CS 7641 Machine Learning Midterm Report

Christine Olds, Christian Lee, Jeongrok (Michael) Yu,
Tae Eun (Harrison) Kwon and Alexandra Pfleegor

I. Introduction/Background

A. Literature Review

Extensive research investigates TED Talks' role in popularizing topics [1] [2] [3]. Nonlinear methods, including Support Vector Regression and Gaussian radial basis functions, enhance the accuracy of predicting a video's popularity [4]. Sentiment Analysis is used to extract a text's emotional tone, classifying it as negative, neutral, or positive [5]. Apriori association rule mining can cluster groups of commonly occurring keywords [6]. Features such as keywords, sentiment, etc. can be used to classify and predict the popularity of TedTalks [7].

Encoder-decoder models, unlike BERT or GPT models, can take in the input sequence and generate embedding using the encoder, then produce an output sequence. This has been pivotal in the advances in numerous fields of Natural Language Processing including text summarization. Most of the state-of-the-art models now use encoder-decoder models like BART and T5 and fine-tune the models with human-annotated summaries [8] [9] [10]. However, there has been a boom in decoder-only models such as the Chat-GPT model, thanks to their scalability in pre-training data and the incorporation of Reinforcement Learning with Human Feedback (RLHF) [11]. We believe that it would be interesting to see how Chat-GPT performs when only incorporating prompt engineering and plan to compare the performance between the two different architectures of models.

B. Data Description and EDA

The dataset comprises data on 5,662 TED Talks from 1972 to 2022 and includes the features shown in Table I for each talk. Exploratory data analysis was performed in a Jupyter Notebook using a Python 3 kernel.

1) Numerical Features: The expected numerical columns are `_id`, `duration`, `likes`, and `views`. A summary of the dataset revealed that the `likes` column was being read as a string rather than an integer. To fix this, we created a function that converts the string into an integer and multiplies it by a factor of 1,000 if there was a "K" at the end of the string or a factor of 1,000,000 if there was an "M" at the end of the string. The four numerical features are now `_id`, `duration`, `likes`, and `views`.

The EDA revealed right-skewed historgrams for views and likes, which we acknowledge can be attributed to the exponential growth YouTube and the fame TED talks have gained over the years as well. The polarity is evident in that the most viewed TED talk within our dataset has over 73 million views while the least-watched one has only 587 views. However, we inspected what were the two outliers within the 'durations' category, as a TED talk over 4 hours long was irregular. One of the outliers was a collection of 10 or more TED Talks talking about climate change and the other one was a countdown video for a live TED conference. We concluded that these talks, while talking about one topic, are not similar to other talks, and thus decided to exclude these two videos. Figures 1, 2, 3 display box and whisker plots of `duration`, `likes`, and `views`, respectively.

TABLE I
Data Set Attributes

Attribute	Description
`_id`	unique dataset identifier for each TED Talk
`duration`	duration in seconds of each TED Talk
`event`	TED event at which the talk was given
`likes`	number of likes of each TED Talk, to a factor as indicated [1]
`page_url`	official link to video on TED website
`published_date`	date that TED Talk was published on official TED website
`recorded_date`	actual date that TED Talk was performed (may differ from published date)
`related_videos`	list of video IDs related to respective TED Talk
`speakers`	list of speakers, including each speaker's name and occupation description
`subtitle_languages`	list of subtitle languages available, including the language as well as a code [2]
`summary`	summary of the TED Talk
`title`	title of the TED Talk
`topics`	a list of topics, including a unique identifier for each topic
`transcript`	a complete transcript of each TED Talk in English
`views`	number of views that each TED Talk has
`youtube_video_code`	link to the YouTube video listing

[1] For example, a "K" indicates thousands, while an "M" indicates millions.
[2] For example, English is listed with its respective code, "en".

