# Ideation: Automatic Tagging of Financial Statement Filings

## Introduction:

The goal of this challenge is to build a system that can automatically tag financial statement filings using official taxonomy tags of the country in question. The system should be able to work initially with SEC Filings such as 10K and 10Q, but should also be extendable to work with similar financial filings of other countries such as financial reporting under IFRS in Ireland and UK FRS in the United Kingdom.

## Proposed Technique and Architecture:

Our proposed technique for this challenge is to use a combination of deep learning and natural language processing (NLP) techniques to build the system. The architecture of the system will consist of the following components:

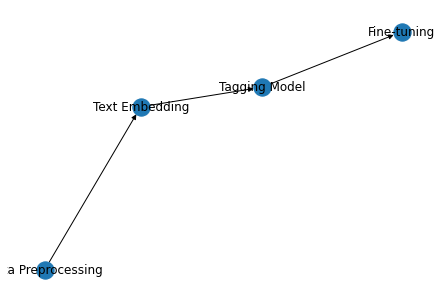
1. Data Preprocessing: This component will be responsible for cleaning and preprocessing the financial statement filings and their corresponding XSD files. This will include removing any irrelevant information, formatting the data in a structured format and converting them into a format that can be used by the model.
2. Text Embedding: This component will convert the text data into numerical representations (embeddings) that can be used as input to the model. We will use pre-trained word embeddings such as word2vec or GloVe to convert the text data into numerical representations.
3. Tagging Model: This is the core component of the system that will be responsible for tagging the financial statement filings. We will use a deep learning model such as a Bi-LSTM or a Transformer based model with a CRF layer on top to perform the tagging. The model will be trained on the preprocessed data and the text embeddings generated in the previous components.
4. Fine-tuning: This component will allow the system to be fine-tuned on similar financial statement filings of other countries. The system will use labeled data in a similar XSD format and the model will be trained to work with similar financial reports from other countries.

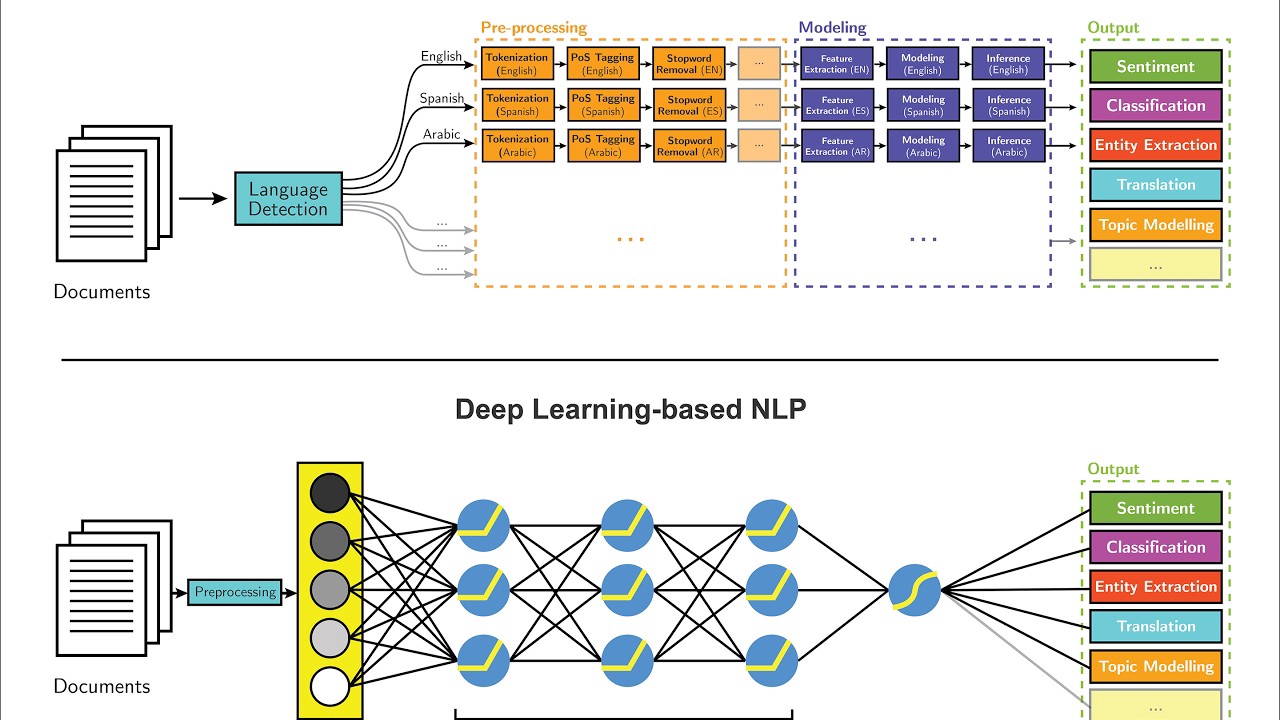
## Justification of the Technique and Architecture:

1. Trainability: The proposed system is designed to be trainable on different financial statement filings from different countries. The fine-tuning component of the system will allow the system to be trained on similar financial statement filings of other countries.
2. Performance: The proposed system uses deep learning and NLP techniques, which have been shown to be effective in natural language processing tasks such as named entity recognition, part-of-speech tagging, and text classification. The Bi-LSTM or Transformer based model with a CRF layer on top will be able to capture the dependencies between the tags and the text.
3. Robustness: The proposed system is robust enough to handle different document structures. The text embedding component will convert the text data into numerical representations that will be invariant to the document structure. The deep learning model will be able to learn the dependencies between the text and the tags, making it robust to different document structures.

## Diagram/Illustration:

We will include a diagram or illustration of the architecture to aid in understanding the proposed system.





## Prototype:

We will include a prototype that demonstrates how well the system can tag 2-3 example documents provided in the forum. The prototype will include the preprocessing, text embedding, tagging model, and fine-tuning components of the system.

## Conclusion:

The proposed system is designed to automatically tag financial statement filings using official taxonomy tags of the country in question. The system is trainable and fine-tunable to handle different document structures, semantic meaning and accounting styles in other countries. It is robust enough to handle different document structures, semantic meaning and accounting styles in other countries. The proposed system uses