



Employ Innovative Algorithms to Assign Weights to YouTube Search Results and Extract Education - Related Data

Jincheng Zhang* and Thada Jantakoon*

Received: May 5, 2024

Revised: December 24, 2024

Accepted: January 3, 2025

* Corresponding Author: Jincheng Zhang, E-mail: zjc1639834588@gmail.com

DOI: 10.14416/j.it.2026.v2.001

Abstract

The rapid advancement of data science, artificial intelligence, and machine learning significantly impacts education, with generative AI technologies like ChatGPT increasingly influencing learners. This study analyzes YouTube video view data to extract educational insights through a weighted approach. We found that videos with over 100,000,000 views received a weight of 30, while those with 10,000 to 99,999 views were assigned a weight of 1. In sentiment analysis using Python, we determined an average sentiment weight of 0.722 for education, indicating a generally positive public perception. Our proposed "Weighted Extraction Weight Addition - TF-IDF" algorithm calculated a TF-IDF value of 0.060 for "education," yielding a final value of 754.16 when weighted with 12,513 views. Furthermore, the "Weighted Extraction Maximum Weight - TF-IDF" algorithm identified "child" as having a maximum weight of 30, highlighting public interest in children's educational content. This analysis aims to enhance educational development by providing valuable information derived from YouTube search results.

Keywords: Innovative Algorithms, Weights to YouTube, Data Extract Education, TF-IDF, Data Science.

1. Background and Significance

The development and application of algorithms to weight YouTube search results and mine educational content are crucial in enhancing the educational utility of online platforms.

The research by Henrique Carlos Fonte Boa Carvalho et al. demonstrates the potential of using text mining and machine learning techniques, specifically the Random Forest classifier, to categorize YouTube videos into educational and non-educational content based on user comments, achieving an accuracy of up to 91.30% [1]. This approach is complemented by Behrouz Minaei-Bidgoli and William F. Punch's work, which shows that combining multiple classifiers and optimizing feature vectors through a Genetic Algorithm can significantly improve the prediction accuracy of educational outcomes [2]. Furthermore, the application of data mining algorithms in mine safety management education, as discussed by Miso Kim, highlights the broader educational implications of algorithmic optimization in specialized fields [3]. The vast catalog of instructional videos on YouTube, as noted by Matthew Fyfield, Michael Henderson, and Michael Phillips, underscores the importance of developing sophisticated algorithms to filter and recommend high-quality educational content [4]. The pedagogical challenges of algorithm literacy, addressed by Hyeon-Seon Jeong, Yeon Ju Oh, and Amie Kim, further emphasize the need for educational programs that unravel the hidden algorithmic processes, thereby enabling learners to critically engage with digital content [5]. The sociability shaping role of algorithms, as discussed by Bernhard Rieder, Ariadna Matamoros-Fernandez, and Òscar Coromina, necessitates empirical approaches to make these algorithms accountable [6]. Sandrine Berges's work on mining massive online financial education resources using the Apriori TIDs algorithm demonstrates the efficiency of

* Faculty of Science and Technology, Rajabhat Maha Sarakham University.

specific algorithms in processing and accurately mining educational content [7]. The study by Michael Knösel, Klaus Jung, and Annalen Bleckmann on the informational value of dentistry videos on YouTube further illustrates the potential of categorizing videos to enhance their educational value [8]. Chenyang Zhao and Junling Wang's personalized recommendation algorithm based on context and trust relationships offers a method to proactively recommend educational videos of interest to users, improving upon traditional collaborative filtering algorithms [9]. Lastly, the research by Lauren Valentino Bryant on the YouTube algorithm's bias towards certain political content underscores the importance of understanding and mitigating unintended consequences of algorithmic recommendations to ensure a balanced and educational online environment [10]. Together, these studies provide a comprehensive overview of the current state and potential of algorithms in weighting YouTube search results and mining educational content, highlighting the importance of accuracy, personalization, and critical engagement in the development and application of these technologies.

YouTube contains a wide range of influential videos related to the topic of education [11]. We employ a Python crawler to acquire video titles and view counts pertaining to education from YouTube search results. Subsequently, we assign significance to the video title information based on the view statistics. Subsequently, we do sentiment analysis on the data and identify popular words using two novel algorithms that we have developed and put forth.

1.1 YouTube Data

YouTube, a highly renowned video platform worldwide [12], hosts a vast collection of educational films. Furthermore, it assumes a significant function within the education sector [13] - [15]. YouTube is commonly utilized for self-directed learning, with a vast array of educational videos available [16] - [18].

1.2 Data Analysis and Mining in Education

In recent years, there has been a significant advancement in data analysis and mining technology [19] - [20]. As a result,

there is a growing trend of utilizing data analysis and mining techniques to gain valuable insights in the field of education [21] - [23]. For instance, individuals employ data analysis and mining techniques to forecast student performance, student attrition rates, student requirements, and the impact of courses, among other factors [24] - [28].

1.3 Research Purpose and Significance

This study employs YouTube video view statistics to give significance to video title data. The objective is to optimize data precision in order to enhance the quality of analytical conclusions. This paper presents the development and proposal of two data analysis algorithms aimed at identifying prevalent words and phrases. The algorithms are referred to as "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF". These two innovative algorithms have the potential to be widely employed in data analysis and mining.

