COMPETENCIES

4030.7.1 : **Constructing Neural Networks**
The graduate builds neural networks in the context of machine-learning modeling.

4030.7.3 : **Natural Language Processing**
The graduate extracts insights from text data using effective and appropriate natural language processing (NLP) models.

INTRODUCTION

Throughout your career as a data analyst, you will assess continuing data sources for their relevance to specific research questions. Organizations use data sets to analyze their operations. Organizations may use these data sets in many ways 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 Data Set" 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, as well as build your network in an interactive development environment.

SCENARIO

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

REQUIREMENTS

Your submission must be your original work. No more than a combined total of 30% of the submission and no more than a 10% match to any one individual source can be directly quoted or closely paraphrased from sources, even if cited correctly. The originality report that is provided when you submit your task can be used as a guide.

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.

*Tasks may **not** be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc., unless specified in the task requirements. All other submissions must be file types that are uploaded and submitted as attachments (e.g., .docx, .pdf, .ppt, .csv).*

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

Part I: Research Question

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

1. Summarize **one** research question that you will answer using neural network models and NLP techniques. Be sure the research question is relevant to a real-world organizational situation and sentiment analysis captured in your chosen data set(s).

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

2. Define the objectives or goals of the data analysis. Be sure the objectives or goals are reasonable within the scope of the research question and are represented in the available data.
3. Identify a 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 data set.

Part II: Data Preparation

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

1. Perform exploratory data analysis on the chosen data set, and include an explanation of *each* of the following elements:
   - presence of unusual characters (e.g., emojis, non-English characters)
   - vocabulary size
   - proposed word embedding length
   - statistical justification for the chosen maximum sequence length
2. Describe the goals of the tokenization process, including any code generated and packages that are used to normalize text during the tokenization process.
3. Explain the padding process used to standardize the length of sequences. Include the following in your explanation:
   - if 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 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 data set.

Part III: Network Architecture

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

1. Provide the output of the model summary of the function from TensorFlow.

2. Discuss the number of layers, the type of layers, and the total number of parameters.
3. Justify the choice of hyperparameters, including the following elements:
   - activation functions
   - number of nodes per layer
   - loss function
   - optimizer
   - stopping criteria
   - evaluation metric

Part IV: Model Evaluation

D. Evaluate the model 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.
   3. Provide visualizations of the model's training process, including a line graph of the loss and chosen evaluation metric.
   4. Discuss the predictive accuracy of the trained network using the chosen evaluation metric from part D3.

Part V: Summary and Recommendations

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

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

G. Recommend a course of action based on your results.

Part VI: Reporting

H. Show your neural network in an industry-relevant interactive development environment (e.g., a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation.

I. Denote specific web sources you used to acquire segments of third-party code that was used 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.