Research Article

Artificial Intelligence for Automated Scoring and Feedback in Chemistry Courses

Alex Rudniy, Ph.D. *Drew University, Madison, NJ*

Structured Abstract

- **Summary:** This work describes design and evaluation of several modern artificial neural network (ANN) models for automated scoring and automated feedback for lab reports in chemistry. Evaluation results demonstrated that these models achieved top performance for the given tasks. The developed models achieved classification accuracy of up to 89% for automated scoring, accuracy of up to 98% for automated splitting of laboratory reports into rubric sections, and accuracy of up to 90% for automated feedback in introductory chemistry courses.
- **Automated Scoring:** This work is aimed at producing automated tools for scoring and feedback to facilitate instructors with the initial phase while leaving the final decision to humans. Therefore, instead of forecasting actual numeric scores, this work intended to predict score buckets ('low score', 'intermediate score', and 'high score'). The automated scoring followed an actual process used by human raters in CHM 2045 General Chemistry Lab I and CHM 2046 General Chemistry Lab II taught at the University of South Florida in spring and fall of 2017. There, lab reports followed a predefined rubric, each section was graded separately, and the final score was set as a sum of all rubric sub-scores. Initially, the divide-and-conquer approach was applied manually to separate lab reports into sections according to the rubric, then to produce a training set for ANN.

  A hybrid ANN combining convolutional and recurrent layers was evaluated on thirteen datasets, one per rubric criterion without agreement in rubric scores—this was done in order to increase the dataset size since neural networks are known to perform better on larger sets of data. Accuracy of up to 89% was achieved.

- **Dividing Into Sections:** This work proceeded with designing an algorithm for dividing lab reports into sections according to the rubric. This task was deemed necessary to further extend the dataset to facilitate production of automated feedback models. A new dataset was constructed in a semi-automated fashion by labeling each line either “Title” or “Not a Title.” The hybrid neural network with convolutional and recurrent layers was applied for identifying section titles, which served as markers for section boundaries. After detecting section titles, dividing lab reports into rubric sections became a trivial task. The model achieved accuracy of 98%.
- **Automated Feedback:** Two ANN architectures were evaluated for automated feedback. First, the task was viewed as a sequence-to-sequence machine translation, where three ANN architectures were evaluated. Second, the task was represented as a machine comprehension problem where a question about a lab report was asked, and the presence, absence, or depth of the identified answer triggered certain feedback. Both these approaches showed unsatisfactory results.

  A search for a new solution led to transforming the task of generating automated feedback to text classification. Preliminary experiments showed low performance due to the appearance of several topics within a single reviewer’s comment. Once again, the divide-and-conquer approach was employed. Texts were re-assigned to several clusters in a way that texts within the same cluster had the same standardized feedback response. Next, deep convolutional ANN was trained and evaluated on this transformed dataset achieving accuracy up to 55% in the Methods sections.

  Another data transformation was done to improve the accuracy of automated feedback. In the dataset, a single section had several different corresponding comments. It was hypothesized that texts with multiple comments confused the algorithm and reduced accuracy. To overcome this issue, 14 new datasets were constructed transforming generating feedback to a binary classification task, i.e., is a particular comment appropriate to a particular text or not. For example, ANN estimated whether “Use Exact Amounts” comment was appropriate for a lab report or not. The accuracy of the newly constructed models improved significantly, reaching 90%..

Keywords: automated scoring, automated feedback, deep artificial neural networks, deep learning

1.0 Background

The United States is experiencing declining global literacy rates, dropping in 2012 from 10th in the world to 20th (Programme for International Student Assessment, 2012). Remarkably, only 24% of graduating students scored at the proficiency level for writing (National Center for Education Statistics, 2012); 57% of SAT takers did not qualify as college ready (College Board, 2013); and 31% of high school graduates failed to meet ACT College Readiness Benchmarks (ACT, 2014). Furthermore, faculty realize feedback is crucial to student learning and writing improvement (Bangert-Drowns, Kulik, C-L. C., Kulik, J.A., & Morgan, 1991; Hattie & Timperley, 2007), yet they avoid assigning writing because the grading process is so time consuming and frustrating (Sun, Harris, Walther, & Baiocchi, 2015). Existing measures of responding to student writing fail to provide students with the timely, helpful, critical feedback they need to improve as writers. Over the past 100 years, researchers have repeatedly criticized teachers' grades as subjective and unreliable measures of students' academic success (Starch & Elliot 1912, 1913a, 1913b; Sax, 1980; Huot, O'Neill, & Moore, 2010; Brookhart et al., 2016) and that teachers' comments on papers often lack helpful, critical commentary (Connors & Lunsford, 1988; Lunsford & Lunsford, 2008; Moxley, 1989, 1992; Schwartz, 1984; Sommers, 1982; Wyatt-Smith, 1997).

Moxley conducted 100 interviews with writing program administrators, STEM and humanities faculty, and students at U.S. universities (Moxley & Walkup, 2016). The interviews identified that chemistry and other STEM gateway courses were staffed almost exclusively by inexperienced graduate students. Administrators were concerned about the quality and fairness of the graduate students' feedback. They said they would love reporting tools that would provide measures of student improvement across drafts and projects, and help them mentor the graduatestudent instructors. Program directors were increasingly responsible for accreditation reporting so predictive tools that harness the intellectual work of instructors and students were appreciated. Furthermore, administrators were responsible to students and they were sensitive to students' complaints regarding random grading, ineffective feedback from instructors or peers, or absence of feedback. Administrators were concerned about retention in STEM, particularly of underrepresented minorities who they say disproportionately lack strong communication and scientific reasoning abilities. In turn, STEM faculty were terribly concerned about what they perceived to be poor communication and scientific reasoning capabilities on the part of students, and they also recognized that more practice writing lab reports and other scientific documents would be helpful, but they did not have the time to grade papers. Some faculty expressed concerns about their ability to provide helpful feedback. The instructors' primary pain point was the exhausting amount of time it took to grade papers. About half of the faculty said they had never tried peer review or had tried it once or twice and found that existing tools did not scaffold feedback processes or facilitate grading and accountability. They said they would be eager to assign more writing and peer reviews if artificial intelligence (AI) could be developed to scaffold and support scientific reasoning and writing. They believe students lack strong scientific reasoning and writing competencies, yet they avoid assigning writing because existing grading methods are deemed time-consuming or ineffective. Some administrators and instructors said they would be willing to use an independent system, but others said integration into their school's LMS tools was essential.

This work describes development of AI models that expedite document markup and writing program assessment, using a corpus of approximately 100,000 scored and commented-on essays. These novel AI technologies are aimed at empowering administrators to research, improve, and demonstrate the efficacy of their curriculum, mentor graduate students, and save their instructors time (by expediting grading and improving peer review). In particular, this work aimed at building AI that can (1) identify in student lab reports the required features; (2) score the features on a scale that matches those scores given by human raters; and (3) suggest comments for student writings.

Overall, AI tools that provide formative feedback, links to resources that clarify the feedback, and accurately score the quality of the feedback can be viewed as either a way of improving quality, decreasing workload, or increasing efficiency, which is needed in virtually every educational institution. Given teachers throughout secondary and postsecondary education avoid assigning writing because grading is time consuming and ineffective, it is reasonable to assume that improving feedback tools, particularly ones that involve AI, could have far reaching benefits. If students were better communicators, researchers, and collaborators, they would be more productive in school and later in their professional careers.

