

# Using Natural Language Processing to Optimize Behavioral Data Analysis in Special Education

Ko, Aly, 12th, [alyssako23@fsha.org](mailto:alyssako23@fsha.org), Flintridge Sacred Heart Academy

Huston, Cassie, 12th, [cassandrahuston23@fsha.org](mailto:cassandrahuston23@fsha.org), Flintridge Sacred Heart Academy;

Advisor Ty Buxman, [tbuxman@fsha.org](mailto:tbuxman@fsha.org)

## Abstract

Special education teachers track the behavior of their students so they can perform a Functional Behavioral Assessment which can ultimately lead to a Behavioral Intervention Plan. These plans can help select what student-specific strategies are best suited for intervention with oppositional behavior. Current methods of tracking can be time-consuming, costly, and difficult to analyze. A natural language processing algorithm has been developed which demonstrates the ability to process text inputs, identify the key topics surrounding oppositional behavior, and assign a positive, negative, or neutral sentiment to the topic. The prototype software is accurate more than 50% of the time.

## 1.0 Introduction

Special education teachers observe a variety of oppositional behavior (uncooperative or defiant behaviors towards peers and/or authority figures) in students. While best practices in the field, and often state and federal regulations, require teachers to document student behavior, one technological barrier some special education teachers face is managing and gathering insight from this behavioral documentation. In an ideal world, educators would have both sufficient time and data analysis skills to look for correlations between the student environment, their interactions, and resulting behaviors, giving insight into potential preventative interventions.

Documentation of student data can be recorded in many formats such as handwritten notes, using a digital notebook (such as google docs), or checkboxes and charts. The documentation format can be determined by individual teacher preference or school/district requirements, and often the format itself can be a hindrance to later data analysis.

## 1.2 Machine Learning

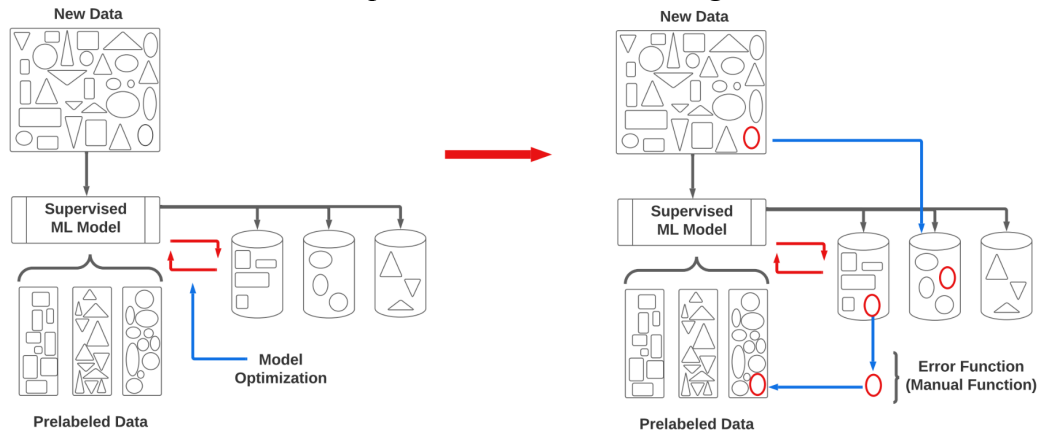
Machine learning (ML) is a type of Artificial Intelligence (AI) that allows a machine to learn from its mistakes to make predictions based on the data it is given. ML enables the machine to become more accurate over time and thus improve its performance and can be used to organize and/or sort data into categories. For example, if a machine was created to distinguish between images of cats and dogs, it would first need to be shown a training dataset (pictures labeled as either cat or dog). This training dataset allows the algorithm to differentiate cats from dogs. After training, the algorithm can then be given new cat and dog pictures for categorization.

