Learning Through Media: Classifying TV & Movie Clips by Academic Subject Area

ClassHook, Inc.

15.572 Analytics Lab: Action Learning Seminar on Analytics, Machine Learning, and the Digital Economy

Fiona Aga
Vojtech Machytka
Aditi Singh
Benjamin Rio

1. Executive Summary

ClassHook is an EdTech startup that allows educators to enhance the classroom experience by using educational TV and movie clips. ClassHook's business model heavily relies on scaling its content curation to address the increasing demand from teachers. Currently, the startup operates through manual search and annotation for clips, which is time-consuming and expensive. Our project directly addresses this bottleneck by employing advanced natural language processing (NLP) techniques to automate the classification of videos based on subtitles.

Our methodology involved investigating three approaches and comparing their performance:

- Baseline models: We developed standard statistical methods to get embeddings from the subtitles, such as Term Frequency-Inverse Document Frequency (TF-IDF). Using classifier learners, such as Multinomial Logistic Regression, allowed us to establish performance baselines. The performance improvement of this method primarily relied on cleaning and preprocessing of the subtitles.
- Zero-Shot Classification: We used pre-trained Transformer models without any further training to use them as Zero-Shot Classifiers. Such a method leverages the understanding of semantics due to attention-based architecture and the extensive (and expensive) pre-training phase. This method achieved similar results to TF-IDF, indicating a high potential for further use of LLMs with training.
- Fine-Tuning LLMs: This was our final model where we fine-tuned the pre-trained Transformer models on ClassHook's data to greatly improve the performance. This step required optimization, accurate data manipulation, and the leveraging of GPUs for training. Our model excels in speed and efficiency, operable via Google Colab without additional installations.

To deliver the greatest impact, we developed an easy-to-use interface, named Hookie, tailored for ClassHook's operational needs. Hookie streamlines the process of academic subject annotation. We anticipate a reduction of up to 70% in the time and effort currently required for video classification. Our work stands out for its innovative use of NLP in EdTech which will lead to significant operational improvements for ClassHook and enrich classroom experiences for teachers all across the country.

2. Business Overview and Issue

ClassHook is an innovative educational technology startup that leverages popular media to create accessible, engaging educational content for K-12 students. By compiling clips from TV shows and movies normally accessible on YouTube, Vimeo, IMDb, or similar websites, ClassHook provides educators with a valuable tool to enhance the classroom learning experience.

Despite the platform's successes, ClassHook is currently facing challenges stemming from content expansion, specifically for more mature students. The startup's manual process required to add new clips—ensuring they meet a specific set of criteria (appropriate length, quality, and educational relevance)—has become a bottleneck. This process demands the employee to watch the full clip, verify the source, and adhere to precise video quality standards and the educational integrity of the content.

Each candidate video undergoes a thorough review using a detailed Clip Curation Template, which requires manual input of data fields such as video URLs, start and end times, relevant educational subjects, engaging titles, and clip descriptions. Moreover, each clip must be independently viewable, requiring no additional context for understanding so it can be used in a classroom experience.

The aforementioned curation process slows the pace at which new content can be added to the repository of videos, negatively influencing the platform's ability to expand. The scarcity of content, especially for older students, is the number one piece of feedback ClassHook's leadership team receives from educators, highlighting the need for a more streamlined approach to content curation.

3. Proposed Solution

We address the content curation bottleneck by automating most of the current annotation process. After the exploration of various model frameworks, we trained and optimized Large Language Models for text classification on the subtitles of each clip. We obtained a model that classifies subtitles into 15 subjects, covering over 90% of all labels, with a micro-F1 score of 0.74. We packaged this workflow into a modular pipeline for maintainability and further scaling.

Secondly, we made the inference available through an API. The inference endpoint accepts two types of inputs: YouTube URLs and text transcripts. It outputs a dataset with the computed probabilities for each subject.

Finally, we propose Hookie, an intuitive web application that helps annotators curate content, from Youtube Links to the subject annotation. This interface allows an annotator to simply copy-paste a YouTube link to get the subtitles and the associated subject classification. The user can also upload a file containing a set of URLs. Hookie parses it, predicts the relevant subjects, and lets the user download a CSV file with the results. We have made the interface easy to set up on any computer. It is even possible to run Hookie by executing only one cell of a Colab Notebook, requiring no installation and environment setup, making it readily available to any annotators, immediately improving their productivity.

4. Methodology

Our process began with data cleansing and preparation. We initially loaded data we received from ClassHook, which included over 7,200 rows of the platform's classified content. The dataset includes the video's name, the TV show/movie, a short description, subtitles, education level, content tags, and associated metadata. After that, we made some initial visualizations to get a clear understanding of our subject distribution as seen in *Fig 1.1*.