1.4 Research Framework, Structure and Research Beneficiaries

The scope of this research mainly includes the following aspects:

We use a Python crawler to obtain education-related video titles and views data from YouTube search results.

Then we weight the video title data based on the number of views.

Perform sentiment analysis on the data.

Use the two new algorithms we invented and proposed "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF" to find popular words.

Utilize the YouTube video view data to assign importance to the video title data, and thereafter employ it for the purposes of data analysis and mining. We have developed and put forward two novel algorithms, namely "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF". The target audience for this study comprises professionals from various fields who require data analysis and mining, such as educators, executives, entrepreneurs, and researchers, among others.

2. Related Theories and Research

Extensive study has been conducted on data analysis and mining prior to the year 18. Data analysis and mining have emerged as prominent study areas in recent years. This field is undergoing rapid development and possesses substantial research potential. Data is highly valuable in the current period [29]. Large corporations dominate a significant quantity of data [30] - [31]. The data accessible to the general public is somewhat restricted, nevertheless we may still extract numerous valuable insights from this limited data and apply them to practical situations. The focus of this research is to utilize innovative methods for conducting weighted analysis of YouTube search results and extracting education-related data using Python crawlers. By leveraging publicly available data, we conduct in-depth analysis and extraction of valuable insights and knowledge pertaining to education. This, in turn, assists individuals in enhancing their educational practices. For instance, individuals can examine fluctuations in emotions by utilizing sentiment analysis outcomes across various time intervals. Additionally, by identifying often used terms, they can gain insights into popular subjects and educational content.

2.1 Theoretical Framework

YouTube is a repository of great educational data. For this research, we employed a Python crawler to acquire the titles and view counts of education-related videos in YouTube search results. The Python web scraper is capable of efficiently and expeditiously retrieving the desired data. However, it may encounter challenges when faced with complex jobs or anti-crawling mechanisms. Collecting data from more sophisticated websites can provide challenges when utilizing Python crawlers. Once we collected the video titles and view data relevant to education from YouTube search results, we assigned a weight to the video title data based on the view data. Weighted data enhances the accuracy of analytical results by including multiple perspectives. Videos of varying nature have distinct effects. Next, we apply sentiment analysis to the data, utilizing machine learning technologies to enhance the accuracy of the sentiment analysis process. And discover

commonly used words using two novel algorithms that we have developed and suggested. The two novel methods are "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF". These two novel methods are derived from the TF-IDF method.

2.2 Key Concepts and Definitions

Data analysis and mining is a technological process that retrieves valuable insights and knowledge from data [32]. It has extensive applications across multiple industries. Automated web crawlers have the ability to gather pertinent data from the Internet [33]. For instance, different search engines employ web crawlers to systematically browse the web pages of diverse websites. Sentiment analysis is capable of analyzing the emotions expressed in textual data [34]. TF-IDF is a method that calculates the significance of words in a document by multiplying the inverse document frequency with the word frequency [35].

2.3 Summarize

The integration of data science, artificial intelligence, machine learning, and other technologies has become extensive across numerous industries [36] - [37]. The realm of education is not exempt from this. Data science, artificial intelligence, machine learning, and other technologies are extensively applied in the subject of education [38]. With the rapid advancement of generative AI, particularly demonstrated by ChatGPT in recent years, individuals are increasingly dependent on these technologies [39]. This study utilized an innovative method to do weighted analysis and extract education-related data from YouTube search results using a Python crawler. By analyzing and mining valuable insights and knowledge pertaining to education, this study aims to assist individuals in their educational endeavors. For instance, individuals can examine fluctuations in emotions by utilizing sentiment analysis outcomes across various time intervals. Additionally, by identifying often used terms, they can gain insights into popular subjects and educational content. Utilize the YouTube video view data to assign importance to the video title data, and thereafter employ it for the purposes of data analysis



and mining. The two novel algorithms, "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF," that we have devised and suggested, possess significant promise for use in the realm of data science.

3. Research Methods

3.1 System Design and Implementation

We design and develop a complete system: using novel algorithms to perform weighted analysis of YouTube search results and mine education-related data through Python crawlers. This complete system implements a complete process from obtaining data to displaying results.

3.2 Algorithm Design and Optimization

We designed several new algorithms, including the following:

We employ YouTube video view statistics to assign weight to the video title data, which we subsequently utilize for data analysis and mining. Weighted data can yield more precise analysis findings by taking into account the varying influences of different movies represented by different views. We have developed and suggested two algorithms, namely "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF", for determining the significance of various words or phrases. These algorithms have a broad spectrum of applications, benefiting practitioners in various fields that require data analysis and mining, such as instructors, managers, business professionals, and researchers. We devised and successfully executed these algorithms.

3.3 Applications

We utilize the comprehensive system that we have developed, along with many novel algorithms, to effectively address practical issues. This system serves as a prime illustration in the realm of education. Conduct a comprehensive study of YouTube search results by assigning weights to different factors and extract education-related data using a Python web crawler. Conduct a thorough analysis and uncover valuable insights and knowledge pertaining to education in order to assist individuals in improving their educational practices. For instance, individuals can examine fluctuations in emotions by utilizing

sentiment analysis outcomes across various time intervals. Additionally, by identifying frequently used terms, they can gain insights into popular subjects and educational content.