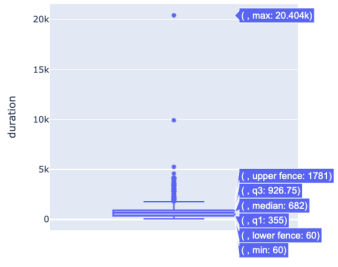

Fig. 1. Box and Whisker Plot for Duration: $\mu = 707.692, \sigma = 521.179$

Next, we checked for correlation between the numerical

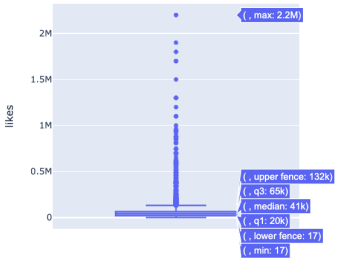

Fig. 2. Box and Whisker Plot for Likes: $\mu = 63,362.765, \sigma = 108,900.280$

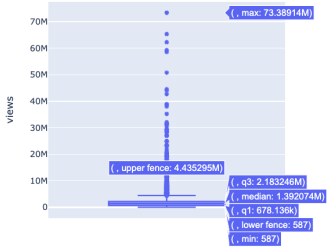

Fig. 3. Box and Whisker Plot for Views: $\mu = 2,130,394.214, \sigma = 3,679,481.018$

columns. As expected, there was a high correlation between likes and views. Figure 4 shows the correlation matrix. The likes and views columns will be used to calculate popularity, eliminating the correlation issue.

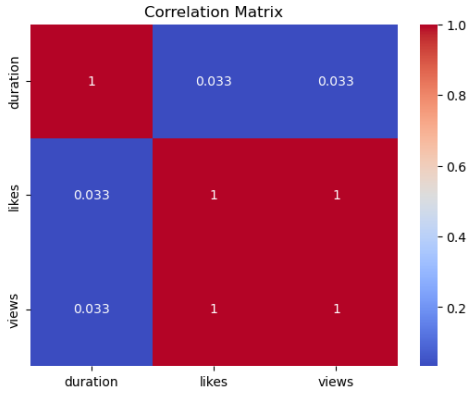

Fig. 4. Correlation Matrix

2) Categorical Features: The main categorical features we were interested in exploring were `speakers`, `subtitle_languages`, and `topics`. The data for these columns were originally in dictionary format. To make the information more accessible, we extracted the values from the dictionaries and created three new columns: `speakers_list`, `sl_list` (subtitle languages list), and `topics_list`.

Using this information, we were able to see that speakers oftentimes return, with the top three speakers being Alex Gendler with 48 videos, Iseult Gillespie with 35 videos, and Matt Walker with 17 videos. It will be interesting to see if these speakers tend to have more 'popular' videos. Additionally, we found that the most commonly presented topics are science with 1267 videos, technology with 1231 videos, and TED-Ed with 878 videos.

3) Transcript: Since we plan to use the calculated sentiment of the transcript in the regression model to predict popularity, the transcript is an integral part of the dataset. All text cleaning will be explained in the data preprocessing section. During EDA, we noticed that 648 videos were missing transcripts. Since we cannot recreate or estimate these transcripts, we will be excluding those videos from the dataset.

4) Time Series: Since we have the recorded and published dates from the data, we plotted the number of talks published per year. This graph is shown in Figure 5.

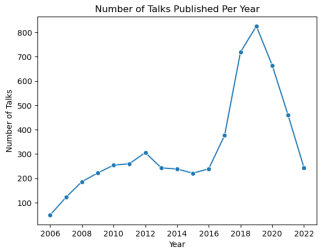

Fig. 5. Number of TED Talks Per Year

C. Dataset Link

The dataset can be accessed on Kaggle using this link.

II. Problem Definition

A. Problem

- Create a regression model to predict the popularity of TED Talks based on association rule mining and sentiment of the transcript
- Compress really long transcripts to summarize and compare with the ground truth

B. Motivation

- Understand how the commonly used words and sentiments of a TED Talk affect its popularity.
- Challenge state-of-the-art text summarization models to encapsulate a long transcript into short sentences.

III. Methods

A. Data Preprocessing

As seen in Figure 6, we utilized several steps to clean and process our data.