This is a simple application of ML, but it shows the three primary components of an algorithm: a decision process, an error function, and a model optimization process (“What is Machine Learning?”, 2020). During training, the decision process includes taking in the data (the labeled data identifying what pictures are cats and which are dogs) and adjusting the algorithm parameters to be able to sort the data correctly. Once trained, a new image is an input to the decision process and the result of the categorization (correct/incorrect) is referred to as the error function. In model optimization, the error function is used to “re-train”, or update the algorithm parameters. This form of model optimization is what separates the ML model from a heuristic model. In a heuristic model, the error function may be used as a new input, adding another image to the “known cat/dog list”. In ML, the error function is used to actually modify the algorithm parameters. Every model optimization cycle in ML results in a different program (a modified “brain” of sorts), improving future performance (identifying cats and dogs more accurately). Machine learning moderately improves its precision by “learning” just as the human brain does through the model optimization process. Two types of machine learning are supervised (typically used in algorithms that are meant to make predictions) and unsupervised (used in algorithms that intend to understand relationships between the data) (Seldon, 2022) (“What is Machine Learning (ML)?”, 2020).

Supervised machine learning models use labeled data sets (already organized by humans) to assure that the data is accurately sorted. When the supervised ML model is shown new data it has never seen before, it is able to categorize it based on prior data that it was given. Error function in the supervised model applies when the model makes a mistake and sorts data into the

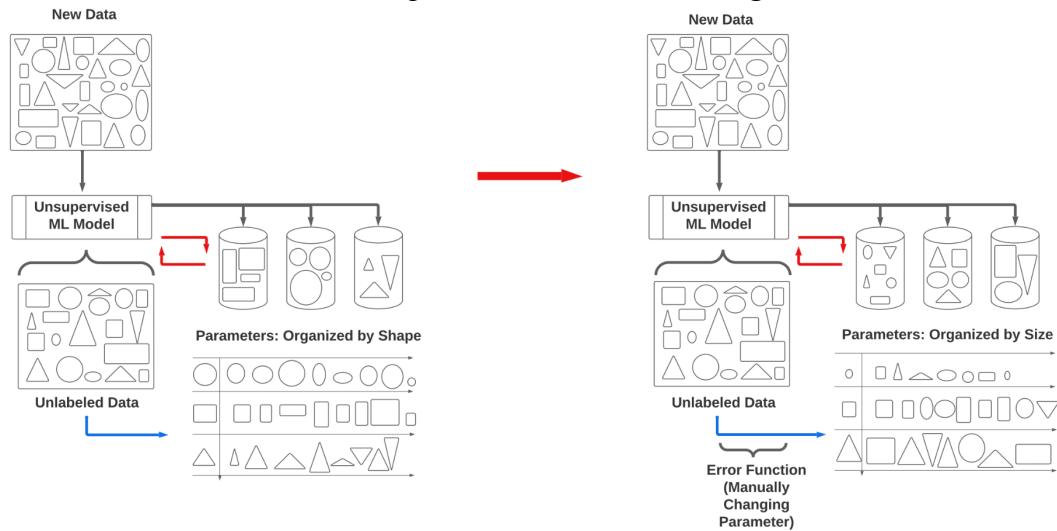
wrong category. In Figure 1, the supervised machine learning model is sorting shapes based on the prelabeled data it has been given as a reference. The red oval shape represents the incorrectly sorted data that now goes through the error function where its placement is manually corrected. The model then incorporates this correction in its prelabeled data. Now the model knows to sort the oval into the correct category when it encounters it again.

*Figure 1: Diagram of Decision Processing, Error Function, and Model Optimization Steps Within a Supervised Machine Learning Model*



The function of unsupervised machine learning is to detect patterns, relationships, and similarities/differences in datasets. Unsupervised machine learning uses unlabeled data sets and organizes them into groups without prior human intervention. In the unsupervised model, the error function is used by manually changing the parameters of the model by which the data is organized. In Figure 2, the parameters of the model are to sort the dataset by type of shape. Manually modifying the parameters in an unsupervised model influences the output of the categorized data. Originally, the images were organized by type of shape, after the model optimization, the images were organized by the size of the object pictured.

Figure 2: Diagram of Decision Processing, Error Function, and Model Optimization Steps Within an Unsupervised Machine Learning Model



Machine learning can be seen in everyday life as it is used in speech recognition, customer service, recommendation engines, and more (“What is Machine Learning?”, 2020).

