Neural Networks and Deep Learning (2024)

Assignment (4):

Persian News Classification Using LSTM+Attention

Shiraz University

Due date: 3rd of Mordad 23:59 [23/July (2024)]

1 Introduction

Natural Language Processing (NLP) is a field of artificial intelligence focused on enabling computers to understand, interpret, and generate human language. This encompasses tasks like language translation, sentiment analysis, text summarization, and question answering. Deep learning, particularly through architectures like Long Short-Term Memory (LSTM) networks, has significantly advanced NLP. LSTMs are designed to capture long-term dependencies in sequences, making them effective for processing and generating human language. They use a gating mechanism to manage the flow of information, overcoming challenges like the vanishing gradient problem found in traditional RNNs.

2 Problem Statement

For this assignment, you will implement an LSTM model with an attention mechanism to classify a Persian news dataset. You will become familiar with different text embeddings and common pre-processing techniques used in NLP.

2.1 Dataset: [DOWNLAOD LINK]

This dataset contains 96,430 samples of different news categories. The provided dataset has the following structure:

Table 1: Structure of Persian news dataset

Column	Description
title	Title of news
subgroup	Class (category) of news
abstract	A short summary of news
Body	Content of news

There are a total of 6 different classes, including: "Political", "Economic", "Sports", "Cultural/Artistic", "Events", and "Scientific" news.

3 Tasks

1. When dealing with textual data, different preprocessing techniques are used. One of these techniques is removing stopwords. Stop words are a set of commonly used words in a language.
    a. Apply necessary preprocessing techniques to your data before feeding it to your model.
    b. Evaluate whether it is necessary to remove stop words or not. You must justify your choice. If you deem it crucial to remove stop words, you can use public repositories such as: "Persian Stopwords"
2. Raw data is often imperfect. You'll notice that this dataset contains many NaN values, and removing these samples carelessly might lead to losing half of the data. Think of a way to overcome this issue. (**Hint**: don't overthink this :) just use basic ideas)
   You are free to choose the final input to your model based on the original dataset, but you should be able to use most of the provided dataset instead of easily discarding all NaN values.
3. There are many different ways to represent your textual data, such as one-hot embeddings, bag of words model, etc. Study these methods and use the most appropriate one.
4. After training your model, save it so that you can use it later instead of training from scratch. You should provide one saved trained model in your final submission files.
   If your saved model is too large to upload on Quera, you can email it to one of the TAs.
   Note: Emails sent after the deadline will not be accepted.

Notes:

- Allowed programming languages: Python, MATLAB
- Any sign of cheating would result in a zero grade for this assignment.
- You should upload your submissions at:
  https://quera.org/course/add_to_course/course/16595/
  All of the files should be in a ZIP file named in this format:
  "FirstNameFamilyName-SudentNumber.zip"
  Ex: "AmirZamani-4023040.zip"
- Your reports should be in a PDF file including: key points of your implementation, explanation of your chosen approach, reports of your final results and answers of assignment questions (if given).