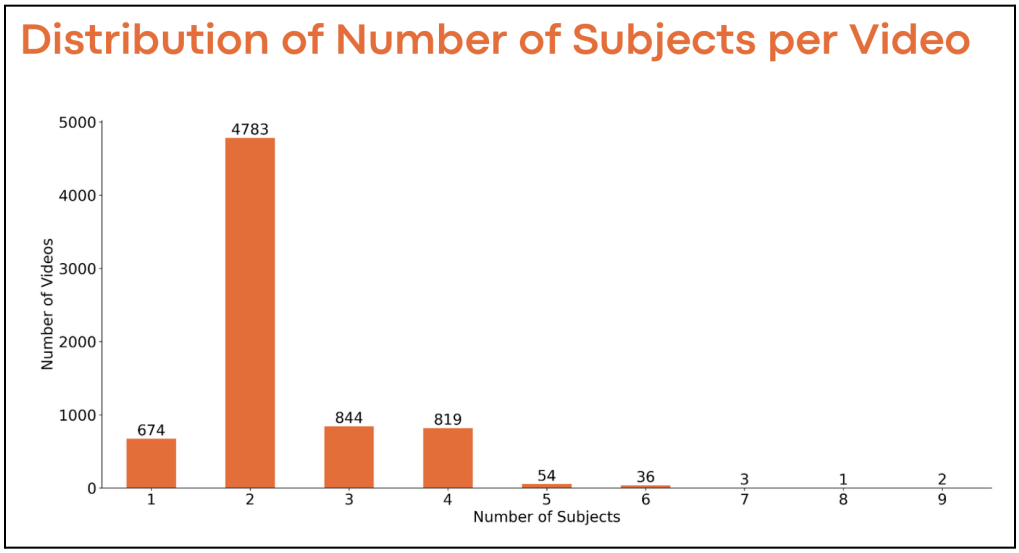

Fig 1.1

In the pre-processing phase, we first removed the non-essential columns to streamline the dataset, focusing on relevant attributes, namely the video's subtitles (removing clips where

subtitles were not available), and classified subjects. A video can be classified into multiple subjects making this a multi-classification problem. The textual data from subtitles was standardized (converted to lowercase and stripped of non-alphanumeric characters), a standard step for NLP use cases, and crucial for ensuring consistency in subsequent tasks. We then re-shape the dataset to implement one-hot encoding so that each subject in the dataset is a binary column where 1 indicates if the video belongs to that subject and 0 if not. Next, we transitioned into model development.

4.1. Model Introduction

In our systematic approach to enhancing ClassHook's platform, we considered an array of model architectures, each representing a broader family of analytical methods. Our investigative process delved into various iterations within each model category, refining and adapting them to align with our objectives.

We implemented TF-IDF for text representation, pairing it with multinomial logistic regression. When developing our Large Language Models (LLMs), we evaluated different configurations, fine-tuning several versions to identify the optimal model for our specific needs.

The selection of the TF-IDF with Logistic Regression, Zero-shot BART, and fine-tuned BERT models reflects the endpoint of our evaluation phase. These models were not chosen in isolation but as representatives of their respective model families, each emerging as the leading solution after a comprehensive and measured exploration. They enabled us to find the most effective tools for ClassHook's content curation and classification ambitions.

4.2. Statistical Baseline Model: TF-IDF with Logistic Regression

In order to classify subtitles into academic subjects, we established a baseline model using Term Frequency-Inverse Document Frequency (TF-IDF) coupled with Logistic Regression within a multi-output framework. This baseline model strategically focused on 15 academic subjects, which were selected for their representation of over 90% of the dataset, ensuring that the model was concentrated on the most prevalent topics within our corpus.

The TF-IDF method transformed the text into a feature vector, spotlighting words that were uniquely significant to each subject, while the Logistic Regression classifier, known for its efficacy in binary and linear separation, was adapted to predict multiple labels per subtitle instance. Although this baseline model attained high precision in certain categories, such as

'Biology' and 'Business', the recall was markedly low, indicating a propensity of the model to make conservative predictions. The F1-score, a metric that balances precision and recall, was around 0.37, reflecting the baseline model's limitations in capturing the full scope of each subject. This highlighted the need for a model that could improve upon these foundational results. Advancing from the baseline, we employed more sophisticated techniques using large language models.

4.3. Zero-Shot Classification: BART

For the zero-shot classification model, we leveraged the BART large model known for its effectiveness in natural language understanding and entailment tasks. This approach was taken to classify video subtitles into academic subjects without prior explicit examples for each class, which is typically necessary for supervised learning models.

