Imam Mohammad Ibn Saud Islamic University

College of Computer and Information Sciences

Computer Science Department

Course Title:	Natural Language Processing
Course Code:	CS365
Course Instructor:	Dr. Fahman Saeed
Project Title:	**Exploring Arabic NLP Tasks - Traditional vs. Modern Approaches**
Semester:	2nd , 24-25
Due Date:	11th week
Marks:	30

Instructions:

Team Structure:

1. Teams of up to Three students are permitted. Collaboration and effective teamwork will be part of the evaluation.

Deadline:

2. The project deadline is the (phase 1: 7th week and phases 2 and 3: 11th week) from the date of assignment.

Technology Requirement:

3. The end output needs to be a **Notebook (.ipynb file) on the Kaggle or Colab** website, and all code needs to be written in **Python**.

Question	CLOs	Question Marks
1	2.1	10
2	2.2	15
3	2.3	5
Total		**30**

Student Name (in English): _____ **Student ID:** _____

Project Overview:
This project explores text classification and summarization on Arabic data from selected Arabic corpora. The project is divided into two phases with a total of 30 marks:

- **Phase 1 (10 marks)**: Implementation using traditional NLP methods
- **Phase 2 (15 marks)**: Implementation using modern deep learning and transformer-based approaches
- **Phase3: Comparative Analysis (5 marks)**: Comprehensive comparison between traditional and modern approaches

Students will compare the effectiveness of traditional versus modern techniques for Arabic NLP tasks, addressing the unique challenges of Arabic language processing.

Team Structure

- Teams of up to **three** students are permitted
- Collaboration and effective teamwork will be part of the evaluation

Deadline

- The project is due by (phase 1: 7th week and phases 2 and 3: 11th week) of the semester

Technology Requirements

- The final deliverable must be a **Jupyter Notebook (.ipynb file) on Kaggle or Colab**
- All code must be written in **Python**
- Total marks: 30 (with potential for 3 bonus marks)

Project Description
Students will select one dataset from the provided options and implement two NLP tasks. For students enrolled in both Digital Image Processing and NLP courses, there are additional multimodal project options available.

Available Datasets

Arabic Text Datasets

1. **KALIMAT Multipurpose Arabic Corpus**
    - Contains newspaper articles across six topics (culture, economy, local news, international news, religion, sports)
    - Access: KALIMAT Corpus on SourceForge
2. **KSUCCA Corpus (King Saud University Corpus of Classical Arabic)**
    - Large corpus of Classical Arabic texts
    - Access: KSUCCA Corpus
3. **AraFacts Corpus**
    - Dataset for Arabic fact-checking and claim verification
    - Access: AraFacts Paper

Multimodal Datasets (For students taking both DIP and NLP courses)

1. **MIMIC-CXR Dataset**
    - A dataset of chest X-rays with associated radiology reports
    - Access: MIMIC-CXR Dataset
2. **PubMed Central Open Access Subset**
    - A dataset of medical articles with images and text
    - Access: PubMed Central Open Access
3. **Flickr8k/Flickr30k Datasets**
    - Datasets of images with captions for image captioning tasks
    - Access: Flickr8k, Flickr30k

4. **MS COCO Captions Dataset**
    - A dataset for image captioning with over 330,000 images and captions
    - Access: MS COCO Captions
5. **Traffic Sign Datasets**
    - German Traffic Sign Recognition Benchmark (GTSRB)
    - Belgium Traffic Sign Dataset
    - Access: GTSRB, BTSD

Available NLP Tasks

Text-Only Tasks

1. **Text Classification**
    - Categorize Arabic text into predefined classes (e.g., topics in KALIMAT)
2. **Text Summarization**
    - Generate concise summaries of Arabic articles

Multimodal Tasks (For students taking both DIP and NLP courses)

1. **Medical Image Analysis Using Images and Text**
    - Combine image analysis with text-based data (e.g., patient reports) to provide a comprehensive diagnostic tool
    - Use preprocessing techniques to enhance both image and text data
2. **Automatic Image Captioning**
    - Generate descriptive captions for images
    - Combine visual features with language modeling
3. **Traffic Sign Recognition with Text Description**
    - Classify traffic signs and generate textual descriptions
    - Combine computer vision with natural language generation

Project Structure

The project is divided into three phases, allowing students to compare traditional and modern approaches:

Phase 1: Traditional NLP Methods (10 marks)

In this phase, students will implement and evaluate traditional machine learning and rule-based approaches for their chosen tasks.