### 1.3 Natural language processing

A specific type of machine learning is Natural Language Processing (NLP), which uses ML algorithms to organize text and audio similar to the way humans process text and audio. NLP can be used in a variety of ways and is a part of some people's daily life. For example, Google Translate and virtual help/chatbots (such as Siri) are examples of NLP (“What is Machine Learning?”, 2020). There are multiple types of NLP tasks such as speech recognition (converts audio to text), part of speech tagging (determines the part of speech of each word), word sense disambiguation (determines the meaning of the word in context when a word has multiple meanings), coreference resolution (determines when words are referring to the same thing such as “she” and “girl”), and natural language generation (puts data into human language) (“What is Machine Learning?”, 2020). A Natural Language Toolkit (NLTK) is a specific library in Python that works with human language data and NLP and has pre-made tools that can be used to access these methods mentioned above. One of these premade tools includes tokenization, which is the process of breaking down the words in the text into individual “tokens” which will help the computer process the text. Another useful tool for this project is stemming and lemmatization, which is the process of breaking words down to their roots. NLP uses all of these tools and methods to process audio or text language, including complex tasks

such as using slang, sarcasm, or idioms (“What is Machine Learning?”, 2020). This type of language can be more easily understood by humans but can be hard to incorporate into an algorithm. With training and calibration, the NLP model is able to accurately identify patterns and sentiments in human language. Since the defining feature of a machine learning algorithm is that it uses model optimization, the more it is used, the more it is able to adjust its parameters and learn from feedback.

An example of a program that uses machine learning is ChatGPT. ChatGPT is a chatbot that uses Reinforcement Learning with Human Feedback (this is an example of the error function, where humans provide feedback, to produce better results) to predict words and produce human-like responses. The generated responses are based on the informational data it was trained on, which makes it an example of supervised learning.

### 1.3.1 Topic Modeling

Topic modeling is a type of trained and unsupervised machine learning that analyzes documental text datasets by clustering words/phrases into groups and detecting their patterns and frequency in order to identify abstract “topics” that encompass the general idea of the documents. These clusters put the words/phrases into categories that the computer has determined as important categories based on what kind of words are used (nouns, adjectives, etc), frequency of use, and word context (for example, when two words have a similar meaning, they fall under the same category). The sorting of the topics is the decision-processing component of the ML application. Topic modeling uses connections between words to sort them into the same category. Using the error function, the topic modeling algorithm improves its performance by constantly updating its parameters based on previous data. One aspect of topic modeling is a group of words called “stopwords.” These are words that are used repeatedly throughout large chunks of text that have no significant contribution to the meaning of the text as a whole. There are some generic stopwords that are removed to filter out unimportant words such as “a,” “that,” “in,” “out”, etc.

A Latent Dirichlet Allocation (LDA) model is also used in topic modeling. The LDA model assumes that there are a fixed number of topics and that the document is a mixture of these distributed topics. By assuming the document is a mixture of topics, the model is able to

identify and distinguish between unique topics before presenting them in the output. It measures the most probable words associated with the topic to generate a collection of words for each topic. The LDA model is used in text classification, information retrieval, and recommendation systems.

### 1.3.2 Sentiment Analysis

Sentiment Analysis is a subset of Natural Language Processing that is trained to analyze a textual dataset and determines whether the sentiment of a piece of text is positive, negative, or neutral. It is used in processing customer reviews to detect positive or negative sentiments in comments and posts. The sentiment analysis algorithm identifies and extracts the words and phrases that indicate certain sentiments. These extracted features are then categorized as positive, negative, or neutral based on association with specific sentiment words or phrases. The sentiment analysis model is a supervised model that is trained using a large dataset containing known sentiment labels. The model improves its accuracy by learning (human identification of error functions) from these sample datasets so that it can predict the sentiment of new data (model optimization). A score is then given to relay the percentage of the text that is positive, negative, and neutral. This is an example of supervised learning.

### 1.4 Literature Review

A research group at Seoul National University used ML and Topic Modeling to explore the behavior patterns of online users with Emotional Eating (EE) Behavior in order to create personalized intervention plans. Data was extracted from the online posts of a subcommunity on the social media platform, Reddit, that focused on weight management. Topic Modeling was used to sort these posts into four model-generated categories: sharing feelings, noticing physical appearance/changes, asking for dietary advice, and sharing dietary advice. Keywords that were frequently associated with these topics were detected (the decision process) and recorded to increase the efficiency of data analysis. The trained model found that for different types of EE, the feedback distribution of the topics showed significant differences between users. The detection of these disparities between the use of linguistics in types of EE confirmed that the model was able to detect abnormal behavior (using the error function). The results were also validated by an expert group of medical professionals in order to train the ML model accurately.