The BART model operates by predicting the likelihood of a sequence of words and is particularly adept at tasks that require understanding the relationship between sentences, which is crucial for zero-shot classification. Its architecture is designed for sequence-to-sequence tasks and can generate a variety of NLP outputs, including zero-shot classification probabilities.

As part of preprocessing, we further refined the labels to include only those that appeared more than a minimum threshold in the dataset, focusing the classification task on the most significant subjects. This refinement resulted in a concentrated set of labels that the model would classify against, optimizing the classification process for relevance and efficiency.

The model was executed in a batch processing mode, which allowed for the classification of a large number of subtitle texts in an efficient manner. This approach was crucial given the computational demands of the BART model and the volume of the dataset.

Upon analysis of the results, the zero-shot classifier achieved a micro F1 score of 0.4598 for the top 4 classes, indicating a moderate level of accuracy in balancing precision and recall. The micro-precision for these classes was 0.4895, and the micro-recall was 0.4336. When examining the precision and recall for individual subjects such as 'History', 'Math', 'Science', and 'Social Studies', the model showed varying degrees of success. 'Math' and 'Science' saw higher precision and recall, whereas 'History' had lower precision, and 'Social Studies' had particularly low recall.

The macro-average scores across all subjects revealed challenges in achieving consistent performance, which is often the case with zero-shot learning due to the absence of training on specific label instances. The weighted average scores provided an adjusted metric that accounted for label imbalance, showing precision and recall of 0.49 and 0.43, respectively.

The zero-shot classification model using BART provides a baseline for subtitle classification in the absence of training data. While the model demonstrates the potential for identifying academic subjects with a reasonable degree of accuracy, the results also highlight the limitations of zero-shot learning, particularly in achieving high recall across a diverse set of subjects. Therefore, our next model included training to enhance the model's performance, particularly in underperforming classes.

4.4. Fine-Tuned Language Model: BERT

The last model used is a BERT-based transformer, specifically fine-tuned for multi-label classification tasks. Unlike the zero-shot model previously utilized, this one has undergone training on a labeled dataset, allowing it to learn from the data and adjust its parameters accordingly. The training data comprised subtitles classified into 15 different academic subjects, and the model was trained over seven epochs.

During the training process, the model showed progressive improvement in its ability to classify subtitles accurately. The F1 score, which harmonizes precision and recall, exhibited a substantial increase, indicating that the model was not only identifying relevant subjects more frequently but was also reducing the number of false positives. Similarly, the ROC/AUC score improved, reflecting the model’s enhanced capability to differentiate between the various subjects correctly. The accuracy metric, while not the most important, also showed significant growth, which implies that the model’s overall ability to label the data correctly was on an upward trend through the epochs.

Upon final evaluation, the model achieved high evaluation metrics with an F1 score of 0.74. These metrics collectively suggest that the model is highly effective at classifying subtitles into the correct academic subjects with a high degree of reliability. A detailed examination of the precision and recall scores for individual subjects shows strong performance across most of the areas, particularly in subjects such as ‘Math’, ‘Science’, and ‘Social Studies’. These results indicate that the model has a robust understanding of these topics as they are represented in the

subtitles. To further validate the performance of our model, we calculated the cross-entropy score, with a loss value of 0.12. These results are highlighted in the two tables below.

	precision	recall	f1-score	support
Biology	1.0000	0.1538	0.2667	26
Business	0.5366	0.6111	0.5714	36
Communication	0.5500	0.3438	0.4231	32
Economics	0.6250	0.2778	0.3846	18
Health	0.8750	0.6667	0.7568	42
History	0.9677	0.9375	0.9524	96
Language Arts	0.5581	0.8421	0.6713	57
Life Skills	0.4307	0.7973	0.5592	74
Math	0.9773	0.8776	0.9247	49
Philosophy and Theology	0.0000	0.0000	0.0000	23
Physics	0.9565	0.6286	0.7586	35
Psychology	1.0000	0.0385	0.0741	26
Science	0.9338	0.8924	0.9126	158
Self-awareness	0.5676	0.7778	0.6562	27
Social Studies	0.7621	0.9611	0.8501	180
Social-Emotional Learning	0.4962	0.8784	0.6341	74
micro avg	0.7082	0.7692	0.7374	953
macro avg	0.7023	0.6053	0.5873	953
weighted avg	0.7467	0.7692	0.7203	953
samples avg	0.7938	0.8036	0.7697	953

