# Social Media Data Analysis: Predicting Video Performance and Quality based on YouTube Thumbnails and Titles

Samuel Rauh

21.09.2025

## 1 Introduction

Understanding how humans make decisions and what captures their attention is fundamental to many fields, including psychology, marketing, and human–computer interaction. In the digital age, social media platforms offer a unique perspective on these processes, as billions of users make decisions about what to click on, watch or ignore. These actions reveal a great deal about the mechanisms of human perception, attention and judgement in online environments.

Among social media platforms YouTube offers a particularly valuable opportunity to investigate these mechanisms. Whereas most platforms today rely almost entirely on algorithmic curation, leaving users with limited control over what they consume [19], YouTube still preserves a crucial element of direct choice. From the set of videos presented, often shaped by recommendations, the final decision remains with the user. This decision is typically made on the basis of only two cues: the thumbnail and the title of a video.

This raises a central question: What drives a user to select one video over another, and what impact do the thumbnail and title have on the performance of a video?

In theory, a good thumbnail and title should be informative about the video's content and catch the user's attention. Some channels abuse the second part by focusing on misleading users with their thumbnails and titles, which encourage them to click on the video. While such improper practices may boost engagement for certain channels, they undermine the user experience, erode trust, and reduce the perceived quality of the content.

Studying thumbnails and titles systematically thus serves several purposes:

1. It deepens our understanding of the cognitive and perceptual factors that drive attention online.

2. It opens pathways for detecting and mitigating clickbait, thereby promoting healthier and more trustworthy platforms.

3. It provides practical guidance for content creators on how to present their work effectively.

This work aims to develop a deep learning model that can predict the performance and quality of YouTube videos based solely on their thumbnails and titles. By focusing on these two key factors influencing user choice, this study contributes to a broader understanding of human decision-making in digital environments, as well as to the development of tools that can enhance the quality of YouTube content.

### 1.1 Background

The influence of thumbnails and titles on video performance has been widely studied in recent years. A large body of work demonstrates that these elements are strong predictors of both engagement and clickbait potential, and several approaches have been proposed to model them.

Both text and images are unstructured data, so they need to be transformed into a numerical format before they can be used to make predictions. Many existing research papers rely on

feature extraction, using algorithms to extract specific information from titles and thumbnails that can later be forwarded to a classical machine learning model. For video titles, this includes features such as title length, the number of uppercase characters, and the presence of question marks. For thumbnails, such features can include image brightness, the presence of text and the detection of faces and other objects [17, 2, 9, 4, 5, 8, 18]. These methods have proven successful, but they rely on a limited selection of parameters made by researchers rather than on a general approach based solely on the data.

Several studies have also expanded their scope beyond thumbnails and titles, incorporating additional metadata such as comments, video descriptions, tags, or even audiovisual features extracted directly from the video content itself [16, 4, 9, 10, 18, 15]. While such information can indeed improve predictive performance and may be valuable for search engine optimization, it does not directly influence the user's click decision. Consequently, these approaches provide less insight into the specific impact of thumbnails and titles on human choice.

Also many of the research papers focus on specific datasets rather than a representative sample of videos. For instance, Mowar et al. [9] use a dataset containing Bollywood-related videos, Koller and Grabner [7] analyze a curated dataset of science-related YouTube channels, and Jang et al. [6] investigate videos from brand-channels.

Closer to the focus of this study, Qin, Wang, and Zhu [12] and Koller and Grabner [7] attempt to predict video performance directly from thumbnail and title information. Both works present the task as a classification problem and produce promising results, showing that thumbnails and titles alone can effectively predict engagement.

Overall, the literature shows that thumbnails and titles play a critical role in a videos performance and its detection of clickbait, yet most prior work relies on hand-crafted features or extends the task with additional metadata. As a result, there remains a need for more general, end-to-end approaches that focus specifically on the perceptual impact of thumbnails and titles in driving human decision-making.

