Assignment 2 report (maximum 5 pages)

Full Name: Jingyun He
Student ID: 530416562

Introduction:
The aim of this study is to build and evaluate the sentiment analysis models on the given dataset. The main goal is to create models that can accurately predict the sentiment of text data.

1. Data Preprocessing (2 points)

In the step of data preprocessing, I applied various preprocessing techniques to clean and prepare the text data:

- Data cleaning
  - Remove punctuation, stopwords, and other special characters from the text.
  - Convert all text to lowercase for consistency.
- Lemmatization
  - Lemmatize the text and reduce the words to their base form, which helps to improve the performance of the models.
- Normalization
  - Use Principal Component Analysis (PCA) to reduce the dimensionality of the dataset and improve computational efficiency.

2. Methodology (6 points)

2.1. Convolutional Neural Network (CNN)

CNN is a supervised deep learning algorithm that uses convolutional layers to identify the local patterns in the input data. It differentiates one from others by assigning importance (weights and biases) to various objects.

2.2. Long Short-Term Memory (LSTM)

LSTM is a supervised deep learning model that uses LSTM layers, a type of recurrent neural network, to capture long-range dependencies in the input data.

2.3. K-Means Clustering

K-Means Clustering is an unsupervised machine learning algorithm that groups the input data into clusters based on similarity.

I also fine-tune the hyperparameters for each model above by techniques like cross-validation, the “Elbow” method, and the “Silhouette” method.

3. Performance, Evaluation & Comparison (8 points)

3.1. Precision

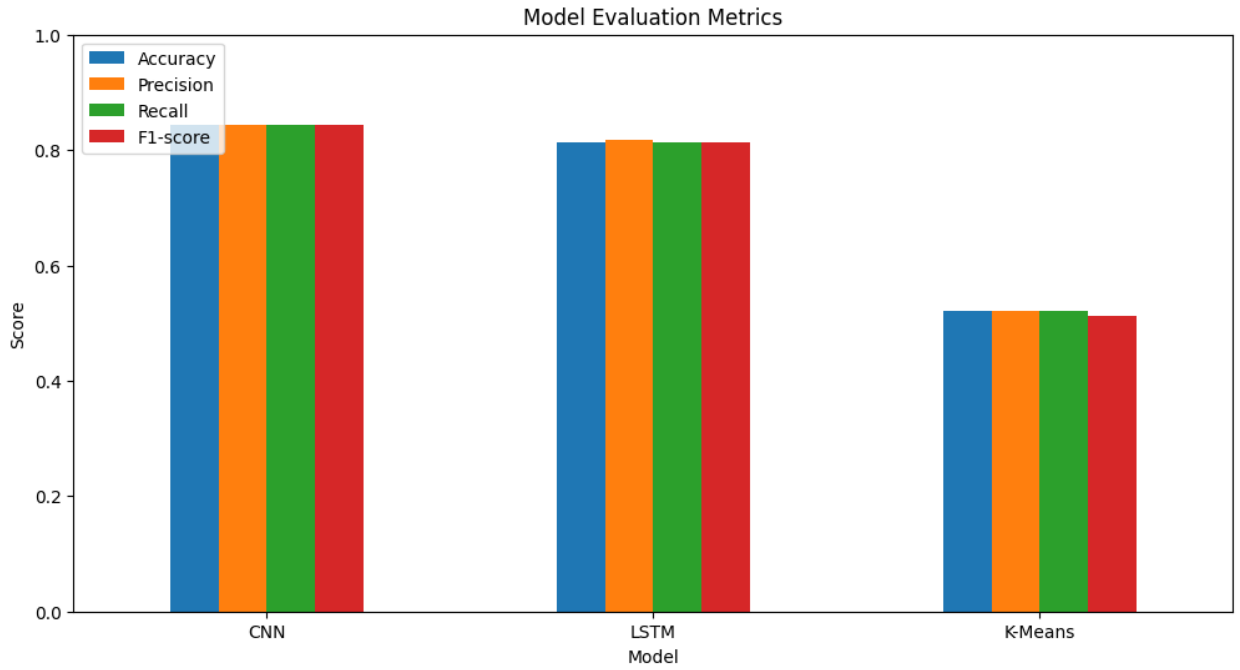
Graph 1 represents the model evaluation metrics, the x-axis represents models, and the y-axis represents scores. Every yellow bar represents the precision score for the corresponding model. Graph 2 is the evaluation result in tabular format. According to these two visualizations, CNN has the highest precision score (0.84), then is LSTM (0.82), and k-means has the lowest precision score (0.52).

3.2. Recall

Graph 1 represents the model evaluation metrics, the x-axis represents models, and the y-axis represents scores. Every green bar represents the recall score for the corresponding model. Graph 2 is the evaluation result in tabular format. According to these two visualizations, CNN has the highest recall score (0.84), then is LSTM (0.81), and k-means has the lowest recall score (0.52).

3.2. F1-score

Graph 1 represents the model evaluation metrics, the x-axis represents models, and the y-axis represents scores. Every red bar represents the F1-score for the corresponding model. Graph 2 is the evaluation result in tabular format. According to these two visualizations, CNN has the highest f1-score (0.84), then is LSTM (0.81), and k-means has the lowest f1-score (0.51).



Graph 1

Model	Accuracy	Precision	Recall	F1-score
CNN	0.84	0.84	0.84	0.84
LSTM	0.81	0.82	0.81	0.81
K-Means	0.52	0.52	0.52	0.51

Graph 2

3.2. Comparison

Based on the visualization Graph1 and Graph2, I observe that the supervised models (CNN and LSTM) outperform the unsupervised model (K-means) in terms of accuracy, precision, recall and f1-score. By these four evaluation methods, CNN has slightly higher scores than LSTM.

4. Conclusion (2 points)

In this report, I have demonstrated the process of building and evaluating sentiment analysis models using various preprocessing techniques (data cleaning, lemmatization and normalization), supervised models (CNN and LSTM), unsupervised models (K-means), and evaluation methods (accuracy, precision, recall and F1-score). The supervised models (CNN and LSTM) perform better than the unsupervised model (K-means) in the experiments.

4. Appendix

Libraries used in the code: numpy, pandas, tensorflow, keras, scikit-learn, seaborn, matplotlib
Install the required libraries using pip:
Copy code

```
pip install numpy pandas tensorflow keras scikit-learn seaborn matplotlib
```