Using IBM's Watson to automatically evaluate student short answer responses

Jennifer Campbell, Katie Ansell, and Tim Stelzer
Department of Physics, University of Illinois Urbana-Champaign, 1110 W. Green St., Urbana, Illinois, 61801, USA

Recent advancements in natural language processing (NLP) have generated interest in using computers to assist in the coding and analysis of students' short answer responses for PER or classroom applications. We train a state-of-the-art NLP, IBM's Watson, and test its agreement with humans in three varying experimental cases. By exploring these cases, we begin to understand how Watson behaves with ideal and more realistic data, across different levels of training, and across different types of categorization tasks. We find that Watson's self-reported confidence for categorizing samples is reasonably well-aligned with its accuracy, although this can be impacted by features of the data being analyzed. Based on these results, we discuss implications and suggest potential applications of this technology to education research.

2022 PERC Proceedings edited by Frank, Jones, and Ryan; Peer-reviewed, doi.org/10.1119/perc.2022.pr.Campbell

I. INTRODUCTION

The analysis of short-answer student responses can be a time-consuming and difficult task for researchers and educators. Industries have adapted to similar tasks by using machine learning, particularly natural language processors (NLP) to quickly analyze text across extremely large sample sizes (see [1, 2] for a list of examples). NLPs are programs founded on machine learning algorithms that process textual or verbal information. Once trained, NLPs can perform analyses such as sentiment, entity detection, and categorization.

The typical process for using a trainable NLP can be broken down into four steps:

1. **Preparation:** Categories are identified and a subset of the data are coded into the categories by humans.
2. **Training:** The coded subset of statements is provided to the NLP with category labels. The NLP uses this subset to establish its categorization model.
3. **Sample analysis:** New statements are sent to the NLP. For each statement the NLP returns confidence values for each category it was trained on.
4. **Categorization:** Humans interpret the confidence scores to assign a category to each statement.

Many NLP tools are now available to the public for low-cost or free use. This technology offers potentially exciting applications to Physics education and Physics Education Researchers to perform text analysis at scales that have previously been prohibitively large; however there is not an established methodology for the use of NLP tools within the PER community.

Investigations into the use and trends of NLP on short-answer text statements for education purposes has grown swiftly since 2015 [3–6]. The results of NLPs' efficacy have varied in STEM research, with Cohen's κ scores ranging from 0.09 to 0.97 [7]. Examples of NLP work done in STEM and PER revolve around using NLPs to grade exam questions or survey other research [8–19]. However, further exploration of NLPs is necessary to determine how they may be utilized for processing research data.

This work aims to add to the community's understanding of NLP as a text analysis tool by using IBM's Watson to categorize short-response samples with varying levels of NLP training and across differing data sets. Watson has the ability to analyze short-answer responses and assign likelihood of belonging to specified categories based on IBM's algorithm and neural network methods, which are not disclosed to the public. Through training and testing with three examples of response categorization, we examine the confidence and accuracy of Watson to characterize its strengths and weaknesses. With this, we propose two methods in which Watson can be used to assist in coding data, one meant for research purposes, and the other for classroom surveys.

II. METHODS

We tested Watson with three different data sources to observe trends of performance and characteristics of the software's algorithms and protocols. Every statement was coded (i.e., labeled with a category) by humans prior to evaluation by Watson. For each data set, a subset of labeled samples was provided to Watson for training. A larger number of samples from the same data set was then provided to the trained Watson model for assignment of confidence scores for each category. For the scope of this paper we adopt a simplified categorization process in which we accept the category with the highest confidence score as the NLP-assigned category.

We evaluate the effectiveness of Watson's categorization by examining the confidence scores of each statement's assigned category and determining the accuracy of Watson's categorization compared to the consensus of the human coders. Comparing across the different data sets allows us to understand Watson's performance for various difficulties of coding, sizes of training set, and mode of categorization. Details for each of the three data sets are described in Table I.

TABLE I. A summary of the data sets used to assess Watson.