# 2 Data

The base data I use origins from the YouNiverse Dataset, which contains metadata of 72.9 million videos from over 136,000 channels published between May 2005 and October 2019. The data was collected between 12 and 17 September 2019 [14]. One reason I chose this dataset is that it offers time-series channel data, which allows me to reconstruct the number of subscribers a channel had at the time of posting a video.

In order to reduce the number of videos, I first selected a specific time period to work with. I chose the first half of 2017. Since all the data was collected at roughly the same time, I could not pick a period too close to 2019. This period strikes a balance between using more recent videos and ensuring that we can compare the number of views, since we can safely assume that this number won't have changed significantly more than two years after a video is published.

To train and test our model we further reduce the number of samples by randomly selecting 50'000 videos with more than 10'000 views. I also filtered out all videos in the 'Music' category, since the performance of music videos largely depends on the popularity of a song outside the YouTube context. Combined with the fact that many music videos have a very high number of views, this would drastically distort the dataset.

The thumbnails can easily be retrieved by using the video ID and calling a specific URL. For this case, I choose a resolution of 480x360 pixels, since the quality shows a decent amount of details, but also keeps the processes for model training and inference to a reasonable level.

Since some of the videos have been taken offline their thumbnails are no longer available. So after filtering out the videos of the music category and accounting for the videos no longer available we are left with a dataset of 35'251 videos which are later split into training (80%), validation (15%), and testing (15%).

## 2.1 Scores

One of the major difficulties in this study was defining the performance and quality scores, largely due to the distorted distribution of the data. In some studies, the researchers worked

with labelled clickbait data, which we did not have in this case. Therefore, we had to rely solely on the number of views, likes, and dislikes that a video received.

### 2.1.1 Performance Score

This score should best measure the performance of a YouTube video. In general, this is represented by the number of views a video receives. However, it is not that simple, since the number of views is largely dependent on the number of subscribers a channel has, as well as on how frequently the channel publishes videos.

To calculate the performance score, I first perform a linear regression analysis, using the logarithm of views as the dependent variable and the logarithm of channel subscribers, as well as the number of videos posted in the previous year, as the independent variables. This results in the log of the expected number of views of a video. Next, I calculate the difference between the absolute expected and actual number of views. To obtain the relative difference, I divide the absolute difference by the minimum of the actual or expected number of views. This provides a relative distribution of the video's performance around the 0 point. Finally, to emphasize the value distribution around the 0 point, we take the inverse hyperbolic sine of the relative difference.

$$\hat{y} = \beta_0 + \beta_1 \log(\text{subs}) + \beta_2 \Delta\text{videos}, \quad D_{\text{total}} = V - \hat{V},$$

$$S = \operatorname{arcsinh}\left(\frac{D_{\text{total}}}{\min\left(V, \hat{V}\right)}\right)$$

### 2.1.2 Quality Score

As we are not working with a labelled clickbait dataset, I have developed a score that reflects the quality of a video. To compute this score, I performed similar calculations to those used for the performance score, but for both likes and dislikes. To calculate the expected number of likes/dislikes, I performed a linear regression analysis, using the logarithm of likes/dislikes as the dependent variable and the logarithm of views as the independent variable. As with the performance score, I calculate the relative difference for likes and dislikes. I then subtract the relative difference in dislikes from the relative difference in likes, and then take the inverse hyperbolic sine again to calculate the final quality score.