1) Popularity Calculation: After some basic data cleaning, we created our popularity metric using the following equation:

$$\texttt{popularity} = \left(\frac{\texttt{likes}}{\texttt{views}}\right) * 100$$

We utilize this ratio to see what percentage of people who viewed each video liked the video. Although this is an imperfect metric, it stands as a good proxy for popularity.

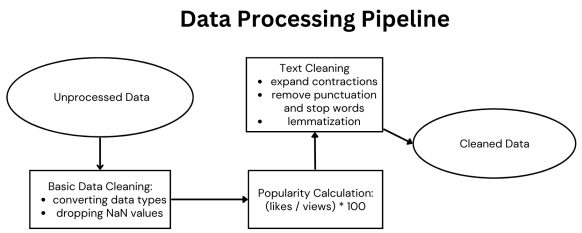

Fig. 6. Data Processing

2) Text Cleaning: We also needed to clean the textual data for our sentiment analysis algorithms. Using the contractions package, we first expand all contractions. For example, we turn 'I'm' into 'I am' and 'can't' into 'can not'. This becomes useful once we remove punctuation and stopwords in the next step. After that, we lemmatize all words in each transcript before putting it all back into one single string.

For the text generation task, we further split the data into sentences because each input is in a sentence form for encoder-decoder models before being tokenized. It was imperative to distinguish acronyms and punctuation so that the model does not confuse periods in the U.S. as the punctuation marker for ending a sentence. Most of the state-of-the-art summarization language models had a maximum token limit for the encoder. Due to the exceeding length of our transcripts, we had to further split one transcript to not exceed the limit and made sure the end of each split was at the end of a sentence. The column with the processed transcript is then called `processed_transcript`.

3) Feature Engineering: In our preprocessing efforts, we have employed prompt engineering through the OpenAI API to extract emotional content from TED Talks, yielding 'detected emotion' data in JSON format. This step has allowed us to capture the nuanced emotional landscape of thses talks. Currently, these detected emotion data have not been incorporated into our SVR model, as the transformation of JSON strings into numerical features is ongoing.

Our forthcoming work will focus on converting these emotional insights into quantifiable features for predictive modeling. The integration of 'detected emotion' feature is expected to enhance our model to predict TED Talks' popularity, based on the hypothesis that the emotional impact of a talk influences its appeal.

B. Machine Learning Algorithms/Models

1) SVR: Our first supervised method is `SVR` (support vector regression model) from `sklearn.svm`. Since we found that other models have had success predicting the popularity of success using the Gaussian radial basis function kernel during our research, we tinkered with the implementation of this kernel, called 'rbf'. Additionally, we also attempted a linear, polynomial, and sigmoid kernel.

2) Text Summarization: Another supervised method we utilized was text summarization. For this, we first had to come up with a clever way to split the long transcripts into smaller chunks, as all the encoder-decoder models we plan to use have a maximum token limit of 1024 tokens. Therefore, we first tokenized the transcripts into tensor embeddings using the BART-large model [9] and determined the splitting points while maintaining the sentence information. Next, we sub-sampled about 10 videos that we all personally watched, and conducted summarization on this small batch before scaling our experiments.

First, we utilized the BART summarization model to summarize the splits of transcripts to output a compressed transcript that was shorter than 1024 tokens. Then, we ran the summarization model again on this compressed transcript to shorten the summary like the one-sentence ground truth summary within our TED dataset. When doing so, because we are running summarization on the output of the model, we often face hallucination, meaning that the model was likely to output the same sentences from the previous steps.

Next, we utilized an extractive summarization method using the Sumy module to compress the transcript. Within Sumy, we integrated LexRank, an extractive summarization algorithm that is based on PageRank algorithm [12]. LexRank first computes the similarity between sentences in the text, constructs a graph representation of the text, iteratively ranks the text based on the similarity of the graph representation, and extracts the highly ranked sentences. Given this extraction, we now feed it to the BART summarization model and obtain more relevant results.

Another model we tried is just using a decoder-only model, ChatGPT, to summarize our long transcript. ChatGPT is advantageous when compared to other models in that it has a much larger maximum token length of 16385 tokens. This means that ChatGPT can take in the transcript at once, without losing any information about the transcript during summarization. An example of our summarization result is shown in Table VI within the Appendix.