3.4 Constructed and Annotated a Complete Dataset

We generated a comprehensive dataset using a thorough CRISP-DM mining procedure to guarantee the trustworthiness and dependability of our research.

4. Collect, Analyze and Mine Data

4.1 Collect Data

For the data collection part, we utilize a comprehensive list of 10,000 English words available from MIT online (<https://www.mit.edu/~ecprice/wordlist.10000>). These 10,000 English words are primarily everyday words. I have successfully saved these 10,000 English words into a file named "10,000 words provided by MIT.csv".

Here are the first 10 words from the file "10,000 words provided by MIT.csv":

Table 1. The first 10 words in "10,000 words provided by MIT.csv".

Word
a
aa
aaa
aaron
ab
abandoned
abc
aberdeen
abilities
ability

Next, we utilize the Python code below to generate the necessary search terms for data collection on YouTube (In this article, pseudocode will be used throughout.):

Start

Read CSV file into data

Define replace_the(row):



Get the first word

Return "intitle:\"education\" \" + first word

Apply replace_the to 'Word' column, store results in 'Search term' column

Save data to new CSV file

End

The first 10 rows of data in the "Generate search terms.csv" file generated by the code are as follows:

Table 2. The first 10 rows of data in the "Generate search terms.csv" file.

Word	Search term
a	intitle:"education" a
aa	intitle:"education" aa
aaa	intitle:"education" aaa
aaron	intitle:"education" aaron
ab	intitle:"education" ab
abandoned	intitle:"education" abandoned
abc	intitle:"education" abc
aberdeen	intitle:"education" aberdeen
abilities	intitle:"education" abilities
ability	intitle:"education" ability

Consider the phrase "intitle:education" as an example. When you perform a search on YouTube using the "intitle:education" a" query, you'll notice that the term "education" appears in the title of the video as well as in the video itself. A video featuring the word "a" in the description text.

Next, we utilize the Python crawler code provided to gather information on video titles and video views from YouTube.

Start

Define create_webdriver()

Set output CSV path

Open CSV and write header

Read search queries from input CSV

For each search_query:

Set retry_count to 0

While retry_count < 3:

Try:

Create WebDriver

Open YouTube

Search for search_query

Wait for results

Get top 3 video titles and second group data

Write results to CSV

Close WebDriver

Break

Catch Timeout:

Wait 2 minutes

Increment retry_count

Catch Exception:

Close WebDriver

Wait 5 seconds

Increment retry_count

End

This Python crawler obtains YouTube video title and video view data. Part of the content in the "Output YouTube data.csv" file output is as follows:

Afterwards, we utilized Python code along with manual filtering to effectively sift through the data. For instance, utilize regular expressions to extract data, eliminate duplicate data, remove data that doesn't meet the criteria, delete outliers, and utilize video playback data to generate weight data for the corresponding video title, among other tasks.

We utilize a combination of coding and manual filtering techniques to efficiently store the filtered data in the designated file, "Data weighted for data mining and analysis.csv".

Here are the first 10 rows of data:

	A	B	C	D	E	F
1	Search Query	Top 3 Video Titles	Top 3 Second Group Data			
2	intitle:"education" a	The Future of Education: A S The Future of Education: A Student's Perspective				
3	intitle:"education" aa	Kisah Teladan Abdurrahman Kisah Teladan Abdurrahman Bin Auf Dan Abu				
4	intitle:"education" aaa	ABC Song Learn ABC Alph ABC Song Learn ABC Alphabet for Children				
5	intitle:"education" aaron	Why e-learning is killing edu Why e-learning is killing education Aaron Bart				
6	intitle:"education" ab	Education, ABC Song Learn Education 来自A.B. Crentsil - Topic 1,632次观看				
7	intitle:"education" abandoned	Abandoned School - A Time (Abandoned School) - A Time Capsule of Educatio				
8	intitle:"education" abc	ABC Song Learn ABC Alph ABC Song Learn ABC Alphabet for Children				
9	intitle:"education" aberdeen	Education at Aberdeen Init Education at Aberdeen Initial Teacher Educat				
10	intitle:"education" abilities	The Power Of Education (Re The Power Of Education (Real Life Story) 来自I				
11	intitle:"education" ability	The Science of Teaching, Effc The Science of Teaching, Effective Education, ar				
12	intitle:"education" able	Railway Station Related Wor Railway Station Related Word रेलवे स्टेशन				
13	intitle:"education" aboriginal	Aboriginal kids embracing th Aboriginal kids embracing their education from				
14	intitle:"education" abortion	Maeve's Abortion Story Sex Maeve's Abortion Story Sex Education Netfli				
15	intitle:"education" about	Speech on Importance of Ed: Speech on Importance of Education 2nd Short				
16	intitle:"education" above	Why More Education Is Not ,Why More Education Is Not Always Better relat				
17	intitle:"education" abraham	Encouragement for Keeping Encouragement for Keeping Secrets Ibn Hibb				
18	intitle:"education" abroad	How to study in the US for FI How to study in the US for FREE Education				
19	intitle:"education" abs	Algorithme Darjaja #EP 10 : I Algorithme Darjaja #EP 10 : Déclaration des vari				
20	intitle:"education" absence	Danish Ministry of Educator Danish Ministry of Education: Absence 来自W				
21	intitle:"education" absent	Missing Pupils: Why are so m Missing Pupils: Why are so many children with s				
22	intitle:"education" absolute	The absolute state of Americ Th The absolute state of American Education 来自I				
23	intitle:"education" absolutely	Lets Talk About Education Lets Talk About Education Jayne Ellsperrman				
24	intitle:"education" absorption	EDUCATION BY ABSORPT EDUCATION BY ABSORPTION 来自Moira 21				
25	intitle:"education" abstract	OnLine Education Abstract, OnLine Education Abstract 来自◆◆◆◆◆◆◆◆◆◆				
26	intitle:"education" abstracts	Education Abstracts simple s Education Abstracts simple search default 来自				
27	intitle:"education" abu	Is Secular Education More In Is Secular Education More Important?? - Sheikh				
28	intitle:"education" abuse	Drug Education Video, Legal Drug Education Video 来自Cormack Abbott 139				
29	intitle:"education" ada	IDEA - Individuals with Disab IDEA - Individuals with Disabilities Education A				

