UNIVERSITI SAINS MALAYSIA

ASSIGNMENT 2: Natural Language Processing Semester I 2025/2026

Title: Stock trend prediction with news sentiment

Objectives

O1: Construct natural language processing and deep learning components in problems involving prediction, classification and sequence modeling in text and speech.

O2: Design solutions using natural language processing and deep learning techniques for problems in text and speech analytics.

Steps and Deliverables

0. Dataset: stock_trend.csv. The dataset shows the news, the time news is published, the name of the public listed company, the stock quote, the stock price before the news are published, and the price after the news are published.

1. Understanding and prepare the Dataset (30%)

- **Task**:
  - Explain the steps to construct the dataset stock_trend.csv in detail. Note: no need to write the program. (20%)
  - Prepare the data (10%): If relative price changes go up more than 10%, then set to uptrend; if relative price changes go down more than 10%, then set to downtrend; else set to flat. Split the data to 70% training, 20% for validation and 10% for test. Print out 10 lines after completing each substeps.
- **Deliverable**:
  - A report to explain the steps to construct the dataset.
  - Jupyter notebook script for preparing the data.

---

2. Stock trend prediction with LSTM (30%)

- **Task**: Create a suitable model to predict stock trend using LSTM. Apply and extend the codes used in the lab session in CPC353. Train the GloVe word embedding. Handle OOV. Evaluate the performance of the model using accuracy, precision, recall and F1.
- **Deliverable**:

- Jupyter notebook script for training, validation and testing.

---

3. Stock trend prediction with pretrained Transformer model. (33%)

- **Task**: Fine-tuned a suitable model to predict stock trend using a pretrained Transformer model. Note: you are allow to use any instruct-based LLM (Gemini, ChatGPT…) to help you to complete this task. Evaluate the performance of the model using accuracy, precision, recall and F1.
- **Deliverable**:
  - Jupyter notebook script for training, validation and testing.

---

4. Report. (7%)

- **Task**: Explain the models evaluated for stock trend prediction. Analyze the results obtained and made suitable conclusions. Suggest one improvement that you think can improve the performance of the model.
- **Deliverable**:
  - A report.

Note:

This assignment is to be carried out by maximum 3 persons in a group. Submit your Jupyter notebooks and report separately. No need to send the preprocessed data, but you need to show top ten lines of your data used in training, testing in the Jupyter notebook. No copying between groups, copying codes from online or using instruct-based LLM except if permitted.

. Submission deadline: Sunday, 18/1/2026 11.59 pm to elearning.

UNIVERSITI SAINS MALAYSIA

School of Computer Sciences

CPC353 Natural Language Processing

Assignment Assessment Rubric

	Excellent (80-100%)	Good (65-79%)	Moderate (40-64%)	Poor (0-39%)	Total
Part 1a: Understanding (O1: 20%)	The steps are explained correctly.	Some of the steps lack details.	Some steps are incorrectly explained.	All steps are incorrectly explained.	
Part 1b: Preprocessing (O1: 10%)	Dataset is correctly prepared	Some of the steps are incorrect.	-	Dataset is not correctly prepared.	
Part 2: LSTM model (O1: 10%)	An LSTM model using GloVe embedding that are able to handle OOV is correctly trained and tested.	An LSTM model is correctly trained and tested. Word embedding train and tested using GloVe embedding but do not handles OOV.	An LSTM model is correctly trained and tested, but without training and using GloVe embedding and handles OOV.	Do not use the codes from the lab session.	
Part 3: Transformer model (O2: 33%)	A pretrained Transformer model is correctly fine-tuned and tested.	Some steps are missing.	Fine-tuning is not correctly done. Some steps are missing.	Do not use a pretrained Transformer model.	
Part 4: Result and discussion(O1: 7%)	Good analysis and clear report.	Analysis is not extensive.	Language used is poor.	No report.	
				Total	