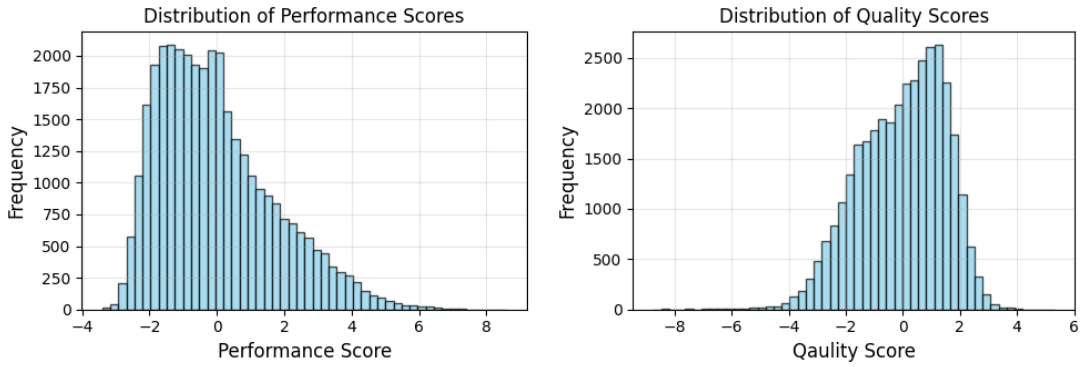


Figure 1: Distribution of target values

Figure 1 shows the distribution of the two target variables. The performance score is positively skewed with many video slightly under-performing and a small number of video strongly over-performing. For the quality score we see the exact opposite. The distribution is negatively skewed with most video having slightly better quality than expected and a smaller number performing much worse.

# 3 Methods

## 3.1 Model

To predict our target variables, we have built a deep learning model that combines the title and thumbnail of a video. The same architecture is used for both performance and quality, similar to that used by [12] and [7].

First, the video titles and thumbnails are fed into the corresponding embedding models. These models have been pre-trained using complex neural networks to extract information from large amounts of unstructured data. The result is a multidimensional vector embedding for each title and thumbnail. The idea is that each dimension of the embedding represents abstract information from the title/thumbnail. Since this process occurs in a 'black box', the specific information held by each dimension is unknown, but similar images and texts are placed closer together in the multidimensional embedding space. For my task, this would mean that some dimensions would reflect a video's performance and quality.

For the titles we are using the Qwen3 0.6B embedding model which embeds them into 1024 dimensions [20]. This model was chosen based on the MMTEB-benchmark since it is a lightweight model but still achieves a good performance [3]. For thumbnails we use the OpenAI CLIP image embedding model which embeds them into 512 dimensions [13].

To gain a quick insight into the title and thumbnail embeddings, I performed a dimensionality reduction, projecting the large number of dimensions down to just two which lets us visualy explore them.