The algorithm used in this study is an example of a supervised ML model that allows the results to become more accurate over time (Hwang, 2020). This study shows how natural language processing can be used to find behavioral patterns by analyzing written text, and is useful as a method guide for the current project.

Another team of software engineers built an unsupervised NLP pipeline to automatically generate a glossary of terms from a body of text (novels, articles, essays, etc). In order to identify the terms accurately, the team used a semantic similarity function to calculate and prioritize the words based on relevance and context. The nouns and adjectives were extracted using the NLTK library before being tokenized. In this way, the final glossary was made up of terms measured by their frequency, context, and complexity. Because this model is unsupervised, the model output was not accurate enough to function alone. This led the team to conclude that evaluation based on the objective scale of precision should be re-evaluated by humans (Mishra, 2022), meaning that at least a semi-supervised system is necessary. The current project will also be using the NLTK library to extract filler words and clean data before it is processed and analyzed by the model and will learn from the recommendation to use semi-supervised algorithms.

### 1.5 Project Statement

Special education teachers need a way to use data analysis to take advantage of recorded student behavior and to identify potential connections between events that lead to oppositional behavioral patterns which have previously gone undetected so they can help prevent oppositional behavior. However, current methods of analysis are inefficient because teachers have to manually analyze the results, which is time-consuming and allows for the possibility of underlying behavioral patterns being overlooked by the human eye.

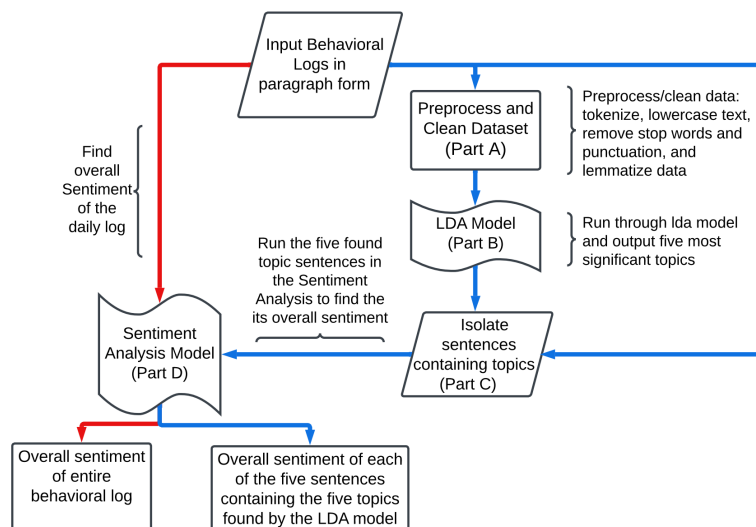
A Machine Learning based analysis algorithm is developed to automatically process special education student behavior logs. Using topic modeling and sentiment analysis, the algorithm will find patterns in behaviors and output connections that would otherwise go unnoticed.

### 2.0 Methods

Two NLP algorithms (a Sentiment Analysis Model and an LDA Model shown in Figure 3) are used to process behavioral logs of children made by education specialists, teachers, and

parents. Behavioral logs are first Preprocessed, which includes tokenization (to break down sentences in the text into smaller units called tokens) and stopwords removal (filtering out the unimportant filler words).

*Figure 3: Diagram of NLP Model Using Topic Modeling and Sentiment Analysis to Identify Main Topics/Ideas from Written Behavior Logs and Their Corresponding Sentiments*



Next, an LDA model (Figure 3B) generates and identifies underlying topics (the ML decision process). The LDA model is able to identify and distinguish between topics before presenting them in the output. The topics generated by the model are formed by a list of words that are frequently associated with each other or found together. The overall topic is then selected by isolating the first term from the list of pulled topics.

Finally, individual sentences containing the topics are isolated (Figure 3C), using a heuristic search. These found sentences from the sentence isolation process are input into a pre-trained sentiment analysis model (Figure 3D). This model is an open-access model used by Twitter to determine the sentiment of tweets. The model gives a positive, negative, and neutral percentage score for each sentence. The sentence's corresponding topic is then labeled with the highest sentiment score given by the model.



