered were excellent, although I wouldn't say the portions were big enough." and the target term is "pizzas", the input should look like:

<s> The pizzas we ordered were excellent, although I wouldn't say the portions were big enough. </s> pizzas </s>

1. **Segment IDs**: RoBERTa does not use segment IDs, so you can skip this step.
2. **Position IDs**: Generate position IDs for the input tokens, ranging from 0 to the maximum sequence length minus 1. Position IDs allow the model to understand the position of each token in the input sequence.
3. **Padding**: Pad the input sequence with <pad> tokens to reach the maximum sequence length used during training. Also, pad the position IDs accordingly with 0s.
4. **Attention mask**: Create an attention mask for the input sequence. The attention mask is a binary vector that indicates whether a token is a <pad> token (0) or not (1).

Once you've prepared the input sequence, including tokens, position IDs, and attention mask, you can feed them into the RoBERTa model to obtain the sentiment prediction for the given target term. Remember to fine-tune the RoBERTa model on your aspect-based sentiment analysis dataset before using it to make accurate predictions for your specific task.

Given that I use Roberta, what can I do if my input sentence contains the term "pizzas" multiple times, but I want to extract the sentiment for only one of them?

In such cases, you can use an approach that highlights the specific occurrence of the target term you're interested in by introducing an additional token. This token will act as a marker to help the model focus on the desired occurrence of the target term. Here's how you can do this:

1. Choose a unique token that does not appear in your dataset. For instance, you can use <tgt> as the marker token.
2. Insert the marker token before the specific occurrence of the target term you want to analyze. For example, if you want to analyze the second occurrence of "pizzas" in the sentence "I love pizzas, but the pizzas at this place were not good.", you would modify the sentence to:

"I love pizzas, but the <tgt> pizzas at this place were not good."

1. Preprocess the input sentence as you would for a regular RoBERTa input (tokenization, adding special tokens, position IDs, padding, and attention mask). Include the target term with the marker token in the input:

<s> I love pizzas, but the <tgt> pizzas at this place were not good. </s> <tgt> pizzas </s>

1. Feed the modified input sequence into the fine-tuned RoBERTa model to obtain the sentiment prediction for the specific occurrence of the target term.

By introducing a marker token, the model can better understand which occurrence of the target term you're interested in and predict the sentiment accordingly. Make sure to fine-tune your RoBERTa model with such modified inputs, where the marker token is used to emphasize the target occurrence during the training process.

and what if my target can include multiple words?

If your target term consists of multiple words, you can still use a similar approach with the marker token. Here's how you can handle multi-word targets with RoBERTa:

1. Choose a unique token that does not appear in your dataset. For instance, you can use <tgt> as the marker token.
2. Insert the marker token before the specific occurrence of the multi-word target term you want to analyze. For example, if you want to analyze the sentiment towards "vegetarian pizza" in the sentence "The vegetarian pizza was delicious, but I didn't like the cheese pizza.", you would modify the sentence to:

"The <tgt> vegetarian pizza was delicious, but I didn't like the cheese pizza."

1. Preprocess the input sentence as you would for a regular RoBERTa input (tokenization, adding special tokens, position IDs, padding, and attention mask). Include the target term with the marker token in the input:

<s> The <tgt> vegetarian pizza was delicious, but I didn't like the cheese pizza. </s> <tgt> vegetarian pizza </s>

1. Feed the modified input sequence into the fine-tuned RoBERTa model to obtain the sentiment prediction for the specific multi-word target term.

By introducing a marker token, the model can better understand which occurrence of the multi-word target term you're interested in and predict the sentiment accordingly. Make sure to fine-tune your RoBERTa model with such modified inputs, where the marker token is used to emphasize the target occurrence during the training process.

and how can I add different aspect categories to this input? for example, if I want to differentiate between the sentiment towards the food quality and the food quantity?