Text Classification with Traditional Methods

- Implement feature extraction techniques:
    - Bag of Words (BoW)
    - Term Frequency-Inverse Document Frequency (TF-IDF)
    - N-gram features
- Apply traditional classifiers:
    - Naive Bayes
    - Support Vector Machines (SVM)
    - Decision Trees or Random Forests
- Implement Arabic-specific preprocessing:
    - Normalization (removing diacritics, standardizing characters)
    - Stemming using Arabic stemmers (e.g., ISRI, Khoja)
    - Stopword removal using Arabic stopword lists

Phase 1 Deliverables

1. **Dataset Preparation and Exploration (2 marks)**
    - Load and explore the chosen dataset
    - Implement appropriate preprocessing for text or multimodal data
    - Visualize dataset characteristics
2. **Traditional Implementation of Task 1 (4 marks)**

- Implement appropriate traditional methods
- Evaluate using relevant metrics
- Analyze results and limitations

3. **Traditional Implementation of Task 2 (4 marks)**
   - Implement appropriate traditional methods
   - Evaluate using relevant metrics
   - Analyze results and limitations

Phase 2: Modern NLP Approaches (15 marks)

In this phase, students will implement deep learning and transformer-based approaches for their chosen tasks(choose two models one from each type).

Task (1) Text Classification with Modern Methods

- Implement deep learning approaches:
  - Recurrent Neural Networks (RNN/LSTM/GRU)
  - Convolutional Neural Networks (CNN) for text
- Implement transformer-based approaches:
  - BERT or multilingual BERT
  - AraBERT or other Arabic-specific pre-trained models
  - Fine-tuning strategies for Arabic text

Task (2) Text Summarization with Modern Methods

- Implement deep learning approaches:
  - Sequence-to-sequence models with attention
  - Pointer-generator networks
- Implement transformer-based approaches:
  - BERT-based extractive summarization
  - T5 or mBART for abstractive summarization
  - Fine-tuning for Arabic summarization

Multimodal Tasks with Modern Methods

- For Medical Image Analysis:
  - Convolutional Neural Networks (CNNs) for image analysis
  - Transformer models for text analysis
  - Multimodal fusion techniques (e.g., attention mechanisms)
- For Image Captioning:
  - CNN-LSTM architectures
  - Transformer-based approaches (e.g., Vision Transformer + GPT)
  - Attention mechanisms for aligning image regions with text
- For Traffic Sign Recognition:
  - Deep CNN architectures for sign classification
  - Transformer models (GPT or DeepSeek) for generating textual descriptions
  - End-to-end multimodal architectures

Phase 2 Deliverables

1. **Modern Implementation of Task 1 (7 marks)**
   - Implement at least one deep learning approach (3 marks)
   - Implement at least one transformer-based approach (3 marks)
   - Evaluate using the same metrics as Phase 1 (1 mark)
2. **Modern Implementation of Task 2 (7 marks)**

- Implement at least one deep learning approach (3 marks)
- Implement at least one transformer-based approach (3 marks)
- Evaluate using the same metrics as Phase 1 (1 mark)

3. **Documentation and Code Quality (1 mark)**
   - Clear, well-structured code with proper comments
   - Reproducible experiments with fixed random seeds

Comparative Analysis (5 marks)

In this section, students will conduct a comprehensive comparison between traditional and modern approaches for both tasks (Note: for **Text Summarization if you don't use traditional method, you can compare CNN vs Transformer models**).

1. **Performance Comparison (2 marks)**
   - Compare traditional vs. modern approaches for both tasks using quantitative metrics
   - Analyze performance differences with statistical significance
   - Create visualizations to illustrate performance differences
2. **Trade-off Analysis (2 marks)**
   - Discuss trade-offs between traditional and modern approaches:
     - Accuracy and performance
     - Computational requirements and training time
     - Model complexity and interpretability
     - Handling of Arabic-specific linguistic features or multimodal challenges
3. **Best Practices and Recommendations (1 mark)**
   - Identify scenarios where each approach excels
   - Provide recommendations for practitioners
   - Suggest potential hybrid approaches that combine strengths of both methods

Interactive Webpage for Model Deployment (Bonus - 3 marks)

- Create a user-friendly interactive webpage (e.g., with Flask or …)
- Allow users to input text and do the task or upload images (for multimodal tasks) and do the task.
- Provide visualization of performance differences between traditional and modern approaches
- Enable real-time comparison between different models

Evaluation Metrics

Text Classification

- Accuracy, Precision, Recall, F1-score
- Confusion matrix
- Classification report

Text Summarization

- ROUGE-1 and BLEU score

Medical Image Analysis

- Classification accuracy for image diagnosis
- BLEU/ROUGE scores for report generation
- Combined metrics for multimodal performance

Image Captioning

- BLEU and ROUGE scores

Project Deliverable Structure
Notebook Structure

1. **Introduction**
    - Project overview and objectives
    - Description of the chosen dataset and tasks
    - Brief background on relevant challenges (Arabic NLP or multimodal processing)
2. **Data Exploration and Preprocessing**
    - Dataset loading and exploration
    - Appropriate preprocessing steps
    - Data visualization and statistics
3. **Phase 1: Traditional Approaches**
    - Traditional implementation of Task 1
    - Traditional implementation of Task 2
    - Evaluation and results analysis
4. **Phase 2: Modern Approaches**
    - Modern implementation of Task 1
    - Modern implementation of Task 2
    - Evaluation and results analysis
5. **Comparative Analysis**
    - Performance comparison across all methods
    - Trade-off analysis
    - Best practices and recommendations
6. **Conclusion**
    - Summary of key findings
    - Limitations and future work
    - References
7. **Bonus: Interactive Demo**
    - Code and interactive interface
    - Instructions for deployment and usage

Student Name (in English): _____ **Student ID:** _____

Evaluation Rubric

Phase 1: Traditional Methods (10 marks)

Component	Criteria	Marks
Dataset Preparation	- Proper loading and preprocessing- Handling of specific challenges (Arabic text or multimodal)- Appropriate data splitting	2
Traditional Implementation of Task 1	- Implementation of appropriate traditional methods- Proper evaluation and analysis- Addressing specific challenges	4
Traditional Implementation of Task 2	- Implementation of appropriate traditional methods- Proper evaluation and analysis- Addressing specific challenges	4

Phase 2: Modern Approaches (15 marks)

Component	Criteria	Marks
Modern Implementation of Task 1	- Implementation of deep learning approaches- Implementation of transformer-based approaches- Proper evaluation and analysis	7
Modern Implementation of Task 2	- Implementation of deep learning approaches- Implementation of transformer-based approaches- Proper evaluation and analysis	7
Documentation	- Code clarity and organization- Reproducibility of experiments	1

Comparative Analysis (5 marks)

Component	Criteria	Marks
Performance Comparison	- Comprehensive comparison of all methods- Statistical analysis of differences- Effective visualizations	2
Trade-off Analysis	- Analysis of computational requirements- Discussion of model complexity and interpretability- Handling of specific challenges	2
Best Practices	- Identification of optimal approaches- Recommendations for practitioners- Potential hybrid solutions	1

Interactive Demo (Bonus - 3 marks)

Component	Criteria	Marks
Functionality	- User-friendly interface- Real-time processing- Error handling	1
Visualization	- Comparison visualization- Performance metrics display- Interactive elements	1
Implementation	- Code quality- Deployment instructions- Documentation	1