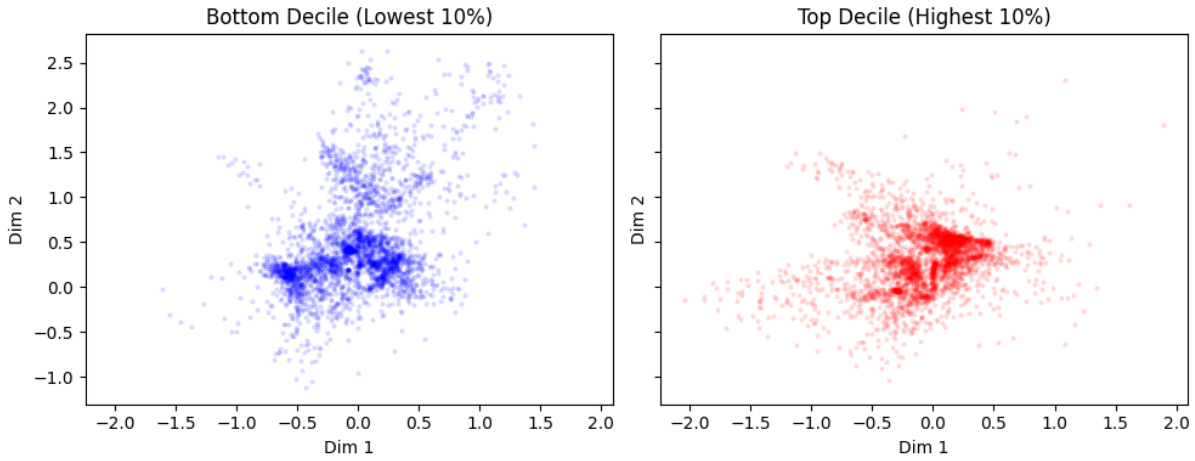


Figure 2: 2D Thumbnail Embeddings: Top vs Bottom Decile of Quality Score

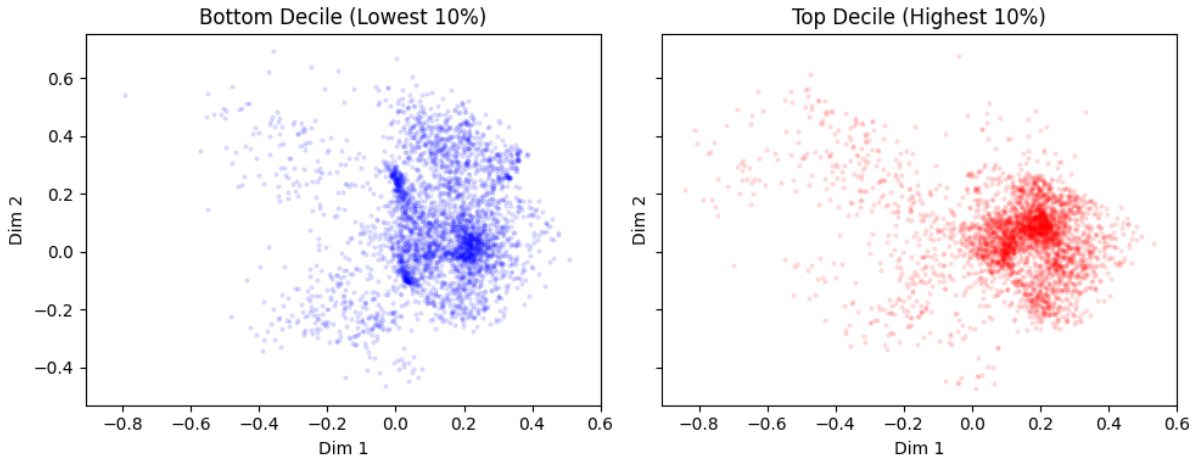


Figure 3: 2D Title Embeddings: Top vs Bottom Decile of Quality Score

Figures 2 and 3 show the two-dimensional projections of the embeddings from the top and

bottom deciles of the quality score. Even at this very simple level of comparison, we can make out some distinctions between the titles/thumbnails with the highest and lowest quality score. I performed the same analysis for the performance score, but the differences were less visible in the two-dimensional visualization.

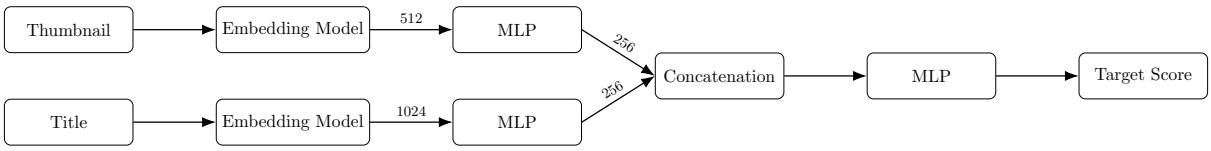


Figure 4: Architecture of the multimodal model.

Once the embeddings for the thumbnails and titles have been created, they are fed into the main deep learning model. Since I am combining two different sources into a single model, a multi-modal neural network is being built. In my case, the model's architecture first feeds each of the two embeddings into a separate multilayer perceptron (MLP) containing only fully connected layers, reducing the number of dimensions down to 256 each. The results of the two MLPs are then concatenated and fed together into a final set of fully connected layers, which ultimately return a single performance and quality score value, as shown in figure 4.

For all the computation I use Python 3.11 and the PyTorch deep learning framework [11].