Figure 1. Screenshot of part of the content in the "Output
YouTube data.csv" file output by the Python
crawler code.

Here is the method we use to generate weight data for video titles:

The "View Count" column is a number from 10000 to 99999, then the weight data of the video title is 1.

The "View Count" column is a number from 100000 to 999999. Then the weight data of the video title is 5.

The "View Count" column is a number from 1000000 to 9999999, then the weight data of the video title is 10.

The "View Count" column is a number from 10000000 to 99999999, then the weight data of the video title is 20.

If the "View Count" column is 100000000 and above, then the weight data of the video title is 30.

4.2 Sentiment Analysis

We utilize the Python code provided in [30] to conduct sentiment analysis. The results of the sentiment analysis are then given weight based on the data from our "Weighted Weight" column.

Start

Read CSV file into data

Drop rows with NaN values

Table 3. The first 10 rows of data in the "Data weighted for data mining and analysis.csv" file.

Video Title	View Count	Weighted Weight
ABC Song Learn ABC Alphabet for Children Education ABC Nursery Rhymes	394138388	30
Fun Learning Names of Fruit and Vegetables Wooden Toys Cutting Fruit Education videos Fun for Kids	183603911	30
Learn Colors With Balls Colours For Kids And Children Learning & Education For Toddlers & Babies	159394061	30
Wild Zoo Animals Toys Baby Mom Learn Animals Names Education Toys for Kids	120478517	30
Safety education Poli theater Don't go anywhere dangerous.	97960316	20
BALDI'S BASICS In Real Life!! FGTEEV goes to School of Education & Learning (Skit)	71020729	20
BALDI's BASICS of FNAF EDUCATION & CLONING MOD + I'M BALDI vs. Principal (FGTEEV Cheat Escape #3)	59494934	20
Learn ABC's with Morphle - Alphabet letters education for kids	51421438	20
BALDI'S BASICS NO EXIT ESCAPE + PRANK & MOD! FGTEEV RAGE Teacher Ending! (Education & Learning #2)	49183913	20
Male reproductive organs #education	44738605	20

Load BERT model and tokenizer

Set device to GPU if available

Move model to device

Define analyze_sentiment(text):

Tokenize text

Predict using BERT model

Get predictions

Determine sentiment (Positive/Negative)

Return sentiment and confidence

Initialize total weighted sentiment and total weight to 0



For each row in data:	Print success message								
Get video title and weighted weight									
Perform sentiment analysis on title	End								
Calculate weighted sentiment	Start								
Print title, sentiment, confidence, weighted weight, and weighted sentiment	Read TF_IDF_output.csv into tfidf_df								
Accumulate total_weighted_sentiment and total_weight	Read Word_Weights_Sorted.csv into weight_df								
If total_weight != 0:	Convert 'Word' column to lowercase in both DataFrames								
Calculate and print weighted average sentiment	Find common words in 'Word' columns of both DataFrames								
Else:									
Print no valid results message	Filter rows with common words in both DataFrames								
End									
The code produces a result that falls within the neutral to slightly positive sentiment range, demonstrating people's recognition of the importance of education. The specific result is "Weighted Average Sentiment: 0.7220842402274824."	Merge tfidf_common and weight_common on 'Word'								
	Calculate new column: Multiply = TF-IDF * Weight								
	Sort merged DataFrame by 'Multiply' in descending order								
	Save sorted DataFrame to Multiplication_Output_Sorted.csv								
	End								
	Initially, we compute the "TF-IDF" score for each word. Next, we compute the "weighted extraction weight addition" score for each word. Finally, we multiply the "TF-IDF" score and the "weighted extraction weight addition" score for each word to obtain the "weighted extraction weight addition-TF-IDF" score.								
	The top 20 "TF-IDF" values we calculated are as follows:								
	Table 4. The top 20 content with the largest "TF-IDF" value we calculated.								
	<table> <tr> <th>Word</th><th>TF-IDF</th></tr> <tr> <td>education</td><td>0.060270366</td></tr> <tr> <td>the</td><td>0.02938045</td></tr> <tr> <td>of</td><td>0.025701289</td></tr> </table>	Word	TF-IDF	education	0.060270366	the	0.02938045	of	0.025701289
Word	TF-IDF								
education	0.060270366								
the	0.02938045								
of	0.025701289								
Read CSV file									
Replace NaN in 'Video Title' with "									
Get video titles									
Initialize TF-IDF vectorizer									
Calculate TF-IDF matrix									
Get words and mean TF-IDF values									
Create and sort DataFrame by TF-IDF									
Save DataFrame to CSV									



Table 4. (continued).