## 2.1 Error Function and Model Optimization

Once the topics from the dataset were extracted and the sentiment was assigned by the model, the results are evaluated by people. As a proof of concept, the algorithm was first used on a history text with students in a European History class as validators. Chapter 20 of an AP European History textbook regarding the Industrial Revolution was read by 27 students. These students then reported the top 5 topics that they found to be the most important topics of the chapter, similar to how an LDA model works. The same chapter was then processed by the NLP algorithm for comparison

## 2.2 ChatGPT Log Generation

While various formats for recording behaviors exist, a record taken in paragraph form is needed to take advantage of the NLP algorithm. Consulting with specialists in the field revealed two primary issues that arise with using real-world behavior logs: student privacy and incorrect log formats. ChatGPT was used to generate behavioral logs from the perspective of a special education teacher. The original prompt for ChatGPT was vague:

*Let's say there is a student named Fiona. Fiona shows signs of oppositional behavior, but does not have ODD. Fiona has autism. Write me a behavioral log that Fiona's special education teacher would write about her day at school in paragraph form.*

The generated responses to this prompt were evaluated by a trained special education professional, resulting in suggestions to seek more details (who was there, what time, etc). The prompt was repeatedly adjusted until the response provided by ChatGPT was realistic as if provided by a trained professional. The final prompt was

*“Let's say there is a student named Velma. Velma has a disability. Write a total of 5 days worth of behavioral logs. Each day should be different and should randomly alternate between a day that has overall positive, negative, or neutral behavior. Be specific about identifying triggers of certain behaviors, such as what directions were given and if they were listened to or not, who was giving the directions, why certain behaviors are occurring, what time of day it is, who was near the student (you can make up characters), etc. Be specific. Write it in a paragraph (at least 7 sentences for each log). Write 1 paragraph per day. Also include specifically what the good, bad, or neutral behaviors were. However, do not write that Velma was in a positive, negative, or neutral mood. Instead say a happy, sad, frustrated, excited, calm, etc mood.”*

The final response to the prompt above can be found in Appendix 7.1, and these logs were used for model validation.

### 2.3 Accuracy Verification

There were two parts of the algorithm that needed to be verified for the final testing - the topic modeling and the sentiment analysis. To test the algorithm, logs were generated by ChatGPT and a human went through them and identified the top 5 most important topics and the sentiment related to the topics. The same logs were then run through the algorithm and the 5 topics found by the algorithm were compared to those found by the human. Of the topics that matched, the sentiments of each were also compared.

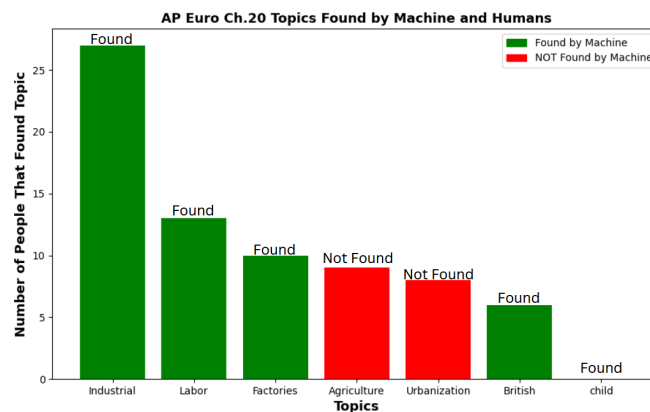
### 3.0 Results

Two tests were initiated to evaluate how well the NLP could identify essential topics as described in 2.1 and 2.3.

#### 3.1 AP European History

For the AP Euro test (see Figure 4), 4/7 (57%) of the topics were found by both the algorithm and humans (industrial, labor, factories, and British). 2/7 were found by the humans and not the algorithm and 1/7 was found by the algorithm and not humans.

*Figure 4: Graph of Validation Data to Determine if the NLP Algorithm was Accurate Enough to Proceed*