The innovation of this work is twofold: designing and training new AI in the form of new deep artificial neural networks (DANN) models for automated grading and commenting of student laboratory reports; and innovative representation of student drafts of laboratory reports and related comments with text features for model input. The term "deep learning" used in this paper refers to the area of artificial intelligence that studies deep artificial neural networks inspired by the structure of the human brain. While there is no strict rule that differentiates shallow vs. deep artificial neural networks, DANNs typically have hundreds or even thousands of layers (LeCun, Bengio & Hinton, 2015). The outcome will set track to designing modules for automated grading and commenting of superior quality, potentially outperforming human graders (Gulshan et al., 2016). The new scoring and commenting features aim to bring eventual benefits to various stakeholders. Student writers can expect real-time advice, peer reviewers will receive auto-suggested feedback, instructors will save time on grading and supervision, and administrators will attain enhanced learning outcomes.

.

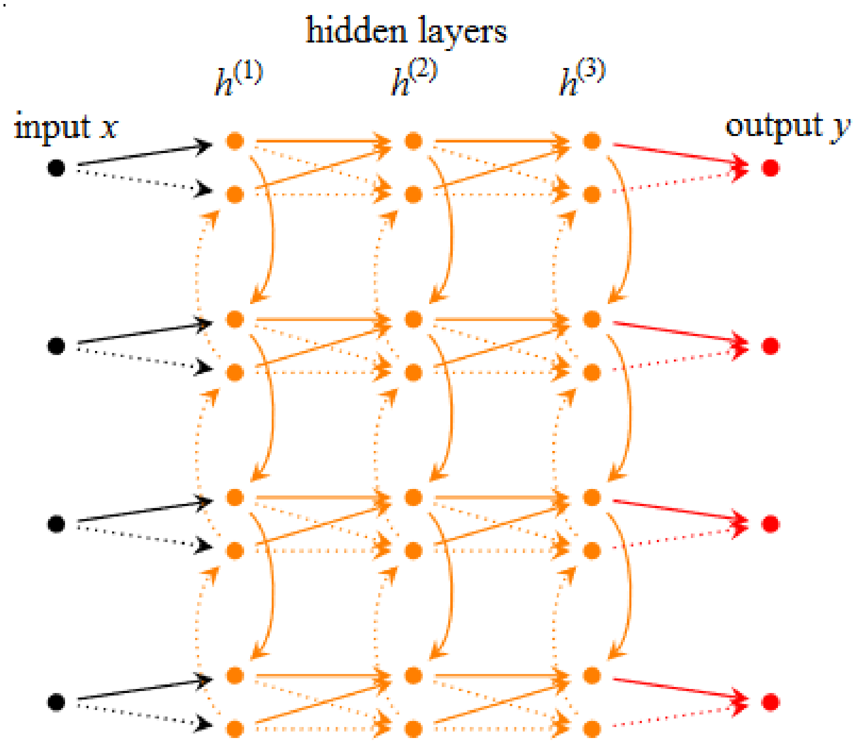
2.0 Literature Review

While bringing an innovation to the field of writing analytics, our work will build upon the latest advances of DANN for natural language processing (NLP) made in recent years. DANNs mimic the activity and structure of the human brain by stacking artificial neurons into multiple hidden and visible layers. By definition, an artificial neuron is a composite data structure with neural synapses modeled as weights, which are multiplied by input values transmitted through neuron synapses and subsequently added together. The neuron body is then presented as an activation function or a threshold indicating that a signal has passed through. Since the seminal work by McCullogh and Pitts (1943), DANNs evolved significantly, overcoming a limited learning ability of single-layer perceptions (Minsky and Papert, 1969) by adapting multi-layer models (Parker, 1985; LeCun, 1986) and finally emerging to complex DANN architectures providing state-of-the-art results in the image recognition and NLP domains (Collobert et al., 2011; Gulshan et al., 2016; Hinton et al., 2012; Kalchbrenner, Grefenstette and Blunsom, 2014).

This work considered the most successful properties of several Convolutional, Recurrent and Long Short-Term Memory (LSTM) DANNs, which demonstrated superior results in sentence classification, sentiment analysis, document summarization, query-based document retrieval, question answering, sentence modeling and general NLP tasks (Kim, 2014; Zhong et al, 2015; Shen et al., 2014; Yih et al., 2014; Kalchbrenner, 2014; Collobert et al., 2011). Figure 1 depicts a multi-layer bi-directional recurring DANN, with each layer feeding the next layer, with the exception of the output. At time-step *t*, each intermediate neuron receives one set of parameters from the previous time-step from the same layer, plus two sets of parameters from the lower layer (Mohammadi et al., 2015).

Figure 1

Multi-layer bi-directional recurring DANN



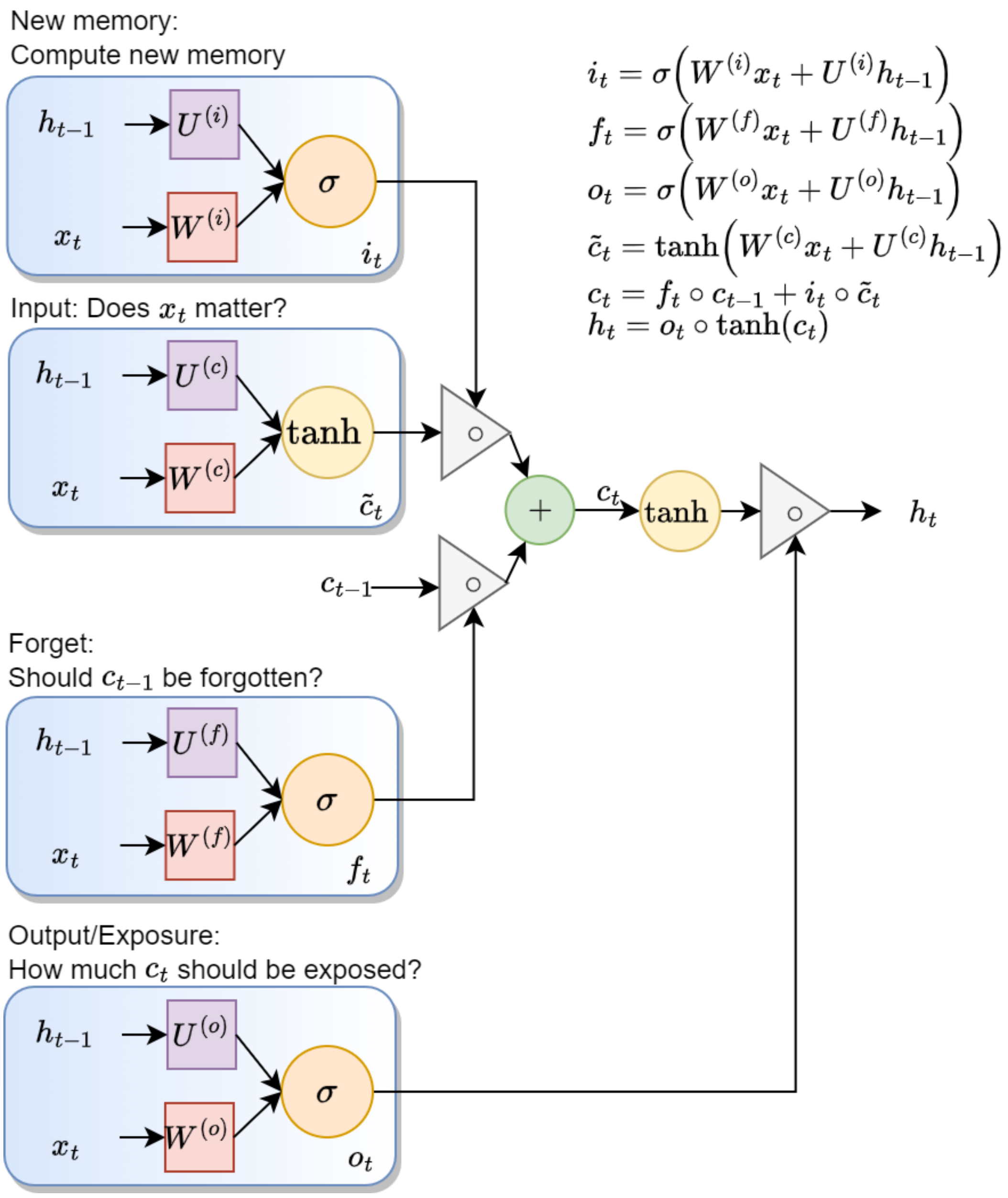
LSTM demonstrated promising results in automated question answering, machine translation, and modeling human conversation by selecting the next sentence.

