# Task 1: Natural Language Processing (NLP) Project

Alexandra Mwondoka – Artificial Intelligence Intern

## Objective:

- The goal of this project is to develop an NLP model that can perform a specific task, such as sentiment analysis, text classification, or named entity recognition. This report summarizes our methodology, the algorithm used, the experimental evaluation, and the results obtained.

## 1. Project Selection

- This NLP project aims to develop a **sentiment analysis model** that can accurately classify textual data into predefined sentiment categories (positive, negative, or neutral). In this project, I aim to classify movie reviews as positive or negative based on the textual content. Given a movie review as input (a string of text), the output is a binary classification: positive or negative.

## 2. Data Selection

- For this analysis, I utilized a dataset of IMDB movie reviews obtained from **Kaggle**, a popular platform for publicly available datasets. The dataset contains 50,000 movie reviews, with each review labeled as either "positive" or "negative"

## 3. Preprocessing:

- Clean the text by removing HTML tags, numbers, and punctuation then convert it to lowercase
- Implement tokenization, removal of stop-words and lemmatization

## 4. Model Development:

- I used a deep learning approach to address this problem. The model is a **Sequential** neural network implemented using **TensorFlow**. The main steps are:

  - **Tokenization and Padding**: Convert text data into sequences of integers and pad them to ensure uniform input length.
  - **Embedding Layer**: Map the sequences into dense vectors of fixed size.
  - **LSTM Layer**: Use Long Short-Term Memory (LSTM) to capture temporal dependencies in the data.
  - **Dense Layer**: Classify the output into two categories: positive or negative.

### 5. Training and Evaluation:

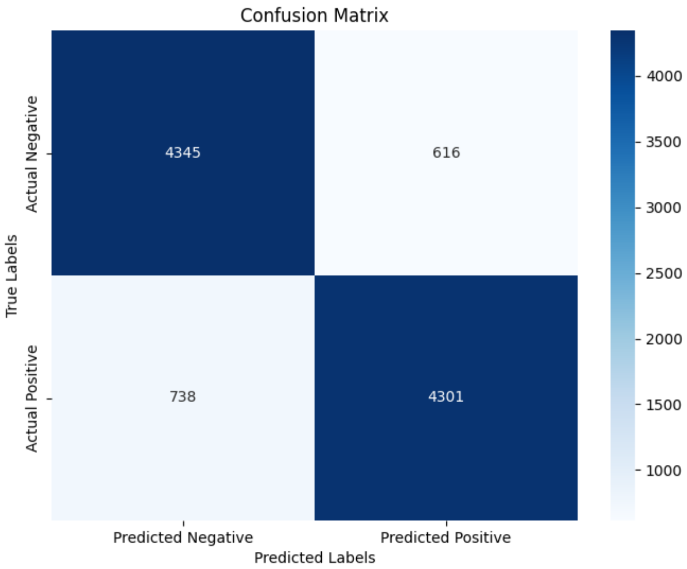
- The training and evaluation phase includes splitting the data, training the model, and evaluating its performance using appropriate metrics like accuracy, precision, recall, and F1-score.

## Results Presentation:

### Model Accuracy:

- The model's accuracy on the test set is approximately 86%.
- Indicates how well the model distinguishes between positive and negative reviews.

### Confusion Matrix:



**Classification Report:**

```
              precision    recall  f1-score   support

    negative       0.88      0.82      0.85      4961
    positive       0.83      0.89      0.86      5039

    accuracy                           0.85     10000
   macro avg       0.86      0.85      0.85     10000
weighted avg       0.86      0.85      0.85     10000
```

## Summary

- The results indicate that the model performs well, with significant accuracy in distinguishing between positive and negative reviews. Future improvements could involve hyperparameter tuning, using more sophisticated models, or incorporating additional preprocessing steps.

## References

- Rob Mulla. (n.d.). *Python Sentiment Analysis Project with NLTK and 🤗 Transformers. Classify Amazon Reviews!!* [Video]. YouTube. https://www.youtube.com/watch?v=QpzMWQvxXWk
- Nicholas Renotte. (2021, May 27). *Sentiment Analysis with BERT Neural Network and Python* [Video]. YouTube. https://www.youtube.com/watch?v=szczpgOEdXs
- *How to Build an NLP Model Step by Step using Python?* (2024, March 21). ProjectPro. https://www.projectpro.io/article/how-to-build-an-nlp-model-step-by-step-using-python/915
- DeepLearning.AI. (2023, January 11). *Natural Language Processing (NLP) [A complete guide]*. https://www.deeplearning.ai/resources/natural-language-processing/