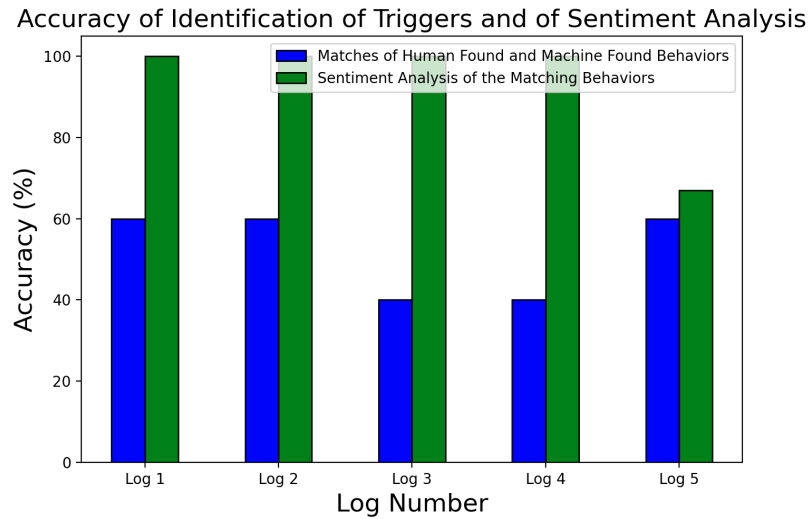


#### 3.2 Simulated Behavioral Log Topic Identification

The average success rate when applying the NLP to ChatGPT-generated simulated behavior logs was 52%. Success is defined as a match between the identification of a topic by the NLP and by a human reading the same log. The average result for the sentimental analysis (of

the topics that both the algorithm and humans found) was 93.2%. The results can be seen in Figure 5. The algorithm only determines the sentiment for topics identified by the LDA

Figure 5: Graph of NLP and Sentiment Analysis using Behavioral Logs Generated by ChatGPT



#### 4.0 Discussion

The AP European History test showed that the algorithm could successfully isolate topics. To improve the accuracy of topic selection, additional training (using a larger data set) would be helpful along with an extended model optimization phase (requiring significant student validator time). This was not done because this test was only a proof of validation, so it was sufficient enough to move on to the next section, which was an application to behavioral logs. The first try of topic modeling was 57% accuracy with limited resources and training. To make an ML algorithm with a higher accuracy rate, large amounts of data are required to train the model, along with lots of feedback to continue the training. To do this would require lots of data, feedback from humans, and time.

Application of the model to the ChatGPT generated “Simulated Behavioral Logs” resulted in a 52% match rate in topic identification. For matched topics, the sentiment analysis had an accuracy rate of 93.2%. This tells us that more training is needed for the NLP algorithm. ML requires large chunks of data to learn from and because this model is in the process of being trained, it will improve in performance as it is given more data and learns the specific formatting

and wording patterns that pertain to the user. The sentiment analysis was pre-trained and this is evident in the accuracy of its results.

The interpretation of “success” for this project is an accuracy rate high enough to be comparable to, if not more accurate than that of humans. However, no testing has been done on the accuracy rate of humans when it comes to topic modeling and sentiment analysis. This is because human results of text analysis are often subjective. Specifically, for this project, a study on finding topics over long periods of time would prove to be most useful as the model will be able to refine its parameters to the user’s unique writing patterns, thus increasing its effectiveness the more it is used.

## 5.0 Conclusion

A machine learning algorithm has been developed using topic modeling and sentiment analysis for application to behavioral logs. This model consistently outputs data on significant topics and their sentiments in an efficient manner. The implementation of stop words allows for the model to become more accurate over use and time. This algorithm will allow teachers and specialists to efficiently process data and detect patterns in behavioral logs over time. The format of the data must be in an organized and consistent manner for the most effective performance of the model.

In the future, other ways of submitting data would be ideal. One limitation right now is that the algorithm can only accept data in paragraph form in a .json file. Another way to improve this project would be to have a human look at the ChatGPT logs, then rate the NLP against the human. More critically, record time and see how long it takes the human compared to your algorithm to obtain quantifiable data. Along with that, being able to detect connections between the topics would give more insight into the findings.

## 6.0 Works Cited

Corsetto, K. (2022, June 11). *How data collection shapes teacher and student success*. n2y.

Retrieved December 12, 2022, from

<https://www.n2y.com/blog/data-collection-class-success/>

Hwang Y, Kim HJ, Choi HJ, Lee J. *Exploring Abnormal Behavior Patterns of Online Users With*

*Emotional Eating Behavior: Topic Modeling Study*, J Med Internet Res 2020;22(3):e15700

Retrieved December 12, 2022 from

<https://ieeexplore.ieee.org/abstract/document/5457456>

IBM Cloud Education. (n.d.). *What is machine learning?* IBM. Retrieved October 26, 2022,

from <https://www.ibm.com/cloud/learn/machine-learning>

IBM Cloud Education. (n.d.). *What is natural language processing?* IBM. Retrieved November