2.1 Complex Structure of the LSTM Neuron.

We turn now to a brief explanation of LSTM modeling, based on Figure 2.

Figure 2

Structure of the LSTM Neuron


New memory:
Compute new memory
h_{t-1}
$U^{(i)}$
x_t
$W^{(i)}$
σ
i_t
Input: Does x_t matter?
h_{t-1}
$U^{(c)}$
x_t
$W^{(c)}$
$\tanh$
$\tilde{c}_t$
c_{t-1}
Forget:
Should c_{t-1} be forgotten?
h_{t-1}
$U^{(f)}$
x_t
$W^{(f)}$
σ
f_t
Output/Exposure:
How much c_t should be exposed?
h_{t-1}
$U^{(o)}$
x_t
$W^{(o)}$
σ
o_t
$\circ$
$+$
c_t
$\tanh$
h_t
$i_t = \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1}\right)$
$f_t = \sigma\left(W^{(f)}x_t + U^{(f)}h_{t-1}\right)$
$o_t = \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1}\right)$
$\tilde{c}_t = \tanh\left(W^{(c)}x_t + U^{(c)}h_{t-1}\right)$
$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$
$h_t = o_t \circ \tanh(c_t)$

In the above scheme, x_t is an input word vector at time t; sigmoid function σ outputs values between 0 and 1; hyperbolic tangent *tanh* outputs values between -1 and 1; matrices U and W store weights; input gate i_t, decides if the input should be discarded or kept based on its value and past hidden state. Gate f_t assesses usefulness of the past memory cells using input and past hidden state. Output/exposure gate o_t does not explicitly exist and is used to separate final memory from the hidden state. New memory cell $\check{c}_t$ uses the input word (or sentence) x_t and the past hidden state h_{t-1} to generate new memory. Final memory generation c_t sums the results of forgetting the past memory c_{t-1} according to forget state f_t with new memory $\check{c}_t$ processed according to the input gate i_t. The final neuron's decision is stored in the neuron output h_t. The purpose of the sigmoid function is to determine parts of the cell to be updated or forgotten while the hyperbolic tangent facilitates in determining the new candidate values for the cell state. Because this form of modeling has never been used for the application we describe, its use constitutes an innovation.

2.2 Avoiding Overfitting.

As an advanced machine learning algorithm, DANNs are prone to overfitting. While only minimal errors are produced during the training phase, errors become large when applying a model to unseen data. This phenomenon is known as overfitting due to situations when a DANN learned training examples but did not generalize to new ones. Two common solutions to overcome this error are (1) to use large training sets (Hinton et al., 2012) and (2) to increase the amount of noise in a training set (Socher, 2014). We will apply both these approaches by using large noisy datasets preserved in the MyReviewers chemistry data warehouse. The noise is understood as disagreement in rubric scores and comments among graders (Moxley & Eubanks, 2015) or as a measurement error. The amount of noise is a controlled parameter, which will be adjusted during the experimental phase.

3.0 Research Methodology

Design of a new DANN is a challenging task involving careful selection of the architecture and parameters of a neural network, e.g., to use a Convolutional NN or Restricted Boltzmann Machines NN, or Recurrent NN, etc.; the number of hidden layers; the number of neurons at each layer—an insufficient number will lead to weak expression power while a large number leads to prolonged running times and noisy outputs; and a learning rate parameter—when it is too high, the network will concentrate on a few last input examples, while a choice of a low rate value would lead to an increase in running time. As noted above, we will use a TensorFlow framework, which simplifies the process of designing a neural network to assembling a graph of nodes and edges, where nodes represent data transformation operations such as mathematical functions, or data input/output. Tensors—dynamic multi-dimensional arrays—are the structures carrying the data along the NN according to the directions specified by edges. This work constructed several DANN models and subsequently evaluated them by computing classification accuracy. Finally, better preserving models were typically kept for future use.

Classification accuracy represents the proportion of accurately classified instances for the entire dataset. Accuracy is commonly presented as a percentage, where a higher percentage indicates better performance. The rationale behind adopting classification accuracy as the performance evaluation metric in this study is multifaceted. Firstly, its simplicity and interpretability make it advantageous for effective communication with the non-technical audience addressed in this paper. Secondly, accuracy proves to be a fitting metric for the text classification tasks undertaken in this research due to the relatively balanced distribution of classes. Thirdly, the selection of accuracy aligns with common practice in evaluating deep learning models for text classification, as noted by Kilimci and Akyokus in their work (2018).

After a DANN is built, the next challenge of training the model comes into play. Usually, training begins with a random initialization of nodes, a process that is prone to reaching a local minimum and may ignore better solutions. With multiple hidden layers, DANN's training may be subject to overfitting, when a model produces correct output on a training set while tremendously degrading on new data (Bengio et al., 2003; Erhan et. al, 2010). A common solution to avoid overfitting is to present a DANN with a large input dataset containing noise, which helps generalization. The abundance of such data is stored in the recently built MyReviewers data warehouse. The presence of noise is explained by some peer graders inflating scores (Moxley & Eubanks, 2015). While not eliminating the noise completely, this work aims to reduce it by selecting papers that were assigned identical scores by instructors and students to create the training set.

3.1 General Research Design

The items below describe a process of designing and evaluating a DANN model.

1. Convert the corpus of essays, related comments, and rubric scores to numeric form readable by ANN.
2. Select configurations of the proposed models by deciding dimensionality of input, number of hidden layers, number of neurons per level, learning rate, and other parameters. Code models using an ANN framework with an appropriate programming language.
3. Perform training using computational resources designated for research in artificial intelligence. Store models for future re-use.
4. Conduct evaluation by applying selected models to testing data in automated fashion. Compare predicted values to actual ones for comments and scores. Calculate evaluation metrics, e.g., classification accuracy, for performance estimation.

3.2 Data Overview

This work used lab reports and corresponding rubric scores and rubric comments produced in University of South Florida CHM 2045/2046 chemistry courses taught in spring and fall of 2017. The following report structure was used by students and human raters (Table 1).

Table 1

Lab report structure and maximum scores per section for General Chemistry Lab Report

N	Report Section	Score
1	Introduction	9
2	Methods Part 1	3
3	Methods Part 2	3
4	Methods: Safety	3
5	Results Part 1	3
6	Results Part 2	3
7	Results: Calculations	3
8	Discussion Part 1	3
9	Discussion Part 2	3
10	Discussion: Sources of Error	3
11	Conclusion	3
12	Research Connection	3
13	References	3
14	Overall Format	5

The dataset was transformed and reshaped for modeling and evaluation. When evaluating ANN methods, several data subsets were produced and utilized in experiments. Certain data processing actions such as division into corpora of initial and final drafts, drafts with agreement in scores among at least two graders, and separation by course (CHM 2045 or CHM 2046) were done programmatically within a data warehouse. On the other hand, extraction of sections from lab report texts according to the structure shown in Table 1 and assigning rubric comments to categories, where each comment frequently contained several points such as “Clarify methods,” “Correct grammar,” and “Use exact amounts,” was done manually.