Word	TF-IDF
in	0.024034098
for	0.020932313
and	0.018791291
to	0.018434961
kids	0.01746805
sex	0.014137011
children	0.013942088
is	0.013930863
how	0.013325934
science	0.01293237
aumsum	0.012458118
what	0.012026109
on	0.01051055
with	0.00897952
why	0.008971687
system	0.008744217
physical	0.008597416

The top 20 "weighted extraction weight addition" values we calculated are as follows:

Table 5. The top 20 content with the largest "weighted extraction weight addition" value we calculated.

Word	Weight
education	12513
	10232
-	4065
the	3163
#education	2953
for	2518

Table 5. (continued).

Word	Weight
of	2410
in	2320
#kids	2107
#aumsum	2101
#science	2018
and	1946
#children	1925
to	1692
videos	1339
a	1338
sex	1317
kids	1231
+	1209
more	1157

The top 20 "weighted extraction weight addition-TF-IDF" values we calculated are as follows:

Table 6. The top 20 content with the largest "weighted extraction weight addition-TF-IDF" value we calculated.

Word	TF-IDF	Weight	Multiply
education	0.060270366	12513	754.163091
the	0.02938045	3163	92.93036371
of	0.025701289	2410	61.9401059
in	0.024034098	2320	55.759107
for	0.020932313	2518	52.70756497
and	0.018791291	1946	36.56785251
to	0.018434961	1692	31.19195413
kids	0.01746805	1231	21.50316938



Table 6. (continued).

Word	TF-IDF	Weight	Multiply
sex	0.014137011	1317	18.61844315
is	0.013930863	1026	14.29306509
how	0.013325934	1007	13.41921592
what	0.012026109	985	11.84571695
videos	0.008298883	1339	11.11220493
more	0.007785355	1157	9.007656219
on	0.01051055	790	8.303334725
why	0.008971687	795	7.132490938
with	0.00897952	750	6.734639994
learning	0.007015413	782	5.486052617
system	0.008744217	597	5.220297541
physical	0.008597416	515	4.427669361

From the 8th data "kids" among the top 20 contents with the largest "weighted extraction weight addition-TF-IDF" value, we can predict that people are very concerned about education about kids.

We search for "math" and "music" respectively in the "Multiplication_Output_Sorted.csv" that calculates the "weighted extraction weight addition-TF-IDF" value output, and rank them from high to low according to the "Weight" value, where "math" is at 188 name, "music" is at number 66. From this we can predict that in terms of education, compared with "math" and "music", people may pay more attention to content related to "music" (this may be related to the fact that many children learn by listening to music on YouTube).

4.4 Weighted Extraction Maximum Weight-TF-IDF

We use the "weighted extraction maximum weight-TF-IDF" algorithm we invented and proposed to find popular words. We use the three data "weighted extraction maximum weight", "TF" and "IDF" to multiply to calculate "weighted extraction maximum weight-TF-IDF". The code used is as follows:

Start

Read TF_IDF_output.csv into tfidf_df

Read Word_Weights_Sorted 02.csv into weight_df

Convert 'Word' column to lowercase in both DataFrames

Find common words in 'Word' columns of both DataFrames

Filter rows with common words in both DataFrames

Merge tfidf_common and weight_common on 'Word'

Calculate new column: Multiply = TF-IDF * Weight

Sort merged DataFrame by 'Multiply' in descending order

Save sorted DataFrame to Multiplication_Output_Sorted 02.csv

End

Initially, we compute the "TF-IDF" score for each word. Next, we compute the "weighted extraction maximum weight" score for each word. Finally, we multiply the "TF-IDF" score and the "weighted extraction maximum weight" score for each word to obtain the "weighted extraction maximum weight-TF-IDF" score.

The following list presents the top 20 contents based on their "weighted extraction maximum weight" values that we have calculated:

Table 7. The top 20 content with the largest "weighted extraction maximum weight" value we calculated.

Word	Weight
abc	30
toddlers	30
videos	30
kids	30

Table 7. (continued).

Word	Weight
colors	30
with	30
balls	30
colours	30
babies	30
toys	30
wild	30
zoo	30
animals	30
baby	30
mom	30
song	30
cutting	30
&	30
wooden	30
rhymes	30

The top 25 contents with the largest "weighted extraction maximum weight-TF-IDF" values we calculated are as follows:

Table 8. The top 25 content with the largest "weighted extraction maximum weight-TF-IDF" value we calculated.

Word	TF-IDF	Weight	Multiply
education	0.060270366	30	1.808110983
of	0.025701289	30	0.771038663
for	0.020932313	30	0.6279694
the	0.02938045	20	0.587609002
and	0.018791291	30	0.563738734
kids	0.01746805	30	0.524041496

Table 8. (continued).