### 3.1.1 Training

While the two models for performance and quality have the same general setup, the hyperparameter as well as the exact composition of the models are determined in an iterative hyperparameter tuning process using the Optuna framework [1]. We train for learning rate, weight decay, batch size as well as the composition of the three MLP parts (one for each sub-model + final MLP combining them) where we make the distinction between a "small", "medium", and "large" model. The parameters which yielded the best results are displayed in table 1.

| | Weight Decay | Learning Rate | Batch Size | Submodel Size | Final MLP Size |
|---|---|---|---|---|---|
| Performance | 1.04e-4 | 1.48e-4 | 4 | large | medium |
| Quality | 6.53e-6 | 7.34e-5 | 32 | medium | medium |

Table 1: Comparison of hyperparameters.

As a loss function we use mean absolute error (MAE) since we already used the inverse hyperbolic sine in the performance score to lay more emphasys on the differences around the 0 point.

For both models the best training results are achieved very quickly. For performance it only took 5 epochs and for quality 8 epochs. The performance model achieves a best validation loss of 1.32 and the quality model on of 0.98.

# 4 Results

Table 2 summarizes the predictive performance of our models on the test set. For both targets, the models outperform the baseline of always predicting the mean, which was estimated using 10,000 bootstrap simulations. Specifically, the model achieves a mean absolute error (MAE) of 1.2983 for performance and 0.9890 for quality, compared to their respective bootstrapped baselines (1.3792 and 1.2423). The 95% confidence intervals of the baseline further confirm that these improvements are statistically robust. Additional permutation tests produce even weaker baselines than bootstrapping does, which strengthens the evidence that our results are not due to chance.

When we compare the distributions of the predicted and actual values in figures 5 and 6, we can see that the predicted values are similarly skewed, just to a much greater extent. There is a particularly high peak for values just under 0 for the performance score, whereas the distribution of the predicted quality score resembles the actual distribution much better.