3.3 Text Representation

Commonly, textual data cannot be used as is as an input of an ANN model in the form of characters or tokens. Instead, numerical representations are used. In this work, we applied two approaches to represent text: document-term matrices (DTM), and word embeddings, known as word vectors.

A corpus of lab reports was represented as a DTM, where a row corresponded to a document, each word appearing at least once in the corpus was represented by a column, and a cell on the intersection of a row and a column contained a non-negative integer number, denoting the number of appearances of a particular word in a particular document.

Word embeddings of three types were used: (1) naïve vectors where each word was assigned a unique fixed-dimension vector, e.g. a vector with 300 elements, (2) word vectors ingested from the laboratory reports corpora using Word2Vec algorithm, and (3) GloVe word embeddings trained on Wikipedia articles combined with the English Gigaword corpus of newswire texts (Pennington, Socher & Manning, 2014). In approach (1), each vector is represented by a vector of random numbers that have no inherent meaning or correlation with the semantics of the words. Approach (2) uses numeric word representations that are tailored to the vocabulary and context of the lab reports corpora. These vectors capture domain-specific nuances and semantics. Approach (3) uses word representations that capture semantic relationships in vast datasets that might not reflect domain-specific nuances or contexts.

3.4 Solving Data Uncertainty with Interrater Agreement

This work tackled another problem related to data uncertainty. Moxley and Eubanks (2015) demonstrated that peers tended to inflate scores as compared to instructor evaluation. It is possible to overcome this issue by employing an interrater agreement of at least two graders, a student and an instructor, to consider a grade to be reliable (Rudniy and Elliot, 2016). DANNs applied to large datasets are known to overcome this type of uncertainty due to (1) the ability of deep learning models to generalize on noisy data and (2) the necessity of noise in the training set for avoiding overfitting.

To achieve interrater agreement, we used the divide-and-conquer approach by extracting those sections of lab reports in CHM 2045 and CHM 2046 which had matching scores from at least two graders. Models designed in this work did not assign exact scores. Instead, each text section was assigned one of three scores: “low,” “intermediate,” or “high.” Table 2 shows distribution of texts per lab report section. Column 1 contains the score as assigned by the rubric used in the lab, column 2 shows the number of texts per score, column 3 indicates the bucket to which texts with certain scores were assigned, and number and percentage of texts per bucket are in columns 4 and 5 respectively.

Table 2

Score assignment to "low," "intermediate," or "high" categories

(a) Introduction

Score	N	Bucket	Adjusted N	%
0	100	Low	122	18%
1	16			
2	6			
3	2	Medium	60	9%
4	14			
5	24			
6	20			
7	80	High	496	73%
8	152			
9	264			
Total			**678**	**100%**

(b) Methods 1

Score	N	Bucket	Adjusted N	%
0	140	Low	150	14%
0.5	10			
1	14	Medium	94	9%
1.5	42			
2	38			
2.5	140	High	798	77%
3	658			
Total			**1042**	**100%**

(c) Methods 2

Score	N	Bucket	Adjusted N	%
0	222	Low	232	22%
0.5	10			
1	10	Medium	84	8%
1.5	38			
2	36			
2.5	100	High	730	70%
3	630			
Total			**1046**	**100%**

(h) Discussion 1

Score	N	Bucket	Adjusted N	%
0	504	Low	518	14%
0.5	14			
1	14	Medium	74	9%
1.5	42			
2	18			
2.5	90	High	458	77%
3	368			
Total			**1050**	**100%**

(i) Discussion 2

Score	N	Bucket	Adjusted N	%
0	602	Low	612	55%
0.5	10			
1	16	Medium	62	5%
1.5	32			
2	14			
2.5	96	High	448	40%
3	352			
Total			**1122**	**100%**

(j) Discussion Sources of Error

Score	N	Bucket	Adjusted N	%
0	542	Low	548	43%
0.5	6			
1	8	Medium	38	3%
1.5	18			
2	12			
2.5	56	High	700	54%
3	644			
Total			**1286**	**100%**

(d) Methods Safety

Score	N	Bucket	Adjusted N	%
0	330	Low	334	28%
0.5	4			
1	4	Medium	70	6%
1.5	40			
2	26			
2.5	82	High	770	66%
3	688			
Total			**1174**	**100%**

(e) Results 1

Score	N	Bucket	Adjusted N	%
0	140	Low	150	14%
0.5	10			
1	14	Medium	94	9%
1.5	42			
2	38			
2.5	140	High	798	77%
3	658			
Total			**1042**	**100%**

(f) Results 2

Score	N	Bucket	Adjusted N	%
0	374	Low	378	34%
0.5	4			
1	0	Medium	40	4%
1.5	24			
2	16			
2.5	70	High	684	62%
3	614			
Total			**1102**	**100%**

(k) Conclusion

Score	N	Bucket	Adjusted N	%
0	520	Low	538	48%
0.5	18			
1	12	Medium	90	8%
1.5	48			
2	30			
2.5	98	High	494	44%
3	396			
Total			**1122**	**100%**

(l) Research Connection

Score	N	Bucket	Adjusted N	%
0	766	Low	776	58%
0.5	10			
1	8	Medium	48	3%
1.5	22			
2	18			
2.5	52	High	520	39%
3	468			
Total			**1344**	**100%**

(m) References

Score	N	Bucket	Adjusted N	%
0	542	Low	542	58%
0.5	0			
1	84	Medium	204	3%
1.5	0			
2	120			
2.5	0	High	674	39%
3	674			
Total			**1420**	**100%**

(g) Results Calculations

Score	N	Bucket	Adjusted N	%
0	610	Low	618	48%
0.5	8			
1	6	Medium	52	4%
1.5	30			
2	16			
2.5	44	High	624	48%
3	580			
Total			**1294**	**100%**

(n) Overall Format

Score	N	Bucket	Adjusted N	%
0	96	Low	116	13%
1	20			
2	18	Medium	360	41%
3	94			
4	248			
5	404	High	404	46%
Total			**880**	**100%**

3.5 Automated Scoring with ANN

Classic ANN proved to be unsuitable for sequential data such as text, due to the inability to preserve information about previous elements in a sequence. On the other hand, a convolutional neural network (CNN) uses fixed-size input and fixed-size output, making it unsuitable for texts while appropriate for image processing. Unlike CNNs, recurrent artificial neural networks (RNN) work with unrestricted lengths of input and output, preserve internal memory, and are commonly applied for textual data. Therefore, an RNN model was used for the automated scoring task.

RNNs' chain-like structure can be viewed as multiple copies of a single network. Advanced RNNs employing Long Short-Term Memory (LSTM) artificial neurons and their simplified version, Gated Recurrent Units (GRU), demonstrated high accuracy for natural language processing tasks.

Following recent advances in text classification (Xiao & Cho, 2016; Sainath, Vinyals, Senior & Sak, 2015), we applied a hybrid architecture combining convolutional and recurrent layers. The model was constructed of an embedding layer, followed by a convolutional, max-pooling, GRU, and softmax layers (Britz, 2016; Zhang, 2017; Chen, Ye, Xing, Chen & Cambria, 2017; Guggilla, Miller & Gurevych, 2016).

