Integrating Natural Language Processing in Middle School Science Classrooms: An Experience Report

Gloria Ashiya Katuka*
University of Florida
Gainesville, FL, USA
gkatuka@ufl.edu

Srijita Chakraburty
Indiana University
Bloomington, IN, USA
srichak@iu.edu

Hyejeong Lee
Indiana University
Bloomington, IN, USA
hle6@iu.edu

Sunny Dhama
University of Florida
Gainesville, FL, USA
sunny.dhama@ufl.edu

Toni Earle-Randell
University of Florida
Gainesville, FL, USA
tearlerandell@ufl.edu

Mehmet Celepkolu
University of Florida
Gainesville, FL, USA
mckolu@ufl.edu

Kristy Elizabeth Boyer
University of Florida
Gainesville, FL, USA
keboyer@ufl.edu

Krista Glazewski
North Carolina State University
Raleigh, NC, USA
kdglazew@ncsu.edu

Cindy Hmelo-Silver
Indiana University
Bloomington, IN, USA
chmelosi@indiana.edu

Tom McKlin
The Findings Group
Atlanta, GA, USA
tom@thefindingsgroup.org

ABSTRACT

With the increasing prevalence of large language models (LLMs) such as ChatGPT, there is a growing need to integrate natural language processing (NLP) into K-12 education to better prepare young learners for the future AI landscape. NLP, a sub-field of AI that serves as the foundation of LLMs and many advanced AI applications, holds the potential to enrich learning in core subjects in K-12 classrooms. In this experience report, we present our efforts to integrate NLP into science classrooms with 98 middle school students across two US states, aiming to increase students' experience and engagement with NLP models through textual data analyses and visualizations. We designed learning activities, developed an NLP-based interactive visualization platform, and facilitated classroom learning in close collaboration with middle school science teachers. This experience report aims to contribute to the growing body of work on integrating NLP into K-12 education by providing insights and practical guidelines for practitioners, researchers, and curriculum designers.

CCS CONCEPTS

• **Social and professional topics → K-12 education**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGCSE 2024, March 20–23, 2024, Portland, OR, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0423-9/24/03…$15.00
https://doi.org/10.1145/3626252.3630881

KEYWORDS

Natural language processing, NLP and AI learning, NLP+Science, middle school science classrooms

ACM Reference Format:
Gloria Ashiya Katuka, Srijita Chakraburty, Hyejeong Lee, Sunny Dhama, Toni Earle-Randell, Mehmet Celepkolu, Kristy Elizabeth Boyer, Krista Glazewski, Cindy Hmelo-Silver, and Tom McKlin. 2024. Integrating Natural Language Processing in Middle School Science Classrooms: An Experience Report. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024), March 20–23, 2024, Portland, OR, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3626252.3630881

1 INTRODUCTION

Natural language processing (NLP) focuses on understanding, interpreting, and generating natural language [18, 23, 40] and is one of the fastest growing sub-fields of artificial intelligence (AI). Recent emergence of large language models such as ChatGPT and GPT-4 [5, 29] have increased the popularity of NLP technologies. In line with this prevailing trend, there has been an increase in initiatives to develop curriculum and learning tools to teach NLP in both K-12 and higher education settings [3, 19, 28]. These efforts include using NLP for teaching tweet classification in social studies [19], involving students in disaster relief area determination using NLP [28], and exploring NLP learning using visualizations [3]. While these efforts are increasing the spread of NLP education in K-12, NLP remains relatively underexplored in middle school classrooms.

To engage middle school students in NLP and AI, researchers are continuously exploring innovative ways to bring AI-based learning experiences to their classrooms [31, 32]. Despite the fact that NLP at its core is a complex field, especially for young learners, introducing its foundational concepts in an interactive and engaging way holds the potential to foster learners' interest. For example, textual analysis and visualization using NLP models such as keyword extraction and sentiment analyzsis provide insights into textual data [3]. Such NLP models have been used to understand public sentiment, for example by tracking the propagation of fear and its impact on the spread of Covid-19 by analyzing Twitter data [33]. Understanding how textual data fuels advanced applications to address real-world issues provides a compelling motivation for students to further explore NLP. Our work aims to expose middle school students to NLP by integrating it into their science classrooms, with the goal of analyzing textual data such as tweets [1] about science-based topics. Recently, researchers are integrating computer science (CS) into core STEM disciplines like mathematics and science to increase interest in STEM and expose students to CS [6, 15]. Similarly, integrating NLP into science holds the potential to provide an authentic and engaging learning experience for middle school learners.

