NLM2 — NLM2 TASK 2: SENTIMENT ANALYSIS USING NEURAL NETWORKS

ADVANCED DATA ANALYTICS — D213
PRFA — NLM2

TASK OVERVIEW SUBMISSIONS EVALUATION REPORT

COMPETENCIES

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

4030.7.3 : Natural Language Processing (NLP)
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 have many possible uses for these data sets to support their decision-making processes.

In your previous coursework, you have explored a variety of supervised and unsupervised data mining models. You have seen the power of using data analytical techniques to help organizations make data-driven decisions and now want to 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 choose a data file from the Web Links section. The available data sets are as follows:

- Amazon Product Data set
- UCSD Recommender Systems Data sets
- UCI Sentiment Labeled Sentences Data set

For this task, you will build a neural network 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.

Note: You cannot use the same data set that was provided in the exemplar for this course.

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).*

Choose **one** dataset from the Web Links section and use it 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 dataset.
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 dataset, and include an explanation of each of the following elements:
   - presence of unusual characters (e.g., emojis, non-English characters, etc.)
   - 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, including 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.
6. Provide a copy of the prepared dataset.

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 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 instead of defining the number of epochs, including a screenshot showing the final training epoch.
2. Provide visualizations of the model's training process, including a line graph of the loss and chosen evaluation metric.
3. Assess the fitness of the model and any measures taken to address overfitting.
4. Discuss the predictive accuracy of the trained network.

Part V: Summary and Recommendations

E. Provide the code 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. Create your neural network using an industry-relevant interactive development environment (e.g., a Jupyter Notebook). Include a PDF or HTML document of your executed notebook presentation.

I. List the web sources used to acquire data or 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, 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
The submission does not provide a summary of 1 research question.	The submission summarizes 1 research question, but the research question is not relevant	The submission summarizes 1 research question, and the research question is relevant to a realistic organizational situation

	to a realistic organizational situation or sentiment analysis or cannot be addressed using the selected dataset, neural network model, or NLP techniques.	and sentiment analysis. The research question can be answered using the selected dataset, neural network model, and NLP techniques.

A2:OBJECTIVES AND GOALS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not define the objectives or goals of the data analysis.	The submission defines the objectives or goals of the data analysis, but 1 or more of the objectives or goals are not clear or reasonable within the scope of the research question or are not represented in the available data.	The submission clearly defines *each* of the objectives or goals of the data analysis. *Each* objective or goal is reasonable within the scope of the research question and represented in the available data.

A3:PRESCRIBED NETWORK

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not identify a type of neural network.	The submission identifies a type of neural network, but the identified network cannot be trained to produce useful text classification predictions on text sequences on the selected dataset. Or the network identified is not industry relevant.	The submission identifies an industry-relevant type of neural network that can be trained to produce useful text classification predications on text sequences on the selected dataset.

B1:DATA EXPLORATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not perform an exploratory data analysis on the chosen data set.	The submission performs exploratory data analysis on the chosen dataset, but 1 or more of the listed elements are not explained. Or the explanation of 1 or more of the elements contains inaccuracies or are not aligned to the chosen data set.	The submission performs exploratory data analysis on the chosen dataset and includes an explanation of *each* of the 4 listed elements. *Each* element is accurate and aligns to the chosen data set.

B2:TOKENIZATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not describe the goals of the tokenization process.	The submission incompletely describes the goals of the tokenization process or does not include *any* code generated or *any* packages that are used to normalize text during the tokenization process. Or the description contains inaccuracies.	The submission completely describes the goals of the tokenization process, including *any* code generated and packages that are used to normalize text during the tokenization process. The description is accurate.

B3:PADDING PROCESS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not explain the padding process used to standardize the length of sequences.	The submission inaccurately explains the padding process used to standardize the length of sequences. Or the explanation does not include where the padding occurs, or the padding process does not align with the chosen data. Or a screenshot of a single padded sequence is not provided.	The submission accurately explains the padding process used to standardize the length of sequences and includes an explanation of where the padding occurs. The padding process aligns with the chosen data, and a screenshot of a single padded sequence is provided.

B4:CATEGORIES OF SENTIMENT

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not identify how many categories of sentiment will be used.	The submission inaccurately identifies how many categories of sentiment will be used. Or the activation function does not align with the number of classes used or is inappropriate for the network.	The submission clearly and accurately identifies how many categories of sentiment will be used and an appropriate fitting activation function for the final dense layer of the network.

B5:STEPS TO PREPARE THE DATA

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT

The submission does not explain the steps used to prepare the data for analysis.	The submission explains the steps used to prepare the data for analysis, but the steps are incomplete or do not include the size of the training, validation, or test split. Or 1 or more of the steps are not related to preparing for neural network models or NLP techniques.	The submission explains the steps used to prepare the data for analysis and accurately includes the size of the training, validation, and test split. The steps relate to preparing for neural network models and NLP techniques.

B6:PREPARED DATASET

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide a copy of the dataset.	The submission provides a copy of a dataset, but the dataset is not fully prepared.	The submission provides a copy of a fully prepared dataset.

C1:MODEL SUMMARY

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide the output of the model summary of the function from TensorFlow.	The submission provides an incomplete output of the model summary of the function from TensorFlow, or the output does not align with the type of network used.	The submission provides the complete output of the model summary of the function from TensorFlow. The output aligns with the type of network used.

C2:NETWORK ARCHITECTURE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss the number of layers, the type of layers, or the total number of parameters.	The submission discusses the number of layers, the type of layers, and the total number of parameters in the network, but 1 or more are incomplete or inaccurate.	The submission completely and accurately discusses the number of layers, the type of layers, and the total number of parameters in the network.

C3:HYPERPARAMETERS

NOT EVIDENT		COMPETENT

The submission does not justify the choice of hyperparameters.	**APPROACHING COMPETENCE** The submission does not logically justify the choice of hyperparameters for 1 or more of the listed elements. Or 1 or more of the listed elements do not align with the network used.	The submission logically justifies the choice of hyperparameters, including *each* of the 6 listed elements, and *each* element aligns with the network used.

D1:STOPPING CRITERIA

NOT EVIDENT The submission does not discuss the impact of using stopping criteria instead of defining the number of epochs.	**APPROACHING COMPETENCE** The submission inaccurately discusses the impact of using stopping criteria instead of defining the number of epochs. Or the submission does not include a screenshot showing the final training epoch.	**COMPETENT** The submission accurately discusses the impact of using stopping criteria instead of defining the number of epochs. A screenshot showing the final training epoch is provided.

D2:TRAINING PROCESS

NOT EVIDENT The submission does not provide *any* visualizations of the model's training process.	**APPROACHING COMPETENCE** The submission provides incomplete visualizations of the model's training process, or the visualizations are not clearly labeled or do not align with the model's training process. Or it does not include a line graph of the loss or the chosen evaluation metric.	**COMPETENT** The submission provides complete visualizations of the model's training process, including a line graph of the loss and the chosen evaluation metric. The visualizations are clearly labeled and align with the model's training process.

D3:FIT

NOT EVIDENT The submission does not assess the fitness of the model.	**APPROACHING COMPETENCE** The submission incompletely or inaccurately assesses the fitness of the model, or the assessment does not include *any* measures taken to address overfitting.	**COMPETENT** The submission completely and accurately assesses the fitness of the model, and the assessment includes *any* measures taken to address overfitting.

D4:PREDICTIVE ACCURACY

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss the predictive accuracy of the trained network.	The submission discusses the predictive accuracy of the trained network but does not use the selected evaluation metric from part D2, or the submission uses an accuracy metric for an incompletely trained model.	The submission discusses the predictive accuracy of the trained network using the selected evaluation metric from part D2.

E:CODE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide the code used to save the trained network within the neural network.	The submission provides the code used to save the trained network within the neural network, but the code is incomplete or contains inaccuracies.	The submission provides the code used to save the trained network within the neural network, and the code is complete and accurate.

F:FUNCTIONALITY

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss the functionality of the neural network.	The submission inaccurately discusses the functionality of the neural network, or it does not discuss the impact of the network architecture. Or the discussion does not align with the research question from part A.	The submission accurately discusses the functionality of the neural network, including the impact of the network architecture. The discussion aligns with the research question from part A.

G:RECOMMENDATIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not recommend a course of action based on results.	The submission recommends a course of action based on results, but the recommendation is not appropriate based on the research question or the results of the data.	The submission recommends an appropriate course of action based on the results as they relate to the research question.

H:REPORTING

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
A neural network is not created. Or the PDF or HTML document of the executed notebook presentation is not provided.	The neural network is not created in an industry-relevant interactive development environment. Or the created network is incomplete, contains inaccuracies, or does not align with the data analysis of the report.	The neural network is created in an industry-relevant interactive development environment and is complete, accurate, and in alignment with the data analysis of the report. A PDF or HTML document of the executed notebook presentation is provided.

I:SOURCES FOR THIRD-PARTY CODE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not list *any* web sources.	The submission lists only *some* of the web sources used to acquire data or segments of third-party code. Or 1 or more of the listed web sources are not reliable.	The submission lists *all* web sources used to acquire data or segments of third-party code, and *all* the listed web sources are reliable.

J:SOURCES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not include both in-text citations and a reference list for sources that are quoted, paraphrased, or summarized.	The submission includes in-text citations for sources that are quoted, paraphrased, or summarized and a reference list; however, the citations or reference list is incomplete or inaccurate.	The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available.

K:PROFESSIONAL COMMUNICATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
Content is unstructured, is disjointed, or contains pervasive errors in mechanics, usage, or grammar. Vocabulary or tone is unprofessional or distracts from the topic.	Content is poorly organized, is difficult to follow, or contains errors in mechanics, usage, or grammar that cause confusion. Terminology is misused or ineffective.	Content reflects attention to detail, is organized, and focuses on the main ideas as prescribed in the task or chosen by the candidate. Terminology is pertinent, is used correctly, and effectively conveys the intended meaning.

Mechanics, usage, and grammar promote accurate interpretation and understanding.

WEB LINKS

Amazon Product Data set

UCSD Recommender Systems Data Sets
You can choose one of the available data sets from this web link.

UCI Sentiment Labeled Sentences Data Set