Thirteen models were designed, trained, and evaluated on thirteen datasets, one for each rubric section. The datasets were constructed by loading lab report sections with corresponding attributes into a relational database. Line breaks were removed from texts in order to fit one text into a single line of a file. Accuracy achieved by the models is listed in Table 3. Words were represented as numeric vectors of size 300, which were initialized with uniformly distributed random numbers (naïve word representations).

The models were coded using TensorFlow, an open-source machine learning framework developed by Google Research, and the Python programming language. It is worth noting that Support Vector Machines (SVM) that are known for their effectiveness in text classification demonstrated better performance compared to the hybrid model for the lab report scoring task.

SVM models were built by splitting the data by class, project, and rubric sections. This approach was not taken with the CNN-RNN-GRU models described in this section since it was known that artificial neural networks required larger dataset sizes for training. The underperformance of the neural network is explained by the relatively small training set.

Table 3

Accuracy for automated scoring with a hybrid neural network

N	Laboratory Report Section	Accuracy on Testing Set
1	Introduction	89%
2	MethodsPart1	88%
3	Methods Part 2	84%
4	Methods Safety	79%
5	Results Part 1	81%
6	Results Part 2	82%
7	Results Calculation	58%
8	Discussion Part 1	67%
9	Discussion Part 2	55%
10	Discussion: Sources of Error	80%
11	Conclusion	76%
12	Research Connection	75%
13	References	71%

3.6 Dividing Texts into Sections

The automated scoring (described above) and automated feedback (described below) approaches require texts split into sections according to the lab report rubric structure. This is not a trivial task since lab reports are stored as continuous unstructured bodies of texts without specific separators marking report sections. This separation had been done manually to establish initial datasets for training and validation. The process of manually splitting lab reports into parts had to be automated in order to overcome streamline scoring and feedback when developing a fully automated solution. Given the intricacy of this undertaking, a deep learning approach was deemed the most suitable for addressing this challenge.

The ANN applied for automated scoring (described above) was also applied to build a new model for automatically dividing texts into sections. To train the network, markers of section boundaries were needed. It was noticed that the first line of a section frequently contained that section’s title.

The following approach was taken: the first line of each lab report section was labeled as "Title," while the rest of the lines were marked as "Not a Title." Then these marked sections were used to reassemble the original lab report texts while preserving "Title" or "Not a Title" labels.

An ANN model was trained aiming at detecting section titles using 90% of data for training, and the remaining 10% for testing. The model achieved an accuracy of 98% on the testing set. After detecting section titles, dividing lab reports into rubric sections became a trivial task.

3.7 Automated Feedback as Sequence-to-Sequence Translation With ANN

This work considered only free-form rubric comments, discarding other types of feedback available within the raw data. A FairSeq neural network framework for sequence-to-sequence neural machine translation designed by Facebook AI Lab demonstrated inspiring results for English to German, English to French, and English to Romanian translation (Gehring, Auli, Grangier & Dauphin, 2016; Gehring, Auli, Grangier, Yarats & Dauphin, 2017). As FairSeq would translate a text in English to a corresponding text in German, this work proceeded with modeling automated feedback as a translation of a lab report in English to a corresponding comment in English. The dataset for this experiment was built using spring 2017 Methods Part 1 texts and feedback. The data were split into a training set (60%, N=1,941), a testing set (20%, N=647), and a validation set (20%, N=646).

The adequacy of the dataset size was substantiated by findings in related research, which established it as a suitable training set size for deep learning models. In the study conducted by Kuang, Dong, and Dong (2022), the accuracy of a deep learning model exhibited a noteworthy progression—from 55% with a training set of 288 examples, to 89% with 2160 examples, and further to 96% with a training set of 3671 examples. Lin, Huang, and Chen (2021) demonstrated a similar trend, showing an increase in accuracy from 97.8% with a 1000-item training set to 98.4% with a 2000-item training set, and eventually reaching 99.01% with a training set consisting of 6000 items. Moreover, Dutta and Gros (2018) illustrated substantial accuracy growth in a deep learning classification model across various domains. For anatomy data, accuracy improved from 74% to 88% as the dataset size expanded from 1450 examples to 29000. In the case of animal and vehicle data, accuracy increased from 47% to 71% with a dataset size escalation from 3000 examples to 60000. Similarly, for fashion data, accuracy rose from 83% to 92% with a dataset expansion from 3000 examples to 60000.

Three neural networks included in the FairSeq toolkit were evaluated for translating text to feedback: a standard bi-directional LSTM model, a fully convolutional sequence-to-sequence model, and a neural network with convolutional encoder and an LSTM decoder. Model parameters were adjusted during the evaluation phase. The trained models were evaluated manually for predicting comments for a lab report section. In all cases, the models generated the same comment for various content (Figure 3).

Figure 3

Sample feedback generated with FairSeq models

Lab Report Section 1 containing 119 tokens: “Methods: Part One Method: In this experiment we found the identity of the unknown salt by testing the substance’s properties. We did this by testing the solubility of the salt in three different solutions: ethanol, water, and acetone. We began by mixing 0.08 grams of the unknown substance into 10 mL of each three solutions. We found that the substance was soluble in water and ethanol, however not acetone. Upon finding this out, we tested the pH and conductivity of the substance in water to narrow in on what the substance might be. After finding these answers, we looked up salts that had similar properties as ours and were able to conclude that the unknown salt was MgSO4.”

Generated Feedback 1: “Strong Methods”

Lab Report Section 2 containing three tokens: “Week 1 Methods:”

Generated Feedback 2: “Strong Methods”

The feedback provided in Figure 3 was typical for the sequence-to-sequence translation approach to feedback demonstrating unsatisfactory outcomes and prompting a search for a new model.

3.8 ANN for Machine Comprehension

An ANN for machine comprehension was evaluated for a question-answering task, where a question was asked according to a laboratory manual, while the model identified an answer which was evaluated for the presence of a concept description as written by a student. To verify this technique, the AllenNLP toolkit of the Bi-Directional Attention Flow neural network was used (Seo, Kembhavi, Farhadi & Hajishirzi, 2016; Gardner et al., 2017). This framework demonstrated top accuracy on the SQuAD dataset (Rajpurkar, Zhang, Lopyrev & Liang, 2016).

A manual evaluation was completed using concepts from the CHM 2046 lab manual, Project 1 Calorimetry, the Concepts and Techniques part corresponding to the Methods section. In particular, terms included in the manual were used (Table 4).