Data set	**No. of labels**	N_{test}	$N_{training}$
News Headlines 169	4	4000	169
News Headlines 1000	4	4000	1000
Lab reasoning	4	479	169
Pre-flight	2	732	398

A. News headline data

The first set of data used in this study is a series of news headlines from different genres of news, provided by Rishabh Misra on Kaggle [20], an online resource for developing skills in computer science. This data set contains around 200,000 news headlines that are labeled with corresponding categories ('Politics,' 'Wellness,' etc.). This large data set offers 41 different categories, making it possible to select a small sample of categories for testing with Watson while still representing a large number of short text samples. Although there is no guarantee that Watson's performance on this data will be indicative of its performance on student responses to physics questions, it demonstrates how the NLP performs in a general categorization case which can then serve as a baseline of effectiveness when investigating its performance in PER.

For each test of Watson with the news headline data, four categories were selected. Short text samples belonging to those four categories were randomly selected to form the training set and the testing set. No samples with other labels were included in the testing set. The intent of filtering the data in this way was to examine how Watson performs in a simplified case where all samples are considered categorizable.

The news headline data were used to evaluate Watson across two dimensions:

- **Varying training set size.** Models were trained on 169 and 1000 samples using the same four categories. The low and high training models were tested with the same set of 4000 samples.
- **Varying tested categories.** Two data subsets were created, each corresponding to a different set of four category labels. For each subset, models were trained on 1000 samples and evaluated on 4000 samples from that subset.

B. Lab reasoning data

The second set of data comes from a lab reasoning experiment in an introductory physics course. In an online assignment, students were asked to explain their reasoning to support a yes/no decision to a lab-related data analysis question. Responses were coded into four categories representing the main type of reasoning that appeared in students' responses. This data set serves as an example of projects whose data can span over multiple semesters, where such large sets can be more approachable with NLP. It also provides insight for how Watson deals with difficult categorization tasks.

This data was prepared by five coders in two stages: First, using a group agreement process to establish coder training, and afterward, in pairs coding together. Interrater reliability (IRR) was checked frequently throughout the coding process and disagreements were resolved by discussion between coders. Prior to resolving conflicts, the overall IRR (using Krippendorff's α [21]) was $\alpha = 0.766$. Generally, researchers in the field aim for an IRR score of at least 0.8. We performed the experiment despite being below the threshold to evaluate Watson's performance on statements across varying levels of difficulty for human coders.

These data were somewhat difficult to code, with approximately 13% of samples ultimately ending up in an 'other' or dual-coded category due to disagreement. The training set that was provided to Watson only contained code-able statements; however a small number of dual-coded statements were necessary to increase representation of certain categories in the training sample.

C. Pre-flight data

The last set is a series of student responses to a pre-flight question in an introductory physics course. Students were asked to use physics to explain what happens in a certain scenario. Their answers were coded with a binary for whether a specific physics concept was present. This data set offers an alternative model in which Watson may flag whether topics are present rather than categorizing statements as discussed in the previous two data sets.

Student statements were coded for the presence of the concept by two coders. The coders had $\alpha = 0.79$, close to the generally accepted value of 0.8.

III. RESULTS

Results of Watson's performance are presented here by each experimental case. For each of the data sets and experiment variations, we perform analysis by arranging the coded statements by Watson's highest confidence score to lowest, binning into groups of 50 for small data sets (lab reasoning, pre-flight) and 200 for large (news headlines), taking the average values for each bin, and plotting onto a graph with error bars representing the standard error on the mean. We then compare Watson's labels to the respective pre-established labels to establish accuracy rates for each bin. Pre-established labels were determined by either the provided label from the news headlines data set or from the group consensus code designated by the human coders. Displaying the data in this way makes it possible to see a distribution of Watson's confidence scores and determine the relationship between confidence and accuracy.

A. News headline experiment

In this section we examine Watson's performance with the news headline data, including comparison of the low- and high-training models and the data sets with differing categories.

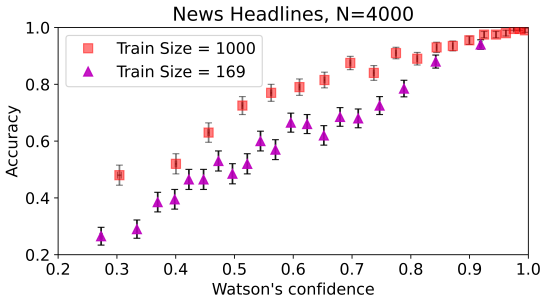

FIG. 1. Watson's accuracy versus confidence for the news data with different training sample sizes. Each point represents 200 samples.

