Project 1: Sentiment Classification

Sentiment analysis of news and social media provides important information for investors to help them make informed business decisions. Your task is to design, implement and evaluate a sentiment classifier for financial news headlines and social media posts. For this Project, I will be working with the FiQA Sentiment Analysis dataset, which contains English news headlines and social media posts related to finance.

I will provide a copy of the data and a 'data_loader_demo' Jupyter notebook containing code for loading the data. Each instance in the dataset has a continuous sentiment score from -1 to 1, which my data loader notebook maps to one of three discrete labels: positive (2), negative (0), or neutral (1). Further information about the data is available on the FiQA website: https://sites.google.com/view/fiqa/home. In the project below, you can use any data you wish, including other data provided by FiQA, to build your sentiment classifier.

1.1. Implement and train a method for automatically classifying texts in the FiQA sentiment analysis dataset as positive, neutral, or negative. Refer to the labs, lecture materials and textbook to identify a suitable method. In your report:

1. Briefly explain how your chosen method works and its main strengths and limitations.
2. Describe the preprocessing steps and the features you use to represent each text instance.
3. Explain why you chose those features and preprocessing steps and hypothesise how they will affect your results.
4. Briefly describe your software implementation.

1.2. Evaluate your method, then interpret and discuss your results. Include the following points:

1. Define your performance metrics and state their limitations.
2. Describe the testing procedure (e.g., how you used each split of the dataset).
3. Show your results using suitable plots or tables.
4. How could you improve the method or experimental process? Consider the errors that your method makes.

1.3. Can you identify common themes or topics associated with negative sentiment or positive sentiment in this dataset?

1. Explain the method you use to identify themes or topics.
2. Show your results (e.g., by listing or visualising example topics or themes).
3. Interpret the results and summarise the limitations of your approach.

Suggested length of report for Project 1: 2.5 pages.

Project 2: Named Entity Recognition

As a first step, your goal is to build a tool for named entity recognition in scientific journal article abstracts. I will be working with the BioNLP 2004 dataset of abstracts from MEDLINE, a database containing journal articles from fields including medicine and pharmacy. The data was collected by searching for the terms ‘human’, ‘blood cells’ and ‘transcription factors’, and then annotated with five entity types: DNA, protein, cell type, cell line, RNA.

I provide a cache of the data and code for loading the data in ‘data_loader_demo’

2.1. Design and implement a method for tagging the five types of named entities in the BioNLP 2004 dataset. Refer to the labs, lecture materials and textbook to identify a suitable method. In your report:

1. Explain how your chosen method works and its main strengths and limitations.
2. Briefly explain how entity spans are encoded as tags for each token in a text.
3. Briefly describe your software implementation.
4. Detail the features you have chosen, why you chose them, and hypothesise how your choice. will affect your results.

2.2. Evaluate your method, then interpret and discuss your results. Include the following points:

1. Explain your choice of performance metrics and their limitations.
2. Describe the testing procedure (e.g., how you used each split of the dataset).
3. Show your results using suitable plots and/or tables.
4. How could you improve the method or experimental process? Consider the errors your
5. method makes.

Suggested length of report for Project 2: 2 pages.