Table 4

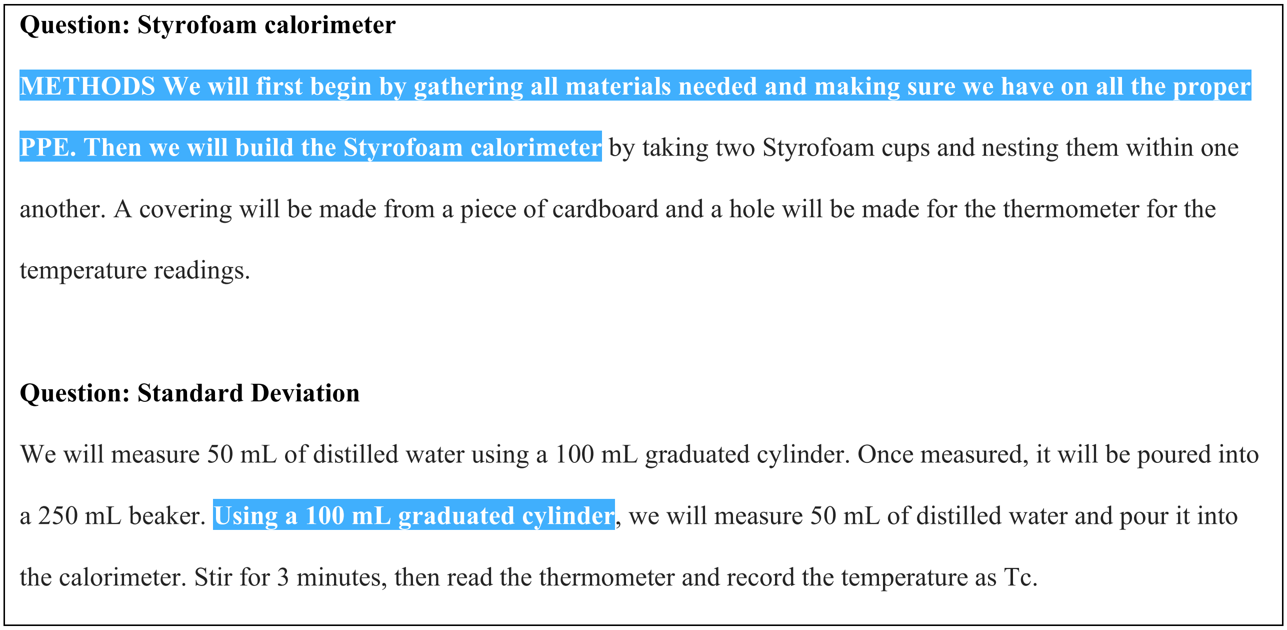
Laboratory Manual II, Project 1 Calorimetry. Template for entering experiment data information

	Heat capacity (C) average	Standard deviation (C)	Enthalpy chemical reaction (ΔH)	Standard deviation (ΔH)	Theoretical value of ΔH
Commercial calorimeter					
Styrofoam calorimeter					

The output produced by the model when supplying different concepts from Table 4 to the same laboratory report section was unsatisfactory as demonstrated in Figure 4. Figure 4 shows the output of the interface of the AllenNLP toolkit, where highlighted text indicates locations that the model identifies as corresponding to the posed questions. The deep learning model was asked to find and highlight text describing a given concept (the questions are shown in bold font) if it was found in the provided lab report excerpt (the paragraph under each question). This was done in an effort to request the artificial neural network to act similarly to a human grader who would scan a lab report to see whether a student elaborated on important concepts.

Figure 4

Sample ANN question answering results



Question: Styrofoam calorimeter

METHODS We will first begin by gathering all materials needed and making sure we have on all the proper PPE. Then we will build the Styrofoam calorimeter by taking two Styrofoam cups and nesting them within one another. A covering will be made from a piece of cardboard and a hole will be made for the thermometer for the temperature readings.

Question: Standard Deviation

We will measure 50 mL of distilled water using a 100 mL graduated cylinder. Once measured, it will be poured into a 250 mL beaker. **Using a 100 mL graduated cylinder**, we will measure 50 mL of distilled water and pour it into the calorimeter. Stir for 3 minutes, then read the thermometer and record the temperature as Tc.

3.9 Dataset Modification

A subsequent analysis conducted to identify an appropriate ANN model revealed that a single reviewer's comment contained multiple logical topics, e.g., a comment could criticize the absence of exact values while praising good references to the related work. To overcome this issue, three new feedback datasets were built. The manual processing included studying the comments, identifying several main topics in each comment, assigning topics to a comment, and re-organizing datasets in a format which could be used as an input for a neural network. Feedback categories containing fewer than 20 texts were removed. The structure showing comments by group with the included number of laboratory reports is shown in Table 5. As an additional point, the generated datasets could serve a dual purpose, not only for training artificial intelligence but also potentially aiding in the training of human graders.

Table 5. Summary of fall 2017 "Methods Part 1" feedback dataset

N	Group	Comment	Number of Texts
1	Amounts	Add amounts.	31
2	Amounts	Use exact amounts.	374
3	Grammar	Correct grammar.	127
4	Grammar	Use appropriate tense.	223
5	Headings	Organize/add headings.	116
6	Materials	Add materials.	29
7	Materials	Clarify materials.	115
8	Methods	Add methods.	92
9	Methods	Clarify methods.	369
10	Methods	Don't include analysis in the Method section.	179
11	Methods	Elaborate on methods.	1,175
12	Methods	Strong methods.	1,420
13	Names	Verify formulas, names, and abbreviations.	51
14	Organization	Format paragraphs.	293
15	Organization	Organize sentences.	189
16	Rubric	Add test.	36
17	Rubric	Follow rubric.	33
18	Rubric	Split Methods into Part 1 and Part 2.	91
19	Third Person	Speak in third person.	266
20	Units	Use proper units.	29
	Total		**5,238**

3.10 Deep Convolutional Neural Network for Feedback Generation

When proceeding with a model for feedback generation, we ensured that a single lab report does not have more than one comment from the same feedback group. For example, a text could be assigned either to “Amounts. Add amounts.” or to “Amounts. Use exact amounts.” but not to both. On the other hand, the same text could have several comments from different groups, e.g. a text could have four comments: “Grammar. Correct grammar,” “Materials. Clarify Materials,” “Methods. Elaborate on methods,” and “Rubric. Follow rubric.”

After conducting initial experiments, it was understood that the cases when a single text possesses several comments, as in the example above, served as noise leading to errors and decreased accuracy in a multi-class classification task. To overcome this problem, one more transformation was done which allowed us to represent feedback generation as a binary classification task. Another dataset was constructed by including a combination of feedback items from the original comment type, such as "Third Person. Speak in third person" (N=266), along with an equal number of comments that fell into other categories. Categories with low number of texts (N<100) were removed. The resulting datasets are listed in Table 6.

Table 6

Datasets for binary classification

N	Group	Comment	Number of Texts
1	Amounts	Use exact amounts.	1,488
2	Grammar	Correct grammar.	398
3	Grammar	Use appropriate tense.	812
4	Headings	Organize add headings.	462
5	Materials	Clarify materials.	476
6	Methods	Add methods.	250
7	Methods	Clarify methods.	1,268
8	Methods	Don’t include analysis in the Methods section.	496
9	Methods	Elaborate on methods.	3,614
10	Methods	Strong methods.	4,862
11	Organization	Format paragraphs.	486
12	Organization	Organize sentences.	544
13	Rubric	Split Methods into Part 1 and Part 2.	274
14	Third Person	Speak in third person.	876

A deep convolutional neural network (Cholett et al., 2015) was used, with texts represented as GloVe word embeddings trained on Wikipedia articles combined with the English Gigaword corpus of newswire texts (Pennington, Socher & Manning, 2014). The model was comprised of the static Keras embedding layer, followed by three pairs of a convolutional layer with Rectified

Linear Unit (ReLU) activation functions (Krizhevsky, Sutskever & Hinton, 2012) connected to a max pooling layer, and a densely-connected layer with ReLUs completing the network. The model used categorical cross-entropy loss function and RMSProp optimizer.

The conducted experiment used 80% of the data for training and 20% for testing, with batch sizes of 64 and 128, and embedding dimensions of 300.

Table 7

Results for the multiclass classification problem

	Training Accuracy			**Testing Accuracy**		
	Spring 2017	**Fall 2017**	**Spring & Fall 2017**	**Spring 2017**	**Fall 2017**	**Spring & Fall 2017**
All Sections	79%	80%	80%	31%	31%	33%
Methods	100%	100%	100%	55%	53%	54%
Amounts, Grammar, Headings, Materials	98%	96%	97%	42%	37%	44%
Organization, Rubric, Third Person	99%	98%	98%	44%	44%	43%

As shown in Table 7, the ANN models achieved high accuracy on the training set while lower accuracy on the test data, indicating the overfitting phenomenon. In the data studied in this work, a lab report section frequently was assigned feedback with multiple focuses. From the machine learning point of view, this may be viewed as noise when the same data instance belonged to several classes. The inadequate performance prompted the exploration of an alternative approach.