Word	TF-IDF	Weight	Multiply
in	0.024034098	20	0.480681957
children	0.013942088	30	0.418262644
to	0.018434961	20	0.368699221
sex	0.014137011	20	0.282740215
with	0.00897952	30	0.2693856
videos	0.008298883	30	0.248966503
what	0.012026109	20	0.240522172
learning	0.007015413	30	0.210462377
on	0.01051055	20	0.210211006
system	0.008744217	20	0.17488434
physical	0.008597416	20	0.171948325
more	0.007785355	20	0.155707108
school	0.007690336	20	0.15380671
video	0.007142199	20	0.142843971
is	0.013930863	10	0.139308627
how	0.013325934	10	0.133259344
science	0.01293237	10	0.1293237
song	0.004257425	30	0.127722759
an	0.006373116	20	0.127462315

From the 23th data "science" among the top 25 contents with the largest "weighted extraction maximum weight-TF-IDF" value, we can predict that people are very concerned about educational content about science.

We search for "computer" and "electrical" respectively in the "Multiplication_Output_Sorted 02.csv" that calculates the "weighted extraction maximum weight-TF-IDF" value output, and rank them from high to low according to the "Weight" value, where "computer" is located 291nd, "electrical" is at 1241th. From this we can predict that in education, compared with "computer" and "electrical", people may pay more attention to content related to "computer".

5. CRISP-DM Data Mining Process

In this study we used the complete process of CRISP-DM for data collection, analysis and mining [40] - [41]. The complete CRISP-DM process specifically includes the following six stages:

5.1 Business Understanding

For this study, it is necessary to do a weighted analysis on YouTube search results and extract education-related data. We require access to educational video titles and view statistics from YouTube search results for the purpose of data analysis and mining. We utilize the view data. Analyze the data related to the video titles based on their weight.

5.2 Data Understanding

We need education-related video title and view data from YouTube search results, and then we weight the video title data by the view data. We collect data on YouTube using search terms like "intitle:"education" a". If you search for "intitle:"education" a" on YouTube, you will find YouTube videos with the word "education" in their titles, and have the word "a" in the video's title or the video's description text.

5.3 Data Preparation

We wrote Python to generate search terms, and used Python crawlers to quickly and efficiently obtain education-related video titles and views data from YouTube search results. We then filtered the data using Python code combined with manual screening. For example, use regular expressions to extract data, delete duplicate data, delete data that does not meet the conditions, delete outliers, use video playback data to generate the weight data of the corresponding video title, etc.

5.4 Modeling

In the modeling stage, we perform corresponding calculations, matching and filtering on the data in "Obtained data first 20 lines.csv" and the data on search results on YouTube. Then use the data on Google Keyword Planner to analyze and filter, and then get the final results. We use code combined with manual filtering to save the filtered data into "Data weighted for data mining and analysis.csv".

5.5 Evaluation

We performed sentiment analysis using the data in

"Data weighted for data mining and analysis.csv" and used two algorithms we invented and proposed to calculate the importance of different words or phrases "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF" calculate popular words. We evaluated the obtained experimental results and concluded that our experimental results are valid.

5.6 Deployment

The output result of the sentiment analysis code is "Weighted Average Sentiment: 0.7220842402274824". 0.7220842402274824 is in the neutral to slightly positive sentiment range. It shows that people currently recognize the role of education. And we also explained the popular words calculated by "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF" accordingly.

6. Results and Analysis

We use this method to weight the YouTube video name data:

The "View Count" column is a number from 10000 to 99999, then the weight data of the video title is 1.

The "View Count" column is a number from 100000 to 999999. Then the weight data of the video title is 5.

The "View Count" column is a number from 1000000 to 9999999, then the weight data of the video title is 10.

The "View Count" column is a number from 10000000 to 99999999, then the weight data of the video title is 20.

If the "View Count" column is 100000000 and above, then the weight data of the video title is 30.

Because different views represent different influences of different videos, more accurate analysis results can be obtained through such weighted data.

We use the BERT model to conduct sentiment analysis on the YouTube video name data weighted by playback volume. The results of the sentiment analysis are positive, indicating that people have confidence in education.

We use our invented and proposed "weighted extraction weight addition-TF-IDF" algorithm to find popular words.

We use the "weighted extraction weight addition" value, the "TF" value, and the "IDF" value to multiply the three data to calculate the "weighted extraction weight addition-TF-IDF" value. We first calculate the "TF-IDF" value of each word, then calculate the "weighted extraction weight addition" value of each word, and then combine the "TF-IDF" value and "weighted extraction weight addition" value of each word Multiply to calculate the "weighted extraction weight addition-TF-IDF" value of the corresponding word.

The calculation method of "weighted extraction weight addition" is as follows. Calculate the "weighted extraction weight addition" values of A B C respectively:

Table 9. Data example for calculating "weighted extraction weight addition".

Video Title	View Count	Weighted Weight
A B	100000001	30
B C	10000001	20
A C	1000001	10

The "weighted extraction weight addition" value of A is:
 $30+10=40$.

The "weighted extraction weight addition" value of B is:
 $30+20=50$.

The "weighted extraction weight addition" value of C is:
 $20+10=30$.

We use the "weighted extraction maximum weight-TF-IDF" algorithm we invented and proposed to find popular words. We use the "weighted extraction maximum weight" value, the "TF" value, and the "IDF" value to multiply the three data to calculate the "weighted extraction maximum weight-TF-IDF" value. We first calculate the "TF-IDF" value of each word, then calculate the "weighted extraction maximum weight" value of each word, and then combine the "TF-IDF" value and the "weighted extraction maximum weight" value of each word Multiply to calculate the "weighted extraction maximum weight-TF-IDF" value of the corresponding word.