This experience report presents our efforts to integrate NLP with science by developing a curriculum and tool to spark the interest of middle school students in using NLP in their science classrooms. We collaborated with five science teachers to design an integrated "NLP+Science" curriculum and develop an NLP-based interactive visualization platform, enabling students to practice NLP concepts and analyze science-based textual data. Our classroom studies with 98 middle school students across two US states indicated great promise for engaging middle school learners in NLP and AI. This experience report aims to provide valuable insights for other researchers and educators who seek innovative ways to teach NLP and integrate AI into middle school classrooms.

2 BACKGROUND AND RELATED WORK

In K-12 education, there have been significant efforts to integrate AI, ML, and NLP into classrooms with the goal of increasing AI literacy among students [22, 25, 36]. These efforts have led to the design and development of curricula and tools to educate and engage young learners[11, 21, 22, 35, 38]. Notably, several tools have been developed to introduce machine learning (ML) and NLP to K-12 classrooms. These tools, such as ML4Kids [1], Cognimates [10], Teachable Machine [2] and StoryQ [13] have proven to be effective in teaching young learners the basics of AI, in a fun and engaging manner, especially for those with minimal programming experience.

With the increasing availability of these AI-based tools and curricula, ML and NLP are finding their way into K-12 classrooms. However, most efforts in developing AI-based curricula have primarily focused on ML [31, 41]. Additionally, existing initiatives in NLP education have been mainly developed for students in high school and higher education [3, 19, 28]. Therefore, there is a need to design and develop curricula and tools that support NLP learning for K-12 classrooms, particularly middle schools.

Another way to promote NLP in K-12 classrooms is through teachers. Recently, researchers are making efforts to co-design or collaboratively design and development of curricula and tools to support AI learning with teachers and learners [24]. Co-design, a collaborative design process involving teachers and researchers, is common in research on K-12 education [34]. Research in CS education has shown that teachers provide valuable expertise and contribute innovative ideas, which have a substantial effect on students' ability to adapt to new curricula and tools [12, 34]. However, involving teachers, especially non-CS teachers, in the initial design of NLP and AI tools and learning activities is less explored [24]. Typically, teachers' contributions to K-12 research are mainly instructional, with limited involvement in early development of curriculum or tools [39]. Meanwhile, studies have found that teachers' confidence and readiness to teach AI predict their intentions to teach AI and are influenced by their perception of AI's relevance [4]. In our work, we extensively relied on teachers' expertise and feedback to integrate NLP and AI learning activities into science classrooms.

3 INTEGRATING NLP WITH SCIENCE

The overarching goal of our project was to integrate AI in science classrooms, which involved developing a curriculum that aligns NLP with inquiry-based Next-Generation Science Standards [8]. To achieve this goal, we collaborated with teachers from the onset of the project. Their expertise and feedback were instrumental in designing the curriculum and developing the platform. In the following sections, we describe the co-design workshops and the development of the tool based on the teachers' feedback.

3.1 Teacher Co-design Workshops

We partnered with two science teachers in Florida and three teachers in Indiana who participated in weekly virtual workshops for a total of nine weeks, six in the Fall of 2022 and three in the Spring of 2023. During these workshops, we introduced AI and NLP concepts to the teachers, with a specific focus on two main NLP tasks: sentiment analysis and keyword extraction. Once the teachers were familiar with the NLP concepts and tasks, we facilitated brainstorming sessions where they shared their ideas on ways we could integrate NLP to their science classrooms. More specifically, our process was guided by an overarching prompt: *How can we create an inclusive, engaging, relevant science inquiry experience that integrates practices from NLP?*

Guided by problem-based learning [16, 17], teachers and researchers brainstormed together about topics and related standards both for science and NLP. Teachers and researchers explored several topics together and arrived at the student-facing problem-based inquiry question: *"When individuals are asked to make choices about single-use plastics, how do they feel about it and does it make a difference?".* We searched social media platforms for discourse related to the topic. We also explored news articles about plastic straws and potential threats to aquatic ecosystems, including specific threats to sea turtles. Then, we discussed prevailing and countervailing ideas about how to influence people's perspectives about single-use plastics and identified potential text data sources such as Twitter, Facebook, Reddit, and YouTube comments that could be used to understand public sentiment.