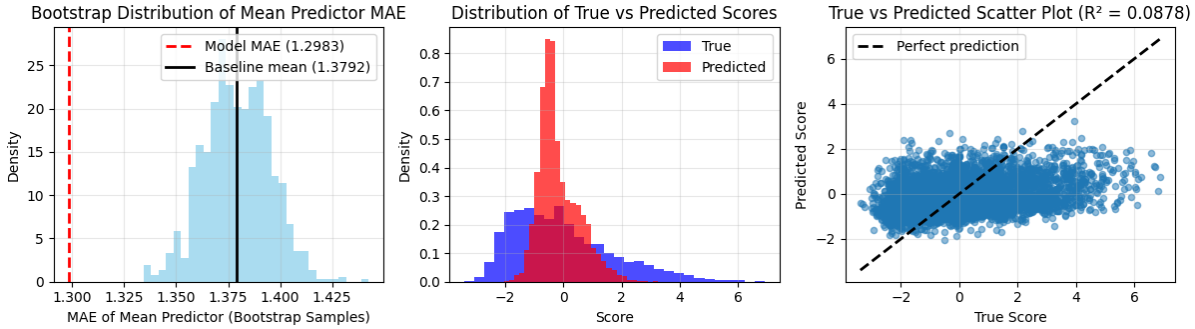


Figure 5: Summary Results: Performance Score

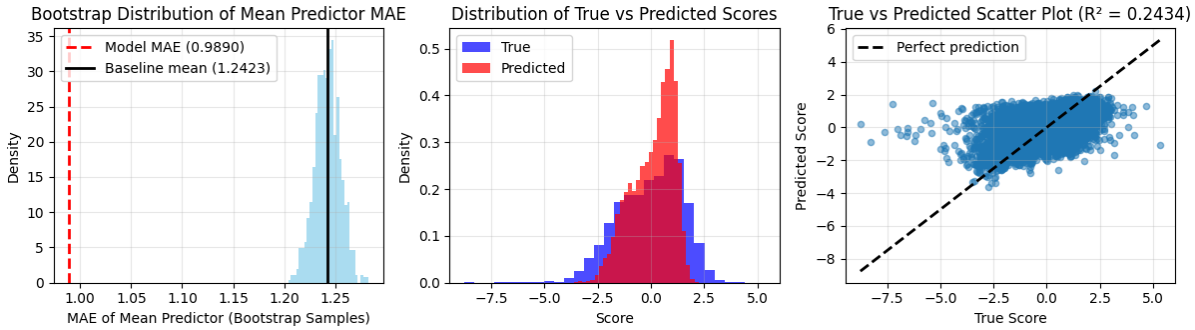


Figure 6: Summary Results: Quality Score

When further comparing the performances of the two models it becomes clear that the prediction of the quality score yield much better results than the one for the performance score. This is visible in the $R^2$ score, which is much higher for quality prediction (0.2434) than for performance (0.0878), consistent with earlier observations from the dimensionality reduction. Overall, these findings suggest that there is a significant relationship between video thumbnails and titles, and the resulting performance and quality of the video. However, given that the target scores were constructed, their absolute values can not be directly interpreted.

Table 2: Model Performance Metrics

| **Target** | **MAE** | **Baseline MAE (95% CI)** | $R^2$ |
|---|---|---|---|
| Performance | 1.2983 | 1.3792 (1.3486 – 1.4096) | 0.0878 |
| Quality | 0.9890 | 1.2423 (1.2172 – 1.2661) | 0.2434 |

# 5 Discussion

The analysis has shown that I was able to successfully develop a method that can identify relationships between unstructured data in thumbnails and titles, and link this information directly to video performance and quality. In the process, I also designed my own performance and quality scores, which allowed me to evaluate these aspects in a consistent way. The results indicate that the models predictions are consistently better than the random baselines. Although the predictive power is not extremely strong, the outcomes clearly demonstrate that there is directly extractable information in video titles and thumbnails that can be used to predict a video's performance and quality without relying on feature extraction. The $R^2$ scores, and even more so the visual analysis of our results, show that especially the quality of a video is reflected in its thumbnail and title.

These findings can have multiple implications. First, the method I developed can serve as a base for qualitative analysis to better understand how humans act online and interact with digital content. By examining which visual and textual cues are most predictive, researchers can gain deeper insight into the mechanisms of attention and decision-making in online environments. Second, the approach can be used to better predict the performance of a video, offering

creators practical guidance on how to present their content more effectively. Finally, since the method allows me to predict the quality of a video from its thumbnail and title, it also has the potential to improve user experience on the platform. By helping to overcome clickbait and highlight genuinely high-quality videos, such a system could contribute to a healthier and more trustworthy online ecosystem.

Although I have successfully identified a relationship between a video's thumbnail and title and its performance, my analysis cannot prove that there is a direct causal influence. While many creators and previous research strongly suggest that such a connection exists, my results only show correlation and not causation.

Another important limitation is that video performance on YouTube is not only influenced by user choice but also by the recommendation algorithm. Even though the user decides which video to click, what they see in the first place is largely determined by YouTube's system. Since I do not know how the algorithm works internally, and it might even evaluate thumbnails and titles itself, it is hard to separate the algorithm's influence from the direct effect of thumbnails and titles.

In addition, performance is tied to trends and changes on the platform. The data I used comes from a specific time period, and the way thumbnails and titles are designed has likely shifted since then. YouTube has also changed as a platform: in the past, users were more focused on the channels they subscribed, while today it is much more algorithm-driven. This makes it difficult to generalize my results to the current situation.

The dataset itself also poses challenges. Most videos receive fewer views, while only a small number achieve very high numbers. This imbalance makes it hard to model performance accurately. Some other studies have focused on specific groups of creators to deal with this problem, and a similar approach might have worked better here.

From a technical perspective, my model relies on pre-trained embedding models for both text and images. These embeddings are powerful but not perfectly tailored to this task. Training or fine-tuning domain-specific models might improve performance. I briefly tried this and found that the extra resource costs were not worth it for the scope of this project.