The calculation method of "weighted extraction maximum

weight" is as follows. Calculate the "weighted extraction maximum weight" values of A B C respectively:

Table 10. Data example for calculating "weighted extraction maximum weight".

Video Title	View Count	Weighted Weight
A B	10000001	20
B C	1000001	10
A C	100001	5

The "weighted extraction maximum weight" value of A is:
 $20 (20>5)$.

The "weighted extraction maximum weight" value of B is:
 $20 (20>10)$.

The "weighted extraction maximum weight" value of C is:
 $10 (10>5)$.

We calculated the popular words using "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF" and explained the output results accordingly.

7. Conclusions and Future Directions

Extensive research has been conducted on the subjects of data analysis and mining in the past. Data analysis and mining have emerged as prominent areas of research and study in recent years. YouTube is a repository of abundant and valuable educational data. This study utilizes YouTube video view data to assign weight to video title data. To enhance the precision of the data and hence improve the accuracy of the analysis outcomes. We have developed and put forward two methods for data analysis that aim to identify common words and phrases. These algorithms are known as "weighted extraction weight addition-TF-IDF" and "weighted extraction maximum weight-TF-IDF". We employ innovative algorithms to utilize Python crawlers. Conduct a comprehensive study of YouTube search results by assigning weights to different factors and extract data relevant to schooling. These algorithms have a broad spectrum of applications, benefiting practitioners in various fields that require data analysis and mining, such as



instructors, managers, business professionals, and academics in related fields. We have devised and successfully executed these algorithms.

By leveraging publicly available data, we conduct analysis and extraction of valuable insights and knowledge pertaining to education, hence assisting individuals in enhancing their educational practices. For instance, individuals can examine fluctuations in emotions by utilizing sentiment analysis outcomes across various time intervals. Additionally, by identifying often used terms, they can gain insights into popular subjects and educational content. In future study, our devised methods and algorithms can be applied to diverse industries including cinema and television, music, art, sports, medicine, and other domains.

The weighted model proposed in this study shows significant advantages in multi-level data analysis, especially when dealing with complex data sets. Compared with existing technologies, our model can automatically adjust weights according to data characteristics by introducing a dynamic weighting mechanism, thereby capturing important information more accurately. In addition, the significant improvement in accuracy of the model makes it more robust and scalable in practical applications.

8. References

- [1] H. C. F. B. Carvalho, F. A. Dorça, C. G. Pitangui, L. P. de Assis, A. V. Andrade, and E. A. C. Trindade. "Classificação automática de vídeos educacionais por meio de comentários apoiada por técnicas de Aprendizado de Máquina: uma análise experimental utilizando o Youtube." *Revista Brasileira de Informática na Educação*, Vol. 30, pp. 419-448, September, 2022.
- [2] B. M. Bidgoli and W. F. Punch. "Using genetic algorithms for data mining optimization in an educational web-based system." *Genetic and Evolutionary Computation Conference (GECCO2003)*, Berlin, Heidelberg, pp. 2252-2263, June, 2003.
- [3] Y. He. "Research on Mining Algorithm and Its Application in Mine Safety Management Education." *EAI International Conference, BigIoT-EDU*, Cham, pp. 121-131, July, 2022.
- [4] M. Fyfield, M. Henderson, and M. Phillips. "Navigating four billion videos: teacher search strategies and the YouTube algorithm." *Learning, Media and Technology*, Vol. 46, No. 1, pp. 47-59, January, 2021.
- [5] H. S. Jeong, Y. J. Oh, and A. Kim. "Critical algorithm literacy education in the age of digital platforms: Teaching children to understand YouTube recommendation algorithms." *In Learning to Live with Datafication*, pp. 153-168, Routledge, London, 2022.
- [6] B. Rieder, A. M. Fernández, and Ò. Coromina. "From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results." *Convergence*, Vol. 24, No. 1, pp. 50-68, 2018.
- [7] Y. Luo. "Mining Algorithm of Massive Online Financial Education Resources Based on Apriori TIDS." *International Conference on E-Learning, E-Education, and Online Training*, pp. 293-304, Cham: Springer Nature Switzerland, July, 2022.
- [8] M. Knösel, K. Jung, and A. Bleckmann. "YouTube, Dentistry, and Dental Education." *Journal of Dental Education*, Vol. 75, No. 12, pp. 1558-1568, December, 2011.
- [9] C. Zhao and J. Wang. "Network Education Video Recommendation Algorithm Based on Context and Trust Relationship." *2013 IEEE 4th International Conference on Software Engineering and Service Science*, pp. 537-540, May, 2013.
- [10] L. V. Bryant. "The YouTube Algorithm and the Alt-Right Filter Bubble." *Open Information Science*, Vol. 4, No. 1, pp. 85-90, January, 2020.
- [11] J. B. Bravo, J. Payumo, and B. Pittendrigh. "Measuring the impact and reach of informal educational videos on YouTube: The case of Scientific Animations Without Borders." *Heliyon*, Vol. 7, No. 12, 2021.