3.2 NLP4Science: Text Analysis Platform

After the teachers had built conceptual knowledge and chosen their topic, we began searching for existing, age-appropriate tools that

[1] Twitter datasets were manually cleaned prior to being used in classrooms to ensure they did not include any inappropriate content

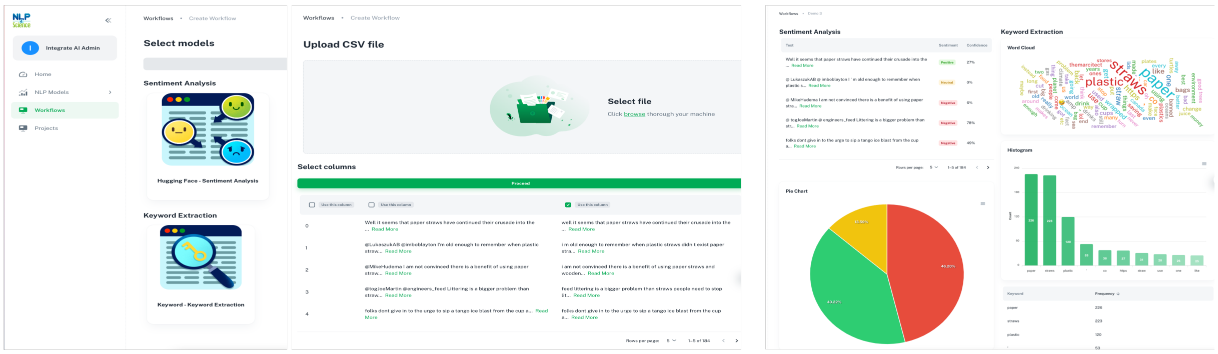


Figure 1: Interactive workflow view to show the combination of keyword extraction and sentiment analysis

could facilitate the integration of NLP techniques in science classrooms. The closest tool we found was MonkeyLearn [27], which was built for commercial use to support textual analyses and visualizations. MonkeyLearn offered a free trial period during which teachers were able to participate in hands-on learning and explore NLP models for keyword extraction and sentiment analysis during the early stages of the Fall 2022 workshop, to gather insights that could inform the development of our future learning platform. While the teachers found the application useful in general, they reported several significant shortcomings, including the complexity of the visualizations, the lack of interactivity between data points and visualizations, and the overall complex pipeline for importing NLP datasets and creating visualizations.

Together, teachers and researchers co-designed mockups of a new NLP platform for middle school science and discussed the best options for taking textual data as input and generating interactive visualizations. Teachers' recommendations emphasized a simple and interactive platform, with an easy-to-follow process for importing NLP datasets and creating visualizations. We took these suggestions and proceeded to develop the first version of the NLP4Science platform (see Figure 1). The platform provides visualizations of analyzed data, including pie charts, histograms, word clouds, bar charts, and tables to display information about the sentiment and keywords of the text [9].

3.2.1 Pilot Classroom Experience. We tested our newly developed tool in a 2-day classroom experience with 8th graders in Indiana in early Spring 2023. During the 2-day period, students were introduced to the basics of AI and NLP and participated in hands-on activities to test the platform. On the first day, students were asked to share their experiences with AI and NLP that were relevant to their daily lives (such as Siri or Grammarly). Then, they were asked to read a news article about the science topic: plastic straws and ocean turtle trauma. Afterwards, students were given a small sample of tweets and asked to manually extract keywords and assign sentiments. On the second day, students used our newly developed tool to analyze tweets about plastic straws in the ocean. During the hands-on activities, students experienced some latency issues when generating visualizations due to the servers not being scaled to their full potential. Also, the students used iPads, which resulted in additional issues with uploading data files, as well as the resolution and scaling for the iPad's screen size. We noted these issues and made adjustments for the full classroom experiences. Overall, the user testing was successful with students generally engaged in class discussions.

4 CLASSROOM EXPERIENCES

After the pilot study, we made some minor changes to the platform and proceeded with further classroom experiences. In the following sections, we present the details of our classroom experiences, including the student population, the structure of the classroom activities, the learning objectives, and details of the main lessons.

4.1 Population

In Spring 2023, we conducted a 5-day classroom experience in four science classrooms with 50 consented middle school students in Florida and in five science classrooms with 48 consented middle school students in Indiana. Among the 48 students who completed the demographics survey in Florida, 22 (46%) were female, 24 (50%) male were male, one (2%) identified as non-binary and one preferred not to report. The distribution of race/ethnicity was 33 (69%) Black or African American, five (10%) Mixed race, three (6%) White or Caucasian, one (2%) Native American or Alaska Native or First Nation, and three (6%) preferred not to report. Among the 48 students who completed the survey in Indiana, 29 (60%) were female, 17 (36%) were male, and two (2%) identified as non-binary. The distribution of race/ethnicity was 44 (91%) White or Caucasian, two (4%) Mixed race, three (6%), and two (4%) Native American or Alaska Native or First Nation.