To differentiate between aspect categories such as food quality and food quantity, you can modify the input format by adding an aspect category token before the target term. This token will help the model recognize which aspect you're interested in when making sentiment predictions. Here's how to do this:

1. Choose unique tokens for each aspect category that do not appear in your dataset. For instance, you can use <quality> for food quality and <quantity> for food quantity.
2. Insert the marker token <tgt> and the corresponding aspect category token before the specific occurrence of the target term you want to analyze. For example, if you want to analyze the sentiment towards the food quality of "pizzas" in the sentence "The pizzas we ordered were excellent, although I wouldn't say the portions were big enough.", you would modify the sentence to:

"The <tgt> <quality> pizzas we ordered were excellent, although I wouldn't say the portions were big enough."

1. Preprocess the input sentence as you would for a regular RoBERTa input (tokenization, adding special tokens, position IDs, padding, and attention mask). Include the target term with the marker token and the aspect category token in the input:

<s> The <tgt> <quality> pizzas we ordered were excellent, although I wouldn't say the portions were big enough. </s> <tgt> <quality> pizzas </s>

1. Feed the modified input sequence into the fine-tuned RoBERTa model to obtain the sentiment prediction for the specific target term and aspect category.

To analyze the sentiment towards the food quantity of "pizzas", you would use the <quantity> token:

<s> The <tgt> <quantity> pizzas we ordered were excellent, although I wouldn't say the portions were big enough. </s> <tgt> <quantity> pizzas </s>

Make sure to fine-tune your RoBERTa model with modified inputs, where the marker token and aspect category tokens are used to emphasize the target occurrence and aspect category during the training process. This will help the model learn to recognize different aspects and make accurate sentiment predictions accordingly.

could I write one natural language question and place it at the beginning of the input sequence instead of using the aspect category token?

Yes, you can use a natural language question to specify the aspect category in the input sequence. This approach can help guide the model to focus on the particular aspect you're interested in when predicting sentiment. Here's how to do this:

1. Write a natural language question that represents the aspect category you want to analyze. For example, if you want to analyze the sentiment towards the food quality of "pizzas", you can use the question "What do you think about the quality of the pizzas?".
2. Insert the marker token <tgt> before the specific occurrence of the target term you want to analyze. For example, if you want to analyze the sentiment towards the food quality of "pizzas" in the sentence "The pizzas we ordered were excellent, although I wouldn't say the portions were big enough.", you would modify the sentence to:

"The <tgt> pizzas we ordered were excellent, although I wouldn't say the portions were big enough."

1. Preprocess the input sentence as you would for a regular RoBERTa input (tokenization, adding special tokens, position IDs, padding, and attention mask). Place the natural language question at the beginning of the input sequence and include the target term with the marker token in the input:

<s> What do you think about the quality of the pizzas? </s> The <tgt> pizzas we ordered were excellent, although I wouldn't say the portions were big enough. </s> <tgt> pizzas </s>

1. Feed the modified input sequence into the fine-tuned RoBERTa model to obtain the sentiment prediction for the specific target term and aspect category.

Make sure to fine-tune your RoBERTa model with modified inputs, where the natural language question is used to specify the aspect category during the training process. This will help the model learn to recognize different aspects based on the questions and make accurate sentiment predictions accordingly.

What marker token would you suggest for BERT?

For BERT, you can use a similar approach as with RoBERTa by introducing a unique marker token that does not appear in your dataset. The choice of token depends on your specific dataset and the pre-trained model's vocabulary.

One option is to use a token that is already in the BERT vocabulary but is unlikely to appear in your dataset, such as [unusedN], where N is a number between 1 and 99, e.g., [unused1].

For example, if you want to analyze the sentiment towards "pizzas" in the sentence "The pizzas we ordered were excellent, although I wouldn't say the portions were big enough.", you would modify the sentence to:

"The [unused1] pizzas we ordered were excellent, although I wouldn't say the portions were big enough."

During the fine-tuning process, you should incorporate this marker token in your training dataset so that the model can learn to recognize it when predicting sentiment for specific occurrences of a target term.