3) Sentiment Analysis: Sentiment Analysis is utilized to find the perspective, view, or attitude of the writer by using computational methods to quantify and categorize a writer's sentiment [5]. For this project, we utilized two different lexicon-based sentiment analysis tools, TextBlob and VADER, as our unsupervised methods. Both TextBlob and VADER assign a score between -1 and 1, where -1 signifies a negative sentiment, 0 signifies a neutral sentiment, and +1 signifies a positive sentiment [13]. After data preprocessing (see data preprocessing section for more information), the TextBlob and VADER algorithms were applied to the processed transcript. New columns containing the scores for each respective algorithm were added to the dataset. See the results section for a discussion on our analysis of the algorithms.

4) Association Rule Mining: Association Rule Mining is used to uncover relationships, or "rules", between variables

in a large dataset. The model uses metrics such as support, confidence, and lift to measure the strength of these rules [6]. For this project, we utilized association rule mining to extract commonly used words and word clusters from TED Talk titles. The support of a rule represents the fraction of all titles that demonstrate that rule. The rules that generated the highest support included single words such as ”life”, ”world”, ”us”, and ”future”. The confidence of a rule measures the strength of the association between its items, that is, how likely an item is to appear given the other item’s presence. The rules that generated the highest confidence included word clusters such as {”brief”, ”history”}, {”solve”, ”riddle”}, and {”climate”, ”action”}.

The original goal in conducting association rule mining was to use the rules it produced and their associated metrics as features in our regression model to predict the popularity of TED Talks. The idea was that extracting common words or strongly associated word clusters may reveal trends of intentional titling of videos for the purpose of attracting the attention of potential viewers (i.e. ”clickbait”). However, a correlation matrix of rule and popularity metrics reveal weak or counterintuitive relationships, as shown in Figure 7. Thus, we concluded not to append the association rule mining metrics to our regression dataset.

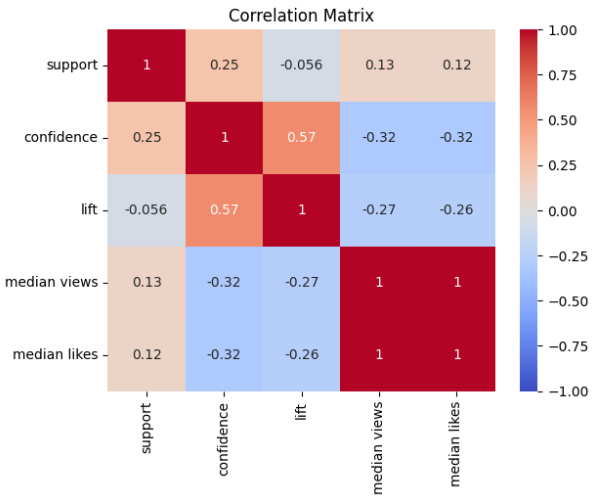

Fig. 7. Correlation Matrix

IV. Results and Discussion

A. Quantitative Metrics and Analysis

1) Regression Analysis: We calculated the mean squared error (MSE) to measure the accuracy of our models. The results for the regressions are shown in Table II.

Comparing the MSE values, we feel the models are very overfit, with the exception of the Sigmoid model. In our next steps, we plan to implement a feature selection model, such as PCA or Lasso, to reduce the amount of overfitting.

TABLE II
Support Vector Regression Models

Kernel	MSE
Linear	0.0056
RBF	0.0059
Poly	0.0057
Sigmoid	0.2674

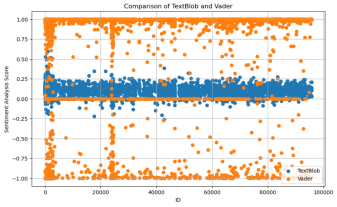

Fig. 8. TextBlob vs. VADER Comparison

2) Sentiment Analysis: After data preprocessing and text cleaning (see data preprocessing section for more information), we applied two sentiment analysis algorithms, TextBlob and VADER, to the transcripts. The disparity in the scores was quite interesting. TextBlob tended to score the transcripts more conservatively as a ’neutral’ sentiment around 0. Table III the summary statistics from the TextBlob algorithm.