4.2 Learning Objectives

One of the goals for these classroom experiences was to understand if NLP techniques can be applied to explore real-world environmental conservation questions, such as, *Are people willing to change away from plastic straw use?* and, *How can we understand people's willingness to change?* We wanted to ensure that students thoroughly understood the basics of NLP. Our teacher workshop (described previously) centered on this goal and built upon the learning objectives from AI4K12 for 6th to 8th grade [36, 37]. We also refined a subset of the learning objectives relevant to our purposes, arriving at the following: *Students should be able to:*

LO1. contrast the unique characteristics of human learning with the ways machine learning systems operate

LO2. use natural language data and identify potential ethical problems with that data when given examples of applications
LO3. illustrate a use case for keyword extraction and describe how the keywords can provide insights on unlabeled text data
LO4. utilize a pre-trained keyword extraction model to extract the most frequent/important words or phrases regarding some topic of science inquiry (such as plastic pollution)
LO5. illustrate a use case for sentiment analysis and describe how the labels are predicted on test data
LO6. utilize a pre-trained sentiment analysis model to predict whether a span of text (such as a tweet) is positive or negative regarding some topic of science inquiry

Table 1: Classrooms experiences

	Classroom I (FL)	Classroom II (IN)
Day 1	Data collection (pre-surveys & pre-assessments) Lesson 1: Introduction to AI, ML and NLP	
Day 2	Lesson 2: Keyword Extraction	
Day 3	Lesson 3: Sentiment Analysis	
Day 4	Lesson 4: Workflows and Pair collaboration	Lesson 4: Data Bias and Ethics
Day 5	Post-surveys, post assessments, Student interviews and partner satisfaction surveys	

4.3 Lessons and Activities

We designed three main lessons to cover our core learning objectives: Introduction to AI, ML and NLP, Keyword Extraction and Sentiment Analysis. In line with these lessons, we worked with teachers during our Fall 2022 and Spring 2023 workshops to co-design a set of hands-on learning activities for NLP in science. Over the course of the teacher workshops, we asked our teachers how much time they could devote to integrating NLP into their classroom lessons without disrupting the existing science curriculum in the current semester. They agreed to a period of 4-6 days, and we planned and structured our study for 5 days, leaving one day as a backup. The first classroom experience in Florida was led by researchers, while the second was led by the teacher. Both experiences followed a similar lesson plan, with slight variations. For the Florida classroom experiences, we centered one of the lessons around a collaborative learning activity where students first analyzed data with a partner, then shared and discussed the insights gained from the data, as well as the potential for data bias and ethics questions. Table 1 shows the lessons covered during the two classroom studies.

On Day 1, students completed a pre-survey and a pre-assessment, then learned about the fundamental concepts of AI, ML and NLP and how they relate to each other. In this lesson, students were introduced to ML and NLP as two common sub-fields of AI, with ML being applied to various types of data and NLP specifically focused on processing textual data. On Day 2, students were introduced to the concept of keyword extraction, which allowed them to extract and visualize the most relevant keywords from a given text. Many common NLP applications, such as text summarization, rely on the foundational principles of keyword extraction [14]. During the lesson, the researcher or teacher explained keyword extraction and stopwords and the students practiced manually extracting keywords and identified stopwords from a sentence. Then, they engaged in hands-on experience using our tool. They used the keyword extraction model to extract the keywords from a sentence and compared these to their manually extracted keywords. They also practiced using the model to extract keywords from news content online (see Figure 2). Next, students learned to upload a textual dataset and analyze it using the keyword extraction model.

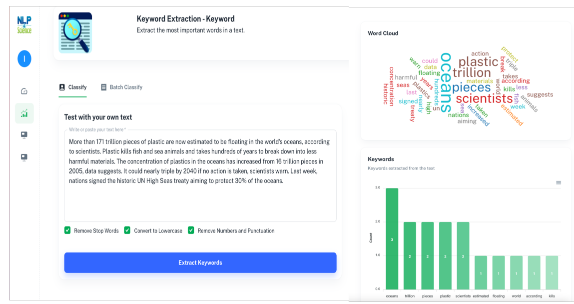


Figure 2: Keyword Extraction

On Day 3, the researcher or teacher discussed environmental conservation, specifically whether plastic, paper, or reusable straws were better for the environment. Then, students were introduced to the concept of sentiment analysis, which typically involves the classification of texts into *positive*, *negative*, or *neutral* sentiments. Sentiment analysis is an important task in NLP and is relevant in many NLP applications such as detecting user's emotion in human-agent interactions [7]. To better understand the concept of sentiment analysis, students revisited extracting keywords from news content, this time to obtain the sentiment of texts (see Figure 3). Afterwards, students learned to upload and analyze a textual dataset using a sentiment analysis model. Then, they discussed the results of the models and shared how the visualizations helped them gain insights about the data.