The results from comparing different training sizes for the same data categories are shown in Fig. 1. For both training models it is clear that Watson's accuracy increases as its confidence increases — in other words, the more confident Watson is, the more likely it has applied a correct label. Using a paired t-test, we find a large, significant increase in Watson's confidence between the 169 sample train size and 1000 sample train size ($M = 0.57, 0.75$, $SD = 0.17, 0.21$ respectively); $t(3999) = -162$, $p < 0.001$, $d = 0.92$. Considering the figure, it is apparent that this increase in confidence is paired with an increase in labeling accuracy. This indicates

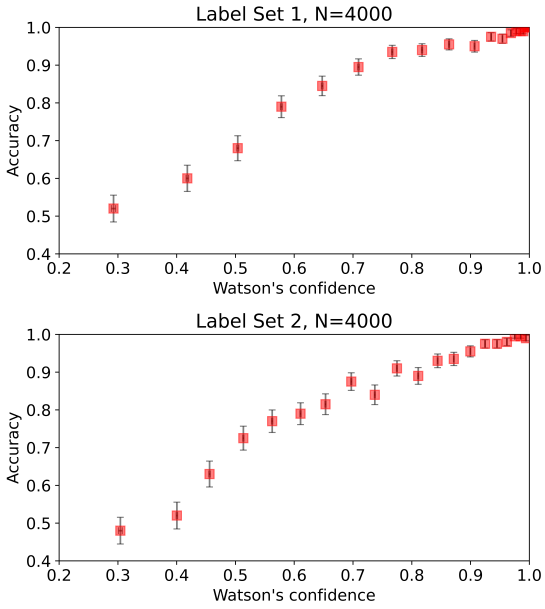

FIG. 2. Watson's accuracy versus confidence for different categories of news data. Each point represents 200 samples.

that increasing the training set from 169 to 1000 significantly improved Watson's accuracy.

Fig. 2 shows the comparison of Watson's performance with two different sets of news headline data, each using 1000 samples to train and categorize 4000 samples, but for different news headline categories. We see that Watson performs well in both cases, with accuracy that matches or exceeds its confidence. There is a statistically significant increase in Watson's confidence from Label Set 1 to 2 ($M = 0.75, 0.82$, $SD = 0.21, 0.21$), $t(3999) = -96.9$, $p < 0.001$, $d = 0.32$. This small to medium effect on the confidence scores indicates that Watson may perform better with certain contexts than others.

B. Lab reasoning experiment

For the lab reasoning data, we consider Watson's confidence and accuracy and compare its performance to the IRR of the human coders. For the 479 statements tested, Watson and the Humans had an α of 0.582, which is much lower than the IRR between human coders.

Fig. 3 compares Watson's accuracy to the humans' IRR scores for the same bins of statements. We again observe the correlation between confidence and accuracy, like with the news headlines experiment. However, we observe that Watson's agreement with humans consistently decreases as its confidence scores decrease, while the human coders maintain high agreement for the same statements. This suggests that humans may be more accurate than Watson for categorizing less clear samples.

C. Pre-flight survey experiment

Fig. 3 shows a comparison between Watson and the human coders for the detection of a topic within responses from the Pre-flight data. As with the previous two experiments, we observe a correlation between Watson's confidence and its accuracy. We also observe that, compared to the lab reasoning data, Watson is both more confident and more accurate overall. This agrees with the observation from the news headlines case that Watson performs better in different contexts.

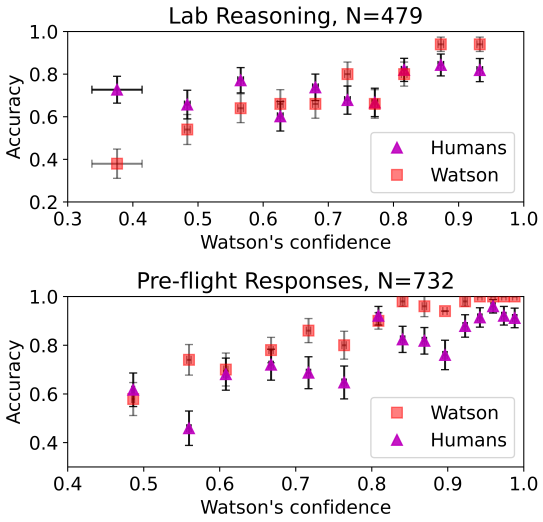

FIG. 3. Watson's accuracy versus confidence when compared to human IRR for the same statements. Each point represents 50 samples.