If you decide to use a custom token not present in the pre-trained BERT's vocabulary, you will need to modify the tokenizer and the pre-trained model accordingly to include the new token. However, this approach requires more effort and can impact the model's performance, so using an existing token like [unusedN] is generally recommended.

What do you think about these questions? (python dictionary structure where key denotes the aspect category and value denotes the question)

Are they written well for Roberta or can they be optimized? If so, please write down your recommended questions in the same structure (using the same aspect categories as keys).

aspect\_in\_nl = { "AMBIENCE#GENERAL": "How is the general ambience of the $T$?", "DRINKS#PRICES": "How are the prices of the drinks called $T$?", "DRINKS#QUALITY": "How is the quality of the drinks called $T$?", "DRINKS#STYLE\_OPTIONS": "How are the style options of the drinks called $T$?", "FOOD#PRICES": "How are the prices of the food called $T$?", "FOOD#QUALITY": "How is the quality of the food called $T$?", "FOOD#STYLE\_OPTIONS": "How are the style options of the food called $T$?", "LOCATION#GENERAL": "How is the location of the restaurant called $T$?", "RESTAURANT#GENERAL": "How is the restaurant called $T$ overall?", "RESTAURANT#MISCELLANEOUS": "How is the restaurant's $T$?", "RESTAURANT#PRICES": "How are the prices of the restaurant called $T$?", "SERVICE#GENERAL": "How is the service called $T$?", }

The provided questions seem well-structured for the most part. However, some of them can be slightly rephrased to sound more natural and better suited for the RoBERTa model. Here are the optimized questions with the same aspect categories as keys:

aspect\_in\_nl = {

"AMBIENCE#GENERAL": "What is the ambience like at the $T$?",

"DRINKS#PRICES": "What are the prices like for the drinks named $T$?",

"DRINKS#QUALITY": "What is the quality like for the drinks named $T$?",

"DRINKS#STYLE\_OPTIONS": "What are the style options like for the drinks named $T$?",

"FOOD#PRICES": "What are the prices like for the food named $T$?",

"FOOD#QUALITY": "What is the quality like for the food named $T$?",

"FOOD#STYLE\_OPTIONS": "What are the style options like for the food named $T$?",

"LOCATION#GENERAL": "What is the location like for the restaurant named $T$?",

"RESTAURANT#GENERAL": "What is the overall experience like at the restaurant named $T$?",

"RESTAURANT#MISCELLANEOUS": "What is the $T$ like at the restaurant?",

"RESTAURANT#PRICES": "What are the prices like at the restaurant named $T$?",

"SERVICE#GENERAL": "What is the quality of the service at $T$?",

}

questions should work well with the RoBERTa model. Just make sure to replace the $T$ placeholder with the appropriate target term in the input sequence when constructing inputs for the model.

Should I include only the target term in the questions or also the marker token?

You should include both the target term and the marker token in the questions. Including the marker token helps the model focus on the specific occurrence of the target term you're interested in when predicting sentiment.

When constructing inputs for the model, replace the $T$ placeholder with the marker token followed by the target term. For example, if you want to analyze the sentiment towards the food quality of "pizzas" in the sentence "The pizzas we ordered were excellent, although I wouldn't say the portions were big enough.", and you are using <tgt> as the marker token, the modified question would be:

"How is the quality of the food called <tgt> pizzas?"

Then, construct the input sequence for RoBERTa, including the question, the review text with the marker token before the target term, and the target term with the marker token:

<s> How is the quality of the food called <tgt> pizzas? </s> The <tgt> pizzas we ordered were excellent, although I wouldn't say the portions were big enough. </s> <tgt> pizzas </s>

Including the marker token in the questions helps the model associate the specific target term occurrence in the review text with the aspect category in the question. Make sure to fine-tune your RoBERTa model with such modified inputs, where the natural language question and marker token are used to specify the aspect category during the training process.