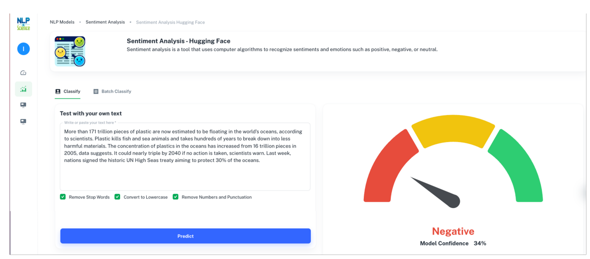


Figure 3: Sentiment Analysis

On Day 4, the Indiana students discussed data bias and ethics as the main lesson; whereas the Florida students participated in a hands-on collaborative activity and discussed the ethics around data collection. In Indiana, students were introduced to the Ethical Matrix [26], and used it to understand the values and biases different stakeholders might have regarding the plastic problem statement. They also explored the possible ethical problems that might arise due to a conflict in these values. Then, they discussed how AI systems designed on biased datasets can lead to similar ethical problems, favoring certain groups of people over others. In Florida, Students collaborated to build a workflow utilizing both keyword extraction and sentiment analysis models on the platform, using a Twitter dataset about plastics in the ocean (Figure 1). Using a worksheet as scaffolding, we prompted students to discuss with one another what their analysis revealed, and they shared their insights with each other. On Day 5, all the students completed a post-survey and post-assessment. Also on Day 5, the students in Florida completed a survey about their partners and were interviewed about their experience with our platform.

5 OUTCOMES

5.1 Student Feedback and Learning Outcomes

We administered pre- and post-surveys and assessments to better understand students' attitudes towards NLP, AI, and STEM, and to gauge their content knowledge.

5.1.1 Student attitude survey. We administered pre-surveys on the first day of the classroom experience and post-surveys on the last day of the classroom experience. We used a subset of survey items from BASIC-SQ [30] and fine-tuned them for relevance to our purposes. Our fine-turned survey focused on capturing the students' interest, identity, and intentions to persist. Some of the survey items related to students' interest include: *I am interested in learning more about NLP*, *I am interested in taking more computer science classes related to NLP in school*, *Using NLP was enjoyable*, *Using NLP to investigate science questions was meaningful to me.*

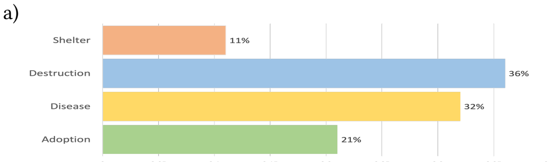
At the end of the classroom studies, we used a paired-samples t-test to determine any significant difference between and pre- and post-survey responses. We found an increase in their enjoyment of using NLP. For the survey item, "Using NLP was enjoyable", the mean response increased in Florida (from 3.41 to 3.85) and Indiana (from 3.22 to 3.61). This was statistically significant with p values of 0.0225 and 0.03, and also had an effect size (Cohen's d) of 0.218 in Florida and 0.35 in Indiana.

5.1.2 Student knowledge assessment. We developed 10 assessment items based on the six main learning objectives. We aligned the assessments with three constructs: Introduction to AI/ML/NLP; Keyword Extraction; and Sentiment Analysis. Table 2 shows three of the assessment items that captured the three constructs. *Item #2* was aimed towards exposing students to some real world applications of NLP (see LO2), Item #3 tested students' conceptual understanding of keyword extraction (see LO3), and Item #5 tested students' ability to identify the appropriate visual representation of sentiment analysis (see LO5).

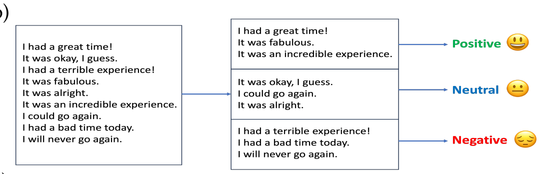
We compared the pre- and post- assessments using a paired-samples t-test and found a significant increase in both the Florida and Indiana classrooms. The mean score increased in Florida (from 3.225 to 4.475) and Indiana (from 4.31 to 7.9). This was statistically significant with both p values below 0.001, and also had an effect size (Cohen's d) of 0.62 in Florida and 1.79 in Indiana.