To overcome this issue, the problem was transformed into a binary classification task, where a model was trained to predict if a particular reviewer's comment belonged or did not belong to a lab report. For example, dataset 1 in Table 6 contained 1,488 texts with the comment "Use Exact Amounts." Subsequently, 1,488 data instances were randomly added from sets 2 to 14 (Table 6), ensuring that texts in newly added instances did not appear in set 1. Thus, the noise was removed from the data, and during the training process the ANN model was given a task to classify whether a text should be assigned "Use Exact Amounts" comment or not.

The accuracy of these models improved significantly as shown in Table 8, reaching 90% for "Add Methods." Nevertheless, accuracy of 58% for "Clarify Methods," 59% for "Clarify Materials," 59% for "Strong Methods," 61% for "Use Exact Amounts," can be improved. Hypothetically, this could be overcome by increasing the corpora sizes, adjusting ANN models' architecture and parameters, or experimenting with a scoring and feedback approach that harnesses the capabilities of machine learning.

Table 8

Binary classification

N	Group	Comment	Training Accuracy	Testing Accuracy
1	Amounts	Use exact amounts.	100%	61%
2	Grammar	Correct grammar.	100%	70%
3	Grammar	Use appropriate tense.	100%	69%
4	Headings	Organize add headings.	100%	65%
5	Materials	Clarify materials.	100%	59%
6	Methods	Add methods.	100%	90%
7	Methods	Clarify methods.	100%	58%
8	Methods	Don’t include analysis in the Methods section.	100%	81%
9	Methods	Elaborate on methods.	100%	62%
10	Methods	Strong methods.	100%	59%
11	Organization	Format paragraphs.	100%	73%
12	Organization	Organize sentences.	100%	62%
13	Rubric	Split Methods into Part 1 and Part 2.	100%	63%
14	Third Person	Speak in third person.	100%	75%

4.0 Discussion and Conclusions

Several multi-faceted outcomes were produced during the course of the project, including text representation models, models for automated scoring, automated feedback, and automated splitting texts into sections; as well as a number of datasets for scoring and feedback. Additionally, a number of high performing artificial neural networks (ANN) were evaluated. Finally, the produced models were constructed with re-usable technology. allowing streamlined integration into applications for use in academia.

4.1 Text Representation Models

This work produced two laboratory report representation models. First, the use of GloVe word embeddings trained on the corpus of spring 2017 and fall 2017 CHM 2045 and CHM 2046 laboratory reports. As an example, artificial intelligence learned from the context that the most similar words for “reaction” are “reactions” (similarity = 0.65), “eaction” (0.49), “rxn” (0.41), “neutralization” (0.40), “equation” (0.39), “exchange” (0.37), “titration” (0.36), and “decomposition” (0.35).

The second representation model consists of document-term matrices constructed from the laboratory reports corpora. Both models may be used in future work.

4.2 Automated Scoring Models

Automated scoring models were designed using word embeddings and a hybrid ANN with convolutional and recurrent layers. These models may be integrated into academic platforms. The models demonstrated relatively high accuracy.

4.3 Datasets for Automated Scoring

Datasets for training machine learning models for the automated scoring task were produced by applying custom programming, extensive analysis, and manual processing for CHM 2045 and CHM 2046: (1) initial drafts of lab reports, (2) final draft of lab reports, (3) thirteen datasets (one per rubric criterion) of initial draft with agreement in rater's scores, (4) 52 datasets of initial drafts (one per rubric criterion, course, and project) with agreement in scores, and (5) lab reports fitted into a single line in column 1 and score bucket in column 2.

4.4. Automated Splitting Into Sections

A dataset for splitting lab reports into sections was constructed of CHM 2045 and CHM 2046 texts with section header lines marked as "Title" and remaining lines marked as "Not a Title." The ANN-based model achieving high accuracy of 98% may be reused in future work.

4.5 Datasets for Automated Feedback

Several datasets were built for training ANNs for the automated feedback prediction: (1) a dataset for sequence-to-sequence translation comprised of training (60%), testing (20%) and validation (20%) sets; (2) twenty datasets of texts and related feedback, where each comment is mapped to one or several standardized categories; and (3) fourteen datasets for binary classification.

4.6 Automated Feedback Models

Two ANNs were evaluated for transforming sequence-to-sequence translation and machine comprehension to the task of producing automated feedback. Two models employing a deep convolutional neural network for feedback generation were built and evaluated, demonstrating high accuracy.

5.0 Directions for Further Research

Several directions for future work were identified. First, investigate the performance of the ANN models applying GloVe word embeddings trained on CHM 2045 and CHM 2046 corpora instead of the custom Word2Vec or pre-trained GloVe representations. Second, reviewers frequently advised a writer against using an example given in the laboratory manual. This type of feedback may be automated with the application of TF-IDF, n-grams, and other common approaches for similarity or plagiarism detection. Third, it is known that ANN performance increases with the size of an input dataset. We propose to retrieve and process a larger body of data and perform additional evaluation, which should also facilitate with avoiding models overfitting. Fourth, instead of keeping lab reports with matching scores from at least two graders, implement another approach that would model ratings as scalars using a random-effects model to account for rater variability as a source of variance orthogonal to "true" scores. Fifth, future research directions could include processing remaining sections, which were excluded from the scope of this work due to the time constraints. Additionally, it could be beneficial to study the effect of ANN architecture modifications, such as adding additional layers, changing layer activation functions, applying different optimizers, etc. Finally, the designed ANN could be deployed in an academic setting as an AI rater additionally to human graders. Feedback on AI performance provided by students and instructors could be used in further studies. On the other hand, there are a number of issues that should be anticipated: the inability of ANNs to generalize well to diverse writing styles; challenges in adjusting grading to varying expectations of different instructors; difficulties in providing nuanced feedback on experimental procedures, scientific reasoning, or analytical methods; an inability to interpret tables, graphs, charts, or diagrams; inflexibility to adapt to an evolving curriculum; and ethical considerations such as the unintentional reinforcement of specific writing styles.

Acknowledgments

This work was sponsored by NSF Grant # 1721749 SBIR Phase I: Artificial Intelligence, Scientific Reasoning, and Formative Feedback: Structuring Success for STEM Students and NSF Grant # 1636511 I-Corps: Investigating the Commercialization of Peer Review Tools and Writing Analytics.

Author Biography

Alex Rudniy, Ph.D. is an Assistant Professor of Computer Science and Co-Director of Data Science at Drew University. He specializes in research focused on artificial intelligence applied to natural language processing. In addition to his research, he teaches courses in data analysis, artificial intelligence, and software engineering.

References

ACT. (2014). *The condition of college and career readiness 2014.* https://www.act.org/content/dam/act/unsecured/documents/CCCR14-NationalReadinessRpt.pdf

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*, 213-238.

Bengio, Y., Ducharme, R., Vincent, P., & Javin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137-1155.

Britz, D. (2016). *Implementing a CNN for Text Classification in Tensorflow.* https://github.com/dennybritz/cnn-text-classification-tf

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research*, 86(4), 803–848. http://www.jstor.org/stable/44668237

Chen, G., Ye, D., Xing, Z., Chen, J. & Cambria, E. (2017). Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. *Proceedings*: *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, 2377-2383.

Chollet, F., Rahman, F., Lee, T., Zabluda, O., Pumperla, M., Santana, E., …Mezzode, S. (2015). *Deep Learning for Humans*. https://github.com/keras-team/keras

College Board, (2013). *2013 SAT Report on College and Career Readiness. College Board Report*. College Board.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research, 12*, 2493–2537.