TABLE III
TextBlob and VADER Results

	Min	Mean	Median	Max
TextBlob	-0.259643	0.095959	0.098674	0.594444
VADER	-0.999900	0.656523	0.996300	1.000000

VADER tended to assign more extreme scores closer to -1 or 1. Table III also shows the summary statistics from the VADER algorithm.

The disparity between the two algorithms is apparent in Figure 8 where blue is the TextBlob algorithm and orange is the VADER algorithm. Since sentiment analysis is an unsupervised learning method, we created a ground truth by manually reviewing 50 videos and assigning a score between -1 and 1. We then compared our scores to the scores assigned by TextBlob and VADER by calculating the MSE.
For the 50 videos we manually scored, the summary of statistics is shown in Table IV.

TABLE IV
Sentiment Analysis Summary Statistics from Subset of Data

	Min	Mean	Median	Max
TextBlob	0.007780	0.127299	0.122397	0.268511
VADER	-0.999800	0.919600	0.999500	0.999900
Manual	-0.500000	0.320000	0.400000	0.900000

The distribution of scores for these 50 videos can be visualized in Figure 9 where blue is the TextBlob score, orange is the VADER score, and green is the manual score. The MSE’s of the TextBlob and VADER algorithms when compared to the manual scores were 0.15 and 0.57, respectively. See Figure 10 for a visualization of the MSE scores, where blue is

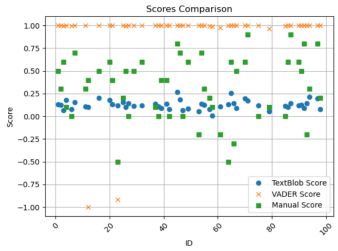

Fig. 9. Comparison Between TextBlob, VADER, and Manual Scores

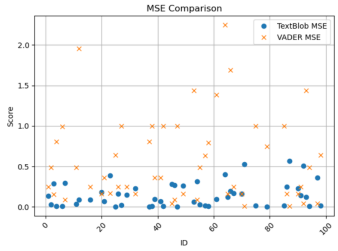

Fig. 10. TextBlob vs. VADER MSE Scores

the TextBlob algorithm and orange is the VADER algorithm. Therefore, based on these 50 videos, TextBlob is a more accurate algorithm compared to VADER. We think this is because VADER is intended for analyzing social media text and reviews of movies or products [13]. According to the mission statement of TedTalks, their goal is "discovering and spreading ideas that spark conversation, deepen understanding and drive meaningful change" [14]. Thus, the fact that the scores of our sentiment analysis algorithms tend to be more positive is expected with the general sentiment of TedTalks.

3) Text Generation: We initially proposed to evaluate our summaries using automatic summarization or generation metrics such as BLEU or ROUGE score [15] [16]. However, we realized that our goal is not to replicate the exact summary but to retain as much information as possible. Therefore, we modified our metrics for the text generation task and plan to measure the similarity of the summaries generated by different models. We plan to utilize embeddings obtained from our summarization models and measure the similarity between the generated summary and the ground truth by computing the cosine similarity between these two texts.

B. Next Steps

To further refine our predictive models, we will explore several key areas. Alongside efforts to reduce overfitting in our SVR model, we plan to delve into the integration of 'detected emotion' features, derived from TED Talk transcripts. This involves transforming emotional data into numerical formats for SVR model enhancement and carefully selecting and preprocessing these new features to ensure optimal model performance. Additionally, we aim to fine-tune our text summarization models, particularly the BART model, to observe any performance improvements. By incorporating emotional content and refining our models, we anticipate gaining deeper insights into factors driving TED Talks' popularity and enhancing our predictive accuracy.

V. Contribution Table

See each team member's contributions in table V.