Table 2: Example NLP-based knowledge assessment items

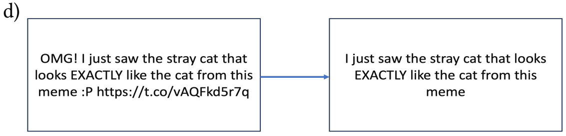
<table>
<tr><th>Assessment Items</th></tr>
<tr><td>Construct 1: Intro to AI/ML/NLP</td></tr>
<tr><td>Which of the following is NOT an example of an NLP application?
a) Spell checking
b) Machine translation
c) Search engine
d) A calculator</td></tr>
<tr><td>Construct 2: Keyword Extraction</td></tr>
<tr><td>What is the goal of keyword extraction in NLP?
a) To identify the most important words in a text.
b) To translate a text from one language to another.
c) To recognize the sentiment of a text.
d) To generate new text.</td></tr>
<tr><td>Construct 3: Sentiment Analysis</td></tr>
<tr><td>Imagine that you have used a sentiment analysis model on a dataset about someone's experience going to an event. Which option below is most likely to be the output of that sentiment analysis model?
a)



b)



c)
Text
I had a great time!
It was okay, I guess.
I had a terrible experience!
It was fabulous.
It was alright.
It was an incredible experience.
I could go again.
I had a bad time today.
I will never go again.
d)



</td></tr>
</table>

5.2 Lessons Learned and Challenges

In this section, we present the challenges we identified while implementing the NLP and science learning activities and the lessons we learned for broader use:

- **Challenge 1: Responding to rapidly evolving NLP landscape.** This project was conceived and begun just months before the widespread of ChatGPT and LLMs. Initially, our curriculum was designed and developed with a 50/50 split between NLP and Science learning objectives. However, with the rapid emergence of ChatGPT and high demand from teachers to include it in the curriculum, it was challenging to quickly design learning objectives and activities dedicated to LLMs. We found that facilitating conversations between students and teachers provided insight into new ideas and topics for future lessons. Moving forward, we need to design learning objectives flexible enough to allow for technological change, and develop learning activities that can be easily adjusted to ensure that students have the opportunity to learn the most relevant content.

- **Challenge 2: Differentiating across grade levels.** Our sixth-grade students found it more challenging to fully grasp the NLP concepts compared to eighth graders. One suggestion made by the sixth-grade teacher was creating a worksheet with the vocabulary to help them better understand the concepts. Another suggestion made by the same teacher was to create a lesson where the students investigate specific scientific practices related to a specific science standard and decide which NLP model will be best to use for the analysis.
- **Challenge 3: Exploring new ways to acquire relevant data due to restricted access to main sources.** Most recently, several data sources, such as the Twitter API, have become restricted or difficult to access. While this presents significant challenges in our efforts to provide real-world data for students to analyze, we also view it as an opportunity to explore newer technologies, such as generating data using ChatGPT. The generated data will supplement the lack of real-world data. This introduces students to more advanced NLP topics such as Generative AI and LLMs, while also facilitating discussions on AI ethics, particularly regarding privacy, data ethics, bias, diversity, and misinformation.
- **Challenge 4: Balancing core subject time with time devoted to learning AI or NLP.** Although our classroom experiences suggest that the "NLP+science" curriculum was successful, one challenge is that providing a more comprehensive NLP-based learning experience would require more class time. In subjects such as science where a mandatory set of learning objectives must be covered, classroom time is limited and all curricula must be designed accordingly.

In addition to the challenges above, we observed the following helpful lessons that point to opportunities.

- **Lesson Learned 1: Non-CS teachers bring valuable insights for teaching NLP.** During the co-design workshops, we had the opportunity to both teach and learn from our middle school science teachers, which greatly influenced the design and development of the learning objectives and tool. All our teachers were non-CS teachers with little to no prior experience in NLP, CS, or AI. As the workshops progressed, we witnessed remarkable growth in their confidence. They contributed numerous ideas, asked insightful questions, and displayed a genuine willingness to learn throughout the entire workshop. These teachers not only provided a unique perspective to our workshops but will likely have a significant impact in sharing their newfound knowledge and experiences with their students. This experience has shown us the importance of engaging with teachers in the early stages of design, as their insights can be invaluable to success in designing and developing novel learning experiences.
- **Lesson Learned 2: NLP can be engaging for middle school students.** Teaching keyword extraction and sentiment analysis proved to be manageable and engaging for students for the purpose of sparking their curiosity in NLP and AI. We provided learning activities that students could easily understand and engage with, such as real-time data analysis, which allowed them to extract text from the internet, explore trending topics, and detect sentiment. We hope to expand upon these studies and incorporate more lessons and hands-on activities in the future.
- **Lesson Learned 3: NLP works well with science, and could work well with other subjects too.** We believed that introducing NLP to learners through science concepts would maintain a familiar setting, easing their transition into NLP concepts and fostering greater engagement. While researchers have found success in integrating CS with science and other disciplines [15], we hope that this experience will follow a similar route and pave the way for future integration of NLP with additional subjects.
- **Lesson Learned 4: Using NLP to analyze real-world science data provides an appropriate bridge between NLP and science.** Using real-world tweets about plastic straws and their impact on the oceans for the data-driven hands-on activities appears to provide an appropriate bridge between NLP and science. Twitter has served as a valuable source of textual data for researchers [20] through its accessible Twitter API. These tweets showcased people's opinions and expressions on social media with the intention of capturing students' attention and exposing them to the overall sentiments of the public opinions surrounding crucial environmental issues. Leveraging our platform, students had the opportunity to analyze the tweets and share their thoughts and ideas on how they could utilize their own social media platforms to bring attention to important issues. We observed that the students were highly engaged in using the tweet datasets to create workflows.