D. Confidence cut-offs to improve accuracy

In this section we provide preliminary analysis to explore applications for Watson in PER. Because the results from all three data sets establish a pattern where Watson's accuracy increases with confidence, it may be possible to apply a "confidence cutoff score" above which Watson's accuracy may be acceptable for certain applications.

Table II provides examples of how results of the three experiments would be affected if we were to only consider statements with confidence scores exceeding 0.5, 0.75, or 0.9. For each of the three experiments we see that increasing the confidence score threshold increases overall accuracy; however, fewer samples are represented in the result. The gains from applying these thresholds vary across data type. For example, we note that for the Lab reasoning data applying a 0.75 threshold raises Watson's accuracy to a value near that of the human coders' IRR, but only 21% of samples are considered.

TABLE II. Average of correct samples and fraction of samples preserved based on confidence score threshold.

Confidence Score	≥ 0.5	≥ 0.75	≥ 0.9
News, 1000 train set			
Accuracy	0.690	0.858	0.956
Frac. Samples	0.618	0.171	0.034
Lab Reasoning			
Accuracy	0.751	0.836	0.937
Frac. Samples	0.87	0.42	0.10
Krippendorff's α	0.623	0.719	0.879
Pre-flight			
Accuracy	0.898	0.959	0.996
Frac. Samples	0.996	0.675	0.372
Krippendorff's α	0.791	0.865	0.909

IV. DISCUSSION

The results from these experiments provide several valuable insights into Watson's performance that can inform future practice. First, across all data sets and trials Watson has a consistent correlation between its confidence and its accuracy. Second, the Lab reasoning experiment and comparison between Concept 1 and Concept 2 in the Pre-flight experiment indicate that Watson is less accurate in situations where human coders also struggle. On the other hand, statements that are easier for humans to categorize are easier for Watson to label accurately. Third, Watson's performance depends on the data provided and the categories it is trained on. As seen with the variance in confidence plots across the experiments, Watson does not work with all categories equally well, and training will also greatly affect its understanding. This suggests that variables such as number of categories, complexity of statements, and coding scheme is more important for Watson's performance than humans'.

A. Watson's confidence as a guide

The correlation between Watson's confidence and accuracy may offer opportunities for semi-automated coding. By arranging the statements Watson codes from lowest score to highest, one could work alongside Watson, treating it as a second coder. In such a model, humans can code statements, train Watson, let Watson process the remaining data, then compare results and resolve disagreements by themselves. When coding in this way the researcher can expect to have to resolve many disagreements for the low confidence scores, but will eventually reach a threshold confidence score at which they deem Watson to be consistent enough to entrust all remaining statements to it.

Pair coding can potentially reduce the amount of work done by coders in situations with less confidence in the NLP. In the optimal case, Watson will perform excellently and workload can be cut down. In the worst case scenario, human coders will have to go through all statements. However, since the time required to train Watson is already necessary when coding all statements, no extra time is needed to run Watson, and therefore there is no net loss.

This scheme is a potentially useful tool for physics education research. Studies that require the coding of large data sets can be processed faster by partnering with trained NLPs.

B. Watson as a classroom analyzer

In situations where Watson does not have to be perfect, the software performs well enough to do categorization for rough analysis. An example of this kind of scenario is similar to the Pre-flight experiment, where Watson was trained to detect if student responses contain discussion of specific topics. In such a case Watson's results could be arranged to provide a statistical breakdown of what students were thinking without needing to worry about the exact number of times a given concept appears.

This form of analysis, when combined with students' responses to other related questions, can provide an efficient and handy synopsis tool for instructors to understand their class' understanding in preparation for lectures. In terms of preparing the NLP for this scenario, once the scripts for training and testing are setup, as long as the format of the questions are the same, nothing should need to be changed for any future analysis.

V. CONCLUSION

In this paper we showcased the capabilities of a state-of-the-art NLP software and potential methods for utilizing it in both PER and classroom settings. While it is not developed enough yet to meet the PER community's expectations of a reliable coder, it is intelligent enough to convey when statements are too difficult for it to handle. By being aware of NLPs' limitations and characteristics, research can be adapted to implement these software to reduce workload or perform rough analysis.