2, 2022, from <https://www.ibm.com/cloud/learn/natural-language-processing>

Kbar, G., & Aly, S. (2014, September). *Smart workplace for persons with disabilities*

(SMARTDISABLE) - researchgate. Research Gate. Retrieved November 3, 2022, from

[https://www.researchgate.net/publication/285996469\\_SMART\\_workplace\\_for\\_persons\\_with\\_DISABiLitiEs\\_SMARTDISABLE](https://www.researchgate.net/publication/285996469_SMART_workplace_for_persons_with_DISABiLitiEs_SMARTDISABLE)

Mishra, P. (2022, August 4). *Automatic glossary and definition extraction from text using NLP*

*techniques*. Paperspace Blog. Retrieved December 12, 2022, from

<https://blog.paperspace.com/adaptive-testing-and-debugging-of-nlp-models-research-paper-explained/>

Seldon. (2022, October 5). *Supervised vs unsupervised learning explained*. Seldon. Retrieved

December 6, 2022, from

<https://www.seldon.io/supervised-vs-unsupervised-learning-explained#:~:text=Supervised%20machine%20learning%20is%20generally,the%20need%20for%20labelled%20data>

“What Does a Career in Special Education Look like?: Bestcollegereviews.” *Best College*

*Reviews*, 18 Oct. 2022,

<https://www.bestcollegereviews.org/education-and-teaching/special-education/day-in-the-life/>.

“5 Ways to Track Behavior Data - the Intentional IEP.” *The Intentional IEP* -, 13 Sept. 2022,  
<https://www.theintentionaliep.com/ways-track-behavior-data/>.

## 7.0 Appendices

### Appendix 7.1: Response of ChatGPT to the Prompt Stated in 2.2.1

#### *“Day 1: Positive Behavior Log*

*Today, Velma had a really positive day at school. She came to class with a smile on her face and a positive attitude. Velma was able to focus on her tasks and complete them in a timely manner. During math class, she was able to follow the teacher's instructions and complete the worksheet independently. She participated in group activities and collaborated with her classmates effectively. Velma even volunteered to read aloud during reading time. During lunch, Velma sat with her friends and engaged in positive conversation. She was able to transition between activities easily and with a positive attitude. Velma's happy mood was contagious and it positively impacted the classroom environment.*

#### *Day 2: Negative Behavior Log*

*Today was a challenging day for Velma. She came to class feeling tired and frustrated. During math class, she struggled to understand the instructions and asked her teacher for help multiple times. Velma became agitated when the teacher suggested she ask a classmate for assistance instead. During recess, Velma became upset when a group of students accidentally knocked over her bookbag. She yelled and became physically aggressive towards the students. When the teacher intervened, Velma continued to be defiant and refused to calm down. Her negative behavior continued throughout the day, and Velma was unable to complete her work due to her frustration and anger.*

#### *Day 3: Neutral Behavior Log*

*Today, Velma had a relatively neutral day at school. She came to class in a calm and relaxed mood. During math class, she completed the worksheet with minimal assistance from her teacher. During recess, Velma sat alone and read a book quietly. When the teacher asked her to participate in group activities, Velma complied but did not engage much. She completed her work independently but did not ask for help when she struggled. Overall, Velma was quiet and reserved today.*

#### *Day 4: Positive Behavior Log*

*Today, Velma had another positive day at school. She came to class with a smile on her face and a positive attitude. During math class, Velma was able to follow the teacher's instructions and complete the worksheet independently. She participated in group activities and collaborated with her classmates effectively. Velma even volunteered to read aloud during reading time. During lunch, Velma sat with her friends and engaged in positive conversation. She was able to transition between activities easily and with a positive attitude. Velma's happy mood was contagious and it positively impacted the classroom environment.*

#### *Day 5: Negative Behavior Log*

*Today, Velma had a difficult day at school. She came to class in a sad mood and struggled to focus on her work. During math class, Velma became overwhelmed and frustrated with the task at hand. She began to cry and refused to continue the worksheet. The teacher tried to comfort her, but Velma continued to cry and became increasingly agitated. During recess, Velma sat alone and did not interact with her classmates. When the teacher tried to engage her in conversation, Velma became angry and refused to speak. Her negative behavior continued throughout the day, and Velma was unable to complete her work due to her emotional distress.”*