BIRZEIT UNIVERSITY

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

ENCS5141 - INTELLIGENT SYSTEMS LAB

Case Study #2

December 4, 2024

Introduction

In this case study, you will work on subjectivity detection task, which is to accurately determine the subjectivity of a given text snippet. Subjectivity refers to the expression of opinions, sentiments, or personal experiences within the text, while objectivity represents factual information without personal biases.

For the case study, we have a dataset where the sentences are represented as a list of images, with each image corresponding to a single character in the sentence. Your task is to extract the text from the sequence of character images and use it for classification.

The dataset is based on over 1,100 tweets in English. The sentences are not provided in a table format; instead, they are stored in a JSON file where each entry has the following attributes:

- **sentence_id**: A unique identifier for each sentence.
- **image_paths**: A list of strings representing the paths to the images for each character in the sentence. For spaces it will be a string with a space " ".
- **label**: A binary label for the sentence (0 for OBJ, 1 for SUBJ).

```
Example:
{
    "sentence_id": "8745d4da-91c9-4538-acee-b0e7b1c413fd",
    "image_paths": [
        "./images/e5b0e1cb52a548e78112a8dff39b9947.png",
        "./images/c7e54e8c43ff4464b6b5d68f085b8595.png",
        "./images/4325830389b1424b8ea6bdeec5008031.png",
        " ",
        "./images/8151b6c60621444f848644fa8bacf644.png",
        "./images/3c4db43560534cc092088b6400845329.png",
        "./images/a6b8c0e9b7bc46d9836f79ab509d45d2.png",
        "./images/83b5ccd8479a456cbe51a3ea14d2a491.png"
    ],
    "label": "SUBJ"
}
```

Dataset Link: `https://drive.google.com/drive/folders/1_b7AGFCBTP2-Qp1Aaak0rJNB764H7FNJ?usp=sharing`.

Task 1: Data Extraction and Preprocessing

For this task you must extract the sentences from the images using different models. You have to train your model on the characters dataset, which has 785 columns where the first one is the label and the next 784 represents the pixels of the image, that is attached with the case study (you can find the mappings of the characters in the file mapping.txt).

After extracting the sentences, do the cleaning process you find best for the next steps. For models you have to use both **Convolutional Neural Network (CNN)** and **Pre-Trained Models**, which is Pre-trained on EMNIST dataset to capture the characters. For **CNN**, you need to search for papers that have architectures for such datasets, do not try many architectures as one is enough.

Task 2: Subjectivity Classification

For the subjectivity detection task, you will implement and train different models:

- **Long Short-Term Memory (LSTM)**
- **Transfer Learning**: Use pre-trained transformers and train them in the dataset.

Hyperparameter Tuning

- Experiment with different hyperparameters such as learning rate, batch size, and number of layers.
- Document the impact of each hyperparameter on the model performance.

Reporting

- Analyze the results of all combination of models and compare their performance for Subjectivity Classification.
- Report the accuracy, precision, recall, and F1-score for each combination.
- Discuss any limitations faced during the data extraction, preprocessing, and model training phases.
- Discuss the strengths and weaknesses of each model based on the dataset characteristics.

Submission Requirements

- Submit the code as a Jupyter Notebook **ONLY**, do not send colab link. You must use PyTorch for training, TensorFlow will not be accepted.
- Include a report (in **PDF** format with **5-8** pages) detailing the steps taken, models implemented, results, and analysis. Other submissions will not be accepted.
- The report should include tables and plots for the model performance metrics.
- You must compress the submission file with all requested files as the format **StudentId_StudentName.rar**.
- Attach the extracted sentence alongside their sentence id as csv or tsv files for training and testing datasets.

BONUS

If you have more than 98% similarity for the extracted sentence files with the original sentences you will get 2 bonus marks. If 100% you will get 4.