This work highlights the benefits that NLPs bring to PER work, though there are methods to this research and application that we did not cover. Future work should analyze alternative ways to use confidence scores to interpret NLP results. There should also be investigations into how NLPs should be trained or how data sets should be prepared to improve model accuracy. With more research done on how to optimize NLPs, they can become an effective assistant for short-answer processing.

ACKNOWLEDGMENTS

Our thanks go to Sean Golinski, Joseph Kuang, and Alex Nickl for all of their help during the human coding process and scripting preparation for Watson.

[1] https://www.appsruntheworld.com/customers-database/products/view/ibm-watson-natural-language-understanding. Retrieved 5/16/2022

[2] https://enlyft.com/tech/products/ibm-watson-natural-language-understanding. Retrieved 5/16/2022

[3] G. Casalino, B. Cafarelli, E. del Gobbo, L. Fontanella, L. Grilli, A. Guarino, P. Limone, D. Schicchi and D. Taibi, in *Proceedings Of The Second Workshop On Technology Enhanced Learning Environments For Blended Education* (Foggia, Italy, 2021).

[4] Yun, E. Review of trends in Physics Education Research Using Topic Modeling. *Journal Of Baltic Science Education*. **19**, 388-400 (2020)

[5] Burrows, S., Gurevych, I. & Stein, B. The eras and trends of automatic short answer grading. *International Journal Of Artificial Intelligence In Education*. **25**, 60-117 (2014)

[6] Zhai, X., Yin, Y., Pellegrino, J., Haudek, K. & Shi, L. Applying machine learning in science assessment: A systematic review. *Studies In Science Education*. **56**, 111-151 (2020)

[7] Zhai, X., Shi, L. & Nehm, R. A meta-analysis of machine learning-based science assessments: Factors impacting machine-human score agreements. *Journal Of Science Education And Technology*. **30**, 361-379 (2020)

[8] A. Çinar, E. Ince, M. Gezer and Ǒ. Yilmaz, Education And Information Technologies 25, (2020).

[9] Butcher, P. & Jordan, S. A comparison of human and computer marking of short free-text student responses. *Computers amp; Education*. **55**, 489-499 (2010)

[10] Cvetkovic, L., Milasinovic, B. & Fertalj, K. A tool for simplifying automatic categorization of scientific paper using Watson API. *2017 40th International Convention On Information And Communication Technology, Electronics And Microelectronics (MIPRO)*. (2017)

[11] Madnani, N., Loukina, A. & Cahill, A. A large scale quantitative exploration of modeling strategies for content scoring. *Proceedings Of The 12th Workshop On Innovative Use Of NLP For Building Educational Applications*. (2017)

[12] Erickson, J., Botelho, A., McAteer, S., Varatharaj, A. & Heffernan, N. The automated grading of student open responses in Mathematics. *Proceedings Of The Tenth International Conference On Learning Analytics amp; Knowledge*. (2020)

[13] Mohler, M. & Mihalcea, R. Text-to-text semantic similarity for automatic short answer grading. *Proceedings Of The 12th Conference Of The European Chapter Of The Association For Computational Linguistics On - EACL '09*. (2009)

[14] Sultan, M., Salazar, C. & Sumner, T. Fast and easy short answer grading with high accuracy. *Proceedings Of The 2016 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies*. (2016)

[15] Süzen, N., Gorban, A., Levesley, J. & Mirkes, E. Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*. **169** pp. 726-743 (2020)

[16] Zehner, F., Sälzer, C. & Goldhammer, F. Automatic coding of short text responses via clustering in educational assessment. *Educational And Psychological Measurement*. **76**, 280-303 (2015)

[17] Nehm, R., Ha, M. & Mayfield, E. Transforming Biology Assessment With Machine Learning: Automated Scoring of written evolutionary explanations. *Journal Of Science Education And Technology*. **21**, 183-196 (2011)

[18] Ha, M., Nehm, R., Urban-Lurain, M. & Merrill, J. Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. *CBEâLife Sciences Education*. **10**, 379-393 (2011)

[19] Liu, O., Rios, J., Heilman, M., Gerard, L. & Linn, M. Validation of automated scoring of science assessments. *Journal Of Research In Science Teaching*. **53**, 215-233 (2016)

[20] https://www.kaggle.com/datasets/rmisra/news-category-dataset. Retrieved 12/15/2021

[21] K. Krippendorff, *Content Analysis: An Introduction To Its Methodology*, 3rd ed. (Thousand Oaks, CA, 2013), pp. 221-250.