Finally, there are also some conceptual issues. The performance and quality score I created are constructed measures, which means its absolute values cannot be interpreted directly. Video success is also not determined in isolation: outside factors such as promotion on other platforms or cultural events can play a major role. On top of that, my method does not separate the influence of thumbnails from the influence of titles, so I cannot say which of the two matters more for the user's decision.

In conclusion, this paper has demonstrated the potential of the applied methods. Further research could build on these methods to improve our understanding of online decision-making processes and the factors that influence people's interaction with digital content.

## References


[1] Takuya Akiba et al. "Optuna: A Next-generation Hyperparameter Optimization Framework". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[2] Geng Cui et al. "Clicks for money: Predicting video views through a sentiment analysis of titles and thumbnails". In: *Journal of Business Research* 183 (Oct. 2024), p. 114849. ISSN: 0148-2963. DOI: 10.1016/j.jbusres.2024.114849. URL: https://www.sciencedirect.com/science/article/pii/S0148296324003539 (visited on 08/07/2025).

[3] Kenneth Enevoldsen et al. "MMTEB: Massive Multilingual Text Embedding Benchmark". In: *arXiv preprint arXiv:2502.13595* (2025). Publisher: arXiv. DOI: 10.48550/arXiv.2502.13595. URL: https://arxiv.org/abs/2502.13595.

[4] Ruchira Gothankar, Fabio Di Troia, and Mark Stamp. "Clickbait Detection for YouTube Videos". en. In: *Artificial Intelligence for Cybersecurity*. Ed. by Mark Stamp et al. Cham: Springer International Publishing, 2022, pp. 261–284. ISBN: 978-3-030-97087-1. DOI: 10.

1007/978-3-030-97087-1_11. URL: https://doi.org/10.1007/978-3-030-97087-1_11 (visited on 08/07/2025).

[5] William Hoiles, Anup Aprem, and Vikram Krishnamurthy. "Engagement and Popularity Dynamics of YouTube Videos and Sensitivity to Meta-Data". In: *IEEE Transactions on Knowledge and Data Engineering* 29.7 (July 2017), pp. 1426–1437. ISSN: 1558-2191. DOI: 10.1109/TKDE.2017.2682858. URL: https://ieeexplore.ieee.org/abstract/document/7879356 (visited on 08/07/2025).

[6] Ha Eun Jang et al. "Visual Attributes of Thumbnails in Predicting YouTube Brand Channel Views in the Marketing Digitalization Era". In: *IEEE Transactions on Computational Social Systems* 11.6 (Dec. 2024), pp. 8169–8177. ISSN: 2329-924X. DOI: 10.1109/TCSS.2023.3289410. URL: https://ieeexplore.ieee.org/abstract/document/10173777 (visited on 08/07/2025).

[7] Thomas Koller and Helmut Grabner. "Who wants to be a Click-Millionaire? On the Influence of Thumbnails and Captions". In: *2022 26th International Conference on Pattern Recognition (ICPR)*. ISSN: 2831-7475. Aug. 2022, pp. 629–635. DOI: 10.1109/ICPR56361.2022.9956202. URL: https://ieeexplore.ieee.org/abstract/document/9956202 (visited on 08/07/2025).

[8] Heba Al-Mamouri and Wadhah R. Baiee. "Maximizing video popularity prediction: A holistic approach utilizing metadata and thumbnail analysis". In: *AIP Conference Proceedings* 3097.1 (May 2024), p. 050011. ISSN: 0094-243X. DOI: 10.1063/5.0209439. URL: https://doi.org/10.1063/5.0209439 (visited on 08/07/2025).

[9] Peya Mowar et al. *Clickbait in YouTube Prevention, Detection and Analysis of the Bait using Ensemble Learning*. arXiv:2112.08611 [cs]. Dec. 2021. DOI: 10.48550/arXiv.2112.08611. URL: http://arxiv.org/abs/2112.08611 (visited on 08/07/2025).