Connors, R.J. & Lunsford, A.A. (1988). Frequency of formal errors in current college writing, or ma and pa kettle do research. *College Composition and Communication, 39*(4), 395-409.

Dutta, S. and Gros, E. (2018). *Evaluation of the impact of deep learning architectural components selection and dataset size on a medical imaging task.* Proceedings SPIE 10579, 1057911-1057914.

Erhan, D., Courville, A., Bengio, Y., & Vincent, P. (2010). Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research, 11*, 625-660.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N.H., Peters, M., Schmitz, M., & Zettlemoyer, L.S. (2017). *A Deep Semantic Natural Language Processing Platform*. http://arxiv.org/pdf/1803.07640

Gehring, J., Auli, M., Grangier, D., & Dauphin Y. (2016). *A Convolutional Encoder Model for Neural Machine Translation*. https://arxiv.org/abs/1611.02344

Gehring, J., Auli, M., Grangier, D., Yarats, D. & Dauphin Y. (2017). *Convolutional Sequence to Sequence Learning*. https://arxiv.org/abs/1705.03122

Guggilla, C., Miller, T., & Gurevych, I. (2016). CNN- and LSTM-based Claim Classification in Online User Comments. *Proceedings: COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2740–2751.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA : The Journal of the American Medical Association, 316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. http://education.qld.gov.au/staff/development/performance/resources/readings/power-feedback.pdf

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaption of future detectors. *arXiv,* 1207.0580. http://arxiv.org/pdf/1207.0580.pdf

Huot, B., O'Neill, P., & Moore, C. (2010). A usable past for writing assessment. *College English, 72*, 495-517.

Kalchbrenner, N., Grefenstette, E. & Blunsom, P. (2014). *A Convolutional Neural Network for Modelling Sentences*. https://arxiv.org/abs/1404.2188

Kilimci, Z. H. & Akyokus, S. (2018). Deep Learning- and Word Embedding-Based Heterogeneous Classifier Ensembles for Text Classification, *Complexity*, vol. 2018, Article ID 7130146, 10 pages. https://doi.org/10.1155/2018/7130146

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP), (pp. 1746–1751). Association for Computational Linguistics.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. In NIPS, pp. 1106–1114.

Kuang, W., Dong, W., & Dong, L. (2022). *The Effect of Training Dataset Size on SAR Automatic Target Recognition Using Deep Learning. 2022 IEEE 12th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 13–16. https://doi.org/10.1109/ICEIEC54567.2022.9835077

LeCun, Y. (1986). Learning processes in an asymmetric threshold network. In E. Bienenstock, F. Fogel-Soulié, F., & G. Weisbuch (Eds)., *Disordered systems and biological organization* (pp. 233-240). Springer-Verlag.

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature, 521*, 436–444. https://doi.org/10.1038/nature14539

Lin, Y.-S., Huang, P.-H., & Chen, Y.-Y. (2021). Deep Learning-Based Hepatocellular Carcinoma Histopathology Image Classification: Accuracy Versus Training Dataset Size. *IEEE Access*, *9*, 33144–33157. https://doi.org/10.1109/ACCESS.2021.3060765

Liu, W., & Chang, W. (2017). Deep learning for extreme multi-label text classification. *Proceedings*: *40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1-10). ACM.

Lunsford, A.A., & Lunsford, K.J. (2008). Mistakes are a fact of life: A national comparative study. *College Composition and Communication, 59*(4), 781-806.

McCulloch, W., and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 5*, 115-133.

Minsky, M., & Papert, S. (1969). *Perceptrons*. MIT Press.

Mohammadi, M., Mundra, R. & Socher, R. (2015). CS 224D: Deep Learning for NLP. Lecture Notes: Part IV. https://cs224d.stanford.edu/lecture_notes/LectureNotes4.pdf

Moxley, J. M. (Ed.). (1989). *Creative writing in America: Theory and pedagogy*. National Council of Teachers of English.

Moxley, J. M. (1992) Teachers' goals and methods of responding to student writing. *Composition Studies, 20*(1), 17-33.

Moxley, J. M., & Eubanks, D. (2015). On keeping score: Instructors' vs. students' rubric ratings of 46,689 essays. *WPA: Writing Program Administration*, 39, 53–80.

Moxley, J., & Walkup, K. (2016). Mapping writing analytics. *Presented at the 9th International Conference on Educational Data Mining* (EDM 2016), Raleigh, NC, USA. http://ceur-ws.org/Vol-1633/ws2-paper1.pdf

National Center for Education Statistics. (2012). *The nation's report card: Writing 2011*. National Assessment of Educational Programs at Grades 8 and 12. http://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

Parker, D. (1985). *Learning logic, technical report TR-87*. Center for Computational Research in Economics and Management Science, MIT.

Pennington, J., Socher, R., & C. Manning (2014). GloVe: global vectors for word representation. *Proceedings: EMNLP 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532-1543). EMNLP.

Programme for International Student Assessment. (2012). *PISA 2012 results*. http://nces.ed.gov/surveys/pisa/pisa2012/index.asp

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, K. (2016). *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. https://arxiv.org/abs/1606.05250

Rudniy, A., & Elliot, N. (2016). Collaborative review in writing analytics: N-gram analysis of instructor and student comments. Workshop Proceedings: Educational Data Mining.

Sainath, T.N., Vinyals, O., Senior, A., & Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks. *Proceedings: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4580–4584). IEEE.

Sax, G. (1980). *Principles of educational and psychological measurement and evaluation*. Wadsworth.

Schwartz, M. (1984). Response to writing: a college-wide perspective. *College English, 46*(1): 55-62.

Seo, M.J., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). *Bidirectional attention flow for machine comprehension*. https://arxiv.org/abs/1611.01603

Shen, Y., He, X. Gao, J. Deng, L. & Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 101–110). ACM.

Socher, R. (2014). *Recursive deep learning for natural language processing and computer vision.* (Doctoral dissertation). http://nlp.stanford.edu/~socherr/thesis.pdf

Sommers, N. (1982). Responding to student writing. *College Composition and Communication, 33*(2), 148-156.

Starch, D. & Elliot, E. (1912). Reliability in grading high school work in english. *School Review, 20*, 442-457.

Starch, D. & Elliot, E. (1913a). Reliability in grading high school work in history. *School Review, 21*, 676-681.

Starch, D. & Elliot, E. (1913b). Reliability in grading high school work in mathematics. *School Review, 22*, 254-257.

Sun, D. L., Harris, N., Walther, G., & Baiocchi, M. (2015). Peer assessment enhances student learning: The results of a matched randomized crossover experiment in a college statistics class. *PLOS ONE, 10*(12), 1-7.

Wyatt-Smith, C. (1997). Teaching and assessing writing: An Australian perspective. *English in Education, 33*(3), 8-22.

Xiao, Y. & Cho, K. (2016). *Efficient character-level document classification by combining convolution and recurrent layers*. https://arxiv.org/abs/1602.00367

Yih, W.-T., He, X., & Meek, C. (2014). Semantic Parsing for Single-Relation Question Answering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 643–648). Association for Computational Linguistics.

Zhang, Jie. (2017). *Classify Kaggle San Francisco crime description into 39 classes*. https://github.com/jiegzhan/multi-class-text-classification-cnn-rnn

Zhong, S., Liu, Y., Li, B., Long, J. (2015). Query-oriented unsupervised multi-document summarization via deep learning model. *Expert Systems with Applications, 42*, 8146-8155.