Fig 1.2

Epoch	Training Loss	Validation Loss	F1	Roc Auc	Accuracy
1	No log	0.237039	0.459646	0.655121	0.218213
2	0.281400	0.204757	0.529161	0.691956	0.298969
3	0.281400	0.194366	0.578135	0.721720	0.374570
4	0.188400	0.191283	0.591405	0.735373	0.400344
5	0.146500	0.187750	0.606713	0.744808	0.419244
6	0.146500	0.187842	0.613971	0.754633	0.432990
7	0.120600	0.189174	0.624319	0.762237	0.448454

Fig 1.3

The incorporation of a training step has substantially enhanced the model's classification capabilities. The fine-tuning on a relevant dataset has allowed the model to overcome some of the zero-shot model's limitations, particularly in dealing with class imbalance and recognizing a broader spectrum of subjects. Future improvements could focus on addressing the underperformance in certain subjects, such as 'History'. Strategies for improvement could include

increasing the representation of underperforming subjects in the training data, applying class-weighted training techniques, or adjusting the classification threshold to achieve a more balanced performance across all subjects.

The training has rendered the model effective and adaptable to the classification of academic subjects in subtitles. Nonetheless, continuous evaluation is essential, especially as the model encounters new and varied datasets, to ensure sustained accuracy and equitable performance across all academic disciplines.

5. Performance Comparison

Below is a table that reflects the tradeoffs between the three resulting models described above. Our solution is modular in that each model's strengths and weaknesses vary across five buckets: F1 score, inference time, training time, computation cost, and robustness.

This table is not a comprehensive list of all evaluation metrics and design considerations but reflects some of our most critical decisions. While the cross-entropy loss was computed and is an important metric, the F1 score was chosen due to the clear tradeoff between precision and recall, which can give important insights into the classification process. The inference time is important when considering operational efficiency, while the training time impacts scalability and deployment for large datasets. Given that ClassHook is a business, it was important to consider the computational cost for financial reasons and robustness to ensure the adaptability of our model.

	F1	**Inference Time**	**Training Time**	**Computational Cost**	**Robustness**
TF-IDF	0.37	Fast	Fast	Minimal resources required	No semantic representation
Zero-Shot LLM: BART	0.46	Slow	No training needed	Moderate resource usage	No learning
Fine-Tuned LLM: BERT	0.74	Fast	Slower	Extensive resource demand	Learning & tailored to the specific training dataset

Fig 1.4

6. Delivering Impact

Our solution represents a significant advancement in ClassHook's content curation process. By integrating advanced NLP models, particularly the fine-tuned BERT model, into the platform we have successfully automated and streamlined the classification of educational clips and reduced the classification time by 70%. This not only relieves the manual workload but also enhances accuracy and consistency in content categorization, setting ClassHook up for success. Another key aspect of our deliverables is Hookie which can easily be integrated into ClassHook's existing workflow through API calls built into Gradio (the platform on which Hookie exists). This ensures that the ClassHook team does not have to spend additional resources interacting with the raw code. To enhance the user experience, Hookie offers two streamlined functionalities. Firstly, users can input a single video URL to obtain subject-based classifications accompanied by probability scores. Secondly, in consideration of scalability, users have the option to upload a file of video URLs, upon which the system will generate and return a CSV file with the aggregated classifications for each video. In consideration of changing business needs, Hookie can be applied to other modalities in addition to short videos, such as podcasts and audiobooks, which broadens the scope of ClassHook's educational resources. As a result, ClassHook is better positioned to meet the dynamic needs of educators and students alike, reinforcing its role in edTech innovation.

7. Summary and Conclusions

Our solution to ClassHook's challenge in expanding its educational video library involved employing machine learning and NLP techniques, with a focus on the three models described above. In the final deliverable, Hookie notably reduced classification time and improved efficiency. The methodology included data preparation and model development for video categorization based on subtitles, with a comparative analysis at the end to determine the most suitable model for ClassHook's needs. The introduction of Hookie, a user-friendly interface, further streamlined the process and allowed ClassHook to easily implement and use Hookie in their day-to-day operations Essentially, our findings suggest that the integration of advanced NLP models with user-centric interfaces can dramatically improve content curation processes, making the process of adding new videos to the platform more efficient and less tedious.

In consideration of next steps, there are several areas for growth. These include exploring the application of these models to other forms of educational content and the potential for real-time classification to allow for instant educational content updates. Lastly, integrating user feedback mechanisms into the Hookie platform would allow for an improvement in its classification accuracy over time. By allowing educators to confirm or correct the classifications, Hookie can continually learn and adapt, ensuring increasingly precise and reliable video categorization for ClassHook's content library. Ultimately, the usability, modularity, and scalability of our solution ensure that the ClassHook team is set up for success.