OHN1 — OHN1 TASK 2: SENTIMENT ANALYSIS USING NEURAL NETWORKS

ADVANCED ANALYTICS — D604
PRFA — OHN1

Preparation Task Overview Submissions Evaluation Report

COMPETENCIES

4164.1.1 : Applies Neural Networks

The learner applies neural networks to solve a business problem.

4164.1.2 : Applies Natural Language Processing

The learner applies natural language processing to solve a business problem.

INTRODUCTION

Throughout your career in data analytics, you will assess continuous data sources for their relevance to specific research questions. An organization may use such datasets to analyze their operations to support their decision-making processes.

In your previous work, you explored a variety of supervised and unsupervised data mining models. You have seen the power of using data analysis techniques to help organizations make data-driven decisions, and you will now extend these models into areas of machine learning and artificial intelligence. In this course, you will explore the use of neural networks and natural language processing (NLP).

In this task, you will use the "UCI Sentiment Labeled Sentences Dataset" in the Web Links section. You will build a neural network that is designed to learn word usage and context using NLP techniques. You will provide visualizations and a report, and you will build your network in an interactive development environment.

SCENARIO

As a data scientist, you will assess continuous data sources for their relevance to specific research questions throughout your career. The organizations related to the given dataset seek to analyze their operations and have collected variables of possible use to support decision-making processes.

Artifacts must be submitted as a report generated using an industry-relevant interactive development environment (e.g., an R Markdown document, a Jupyter Notebook). Inc the PDF or HTML document of your executed notebook presentation.

Note: The virtual IDE for this assessment is either Anaconda or RStudio, depending on which language you decide to use to complete the task. Please use the "WGU Virtual Lab Environment" web link below.

REQUIREMENTS

Your submission must represent your original work and understanding of the course material. Most performance assessment submissions are automatically scanned through the WGU similarity checker. Students are strongly encouraged to wait for the similarity report to generate after uploading their work and then review it to ensure Academic Authenticity guidelines are met before submitting the file for evaluation. See Understanding Similarity Reports for more information.

Grammarly Note:
Professional Communication will be automatically assessed through Grammarly for Education in most performance assessments before a student submits work for evaluation. Students are strongly encouraged to review the Grammarly for Education feedback prior to submitting work for evaluation, as the overall submission will not pass without this aspect passing. See Use Grammarly for Education Effectively for more information.

Microsoft Files Note:
Write your paper in Microsoft Word (.doc or .docx) unless another Microsoft product, or pdf, is specified in the task directions. Tasks may not be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc. All supporting documentation, such as screenshots and proof of experience, should be collected in a pdf file and submitted separately from the main file. For more information, please see Computer System and Technology Requirements.

You must use the rubric to direct the creation of your submission because it provides detailed criteria that will be used to evaluate your work. Each requirement below may be evaluated by more than one rubric aspect. The rubric aspect titles may contain hyperlinks to relevant portions of the course.

Note: Written responses need to be submitted through EMA.

Use the "UCI Sentiment Labeled Sentences Dataset" web link to complete the following:

Part I: Research Question

Note: Your responses to the task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will not be accepted.

A. Describe the purpose of this data analysis by doing the following:

1. Summarize **one** research question that you will answer using a neural network model and NLP techniques. Be sure the research question is relevant to a real-world organizational situation and sentiment analysis captured in your chosen dataset or datasets.

   Note: If you choose to use more than one dataset, you must concatenate them into one dataset for parts II and III.

2. Define the objectives or goals of the data analysis. Be sure *each* objective or goal is reasonable within the scope of the research question and is represented in the available data.
3. Identify an industry-relevant type of neural network capable of performing a text classification task that can be trained to produce useful predictions on text sequences on the selected dataset.

Part II: Data Preparation

B. Summarize the data cleaning process by doing the following:

1. Perform exploratory data analysis on the chosen dataset, and include an explanation of *each* of the following elements:
   - presence of unusual characters (e.g., emojis, non-English characters)
   - vocabulary size
   - word embedding length
   - statistical justification for the chosen maximum sequence length
2. Describe the goals of the tokenization process, including *any* code generated and *any* packages that are used to normalize text during the process.
3. Explain the padding process used to standardize the length of sequences. Include the following in your explanation:
   - whether the padding occurs before or after the text sequence
   - a screenshot of a single padded sequence
4. Identify how many categories of sentiment will be used and provide an activation function for the final dense layer of the network.
5. Explain the steps used to prepare the data for analysis, including the size of the training, validation, and test set split based on the industry average.
6. Provide a copy of the prepared dataset.

Part III: Network Architecture

Note: Your responses to the task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will not be accepted.

C. Describe the type of neural network model used by doing the following:

1. Provide the output of the model summary of the function from TensorFlow or PyTorch.
2. Discuss the number of layers, the type of layers, and the total number of parameters.
3. Justify the choice of hyperparameters, including *each* of the following elements:
   - activation functions
   - number of nodes per layer
   - loss function
   - optimizer
   - stopping criteria

Part IV: Neural Network Model Evaluation

Note: Your responses to the task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will not be accepted.

D. Evaluate the model's training process and its relevant outcomes by doing the following:

1. Discuss the impact of using stopping criteria to include defining the number of epochs, including a screenshot showing the final training epoch.

2. Assess the fitness of the model and *any* actions taken to address overfitting or underfitting.
3. Provide clearly labeled visualizations of the model's training process and show the loss and accuracy metric.
4. Discuss the predictive accuracy of the trained model using the chosen evaluation metric from part D3.
5. Explain how the analysis complies with artificial intelligence (AI) global ethical standards and mitigates bias.

Part V: Summary and Recommendations

Note: Your responses to the task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will not be accepted.

E. Provide the code you used to save the trained model within the neural network.

F. Discuss the functionality of your model, including the impact of your choice of network architecture.

G. Recommend a course of action based on your results as they relate to the research question.

Part VI: Reporting

Note: Your responses to the task prompts must be provided in a document file. Unless otherwise specified, responses to PA requirements that are included in a Python or RStudio notebook will not be accepted.

H. Submit a copy of your code and output in a PDF or HTML format.

I. Submit a list of *all* the specific web sources you used to acquire segments of third-party code to support the application.

J. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

K. Demonstrate professional communication in the content and presentation of your submission.

File Restrictions

File name may contain only letters, numbers, spaces, and these symbols: ! - _ . * ' ( )
File size limit: 200 MB
File types allowed: doc, docx, rtf, xls, xlsx, ppt, pptx, odt, pdf, csv, txt, qt, mov, mpg, avi, mp3, wav, mp4, wma, flv, asf, mpeg, wmv, m4v, svg, tif, tiff, jpeg, jpg, gif, png, zip, rar, tar, 7z

RUBRIC

A1:RESEARCH QUESTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT