A COMPARISON BETWEEN LOGISTIC REGRESSION AND NEURAL NETWORKS IN A CONSTRUCTED RESPONSE ITEM STUDY

by

MINHO KWAK

(Under the Direction of Cheolwoo Park)

ABSTRACT

The purpose of the study is to demonstrate the prediction quality of logistic regression and artificial neural networks. The main results of the study are the comparisons of the accuracy of both methods. The response variable of the model is a comment assignment by a human rater, and the four predictors are topic proportions estimated from latent Dirichlet allocation. The constructed models for both analyses are mainly concerned with predicting the comment assignment by using the topic proportions as the predictors. The results show that the accuracy of the test data set is generally higher than the accuracy of the cross-validation quality of the logistic regression, and these results are well matched with previous empirical studies. Also, although the use of this accuracy for practical purposes remains still questionable, the results reveal the potential utility the neural network if a larger sample size is available in the future.

INDEX WORDS: Artificial neural networks, Latent Dirichlet allocation, Logistic regression

A COMPARISON BETWEEN LOGISTIC REGRESSION AND NEURAL NETWORKS IN A CONSTRUCTED RESPONSE ITEM STUDY

by

MINHO KWAK

BA, SEOUL NATIONAL UNIVERSITY, SOUTH KOREA, 2010

MA, SEOUL NATIONAL UNIVERSITY, SOUTH KOREA, 2012

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2019

A COMPARISON BETWEEN LOGISTIC REGRESSION AND NEURAL NETWORKS IN A CONSTRUCTED RESPONSE ITEM STUDY

by

MINHO KWAK

Major Professor:	Cheolwoo Park
Committee:	Nicole Lazar
	Seock-Ho Kim

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2019

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Statement of the Problem

Artificial intelligence (AI), which is intelligence governed by machines, is one of the newest areas in science, and is constantly making advances towards the construction of intelligent entities (Russell & Norvig, 2016). With the recent availability of big data, machine learning has been introduced as a promising approach to developing AI, and one of the most popular machine learning algorithms is the artificial neural network (ANN; Sargent, 2001). ANN is modeled after the neuron system in the human brain. The mathematical representation of this system was first suggested by McCulloch and Pitts (1943), and since then computational neuroscience has become a well-established academic discipline, developing much more realistic and detailed models from these early, simple representations (Sargent, 2001).

Previous studies consistently reported that ANNs exhibit high performance compared to traditional statistical methods when applied to classification tasks (e.g., logistic regression), even though they carry a risk of over-fitting (Russell & Norvig, 2016). Especially, ANNs are widely used in not only computer science but also in other disciplines, including the medical field (Agatonovic-Kustrin & Beresford, 2000; Zhou, Wu, & Tang, 2002; Rautaray & Agrawal, 2015; Liang & Hu, 2015). In educational fields, although the use of this method remains limited, it has received substantial attention from educational researchers (von Davier, 2018).

Meanwhile, the constructed response (CR) item has been widely used in various tests in education, and it has an advantage over the selected response (SR) item because it can measure students' knowledge regarding the test with greater depth compared to SR items (Attali, 2014).

However, use of CR items is still restricted because evaluation of these items is difficult compared to evaluation of SR items, and furthermore, formative assessment of CR items is much more restricted because their evaluation entails additional costs (Lee, 2011). These additional cost are incurred because trained raters must manually provide appropriate feedback for the formative purpose.

Recently, topic modeling was introduced into the evaluation of CR items, and it produces the topic proportion, which reflects the semantic structure of the students' answer (Kim et al., 2017 & Kwak et al., 2017). If the topic proportions are related to the feedback comments given by the raters, the appropriate feedback could be automatically assigned to the students' writing based on these models.

Thus, logistic regression (i.e., using the topic proportions as predictors) could be suggested as a traditional approach to predicting the assignment of feedback. At the same time, an ANN could also be suggested as an alternative method, and comparisons of the performance of both analyses could show the advantages and disadvantages of ANNs.

1.2 Purpose of the Study

The purpose of the study is to demonstrate the prediction quality of logistic regression and ANNs. The main results of the study are the comparisons of the accuracy of both methods. Specifically, Chapter 2 describes the background, regrading ANN and gradient descent with

backpropagation. Chapter 3 mainly discusses data and analysis. Chapter 4 explains the results of the study, and Chapter 5 concludes this work.

CHAPTER 2

BACKGROUND ON ARTIFICIAL NEURAL NETWORK

2.1 Artificial Neural Network

The neural network analysis also called an ANN is an intelligent system that can solve various tasks, including pattern recognition, optimization, prediction, and so forth (Jain, Mao, & Mohiuddin, 1996). Since the system basically mimics the human brain, which uses experiences to solve tasks, it is closely related to artificial intelligence (Agatonovic-Kustrin & Beresford, 2000). The relationship between the neural network and artificial intelligence can be explained as one in which the neural network is one of the tools used to develop artificial intelligence (Goodfellow, Bengio, Courville, & Bengio, 2016). It is well represented in Figure 1, as suggested by Goodfellow and his colleagues.

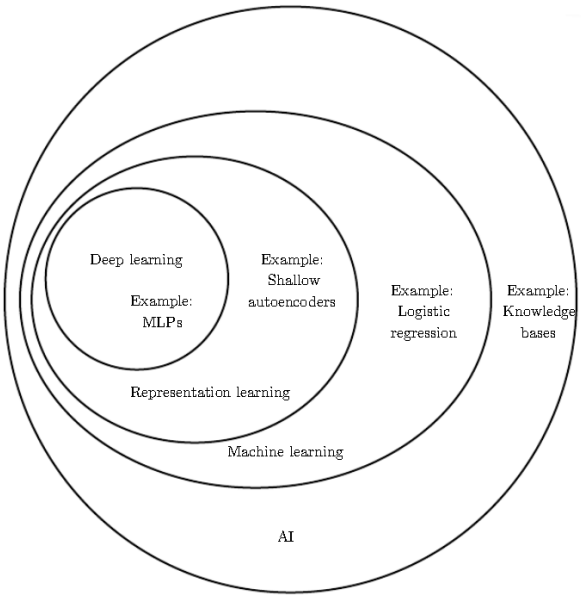

Figure 1 The Relationship between the Neural Network and Artificial Intelligence.

Note. AI is the broadest concept, and machine learning is a specific method to develop the AI. Deep learning is also a particular type of machine learning, and neural network analysis is an approach to performing deep learning (from Goodfellow et al., 2016).

A neural network can be defined as a network consisting of elements that perform simple processing on their local data and communicate with other elements, and can be more easily explained by stressing its similarity to the structure of a real brain (Svozil, Kvasnicka, & Pospichal, 1997). Specifically, in the human brain, a sensory organ receives an external signal and transforms it into an electronic impulse, and this impulse travels through multiple neurons, all connected to each other. In this step, the neurons can be activated or not, depending on whether the impulse exceeds a specific level, called as a threshold.

For example, let us assume a person is staring at a dog. The features of the dog, including its shape, color, the texture of its fur, and its motions, are observed through the sensory organs of the person (e.g., eyes, ears, and skin). Then, a stimulus is transformed into an electronic pulse, and it travels to the related neurons. If the pulse exceeds the neurons' threshold, the pulse is delivered to the brain, and the person can determine whether the object is a dog or not. In an ANN, a similar delivery system is applied to the model. First, each neuron is represented as a unit. The unit is connected with several other units, as a neuron does. Also, the connections (called nodes) among the units are evaluated by the weight coefficient, which refers to the importance of the connection in the network. Typically, an ANN consists of the multiple layers, and each layer also consists of multiple units connected with each other through the nodes (see Figure 2) (Goodfellow et al., 2016).

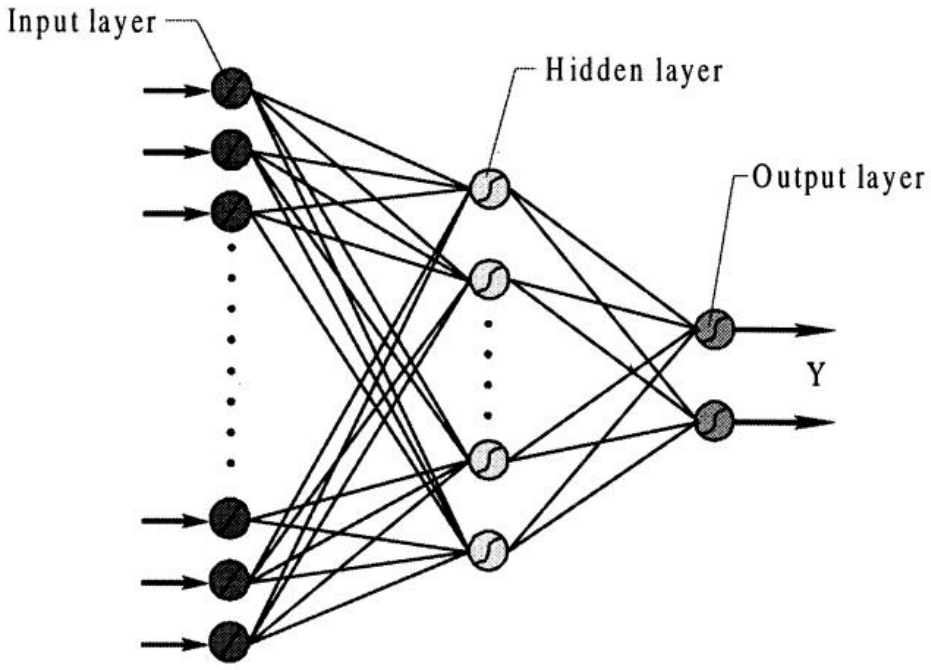

Figure 2. The General Framework of the Multilayer Perceptron (MLP).

Note. Conceptually, the MLP is composed of three layers: the output layer, the hidden layer, and the input layer. However, more than one hidden layer may be present, and if the number of hidden layers increases, the network can be considered deep. Thus, machine learning based on a network that has numerous hidden layers can be called deep learning (from Lek & Guégan, 1999).

As noted above, the ANN was motivated by neuroscience, and from the perspective of neuroscience, human mental activity is a result of the electrochemical activities in neurons (Russell & Norvig, 2016). Since the general structure of the ANN was summarized in previous paragraphs, the more rigorous derivations of the mathematical representation need to be discussed in this section. The mathematical representation of a neuron is shown in Figure 3.

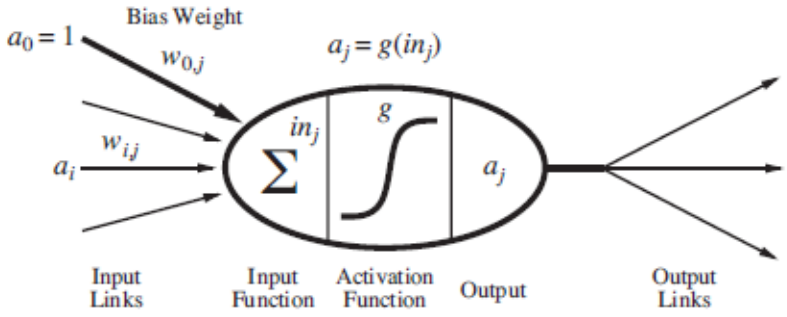

Figure 3. Mathematical Model of a Neuron.

Note. Russell & Norvig (2016) proposed a mathematical model of a neuron. The input signal a_i is delivered from unit *i* to the present unit *j*. The delivered signal a_i is combined with the $w_{i,j,}$which represents the weight of the node between unit *i* and unit *j*, and transformed by *g*, the location of the activation function, which will be explained in the next section. Finally, from $a_j = g(in_j)$, the output activation of unit *j* can be obtained through the network (from Russell & Norvig, 2016).

The main goal of neural network analysis is to learn a target function $f^*(x)$. Generally, since the target function is very complex, the learning algorithm is applied to approximate the target function as closely as possible, and the learning process is called *training neural networks* in common practice.

The critical concept of training aims to minimize loss function between the target function and the approximated one. The mean square error (MSE) loss function can be defined as Equation (1):

$$J(\theta) = \sum_{x \in X}(f^*(x) - f(x;\theta))^2 \qquad (1)$$

where $f^*(x)$ denotes target function, and $f(x;\theta)$ indicates a model parameterized by θ. Also, *x* is an individual data point, and *X* is a set of the data points. If the model is a linearly parameterized by θ consisting of *W* and *b*, it can be represented as Equation (2):

$$f(x; W, b) = x^T W + b \qquad (2)$$

where W is the weights (coefficients) matrix and b is constant (bias) vector. Usually, since the model consists of multiple layers, $f(x; W, b)$ is a composition of the multiple functions chained to each other through the hidden unit h (see Figure 4).

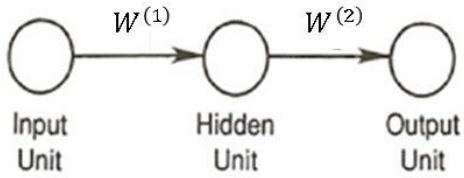

Figure 4 The relationships among Input Unit, Hidden Unit, and Output Unit

Note. The input unit implies the data, and the output unit is the classification results of the network. The hidden unit is a bridge between the input unit and the output unit. In the figure, since only one hidden layer is constructed, the input data are transformed with the activation of the hidden unit. Consequently, the activation of the hidden unit is used as the input of the output unit (from Russell & Norvig, 2016).

For example, the activation of the first hidden layer is used as the second input of the second hidden layer, and so on. Thus, the activation of the hidden layer just before the final layer is used as the input for the final layer. Specifically, the activation of the first layer can be represented as Equation (3):

$$h^{(1)} = f^{(1)}\left(x; W^{(1)}, b^{(1)}\right) \qquad (3)$$

where $W^{(1)}$ indicates the weight matrix for the first layer, and b means the constant vector for the first layer. As mentioned above, since $h^{(1)}$ is input of the second layer, it can be represented as Equation (4):

$$h^{(2)} = f^{(2)}\left(h^{(1)}; W^{(2)}, b^{(2)}\right) \quad (4)$$

where $W^{(2)}$ indicates the weight matrix for the second layer, and b means the constant vector for the second layer. Thus, if the *L*th layer model is developed, the complete model can be represented as Equation (5):

$$f(x; W^{(1)}, b^{(1)}, \dots, W^{(L-1)}, b^{(L-1)}, W^{(L)}, b^{(L)}) = f^{(L)}(f^{(L-1)}f^{(\dots)}(f^{(1)}(x)) \qquad (5)$$

In common practice, the nonlinear function called the activation function is used for modeling *f* to prevent the complete model from being just a product of the weights by ignoring the constant. Thus, the activation function *g* is used to transform the linear combination $x^T W^{(l)} + b^{(l)}$. Within this framework, the hidden units $h^{(l)}$ can be represented as Equation (6):

$$h^{(l)} = g(x^T W^{(l)} + b^{(l)}) \qquad (6)$$

where *g* denotes the activation function. Although the various activation functions can be applied to the transformation, the most commonly used one is the sigmoid function. The function is defined as $(z) = \frac{1}{1+e^{-z}}$, where *z* denotes $x^T W^{(l)} + b^{(l)}$ and is represented in Figure 5.

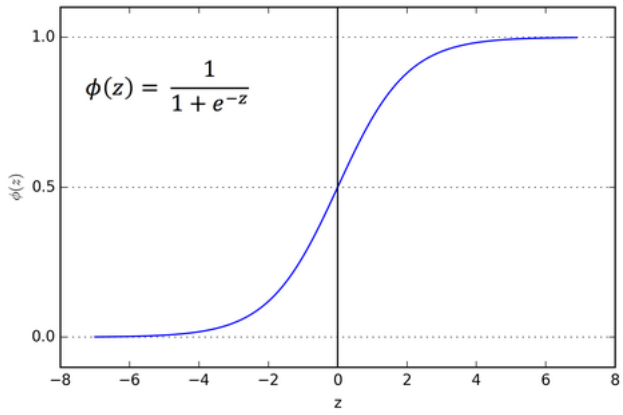

Figure 5 The Sigmoid Function

Note. The y-axis represents the calculated hidden unit value, and the x-axis represents z, which produces the linear combination $x^T W^{(l)} + b^{(l)}$. The main advantage of the sigmoid function is that the function is differentiable (from Goodfellow et al., 2016).

Therefore, the complete network consisting of *L* layers can be represented as Equation (7):

$$f\left(x; W^{(1)}, b^{(1)}, \ldots, W^{(L-1)}, b^{(L-1)}, W^{(L)}, b^{(L)}\right) = W^{(L)^T}\left\{0, W^{(L-1)^T} h^{(L-1)} + b^{(L-1)}\right\} + b^{(L)} \quad (7)$$

where $h^{(L-1)}$ means the value of (L-1)th hidden unit, and can be represented as Equation (8):

$$h^{(L-1)} = W^{(L-1)^T}\left\{0, W^{(L-2)^T} h^{(L-2)} + b^{(L-2)}\right\} + b^{(L-1)}. \quad (8)$$

In a similar way, the complete network can be considered as a set of activation functions with *W* and *b* chained through hidden unit *h*.

2.2 Gradient Descent and Backpropagation

In the previous section, the MSE loss function $J(\theta)$ was introduced. Before discussing the main idea of the Gradient Descent and Backpropagation, it is necessary to clarify the ultimate goal of training the neural network. The main goal of the training is to obtain the solution that minimizes the loss function, and this solution is weight matrices, which produce minimum loss.

The minimum of the loss function can be obtained at the point where the gradient of the function equals 0. However, the point where the gradient is 0 is analytically difficult to calculate. Thus, the approximation approach consisting of gradient descent and back-propagation (Bryson and Ho, 1969) has been suggested as an alternative method. This algorithm has been applied to many learning problems in computer science and psychology (Goodfellow et al., 2016).

First, the gradient descent is updating the weights based on the calculated errors (loss) through a recursive method. It can be represented as Equation (9):

$$\theta = \theta - \alpha \frac{\partial}{\partial \theta_u^{(l)}} J(\theta) \quad (9)$$

where θ_u^l denotes the parameter consisting of the weights and constant of the *u*th unit in the *l*th layer of the network, $J(\theta)$ indicates the loss function, and α is learning rate. As shown in the equation, updating the weights is closely related to the change of the partial derivative. Specifically, since the error function is convex, the partial derivatives of the weights are going to be 0 as the solution progresses closer to the optimal point. For example, it can be assumed that a loss function is governed by two weights, as shown in Figure 6. The figure shows that as the partial derivatives of the weights are going to be smaller, the weight is going to be closer to the optimal position that produces the minimum loss. Thus, the change of the value of the partial derivatives informs how the estimated weights are similar to the optimal point. In other words, if the change of the partial derivatives is small, the point can be considered as the acceptable approximated value of the optimal point.

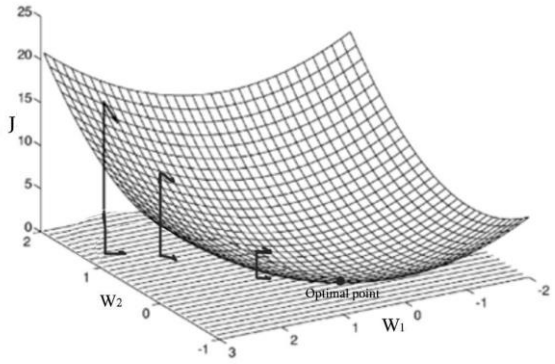

Figure 6 The relationship between the weights updating and the change of the partial derivative. *Note.* The vertical axis represents the loss and two horizontal axes represent two different weights. When the point is moving to $\widehat{w}$, the optimal point, the loss is going to decrease. Specifically, the partial derivatives for both weights are going to be 0 when the point is closer to the optimal point (from Pietersma, 2010).

Although the above example only shows two weights, the calculation of the derivative involves numerous weights because the network consists of multiple layers with multiple units in common practice.

As outlined above, partial derivatives of the target function $\frac{\partial}{\partial\theta}J(\theta)$ are critical to obtain the optimal parameter in the gradient descent method, and backpropagation is widely used in common practice to obtain the derivatives (Jain & Mohiuddin, 1996). The specific algorithm can be summarized as follows (for simplicity, the derivation regarding only the weight matrix, and excluding the constant, is summarized): The initial values for the parameters are specified as certain values. For a training set that is the size of *M*, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_M, y_M)\}$ will be used to train the network. When the *m*th input data is put into the network, the output will be produced based on the final layer. Also, the difference between the output $\hat{y}_m$ and target y_m can be represented as $\delta_m = \hat{y}_m - y_m$. The δ_m implies the error of the prediction of the trained network for the *m*th training observation, and the algorithm backpropogates δ_m into the nodes in the network.

Specifically, $\delta_m^{(l)}$ is used to calculate $\delta_m^{(l-1)}$, which means an error of the *l*-1th layer. Equation (10) shows the derivation of $\delta_m^{(l-1)}$ from $\delta_m^{(l)}$, using weight and the derivative of the activation function with respect to *z*:

$$\delta_m^{(l-1)} = (W^{(l-1)})^T \delta_m^{(l)} * \frac{\partial(g(z^{(l-1)}))}{\partial z^{(l-1)}} \quad (10)$$

where z denotes $x^T W^{(l)} + b^{(l)}$, which means a linear combination of weights and constants. The derivative of activation function $g'(z)$ depends on the form of the activation function. If the activation function is the sigmoid function, the derivative will be $g(z) * (1 - g(z))$.

Also, the gradient of error regarding the weight can be expressed by the activation $a_m^{(l)}$ and the difference $\delta_m^{(l+1)}$ as Equation (11):

$$\frac{\partial}{\partial w_{ij}^{(l)}} J(W; x, y) = a_m^{(l)} \delta_m^{(l+1)}. \qquad (11)$$

When the input data is put into the network, forward propagation is performed. The results of the forward propagation is activation $a^{(l)}$ for the every layer. The activation can be defined as Equation (12):

$$a^{(l)} = g\left(z^{(l)}\right) = g(x^T W^{(l)} + b^{(l)}). \qquad (12)$$

Finally, the weight updating is performed based on Equation (13):

$$w_{ij}^{(l)} = w_{ij}^{(l)} - \alpha \frac{\partial}{\partial w_{ij}^{(l)}} J(w; x, y) = w_{ij}^{(l)} - \alpha a_m^{(l)} \delta_m^{(l+1)} \qquad (13)$$

where $W_{ij}^{(l)}$ denotes the weight between the *i*th unit and the *j*th unit and α denotes the learning rate. For the constant *b*, a similar approach is applied, and the updating is performed based on Equation (14):

$$b_{ij}^{(l)} = b_{ij}^{(l)} - \alpha \frac{\partial}{\partial b_{ij}^{(l)}} J(b; x, y) = b_{ij}^{(l)} - \alpha \delta_m^{(l+1)}. \qquad (14)$$

2.3 Resilient Backpropagation

Resilient backpropagation (RPROP; Anastasiadis, Magoulas, & Vrahatis, 2005) is one of the commonly used variations of backpropagation, and it differs from the traditional approach in that the learning rate is updated in a different manner (Prasad, Singh, & Lal, 2013; Anastasiadis, Magoulas, & Vrahatis, 2005).

The learning rate α of the backpropagation algorithm is a crucial value for the training process and needs to be specified carefully. Thus, resilient backpropagation is preferred because the learning rate is adopted during the training process (Günther & Fritsch, 2010).

A sign-based approach is used to perform the updating of the learning rate based on the direction of the partial derivatives. If the sign of the partial derivate changes between two iterations, indicating the step between these iterations to be too large, the updated step decreases by the additional factor η. On the other hand, if the sign changes, meaning the step size is too small, the step increases. Otherwise, the learning rate is going to be the same with the previous step (see Equation (15)):

$$\alpha^{(l)} = \begin{cases} \alpha^{(l-1)} + \eta \; if \; \frac{\partial J}{\partial w_{ij}^{(l)}} \frac{\partial J}{\partial w_{ij}^{(l-1)}} > 0, \\ \alpha^{(l-1)} - \eta \; if \; \frac{\partial J}{\partial w_{ij}^{(l)}} \frac{\partial J}{\partial w_{ij}^{(l-1)}} < 0, \\ \alpha^{(l-1)} \qquad otherwise. \end{cases} \tag{15}$$

2.4 Comparison between Neural Network Analysis and Logistic Regression

Tu (1996) performed a comparison study between neural network analysis and logistic regression. He summarized the advantages and disadvantages of using an ANN for predicting the medical output. The advantages can be summarized as follows: 1) A less formal statistical training process is required; 2) It has the ability to identify the complex nonlinear relationships underlying the observed variables; 3) It can detect all possible interactions among predictor variables; and 4) Multiple different training algorithms are applied to developing the network.

The disadvantages can be summarized as follows: 1) The interpretation of the model is restricted because it is hard to identify the relationships among the variables, due to the numerous parameters in the model; 2) Use of the models might be limited in the field because it

require rigorous statistical knowledge; 3) It requires greater computational resources; 4) It is likely to be overfitting; and 5) There is no theoretically grounded approach to developing the network.

Nefeslioglu, Gokceoglu, & Sonmez (2008) also performed both analyses to produce landslide susceptibility maps, which charted geological information regarding landslides in order to prevent future disasters. Their results indicated that the ANN shows remarkably high prediction qualities compared to those of logistic regression.

Sahin & Duman (2011) performed a study to detect credit card fraud using an ANN and logistic regression. In this study, they showed the performance of classifiers with respect to accuracy. The results showed that the performance of the logistic regression was similar to the performance of the ANN, with accuracy values at 90%. However, while the accuracy of the logistic regression for the cross-validation was around 70%, the accuracy of ANN was found to be around 90%.

On the other hand, another study produced a different result (Heazlewood, Walsh, Climstein, Kettunen, Adams, & DeBeliso, 2016). Specifically, the classification accuracies based on logistic regression (n=7,175), and discriminant analysis in combination with logistic regression, were very similar in outcome for both the classification of gender and combined classification accuracy. None of the classification techniques based on neural network analyses and multivariate methods of discriminant analysis and logistic regression were overtly superior to each other.

CHAPTER 3

DATA ANALYSIS

3.1 Data

The data set consists of two parts. The first part contains the information regarding the topic proportions that are used as the predictors. The second part contains the comment codes assignment that are used as the response variable.

The topic proportions are estimated based on LDA (Latent Dirichlet Allocation; Blei, Ng, & Jordan, 2003), one of the most popular text mining techniques. The model provides the topic proportions for the documents in the corpus. In this study, four topics are extracted from the corpus obtained from ER (extended constructed response) items in a 7th grade informational test (see Table 1), and the interpretation and structure of the topics will be discussed in following paragraph (see Table 2).

Table 1 *Descriptive Statistics for the Corpus*

Before preprocessing			
Number of Documents	Number of Unique word	Number of total word	Average length
10,292	3,219	1,712,913	166
After preprocessing			
Number of Documents	Number of Unique word	Number of total word	Average length
7,431	2,329	1,101,241	148

First, topic 1 contained terms such as *you, they, get,* and *do not* that are taken to reflect the everyday words. The correlation with the score was weakly negative ($r=-0.130$, $p<.001$). On the other hand, topic 2 had a weakly positive correlation with the score ($r=0.290$, $p<.001$) and contained words such as *children, disease, consequence, school,* and *paragraph*. The correlation

indicates that these words reflect appropriate integrative borrowing. Topic 3 contained words such as *oil, spill, waterway, animal,* and *plant*. The correlation with topic 3 was positively correlated with the score (r=0.035, p=.004). Topic 4 contains words such as *dry, hole, waste,* and *citizen*. It exhibits weak and negative correlation with the score (r=-0.150, p<.001). Words in both topics 3 and 4 appear to be simply borrowed from the passage rather than used appropriately to answer the item.

Table 2 *Top 30 Words and Corresponding Probabilities by Topics with the Correlations for Grade 7 Informational*

	Topic1 Everyday words ($r = -0.130$, $p < .001$)		Topic2 Integrative borrowing words ($r = 0.290$, $p < .001$)		Topic3 Directly borrowing words ($r = 0.035$, $p = .004$)		Topic4 Directly borrowing words ($r = -0.150$, $p < .001$)	
1	you	0.126	children	0.080	oil	0.073	dry	0.094
2	they	0.077	consequence	0.080	spill	0.058	hole	0.082
3	get	0.059	school	0.032	waterway	0.045	waste	0.034
4	do not	0.046	unclean	0.032	animal	0.037	citizen	0.024
5	their	0.030	state	0.032	the	0.036	surface	0.020
6	kid	0.024	illness	0.029	plant	0.034	fill	0.019
7	just	0.018	miss	0.024	Nigeria	0.030	riverbed	0.018
8	would	0.017	disease	0.023	into	0.022	develop	0.017
9	not	0.016	result	0.023	contaminate	0.019	solid	0.016
10	even	0.016	time	0.021	dump	0.015	find	0.015
11	walk	0.015	education	0.020	damage	0.011	most	0.014
12	school	0.014	from	0.020	kill	0.010	mile	0.014
13	them	0.014	this	0.017	ton	0.009	rock	0.013
14	family	0.013	hepatitis	0.016	delta	0.009	bacteria	0.011
15	need	0.013	typhoid	0.016	fertilizer	0.009	state	0.011
16	but	0.013	show	0.015	urbanization	0.009	available	0.011
17	think	0.011	fever	0.014	fish	0.008	debris	0.011
18	live	0.011	passage	0.013	supply	0.008	hot	0.011
19	thing	0.011	many	0.012	farm	0.007	bug	0.010
20	have	0.010	death	0.012	float	0.007	muddy	0.010
21	take	0.009	paragraph	0.012	and	0.007	children	0.009
22	mile	0.009	country	0.010	year	0.007	under	0.009
23	good	0.008	common	0.010	ocean	0.006	small	0.009
24	really	0.008	kid	0.009	food	0.006	less	0.009
25	every	0.008	their	0.009	last	0.006	water	0.009
26	day	0.007	most	0.008	agriculture	0.006	source	0.009
27	bad	0.007	through	0.007	factor	0.006	formation	0.008
28	time	0.007	pollute	0.007	contribute	0.006	country	0.008
29	problem	0.007	today	0.007	increase	0.006	dug	0.007
30	that	0.007	trip	0.006	state	0.006	accept	0.007

The comment assignment is provided by the trained raters, and the comments imply the quality of the students' answers. The interpretation of the comment codes differs across the scores, and these disparities will be explained in the following paragraph.

There are three comment codes: A, B, and C. However, even within the same code, the interpretation of the codes differs according to the scores. For example, code A has different meanings for score 1 and score 4. Specifically, the comment code A of score 1 reads, "You mostly wrote a summary of what you read in the reading passages. Next time, be sure to focus more on answering the question." However, the same comment of score 4 explains, "You did an excellent job including relevant details from the passages AND explaining how these details illustrate the consequences of the water crisis in parts of Africa." Thus, the comments have various interpretations according to which score is assigned (see Table 3).

Table 3 *The Comment Code Matrix*

Score	code A	code B	code C
0	Blank	Copied	Too limited to score, illegible, or incomprehensible
1	You mostly wrote a summary of what you read in the reading passages. Next time, be sure to focus more on answering the question.	You attempted to address the question, but you did not use relevant details from the passages to develop your essay.	Your response contains little original writing.
2	You started to answer the question, but you included limited details from the passages to develop your essay.	Many of the details you included from the passages do not clearly describe the consequences of the water crisis in parts of Africa.	You included relevant details in your essay, but you need to elaborate more on these details to answer the question.
3	You wrote a complete essay. It would be even better if you explained more consistently how the details you cited illustrate the consequences of the water crisis in parts of Africa.	You forgot to include an introduction, or the introduction could be clearer.	If you included more transitions, your essay would be better organized.
4	You did an excellent job including relevant details from the passages AND explaining how these details illustrate the consequences of the water crisis in parts of Africa.	You used many types of transitions to help organize your ideas. Keep it up!	Your introduction was clear and effective. Keep it up!

3.2. Analysis

Since the structures of ANNs are various, there are no strict guidelines to construct the network (Goodfellow et al., 2016). Thus, in this study, MLP was applied, and it consisted of three layers. The three layers consist of 5, 15, and 30 units respectively.

In this study, an R package neuralnet (Fritsch, Guenther, & Suling, 2012) was used to perform the analysis. Also, since RPROP is applied, the learning rate is automatically specified

through the algorithm. There are several criteria to evaluating the convergence of the training as follows: 1) training error; 2) gradient error; and 3) cross-validation. The second criterion, based upon the gradient error of the network, is one of the more popular approaches (Basheer & Hajmeer, 2000). Specifically, training quality is determined according to whether the gradient of the network is practically close to 0 or not. Furthermore, although when a gradient value of 0 indicates optimal training, this training might be ineffective if it requires too much calculation. Thus, the appropriate stopping rule needs to be applied to perform efficient training in common practice (Prechelt, 1998).

The package used in this study uses the value of the gradient as the stopping criteria. For example, a value 0.01 means that the training is stopped when the gradient is smaller than 0.01. The package suggested 0.01 as the default value. However, since the convergence failed with the suggested value, the larger values were attempted sequentially. Specifically, 10 values ranging from 0.01 to 1.0 with 0.01 interval were tried, and 0.50 is the smallest threshold that allows for the convergence of the network. Thus, 0.50 was used as the stopping criteria in this analysis. For the activation function, the sigmoid function was used, and the loss function is calculated as the sum of the square.

CHAPTER 4

RESULTS

4.1 Descriptive Statistics of Topic Proportion and Comment Code

The descriptive statistics regarding the topic proportions as estimated from LDA are shown in Table 4. Since the proportion is bounded with 0 and 1, the arcsine transformation was performed. The arcsine transformation is demonstrated in Equation (16).

$$p' = arcsin(p) \qquad (16)$$

where p is the topic proportion, arcsin(·) is arcsine function, and p' denotes the transformed value ranging from 0 to infinity (Prepas, 1984).

The most commonly used topics are topic 2 and topic 3 (0.57), with topic 4 emerging as the least frequent (0.30). Since the distribution of the transformed proportions rarely follow the normal distribution, the quantile values are also summarized in Table 4.

Table 4 *Descriptive Statistics of the Transformed Topic Proportion*

	Topic 1	Topic 2	Topic 3	Topic 4
Mean (SD)	0.32 (0.342)	0.57 (0.414)	0.57 (0.363)	0.30 (0.335)
Min	0.00	0.00	0.00	0.00
1st Quantile	0.00	0.20	0.32	0.00
Median	0.23	0.59	0.58	0.19
2nd Quantile	0.56	0.85	0.80	0.53
Max	1.57	1.57	1.57	1.57

The relationships among the topic proportions (for simplicity, the transformed topic proportion will be labeled as a just topic proportion from this section forward) are summarized in Table 5. Since the sample size is large ($n = 7{,}431$), the correlations among the topic proportions are significant. There are negative relationships among the topic proportions, and it indicates that the use of a specific topic has a negative impact on the use of the other topic. However, this is to be expected because the sum of the topic proportion was originally constrained to 1. Although the topic proportion is transformed, the order of the topic proportion remains. Thus, they still have negative relationships with each other. The most significant point is that topic 2 shows a relatively high negative correlation with the other topics.

Table 5 *Correlation Matrix of Topic Proportion*

	Topic 1	Topic 2	Topic 3	Topic 4
Topic 1	-			
Topic 2	-.396 (<.001)	-		
Topic 3	-.203 (<.001)	-.489 (<.001)	-	
Topic 4	-.207 (<.001)	-.332 (<.001)	-.248 (<.001)	-

Note The values in the parenthesis denotes the p-value

The frequencies of the comment codes are described in Table 6. When the students scored a 0, most students (84.30%) received code B, which means they copied their answers. For score 1, almost half (46.96%) of the students received code A, meaning the answer is mostly a summary of the reading passages. Code C, indicating too little original writing, was provided for 36.92% of the students. Similarly, for the students who scored a 2, most received code A (36.97%) and code C (51.78%), and these codes indicate that the students provided too few

details in their answers. For score 3, most students (95.42%) received code A, which means that the general organization of the essay was good, but that more details were required. For the students who received a score of 4, code A, indicating excellent work, was provided for a vast majority (90.10%). Thus, the comment code assignments are unbalanced when the scores are extremely low or high (i.e., 0, 3, and 4). This extremely unbalanced frequency of the response variables may impact the classification accuracy of the models (e.g., logistic regression and neural network analysis).

Table 6 *The Frequencies of the Codes According to the Scores*

Score	code A	code B	code C	Total
0	2 (1.16)	153 (84.30)	31 (14.53)	186 (100.00)
1	1092 (46.96)	379 (16.11)	869 (36.92)	2,340 (100.00)
2	1124 (36.97)	342 (11.25)	1574 (51.78)	3,040 (100.00)
3	771 (95.42)	5 (0.62)	32 (3.96)	808 (100.00)
4	91 (90.10)	10 (9.90)	0 (0.00)	101 (100.00)
				6,475

Note. The values in the parentheses represent the percentile.

4.2 The Relationship between Comment and Topic Structure

A correlation analysis was performed to identify the relationships between the topic proportions and the comments assignments. Specifically, biserial correlation analysis was applied because the comment assignment is the binary variable (i.e., assigned=1 and not assigned=0) with the continuous topic proportions (see Table 7).

The results show that most of the values are lower than 0.3, indicating a weak relationship. Particularly, when the scores are 3 or 4, the correlations are smaller than when the scores are 0, 1 and 2. Also, most of the correlations are significant because the sample size is relatively large even though the correlation values are low.

Table 7 *The Biserial Correlation Between the Topic Proportion and the Feedback*

Score	code	Topic1	Topic2	Topic3	Topic4
0 (n=186)	code A (n=2)	-.052 (0.480)	-.006 (.940)	.001 (.990)	.084 (.250)
	code B (n=153)	.100 (.170)	-.130 (.067)	-.490 (<.001)	.370 (<.001)
	code C (n=31)	-.090 (.220)	.140 (.058)	.500 (<.001)	-.400 (<.001)
1 (n=2,340)	code A (n=1,092)	-.089 (<.001)	-.110 (.040)	.060 (.003)	.050 (.016)
	code B (n=379)	-.074 (<.001)	-.046 (.025)	.250 (<.001)	-.180 (<.001)
	code C (n=869)	-.030 (.150)	.150 (<.001)	-.250 (<.001)	.087 (<.001)
2 (n=3,040)	code A (n=1,124)	.026 (<.001)	-.083 (<.001)	.120 (<.001)	-.096 (<.001)
	code B (n=342)	.120 (<.001)	-.160 (<.001)	.058 (.002)	.110 (<.001)
	code C (n=1,574)	-.100 (<.001)	.180 (<.001)	-.150 (<.001)	.022 (.22)
3 (n=808)	code A (n=771)	-.066 (.060)	-.014 (.690)	.062 (0.078)	.008 (.820)
	code B (n=5)	.015 (.680)	.016 (.640)	-.028 (.430)	-.033 (.350)
	code C (n=32)	.065 (.065)	.008 (.810)	-.055 (.120)	.004 (.900)
4 (n=101)	code A (n=91)	.061 (.550)	-.059 (.560)	-.080 (.420)	.062 (.540)
	code B (n=10)	-.061 (.550)	.059 (.560)	.080 (.420)	-.062 (.540)
	code C (n=0)	NA	NA	NA	NA

4.3 The Prediction Quality of Logistic Regression

Evaluating the prediction quality of the logistic regression can be performed with cross-validation. Before the performing the cross-validation, the data set needs to be divided into a training data set and test data set (see Table 8). In common practice, there is no grounded guideline (i.e., the split ratio between the training data set and the test data set) to split the data set, except that the data sets should be exclusive. In this study, 80% of the data set is randomly chosen as the training data set, and rest of the data set is used as the test data set (Garson, 1998, p. 103).

Table 8 *The Results of Training and Test Data Sets Split*

	Training sample size	Test sample size	Total
0	148	38	186
1	1,872	468	2,340
2	2,432	608	3,040
3	646	162	808
4	80	21	101
total	5,786	3,121	8,907

Since the comment codes consist of more than 2 categories (e.g., Code A, Code B, and Code C), the multinomial logistic regression model seems to be natural to fit the data. However, the analysis failed to produce stable estimation results because of non-convergence of the algorithm. Also, the literature suggested the multiple binary logistic regression compared to the multinomial logistic regression when the underlying continuum is obscure (Agresti, 2003). Thus, the binary logistic regression models are applied to fit the data. The model used in this analysis is given as below Equation. Specifically, as the response variable is coded as the binary variable indicating

whether the comment is assigned or not, binary logistic regression models are used for three different comment codes assignments. The models are represented in Equation (17).

$$logit(p_A) = \beta_0 + \beta_1(t_1) + \beta_2(t_2) + \beta_3(t_3) + \beta_4(t_4)$$
$$logit(p_B) = \beta_0 + \beta_1(t_1) + \beta_2(t_2) + \beta_3(t_3) + \beta_4(t_4)$$
$$logit(p_c) = \beta_0 + \beta_1(t_1) + \beta_2(t_2) + \beta_3(t_3) + \beta_4(t_4) \quad (17)$$

where p_A, p_B, and p_C denote the probabilities of assignment of comment A, B, and C, respectively. Also, t_1, t_2, t_3, and t_4 denote the four topic proportions. The R program was used to perform the analysis (R Core Team, 2018). Since the comment interpretations are different across scores, separate analyses were performed for each score.

4.3.1 Score 0

The results of three binary logistic regressions for score 0 ($n = 148$) are summarized in Table 9. For the assignment of comment A, none of the proportions are significant. For the assignment of comment B, topic 3 and topic 4 proportions are significant. While the topic 4 proportion shows a positive effect on the assignment of the comment, topic 3 shows a negative effect on the assignment. In particular, the topic 4 proportion shows the strongest effect on the assignment. For the assignment of comment C, only the topic 4 proportion is significant, and this shows a negative effect on the assignment of the comment.

Since the unit of the predictors is the odds ratio of the transformed proportions, it might be difficult to interpret the magnitude of the coefficient in meaningful way. Alternatively, the direction of the effect might be useful in interpreting the effect of the topic proportion on the comment assignment. Specifically, the positive effect of a topic proportion on a comment assignment implies that a rater is more likely to assign the comment to a document dominated by the topic proportion. For example, if the topic 2 proportion shows a negative effect on the

assignment of the comment B, the rater is more likely to assign comment A to the document mostly composed of the topic 1.

Table 9 *The Results of the Binary Logistic Regression Models When the Score is 0*

		B	SE	df	*p*	Odds ratio	95% CI	
							Lower	Upper
Comment A	constant	6.131	5.243	1	.242	459.685		
	topic 1	1.137	0.862	1	.187	3.117	0.576	16.878
	topic 2	1.401	0.947	1	.139	4.058	0.635	25.949
	topic 3	1.291	0.831	1	.120	3.637	0.713	18.541
	topic 4	3.290	2.183	1	.132	26.850	0.372	1937.5
Comment B	constant	1.516	1.554	1	.329	4.554		
	topic 1	0.162	0.180	1	.368	1.175	0.827	1.671
	topic 2	-0.030	0.187	1	.874	0.971	0.673	1.400
	topic 3	-0.370	0.186	1	.046	0.691	0.480	0.994
	topic 4	0.431	0.167	1	.010	1.539	1.109	2.136
Comment C	constant	-2.090	1.695	1	.217	0.124		
	topic 1	-.0185	0.197	1	.348	0.831	0.565	1.223
	topic 2	0.003	0.205	1	.987	1.003	0.672	1.498
	topic 3	0.361	0.201	1	.073	1.434	0.967	2.127
	topic 4	-0.542	0.181	1	.003	0.581	0.408	0.829

The training classification results from the above model are summarized in Table 10. When the score is 0, the overall accuracy for the comment A is around 80%, which indicates relatively high performance. However, for the interpretation of the results, the imbalance between the response variable needs to be considered because most of the assignments are assigned as B. For example, since 120 students received comment B out of 148 students, even the model is not precise and just classifies every observation as a non-assignment, resulting in accuracy of more than 70%.

Table 10 *The Training Classification Results When the Score is 0*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	126	0	126	100.00
	1.00	2	0	2	0.00
Total		148	0	148	82.43
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	21	7	28	75.00
	1.00	89	31	120	25.83
Total		110	38	148	35.14
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	31	91	122	25.41
	1.00	7	19	26	73.08
Total		38	110	148	33.78

The test classification results when the score is 0 are summarized in Table 11, and the accuracy for three comments is similar with the values obtained in the training data set. The levels of accuracy for B and C comments are lower than the training classification results. Specifically, the accuracy values of comment A, B, and C are 100.00%, 31.57%, and 31.58%, respectively.

Table 11 *The Test Classification Results When the Score is 0*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	38	0	38	100.00
	1.00	0	0	0	0.00
Total		38	0	38	100.00
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	4	1	5	80.00
	1.00	25	8	33	24.24
Total		29	9	38	31.57
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	8	25	33	24.24
	1.00	1	4	5	80.00
Total		9	29	38	31.58

4.3.2 Score 1

The results of three binary logistic regressions (n=1,872) when the score is 1 are summarized in Table 12. For the assignment of comment A, all topic proportions show significantly positive effects on the assignment. For the assignment of comment B, topic 1, topic 2, and topic 4 proportions are significant. All topic proportions show negative effects on the assignment of the comment. Especially, the topic 4 proportion shows the strongest effect on the assignment. For the assignment of comment C, all topic proportions show significantly positive effects on the assignment.

Table 12 *The Results of the Binary Logistic Regression Models When the Score is 1*

		B	SE	df	*p*	Odds ratio	95% CI	
							Lower	Upper
Comment A	constant	2.559	0.276	1	.000	12.924		
	topic 1	0.327	0.034	1	.000	1.386	1.296	1.483
	topic 2	0.230	0.036	1	.000	1.259	1.174	1.350
	topic 3	0.360	0.038	1	.000	1.433	1.330	1.543
	topic 4	0.305	0.034	1	.000	1.356	1.268	1.450
Comment B	constant	-2.820	0.394	1	.000	0.060		
	topic 1	-0.117	0.048	1	.014	0.890	0.811	0.977
	topic 2	-0.065	0.051	1	.203	0.937	0.848	1.036
	topic 3	0.170	0.059	1	.004	1.185	1.056	1.329
	topic 4	-0.268	0.049	1	.000	0.765	0.696	0.842
Comment C	constant	-2.598	0.293	1	.000	0.074		
	topic 1	-0.242	0.036	1	.000	0.785	0.732	0.842
	topic 2	-0.171	0.038	1	.000	0.843	0.783	0.907
	topic 3	-0.431	0.039	1	.000	0.650	0.602	0.701
	topic 4	-0.156	0.035	1	.000	0.856	0.799	0.917

The training classification results from the above model are summarized in Table 13. When the score is 1, the accuracies for three comments vary from 27.46% to 51.87%. Most of the assignments are A or C. Approximately, almost half of the assignments are A, and a similar amount of the assignment is C.

Table 13 *The Training Classification Results When the Score is 1*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	948	39	987	96.05
	1.00	862	23	885	2.60
Total		1,810	62	1,872	51.87
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	244	1328	1,572	15.52
	1.00	30	270	300	90.00
Total		274	1,598	1,872	27.46
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	34	1,151	1,185	2.87
	1.00	39	648	687	94.32
Total		73	1,799	1,872	36.43

The test classification results when the score is 1 are summarized in Table 14, and the accuracies for three comments are similar with the values obtained in the training data set. The levels of accuracy for all comments are higher than those present in the training results. Specifically, the accuracy values of comments A, B, and C are 54.91%, 27. 78%, and 38.46%, respectively.

Table 14 *The Test Classification Results When the Score is 1*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	253	8	261	96.93
	1.00	203	4	207	1.93
Total		456	12	468	54.91
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	59	330	389	15.17
	1.00	8	71	79	89.87
Total		67	401	468	27.78
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	6	280	286	2.10
	1.00	8	174	182	95.60
Total		14	454	468	38.46

4.3.3 Score 2

The results of three binary logistic regressions (n=2,432) when the score is 2 are summarized in Table 15. For the assignment of comment A, the topic 1, 2, and 4 proportions show significantly a negative effect on the assignment. The topic 4 proportion in particular shows the strongest effect on the assignment. For the assignment of comment B, topic 1, 3 and topic 4 proportions are significant, and these proportions show positive effects on the assignment of the comment. Especially, the topic 1 proportion shows the strongest effect on the assignment. For the assignment of comment C, the topic 2 and 4 proportions are significant. Specifically, both topic proportions show positive effects on the assignment of the comment.

Table 15 *The Results of the Binary Logistic Regression Models When the Score is 2*

		B	SE	df	*P*	Odds ratio	95% CI	
							Lower	Upper
Comment A	constant	-1.553	0.234	1	.000	0.212		
	topic 1	-0.077	0.032	1	.015	0.926	0.870	0.985
	topic 2	-0.179	0.039	1	.000	0.836	0.774	0.903
	topic 3	-0.006	0.038	1	.866	0.994	0.923	1.070
	topic 4	-0.199	0.032	1	.000	0.820	0.770	0.873
Comment B	constant	0.109	0.344	1	.751	1.115		
	topic 1	0.359	0.051	1	.000	1.432	1.295	1.583
	topic 2	0.050	0.055	1	.358	1.052	0.945	1.171
	topic 3	0.287	0.065	1	.000	1.332	1.172	1.515
	topic 4	0.326	0.051	1	.000	1.386	1.255	1.531
Comment C	constant	0.276	0.224	1	.219	1.318		
	topic 1	-0.045	0.031	1	.138	0.956	0.900	1.015
	topic 2	0.168	0.039	1	.000	1.182	1.096	1.276
	topic 3	-0.067	0.037	1	.065	0.935	0.870	1.004
	topic 4	0.081	0.031	1	.009	1.084	1.020	1.152

The training classification results from the above model are summarized in Table 16. When the score is 2, the accuracies for three comments vary from 37.09% to 44.90%. Most of the assignments are A or C, each comprising approximately half of the total number.

Table 16 *The Training Classification Results When the Score is 2*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	36	1,499	1,535	2.35
	1.00	31	866	897	96.54
Total		67	2,365	2,432	37.09
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	2,165	0	2,165	100.00
	1.00	267	0	267	0.00
Total		2,432	0	2,432	47.86
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	958	206	1,164	82.30
	1.00	1,134	134	1,268	10.57
Total		2,092	340	2,432	44.90

The test classification results when the score is 2 are summarized in Table 17, and the accuracy for three comments is similar with the values obtained in the training data set. The levels of accuracy for all comments are higher than the training results. Specifically, the accuracy values of comments A, B, and C are 37.34%, 62.66%, and 46.55%, respectively.

Table 17 *The Test Classification Results When the Score is 2*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	6	375	381	1.57
	1.00	6	221	227	97.36
Total		12	596	608	37.34
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	533	0	533	100.00
	1.00	75	0	75	0.00
Total		608	0	608	62.66
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	257	45	302	85.10
	1.00	280	26	306	8.50
Total		537	71	608	46.55

4.3.4 Score 3

The results of three binary logistic regressions (n=646) when the score is 3 are summarized in Table 18. For the assignment of three comments, none of the proportions are significant.

Table 18 *The Results of the Binary Logistic Regression Models When the Score is 3*

		B	SE	df	*P*	Odds ratio	95% CI	
							Lower	Upper
Comment A	constant	1.850	1.081	1	.087	6.360		
	topic 1	-0.282	0.160	1	.079	0.755	0.551	1.033
	topic 2	-0.282	0.295	1	.339	0.755	0.424	1.344
	topic 3	0.033	0.195	1	.867	1.033	0.705	1.514
	topic 4	-0.073	0.155	1	.639	0.930	0.687	1.259
Comment B	constant	-7.076	2.910	1	.015	0.001		
	topic 1	-0.040	0.389	1	.917	0.960	0.448	2.057
	topic 2	-0.230	0.644	1	.721	0.795	0.225	2.809
	topic 3	-0.345	0.448	1	.441	0.708	0.294	1.703
	topic 4	-0.437	0.431	1	.311	0.646	0.278	1.503
Comment C	constant	-1.478	1.171	1	.207	0.228		
	topic 1	0.339	0.176	1	.054	1.404	0.994	1.982
	topic 2	0.379	0.329	1	.249	1.461	0.767	2.784
	topic 3	0.031	0.215	1	.885	1.032	0.677	1.571
	topic 4	0.158	0.168	1	.347	1.171	0.843	1.626

The training classification results from the above model are summarized in Table 19. When the score is 3, the overall accuracy is around 10%, indicating relatively low performance. Specifically, the accuracy values of comments A, B, and C are 10.84%, 1.55%, and 13.47%, respectively.

Table 19 *The Training Classification Results When the Score is 3*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	32	2	34	94.12
	1.00	574	38	612	6.21
Total		606	40	646	10.84
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	5	636	641	0.78
	1.00	0	5	5	100.00
Total		5	641	646	1.55
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	61	556	617	9.89
	1.00	3	26	29	89.66
Total		64	582	646	13.47

The test classification results when the score is 3 are summarized in Table 20, and the accuracy for three comments is similar with the values obtained in the training data set. Specifically, the accuracy values of comments A, B, and C are 5.56%, 1.85%, and 8.64%, respectively.

Table 20 *The Test Classification Results When the Score is 3*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	3	0	3	100.00
	1.00	153	6	159	3.77
Total		156	6	162	5.56
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	3	159	162	1.85
	1.00	0	0	0	0.00
Total		3	159	162	1.85
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	11	148	159	6.92
	1.00	0	3	3	100.00
Total		11	151	162	8.64

4.3.5 Score 4

The results of three binary logistic regressions ($n = 80$) when the score is 4 are summarized in Table 21. For the assignment of three comments, none of the proportions are significant.

Table 21 *The Results of the Binary Logistic Regression Models When the Score is 4*

		B	SE	df	*P*	Odds ratio	95% CI	
							Lower	Upper
Comment A	constant	-2.903	3.454	1	.401	0.055		
	topic 1	-0.446	0.425	1	.294	0.640	0.278	1.473
	topic 2	-1.822	1.316	1	.166	0.162	0.012	2.131
	topic 3	-1.507	1.009	1	.135	0.222	0.031	1.602
	topic 4	-0.486	0.415	1	.241	0.615	0.273	1.386
Comment B	constant	2.903	3.454	1	.401	18.220		
	topic 1	0.446	0.425	1	.294	1.562	0.679	3.594
	topic 2	1.822	1.316	1	.166	6.183	0.469	81.478
	topic 3	1.507	1.009	1	.135	4.512	0.624	32.612
	topic 4	0.486	0.415	1	.241	1.627	0.721	3.667

The training classification results from the above model are summarized in Table 22. When the score is 4, the overall accuracy is around 87.5%, indicating relatively high performance. However, for the interpretation of the results, the imbalance between the response variables needs to be considered, because the most commonly assigned comment code is A. For example, since 70 out of 80 students received comment A, even if the model were not precise and just classified every observation as a non-assignment, the accuracy would still be extremely high.

Table 22 *The Training Classification Results When the Score is 4*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	0	10	10	0.00
	1.00	0	70	70	100.00
Total		0	80	80	87.50
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	70	0	70	100.00
	1.00	10	0	10	0.00
Total		80	0	80	87.50

The test classification results when the score is 3 are summarized in Table 23, and the accuracy for three comments is similar with the values obtained in the training data set. Specifically, both accuracy values of comments A and B equal 100.00%. It is important, moreover, to carefully interpret the results, because there are no observations for the comment B.

Table 23 *The Test Classification Results When the Score is 4*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	0	0	0	100.0
	1.00	0	21	21	100.0
Total		0	21	21	100.0
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	21	0	21	100.0
	1.00	0	0	0	100.0
Total		21	0	21	100.0

4.4 The prediction quality of a neural network

To evaluate the prediction quality of the neural network also requires cross-validation because the model is typically over-fitted. The same training and test data sets shown in Table 8 were used.

4.4.1 Score 0

The training classification results when the score is 0 are summarized in Table 24. When the score is 0, the overall accuracy is around 90%, indicating relatively high performance. However, for the interpretation of the results, the imbalance between the response variables needs to be considered because B is the most commonly assigned code. For example, since 120 out of 134 students received comment B, even were the model not precise and simply classified every observation as a non-assignment, the accuracy would still be 99.0%.

Table 24 *The Training Classification Results When the Score is 0*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	146	0	146	100.00
	1.00	2	0	2	0.00
Total		148	0	148	98.65
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	13	15	28	46.43
	1.00	1	119	120	99.17
Total		14	134	148	89.19
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	122	0	122	100.00
	1.00	26	0	26	0.00
Total		148	0	148	82.43

The test classification results when the score is 0 are summarized in Table 25. When the score is 0, the overall accuracy is around 90%, which indicates relatively high performance, and this might be caused by the imbalance of the observations in the response variable.

Table 25 *The Test Classification Results When the Score is 0*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	38	0	38	100.00
	1.00	0	0	0	0.00
Total		38	0	38	100.00
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	3	2	5	60.00
	1.00	1	32	33	96.97
Total		4	34	38	92.11
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	33	0	33	100.00
	1.00	5	0	5	0.00
Total		38	0	38	86.84

4.4.2 Score 1

The training classification results when the score is 1 are summarized in Table 26. When the score is 1, the accuracy for three comments varies from 63.73% to 83.97%. Most of the assignments are A or C, with each code assigned to approximately half the responses. Thus, since B is assigned in only the remaining minority of cases, the accuracy of the comment B is relatively high (83.97%), because of this imbalance in the data. The classification results regarding other comments (A and C) are 63.73% and 67.79%, respectively.

Table 26 *The Training Classification Results When the Score is 1*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	648	346	994	65.19
	1.00	333	545	878	62.07
Total		981	891	1,872	63.73
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	1,572	0	1,572	83.97
	1.00	300	0	300	16.03
Total		1,872	0	1,872	83.97
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	980	205	1,185	63.30
	1.00	398	289	687	36.70
Total		1,378	494	1,872	67.79

The test classification results when the score is 1 are summarized in Table 27, and the accuracy for three comments is similar with the values obtained in the training data set.

Table 27 *The Test Classification Results When the Score is 1*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	158	96	254	62.20
	1.00	86	128	214	59.81
Total		244	224	468	61.11
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	389	0	398	100.00
	1.00	79	0	79	0.00
Total		468	0	468	83.12
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	242	44	286	84.62
	1.00	107	75	182	41.21
Total		349	119	468	67.74

4.4.3 Score 2

The training classification results when the score is 2 are summarized in Table 28. When the score is 2, the accuracies for three comments range from 76.40% to 94.98%. Most of the assignments are A or C, with each code assigned in approximately half of the responses. Thus, only the small number of remaining responses are assigned comment B. Because of this imbalance in the date, the accuracy of the comment B is relatively high (94.98%). The accuracy results regarding other comments (A and C) are 76.40% and 77.18%, respectively.

Table 28 *The Training Classification Results When the Score is 2*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	1,363	172	1,535	88.79
	1.00	402	495	897	55.18
Total		1,765	667	2,432	76.40
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	2,153	10	2,163	99.54
	1.00	112	157	269	58.36
Total		2,265	167	2,432	94.98
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	903	261	1,164	77.58
	1.00	294	974	1,268	76.81
Total		1,197	1,235	2,432	77.18

The test classification results when the score is 2 are summarized in Table 29, and the accuracies for three comments are similar with the values obtained in the training data set. The levels of accuracy for three comments are lower than those found in the training results. Specifically, the accuracy values of comments A, B, and C are 59.21%, 81.41%, and 48.19%, respectively.

Table 29 *The Test Classification Results When the Score is 2*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	285	96	381	74.80
	1.00	152	75	227	33.04
Total		437	171	608	59.21
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	487	48	535	91.03
	1.00	65	8	73	10.96
Total		552	56	608	81.41
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	212	90	302	70.20
	1.00	225	81	306	26.47
Total		437	171	608	48.19

4.4.4 Score 3

The training classification results when the score is 3 are summarized in Table 30. When the score is 3, the values of the accuracy are ranged from 70.27% to 99.23 %. Since most of the assignments are A, the training accuracy for comments B and C is almost 100%.

Table 30 *The Training Classification Results When the Score is 3*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	10	24	34	29.14
	1.00	168	444	612	72.55
Total		178	468	646	70.27
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	641	0	641	100.00
	1.00	5	0	5	0.00
Total		646	0	646	99.23
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	617	0	617	100.00
	1.00	29	0	29	0.00
Total		646	0	646	95.51

The test classification results when the score is 3 are summarized in Table 31, and the accuracies for three comments are similar with the values obtained in the training data set. The levels of accuracy for three comments are lower than those from the training results. Specifically, the accuracy values of comment A, B, and C are 70.99%, 100.00%, and 98.15%, respectively.

Table 31 *The Test Classification Results When the Score is 3*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	2	1	3	66.67
	1.00	46	113	159	71.07
Total		48	114	162	70.99
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	162	0	162	100.00
	1.00	0	0	0	100.00
Total		162	0	162	100.00
		predicted comment C		Total	correct %
		0.00	1.00		
observed comment C	0.00	159	0	159	100.00
	1.00	3	0	3	0.00
Total		162	0	162	98.15

4.4.5 Score 4

The training classification results when the score is 4 are summarized in Table 32. When the score is 4, the values of the accuracy all equal 87.50%.

Table 32 *The Training Classification Results When the Score is 4*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	0	10	10	0.00
	1.00	0	70	70	100.00
Total		0	80	80	87.50
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	70	0	70	100.00
	1.00	10	0	10	0.00
Total		80	0	80	87.50

The test classification results when the score is 4 are shown in Table 33. Although the values of accuracy are 100.00%, careful interpretation is necessary because the data set is extremely unbalanced.

Table 33 *The Test Classification Results When the Score is 4*

		predicted comment A		Total	correct %
		0.00	1.00		
observed comment A	0.00	0	0	0	100.00
	1.00	0	21	21	100.00
Total		0	21	21	100.00
		predicted comment B		Total	correct %
		0.00	1.00		
observed comment B	0.00	21	0	21	100.00
	1.00	0	0	0	100.00
Total		21	0	21	100.00

CHAPTER 5

CONCLUSION

The present study compared the performance of neural network analysis with the performance of logistic regression. The response variable of the model is a comment assignment by a human rater, and the four predictors are topic proportions estimated from LDA. The constructed models for both analyses are mainly concerned with predicting the comment assignment by using the topic proportions as the predictors. The main goal of the study was to compare the performance of the two methods, and this performance was evaluated by examining each technique's accuracy-- the ratio of correct classification.

The comment codes indicate the quality of the students' writing. Furthermore, interpretations can vary even within the same comment code depending on the assigned score. To account for this variance, five separate analysis for each score (i.e., 0, 1, 2, 3, and 4) were performed.

First, preceding the main analysis, the relationships among the variables were identified based on biserial correlations. The results show that the biserial correlations between the comment code and the score are relatively substantial when the score is 0, 1, and 2. However, the relationship becomes unclear when the scores rise to 3 or 4.

With regard to logistic regression, the results demonstrated the unclear relationship between the score and the interpretation of the comment code. For example, when the score is 3 and 4, none of the predictors are significant. In particular, classification accuracy was only acceptable when the comment code assignment was extremely unbalanced. For example, the

high correct classification accuracies are only observed when most of the students either did not receive or received the comment (e.g., score 4).

For neural network analysis, the accuracy of the test data set is generally higher than the accuracy of the cross-validation quality of the logistic regression, and these results are well matched with previous empirical studies (Nefeslioglu, Gokceoglu, & Sonmez, 2008; Yilmaz, 2010). Also, the accuracy of both models is relatively high when the scores are 0, 3, and 4. These results might be contradictory with the results of the biserial correlation because the magnitude of the correlations is relatively low when the score is higher than 2. Thus, the most likely conclusion is that the high performances of both models are caused by the imbalance of the observations in the response variable. In other words, when scores are 1 and 2, indicating a relatively large sample size, the results which show lower accuracy seem to be more reliable because the response variable (comment assignment) can be considered more balanced (i.e., the data is still unbalanced but it is relatively balanced compared to the other scores). Specifically, when the sample size is small, most of the observations are just going to assign 0 for a specific comment code, and the remaining observations are given the other comment code.

Also, although the general accuracy of the neural network is not remarkably higher than the accuracy of logistic regression, the salient point is that the neural network tends to show higher accuracy for both responses (i.e., assignment and non-assignment). For example, when the score is 2, most of the comment codes are A and C. Thus, the accuracy ranges from 0.00% to 100.00% when using logistic regression. On the other hand, the accuracy of the neural network ranges from 36.7% to 99.5%. Since the data used in this study is unbalanced, it can be concluded that the neural network yields better results for unbalanced data. However, these results may contradict a previous study (Crone & Finlay, 2012) because the previous study pointed out that

logistic regression is more robust than neural network analysis when the data is unbalanced. These contrasting results might be caused because the predictors have unclear relationships with the response variable, which can be inferred by the results of the biserial correlation. Specifically, although most of the coefficients are going to be significant in the logistic regression model, the classification results indicated that the models seemed to be unable to predict the response when the sample size was small or unbalanced. On the other hand, neural networks produced better classification results than logistic regression, even though the use of this accuracy for practical purposes remains still questionable.

However, the results reveal the potential utility the neural network if a larger sample size is available in the future. Previous studies (Zhang, Hu, Patuwo, & Indro, 1999; Olden, Joy, & Death, 2004; Chojaczyk, Teixeira, Neves, Cardoso, & Soares, 2015) have consistently reported that the performance of the model is improved when the sample size is increased. Also, when the sample size is sufficiently large, the ANN model tends to overcome functional model misspecification (Nghiep, & Al, 2001). In other words, classification performance might be improved even though the relationship between the variables is unclear. Additionally, compared to the other empirical study, which included a sample size of more than 50,000 subjects (Ramesh, Baskaran, Krishnamoorthy, Damodaran, & Sadasivam, 2018), the present sample may be too small to develop a stable neural network. This implies that neural network analysis is able to produce better results when the sample size is increased, and this is practically meaningful because the test data used in this study is state-level test data, which can potentially be cumulative.

Moreover, this study's the training quality can be considered lower than that of previous studies, which demonstrated accuracies of over 90% (Subasi & Ercelebi, 2005; Sahin & Duman,

2011; Shi, Lee, Ho, Sun, Wang, & Chiu, 2012). This might be caused by a lower threshold. In this study, a threshold of 0.5 was used, and this is larger than the default setting (0.01). As discussed in the method section, the model with the other setting failed to achieve convergence. Thus, the training quality also can be increased if the sample size or the relationship between the variables are improved.

As Tu (2005) pointed out, neural network model development is empirical, and many methodological issues remain to be resolved. Thus, in future study, another architecture, such as bridged multilayer perceptron (BMLP), might be used as an alternative method because it has been advanced as a more powerful architecture than the multilayer perceptron architectures used in this study (Wilamowski, 2009). Also, since the data are unbalanced, a novel loss function called mean false error, together with its improved version, mean squared false, could be used (Wang, Liu, Wu, Cao, Meng, & Kennedy, 2016).

Reference

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of pharmaceutical and biomedical analysis, 22*(5), 717-727.

Agresti, A. (2003). *Categorical data analysis*. Somerset, NJ: John Wiley & Sons.

Anastasiadis, A. D., Magoulas, G. D., & Vrahatis, M. N. (2005). New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing, 64*, 253-270.

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement, 74*, 795-808.

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. *Journal of microbiological methods, 43*(1), 3-31.

Bengio, Y. (2011, June). Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning* (pp. 17-36).

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade* (pp. 437-478). Springer, Berlin, Heidelberg.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*, 993-1022.

Chojaczyk, A. A., Teixeira, A. P., Neves, L. C., Cardoso, J. B., & Soares, C. G. (2015). Review and application of artificial neural networks models in reliability analysis of steel structures. *Structural Safety, 52*, 78-89.

Crone, S. F., & Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting, 28*(1), 224-238.

Fritsch, S., Guenther, F., & Suling, M. (2012). *neuralnet*, Training of Neural Networks, R package version 1.32.

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists*. Thousand Oaks, CA: Sage.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning*. Cambridge, MA: MIT press.

Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R journal, 2*(1), 30-38.

Heazlewood, I., Walsh, J., Climstein, M., Kettunen, J., Adams, K., & DeBeliso, M. (2016). A comparison of classification accuracy for gender using MLP, RBF procedures compared to discriminant function analysis and logistic regression. In *Proceedings of the 10th International Symposium on Computer Science in Sports (ISCSS)* (pp. 93-101). Springer, Cham.

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer, 29*(3), 31-44.

Jarrett, K., Kavukcuoglu, K., & LeCun, Y. (2009, September). What is the best multi-stage architecture for object recognition?. In *Computer Vision, 2009 IEEE 12th International Conference* on (pp. 2146-2153). IEEE.

Kim, S., Kwak, M., Cardozo-Gaibisso, L., Buxton, C. & Cohen, A.S. (2017). Statistical and qualitative analyses of students' answers to a constructed response test of science inquiry knowledge. *Journal of Writing Analytics, 1*, 83-102.

Kwak, M., Kim, S., & Cohen, A. (2017, January) Mining students constructed response answers. *Paper presented at the International Conference on Writing Analytics*, Tampa, FL.

Lee, I. (2011). Formative assessment in EFL writing: An exploratory case study. *Changing English, 18,* (1), 99-111.

Lek, S., & Guégan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological modelling, 120*(23), 65-73.

Liang, M., & Hu, X. (2015). Recurrent convolutional neural network for object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3367-3375).

Nefeslioglu, H. A., Gokceoglu, C., & Sonmez, H. (2008). An assessment on the use of logistic regression and artificial neural networks with different sampling strategies for the preparation of landslide susceptibility maps. *Engineering Geology, 97*(3-4), 171-191.

Nghiep, N., & Al, C. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks. *Journal of real estate research, 22*(3), 313-336.

Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling, 178*(3-4), 389-397.

Pietersma, A. D. (2010). *Feature space learning in support vector machines through dual objective optimization* (Unpublished master's thesis). University of Groningen, The Netherlands.

Prasad, N., Singh, R., & Lal, S. P. (2013, September). Comparison of back propagation and resilient propagation algorithm for spam classification. In *Computational Intelligence,*

Modelling and Simulation (CIMSim), 2013 Fifth International Conference on (pp. 29-34). IEEE.

Prechelt, L. (1998). Early stopping but when?. In *Neural Networks: Tricks of the trade* (pp. 55-69). Springer, Berlin, Heidelberg.

Prepas, E. E. (1984). Some statistical methods for the design of experiments and analysis of samples In J. A. Downing, & F. H. Rigler (Eds.). *A manual on methods of the assessment of secondary productivity in fresh waters* (pp. 266-335). Oxford, England: Blackwell Scientific Publications.

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

Ramesh, V. P., Baskaran, P., Krishnamoorthy, A., Damodaran, D., & Sadasivam, P. (2018). Back propagation neural network based big data analytics for a stock market challenge. *Communications in Statistics-Theory and Methods, 0*(0), 1-19.

Rautaray, S. S., & Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, *43*(1), 1-54.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*, *60*(5), 503-520.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited.

Sahin, Y., & Duman, E. (2011, June). Detecting credit card fraud by ANN and logistic regression. *In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on* (pp. 315-319). IEEE.

Sargent, D. J. (2001). Comparison of artificial neural networks with other statistical approaches: results from medical data sets. *Cancer: Interdisciplinary International Journal of the American Cancer Society, 91*(S8), 1636-1642.

Shi, H. Y., Lee, K. T., Lee, H. H., Ho, W. H., Sun, D. P., Wang, J. J., & Chiu, C. C. (2012). Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. *PloS one, 7*(4), e35781.

Subasi, A., & Ercelebi, E. (2005). Classification of EEG signals using neural network and logistic regression. *Computer methods and programs in biomedicine, 78*(2), 87-99.

Svozil, D., Kvasnicka, V., & Pospichal, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemometrics and intelligent laboratory systems*, 39(1), 43-62.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology, 49*(11), 1225-1231.

von Davier, M. (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika*, 1-11.

Wang, S., Liu, W., Wu, J., Cao, L., Meng, Q., & Kennedy, P. J. (2016, July). Training deep neural networks on imbalanced data sets. In *Neural Networks (IJCNN), 2016 International Joint Conference on* (pp. 4368-4374). IEEE.

Wilamowski, B. M. (2009). Neural network architectures and learning algorithms. *IEEE Industrial Electronics Magazine, 3*(4).

Wythoff, B. J. (1993). Backpropagation neural networks: a tutorial. *Chemometrics and Intelligent Laboratory Systems, 18*(2), 115-155.

Xie, Y., Lord, D., & Zhang, Y. (2007). Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis. *Accident Analysis & Prevention, 39*(5), 922-933.

Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European journal of operational research, 116*(1), 16-32.

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence, 137*(1-2), 239-263.