- [12] F. Pires, M. J. Masanet, and C. A. Scolari. "What are teens doing with YouTube? Practices, uses and metaphors of the most popular audio-visual platform." *Information, Communication and Society*, Vol. 24, No. 9, pp. 1175-1191, 2021.
- [13] R. Rahmatika, M. Yusuf, and L. Agung. "The effectiveness of YouTube as an online learning media." *Journal of Education Technology*, Vol. 5, No. 1, pp. 152-158, 2021.
- [14] E. T. Maziriri, P. Gapa, and T. Chuchu. "Student perceptions towards the use of YouTube as an educational tool for learning and tutorials." *International Journal of Instruction*, Vol. 13, No. 2, pp. 119-138, 2020.
- [15] A. Alobaid. "Smart multimedia learning of ICT: role and impact on language learners' writing fluency- YouTube online English learning resources as an example." *Smart Learning Environments*, Vol. 7, No. 1, pp. 24, September, 2020.
- [16] I. Dubovi and I. Tabak. "An empirical analysis of knowledge co-construction in YouTube comments." *Computers & Education*, Vol. 156, pp. 103939, October, 2020.
- [17] S. Kohler and T. C. Dietrich. "Potentials and limitations of educational videos on YouTube for science communication." *Frontiers in Communication*, Vol. 6, May, 2021.
- [18] M. M. Neumann and C. Herodotou. "Evaluating YouTube videos for young children." *Education and Information Technologies*, Vol. 25, No. 5, pp. 4459-4475, 2020.
- [19] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi. "Text mining in big data analytics." *Big Data and Cognitive Computing*, Vol. 4, No. 1, pp. 1-34, January 2020.
- [20] V. Grover, R. H. Chiang, T. P. Liang, and D. Zhang. "Creating strategic business value from big data analytics: A research framework." *Journal of Management Information Systems*, Vol. 35, No. 2, pp. 388 - 423, April, 2018.
- [21] H. Luan, et al. "Challenges and future directions of big data and artificial intelligence in education." *Frontiers in Psychology*, Vol. 11, Article 580820, October, 2020.
- [22] C. Fischer, et al. "Mining big data in education: Affordances and challenges." *Review of Research in Education*, Vol. 44, No. 1, pp. 130-160, March, 2020.
- [23] Z. Alamgir, H. Akram, S. Karim, and A. Wali. "Enhancing Student Performance Prediction via Educational Data Mining on Academic Data." *Informatics in Education*, Vol. 23, No. 1, pp. 1-24, March, 2024.
- [24] S. Sucipto, D. D. Prasetya, and T. Widiyaningtyas. "Educational Data Mining: Multiple Choice Question Classification in Vocational School." *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, Vol. 23, No. 2, pp. 379-388, March, 2024.
- [25] M. Salari, R. Radfar, and M. Faghihi. "Predicting students' performance using machine learning algorithms and educational data mining (a case study of Shahed University)." *Business Intelligence Management Studies*, Vol. 12, No. 47, pp. 315-366, March, 2024.
- [26] H. Zeineddine, U. Braendle, and A. Farah. "Enhancing prediction of student success: Automated machine learning approach." *Computers and Electrical Engineering*, Vol. 89, January, 2021.
- [27] A. Khan and S. K. Ghosh. "Student performance analysis and prediction in classroom learning: A review of educational data mining studies." *Education and Information Technologies*, Vol. 26, No. 1, pp. 205-240, January, 2021.
- [28] Q. Wang and W. Zhang. "Research on the Application of Data Mining in Educational Big Data." *Advances in Educational Technology and Psychology*, Vol. 8, No. 4, pp. 164-169, August, 2024.
- [29] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2006.



- [30] C. Rikap and B.-Å. Lundvall. "Big tech, knowledge predation and the implications for development." *Innovation and Development*, Vol. 12, No. 3, pp. 389-416, September, 2022.
- [31] K. Pistor. "Rule by data: The end of markets?." *Law and Contemporary Problems*, Vol. 83, pp. 101-124, 2020.
- [32] I. H. Sarker. "Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective." *SN Computer Science*, Vol. 2, No. 5, pp. 377, September, 2021.
- [33] S. Brin and L. Page. "The anatomy of a large-scale hypertextual Web search engine." *Computer Networks and ISDN Systems*, Vol. 30, pp. 107-117, April, 1998.
- [34] B. Pang and L. Lee. "Opinion mining and sentiment analysis." *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1 - 2, pp. 1-135, July, 2008.
- [35] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [36] S. Raschka, J. Patterson, and C. Nolet. "Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence." *Information*, Vol. 11, No. 4, pp. 193, April, 2020.
- [37] G. Siemens and P. Long. "Penetrating the fog: Analytics in learning and education." *Educause Review*, Vol. 46, No. 5, pp. 30-32, 2011.
- [38] F. F.-H. Nah, et al. "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration." *Journal of Information Technology Case and Application Research*, Vol. 25, No. 3, pp. 277-304, July, 2023.
- [39] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *North American Chapter of the Association for Computational Linguistics*, 2019.
- [40] C. Shearer. "The CRISP-DM model: the new blueprint for data mining." *Journal of Data Warehousing*, Vol. 5, No. 4, pp. 13-22, 2000.
- [41] P. Chapman, et al. "CRISP-DM 1.0: Step-by-step data mining guide." *SPSS White Paper*, pp. 1-8, 2000.