6 CONCLUSION

In this paper, we have shared our experience in co-designing and developing a curriculum and tool to introduce NLP to middle school students in science classrooms. By integrating NLP into middle school science classrooms, we aimed to equip young learners with foundational knowledge in NLP and AI. We collaborated with teachers to iteratively design a text analysis and visualization tool to enable students to interact with NLP models and gain insights about real-world science topics. Overall, we learned that students were highly interested in using NLP, and hands-on activities were vital in fostering their engagement and interests.

This work shows the possibility of integrating NLP into science classrooms and beyond. In the future, we hope to provide public access to our tool and curriculum for use by other schools. Additionally, there is a need to offer more learning activities related to data sources and data pre-processing, as well as data ethics, bias, and privacy. Finally, this paper highlights the need for further research on better ways to integrate and teach NLP and AI in middle school classrooms.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation through grant DRL-2147810 and DRL-2147811. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the authors, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

REFERENCES

[1] 2017. Machine Learning for Kids.

[2] 2017. Teachable Machine. https://teachablemachine.withgoogle.com/

[3] Cecilia O. Alm and Alex Hedges. 2021. Visualizing NLP in Undergraduate Students' Learning about Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[4] Musa Adekunle Ayanwale, Ismaila Temitayo Sanusi, Owolabi Paul Adelana, Kehinde D Aruleba, and Solomon Sunday Oyelere. 2022. Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence* 3 (2022), 100099.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Mehmet Celepkolu, David Austin Fussell, Aisha Chung Galdo, Kristy Elizabeth Boyer, Eric N Wiebe, Bradford W Mott, and James C Lester. 2020. Exploring middle school students' reflections on the infusion of cs into science classrooms. In *Proceedings of the 51st ACM technical symposium on computer science education*. 671–677.

[7] Chloe Clavel and Zoraida Callejas. 2015. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on affective computing* 7, 1 (2015), 74–93.

[8] National Research Council et al. 2013. Next generation science standards: For states, by states. (2013).

[9] S Dhama, G Katuka, M Celepkolu, KE Boyer, K Glazewski, and C Hmelo-Silver. 2023. NLP4Science: Designing a Platform for Integrating Natural Language Processing in Middle School Science Classrooms. In *2023 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 269–273.

[10] Stefania Druga. 2018. Cognimates. http://cognimates.me/home/

[11] Stefania Druga. 2018. *Growing up with AI: Cognimates: from coding to teaching machines*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[12] Eva Durall, Merja Bauters, Iida Hietala, Teemu Leinonen, and Evangelos Kapros. 2019. Co-creation and co-design in technology-enhanced learning: Innovating science learning outside the classroom. *IxD&A* 42 (2019), 202–226.

[13] William Finzer, Jie Chao, Carolyn Rose, and Shiyan Jiang. 2022. StoryQ—an Online Environment for Machine Learning of Text Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12860–12860.

[14] Nazanin Firoozeh, Adeline Nazarenko, Fabrice Alizon, and Béatrice Daille. 2020. Keyword extraction: Issues and methods. *Natural Language Engineering* 26, 3 (2020), 259–291.

[15] Michelle Friend, Andrew W Swift, Betty Love, and Victor Winter. 2023. A Wolf in Lamb's Clothing: Computer Science in a Mathematics Course. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 256–262.

[16] Krista D Glazewski and Peggy A Ertmer. 2020. Fostering complex problem solving for diverse learners: Engaging an ethos of intentionality toward equitable access. *Educational Technology Research and Development* 68 (2020), 679–702.