TABLE V
Contribution Table

Name	Contributions
Harrison	Prompt Engineering to Extract 'detected emotion' Data from transcripts, requirements.txt file for project dependencies
Michael	Text Summarization Code with corresponding report sections
Christian	Association Rule Mining Code with corresponding report sections
Christine	EDA, Sentiment Analysis and SVR Code with corresponding report sections
Alexandra	Data Processing Code with corresponding report sections, Report editing/formatting, GitHub Page
All	Manually scoring sentiments of videos

References

[1] G. S. di Carlo, "The role of proximity in online popularizations: The case of ted talks," *Discourse Studies*, vol. 16, no. 5, pp. 591–606, 2014. [Online]. Available: https://doi.org/10.1177/1461445614538565

[2] K. MacKrill, C. Silvester, J. W. Pennebaker, and K. J. Petrie, "What makes an idea worth spreading? language markers of popularity in ted talks by academics and other speakers," *Journal of the Association for Information Science and Technology*, vol. 72, no. 8, pp. 1028–1038, 2021. [Online]. Available: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24471

[3] C. R. Sugimoto, M. Thelwall, V. Larivière, A. Tsou, P. Mongeon, and B. Macaluso, "Scientists popularizing science: characteristics and impact of ted talk presenters," *PloS one*, vol. 8, no. 4, 2013. [Online]. Available: https://doi.org/10.1371/journal.pone.0062403

[4] T. Trzciński and P. Rokita, "Predicting popularity of online videos using support vector regression," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2561–2570, 2017.

[5] N. Rathee, N. Joshi, and J. Kaur, "Sentiment analysis using machine learning techniques on python," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2018, pp. 779–785. [Online]. Available: https://doi.org/10.1109/ICCONS.2018.8663224

[6] C. M. Rahman, F. A. Sohel, P. Naushad, and S. M. Kamruzzaman, "Text classification using the concept of association rule of data mining," *CoRR*, vol. abs/1009.4582, 2010. [Online]. Available: http://arxiv.org/abs/1009.4582

[7] R. Obiedat, "Predicting the popularity of online news using classification methods with feature filtering techniques," *Journal of Theoretical and Applied Information Technology*, vol. 98, p. 8, 04 2020. [Online]. Available: https://www.researchgate.net/publication/359214966_PREDICTING_THE_POPULARITY_OF_ONLINE_NEWS_USING_CLASSIFICATION_METHODS_WITH_FEATURE_FILTERING_TECHNIQUES

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *CoRR*, vol. abs/1910.10683, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[9] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019. [Online]. Available: http://arxiv.org/abs/1910.13461

[10] J. Phang, Y. Zhao, and P. J. Liu, "Investigating efficiently extending transformers for long input summarization," *ArXiv*, vol. abs/2208.04347, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:251442728

[11] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165
[12] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking : Bringing order to the web," in *The Web Conference*, 1999. [Online]. Available: https://api.semanticscholar.org/CorpusID:1508503
[13] V. Bonta, N. Kumaresh, and J. Naulegari, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian Journal of Computer Science and Technology*, vol. 8, pp. 1–6, 03 2019.
[14] [Online]. Available: https://www.ted.com/about/our-organization
[15] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. Association for Computational Linguistics, 2002, p. 311–318. [Online]. Available: https://doi.org/10.3115/1073083.1073135
[16] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013

APPENDIX

TABLE VI
EXAMPLE OF TEXT SUMMARIZATION

Ground Truth: Sir Ken Robinson makes an entertaining and profoundly moving case for creating an education system that nurtures (rather than undermines) creativity.
BART → BART: "I have a big interest in education, and I think we all do" There isn't an education system on the planet that teaches dance every day to children. Art and music are normally given a higher status in schools than drama and dance.
SUMY → BART: One is the extraordinary evidence of human creativity in all of the presentations that we've had. If you're not prepared to be wrong, you'll never come up with anything original. You'd think it would be otherwise, but it isn't.
ChatGPT: The speaker highlighted the significance of creativity in education, asserting its equal importance to literacy. He critiqued the conventional hierarchy of subjects, urging a reevaluation of how we nurture diverse forms of intelligence and creativity in children to prepare them for an uncertain future.