[10] Meher UN Nisa et al. "Optimizing Prediction of YouTube Video Popularity Using XGBoost". en. In: *Electronics* 10.23 (Jan. 2021). Number: 23 Publisher: Multidisciplinary Digital Publishing Institute, p. 2962. ISSN: 2079-9292. DOI: 10.3390/electronics10232962. URL: https://www.mdpi.com/2079-9292/10/23/2962 (visited on 08/07/2025).

[11] Adam Paszke et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. arXiv:1912.01703 [cs]. Dec. 2019. DOI: 10.48550/arXiv.1912.01703. URL: http://arxiv.org/abs/1912.01703 (visited on 09/21/2025).

[12] Jie Qin, Bei'an Wang, and Tianyu Zhu. "Predicting video popularity based on video covers and titles using a multimodal large-scale model and pipeline parallelism". en. In: *Applied and Computational Engineering* 41 (Feb. 2024), pp. 182–189. ISSN: 2755-273X, 2755-2721. DOI: 10.54254/2755-2721/41/20230741. URL: https://www.ewadirect.com/proceedings/ace/article/view/10296 (visited on 08/08/2025).

[13] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:2103.00020 [cs]. Feb. 2021. DOI: 10.48550/arXiv.2103.00020. URL: http://arxiv.org/abs/2103.00020 (visited on 09/17/2025).

[14] Manoel Horta Ribeiro and Robert West. *YouNiverse: Large-Scale Channel and Video Metadata from English YouTube*. en. Apr. 2021. DOI: 10.5281/ZENODO.4650046. URL: https://zenodo.org/record/4650046 (visited on 08/07/2025).

[15] Lanyu Shang et al. "Towards reliable online clickbait video detection: A content-agnostic approach". In: *Knowledge-Based Systems* 182 (Oct. 2019), p. 104851. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2019.07.022. URL: https://www.sciencedirect.com/science/article/pii/S0950705119303260 (visited on 08/07/2025).

[16] Deepika Varshney and Dinesh Kumar Vishwakarma. "A unified approach for detection of Clickbait videos on YouTube using cognitive evidences". en. In: *Applied Intelligence* 51.7 (July 2021), pp. 4214–4235. ISSN: 1573-7497. DOI: 10.1007/s10489-020-02057-9. URL: https://doi.org/10.1007/s10489-020-02057-9 (visited on 08/07/2025).

[17] Agastya Vitadhani, Kalamullah Ramli, and Prima Dewi Purnamasari. "Detection of Clickbait Thumbnails on YouTube Using Tesseract-OCR, Face Recognition, and Text Alteration". In: *2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST)*. June 2021, pp. 56–61. DOI: 10.1109/ICAICST53116.2021.9497811. URL: https://ieeexplore.ieee.org/abstract/document/9497811 (visited on 08/07/2025).

[18] Savvas Zannettou et al. "The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube". In: *2018 IEEE Security and Privacy Workshops (SPW)*. May 2018, pp. 63–69. DOI: 10.1109/SPW.2018.00018. URL: https://ieeexplore.ieee.org/abstract/document/8424634 (visited on 08/07/2025).

[19] Brahim Zarouali, Sophie C. Boerman, and Claes H. de Vreese. "Is this recommended by an algorithm? The development and validation of the algorithmic media content awareness scale (AMCA-scale)". In: *Telematics and Informatics* 62 (Sept. 2021), p. 101607. ISSN: 0736-5853. DOI: 10.1016/j.tele.2021.101607. URL: https://www.sciencedirect.com/science/article/pii/S0736585321000460 (visited on 09/19/2025).

[20] Yanzhao Zhang et al. "Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models". In: *arXiv preprint arXiv:2506.05176* (2025).

# Appendix

GitHub Repository: https://github.com/samrauh/smda-project-public
The data can unfortunately not be made available since it is too large.
The data can be retrieved from the website https://zenodo.org/records/4650046.