[17] Krista D Glazewski and Cindy E Hmelo-Silver. 2019. Scaffolding and supporting use of information for ambitious learning practices. *Information and Learning Sciences* 120, 1/2 (2019), 39–58.

[18] Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science* 349, 6245 (2015), 261–266.

[19] Arthur Hjorth. 2021. NaturalLanguageProcesing4All: -A Constructionist NLP tool for Scaffolding Students' Exploration of Text. In *Proceedings of the 17th ACM Conference on International Computing Education Research*. 347–354.

[20] Amir Karami, Morgan Lundy, Frank Webb, and Yogesh K Dwivedi. 2020. Twitter and research: A systematic literature review through text mining. *IEEE access* 8 (2020), 67698–67717.

[21] Amit Kumar, Xiaoyi Tian, Mehmet Celepkolu, Maya Israel, and Kristy Elizabeth Boyer. 2022. Early Design of a Conversational AI Development Platform for Middle Schoolers. In *2022 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 1–3.

[22] Irene Lee, Safinah Ali, Helen Zhang, Daniella DiPaola, and Cynthia Breazeal. 2021. Developing middle school students' AI literacy. In *Proceedings of the 52nd ACM technical symposium on computer science education*. 191–197.

[23] Elizabeth D Liddy. 2001. Natural language processing. (2001).

[24] Phoebe Lin and Jessica Van Brummelen. 2021. Engaging teachers to co-design integrated AI curriculum for K-12 classrooms. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–12.

[25] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.

[26] Ben Mepham, Matthias Kaiser, Erik Thorstensen, Sandy Tomkins, and Kate Millar. 2006. *Ethical matrix manual*. Technical Report. LEI, onderdeel van Wageningen UR.

[27] MonkeyLearn Team. 2022. MonkeyLearn: Text Analysis with Machine Learning. https://www.monkeylearn.com

[28] Narges Norouzi, Snigdha Chaturvedi, and Matthew Rutledge. 2020. Lessons Learned from Teaching Machine Learning and Natural Language Processing to High School Students. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (Apr. 2020), 13397–13403. https://doi.org/10.1609/aaai.v34i09.7063

[29] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[30] Outlier Research & Evaluation. 2017. BASICS Study ECS Student Implementation and Contextual Factor Questionnaire Measures [Measurement Scales]. http://outlier.uchicago.edu/basics/resources/Measures-StudentImplementation/.

[31] Tejal Reddy, Randi Williams, and Cynthia Breazeal. 2021. Text Classification for AI Education. In *SIGCSE*. 1381.

[32] Alpay Sabuncuoglu. 2020. Designing one year curriculum to teach artificial intelligence for middle school. In *Proceedings of the 2020 ACM conference on innovation and technology in computer science education*. 96–102.

[33] Jim Samuel, GG Md Nawaz Ali, Md Mokhlesur Rahman, Ek Esawi, and Yana Samuel. 2020. Covid-19 public sentiment insights and machine learning for tweets classification. *Information* 11, 6 (2020), 314.

[34] Samuel Severance, William R Penuel, Tamara Sumner, and Heather Leary. 2016. Organizing for teacher agency in curricular co-design. *Journal of the Learning Sciences* 25, 4 (2016), 531–564.

[35] David Touretzky, Christina Gardner-McCune, Cynthia Breazeal, Fred Martin, and Deborah Seehorn. 2019. A year in K-12 AI education. *AI Magazine* 40, 4 (2019), 88–90.

[36] David Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. Envisioning AI for K-12: What should every child know about AI?. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9795–9799.

[37] David S Touretzky, Christina Gardner-McCune, Fred Martin, and Deborah Seehorn. 2019. K-12 guidelines for artificial intelligence: what students should know. In *Proc. of the ISTE Conference*, Vol. 53.

[38] Randi Williams, Hae Won Park, Lauren Oh, and Cynthia Breazeal. 2019. Popbots: Designing an artificial intelligence curriculum for early childhood education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9729–9736.

[39] King Woon Yau, CS Chai, Thomas KF Chiu, Helen Meng, Irwin King, and Yeung Yam. 2023. A phenomenographic approach on teacher conceptions of teaching Artificial Intelligence (AI) in K-12 schools. *Education and Information Technologies* 28, 1 (2023), 1041–1064.

[40] Tianchang Zhao and Kyusong Lee. 2020. Talk to papers: Bringing neural question answering to academic search. *arXiv preprint arXiv:2004.02002* (2020).

[41] Abigail Zimmermann-Niefield, Makenna Turner, Bridget Murphy, Shaun K. Kane, and R. Benjamin Shapiro. 2019. Youth Learning Machine Learning through Building Models of Athletic Moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*.