Facial Emotion and Engagement: A Multimodal Analysis of YouTube Video Performance

Seminar Paper

Authors:

Göbeler, Christopher (Student ID: 7376835)

Struensee, Fabian (Student ID: 7431362)

Hsien-Pang, Hsieh (Student ID: 7419071)

Supervisor: Christopher Coors

Business Analytics and Econometrics

Faculty of Management, Economics and Social Sciences

University of Cologne

May 17, 2025

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

Göbeler, Christopher; Struensee, Fabian; Hsien-Pang, Hsieh

Köln, den 17.05.2025

Abstract

This study investigates how facial emotional expressions, low-level visual features, and rhetorical tone in video titles influence audience engagement on YouTube. Using a multimodal dataset of 128 videos from a single high-profile creator, we combine frame-level computer vision techniques (OpenCV, DeepFace) with zero-shot natural language classification to extract emotional, aesthetic, and textual features.

Exploratory and correlation analyses reveal that video duration, exaggerated title tone, and the frequency of on-screen faces positively correlate with engagement, while some positive emotional signals (e.g., happiness) show low negative associations. Tree-based ensemble models (Random Forest, XGBoost) and SHAP values highlight the relative importance of structural and affective cues, particularly video length and expressions of anger. Linear regression models further support these findings, with statistically significant effects observed for facial presence and video duration.

Together, the results suggest that engagement on YouTube is shaped not by aesthetics alone, but by a complex interaction of emotional intensity, human presence, and rhetorical framing. These insights offer some practical implications for creators and researchers seeking to understand and optimize content performance on algorithm-driven platforms. But may not be generalized to other creators.

Contents

List of Figures

List of Tables

1 Introduction

With the rapid expansion of online video platforms, YouTube has emerged as the main channel for communication, entertainment, and information sharing. Consequently, understanding the elements that drive user engagement, assessed through metrics such as views, likes, and comments, has become a key area of focus in media analytics and recommender system research. Previous research has explored various influencing factors, including video content, metadata characteristics, and audience interaction patterns (Stappen, Baird, Lienhart, Bätz, & Schuller, 2021; Shen, 2024). Some studies have also considered emotional signals in video narratives or the psychological effects of emotionally charged content on viewer behavior (Hasan, 2020). However, the studies of the role of the producer's facial emotional expressions, which, although visually present in the video, are rarely analyzed as a measurable predictor of engagement. This study aims to fill that gap by systematically assessing facial emotion signals from video creators as part of a comprehensive multimodal analysis.

This study adopts a multimodal methodology to deconstruct video engagement, integrating frame-level facial emotion analysis, the extraction of elementary visual features, and the natural language-based classification of video titles. Utilizing computer vision tools, specifically OpenCV and DeepFace, we extract both emotional and aesthetic indicators from video frames. To evaluate rhetorical tone, we employ zero-shot classification models from the Hugging Face Transformers library, drawing from prior research in dataless classification, which facilitates label prediction without necessitating task-specific training data.

By applying these techniques to a curated dataset of videos from a leading YouTube content creator, we develop a comprehensive feature set for each video. We then performed a regression analysis to determine the key predictors of user engagement. The findings of this study augment the expanding body of literature on computational media analysis and provide actionable insights for content creators and platform developers aiming to enhance viewer interaction.

2 Related Work

2.1 Deep Learning for Facial Emotion Recognition

Facial emotion recognition has dramatically advanced with deep learning techniques, especially convolutional neural networks (CNNs). A landmark example is Facebook's DeepFace system (Taigman, Yang, Ranzato, & Wolf, 2014), which employed a 3D face alignment step and a deep neural network to learn face representations. The DeepFace architecture was a 9-layer CNN (120 million parameters, including locally connected layers) trained on 4 million facial images; it achieved 97.35% accuracy on the LFW face recognition benchmark, nearly closing the gap to human performance. This pioneering work showed that a deep network could effectively encode facial features for high-accuracy recognition. Subsequent models quickly surpassed this level: for instance, Google's FaceNet (Schroff, Kalenichenko, & Philbin, 2015) introduced a unified embedding approach and reached 99.63% LFW accuracy with 128-dimensional face embeddings. Similarly, the VGG-Face model (Parkhi, Vedaldi, & Zisserman, 2015) built on the VGG deep CNN architecture and achieved 97.8% on LFW. Other state-of-the-art face recognition models include DeepID (Sun et al., 2014), one of the first CNN-based systems to surpass human-level face verification (e.g. DeepID2 scored 99.15% on LFW), and ArcFace (Deng et al., 2022), which further improved performance beyond human levels. These advances primarily focused on identity recognition, but they established the CNN architectures and training paradigms also used for expression recognition. In fact, researchers have found that deep learning methods excel at facial expression recognition as well – CNNs automatically learn salient facial features (such as action units or muscle movements) that correlate with emotional states, leading to highly accurate emotion classification.

DeepFace Library

For practical applications, the open-source DeepFace library by Serengil and Ozpinar among others, has made modern face analytics accessible in Python. This library wraps many of the aforementioned models (including VGG-Face, FaceNet, OpenFace, Facebook's DeepFace model, DeepID, Dlib, and ArcFace) under a common API. By default it uses VGG-Face, but developers can easily switch to other backends. The DeepFace framework performs facial attribute analysis in addition to identity recognition – it can estimate age, gender, race, and emotion from an input face. Notably, while its face recognition module relies on pretrained state-of-the-art models, the emotion recognition module uses a dedicated CNN trained for classifying facial expressions. Typically, such emotion

models are trained on datasets like FER2013 or AffectNet and predict basic expressions (happy, sad, surprised, etc.). In summary, deep learning now underpins most facial emotion recognition systems, offering significantly higher accuracy than earlier hand-crafted approaches. Recent surveys confirm that CNN-based methods dominate the field of automatic facial expression recognition, due to their ability to learn and encode subtle facial features that distinguish emotional states.

2.2 Emotional Expressions in Visual Media and Engagement

A growing body of work examines how the emotions displayed in visual content relate to user engagement on social media. Human faces – and their expressions – appear to be powerful drivers of audience response. For example, an early large-scale study on Instagram found that photos containing human faces were significantly more popular than those without: such images were 38% more likely to get "likes" and 32% more likely to receive comments. This indicates that viewers are inherently drawn to content with faces, possibly due to the social and emotional information they convey. Bakhshi et al. (2014) further noted that beyond mere presence of a face, a positive facial expression (such as a smile) can boost engagement – images of people smiling tended to receive more likes and comments on Instagram. In a similar vein, (Jaakonmäki, Müller, & vom Brocke, 2017) observed that Instagram posts featuring people displaying positive emotions correlated with higher consumer engagement. A smile is often described as the "most universal" positive expression, and its presence in a post provides an immediate, relatable emotional signal to viewers. These findings align with broader media studies showing that content evoking strong emotional valence (especially strongly positive or negative sentiments) triggers more user interaction than emotionally neutral content. (Berger & Milkman, 2012), for instance, reported that articles eliciting high-arousal emotions (whether positive or negative) were more likely to go viral than those with a neutral tone. In essence, emotions act as a catalyst for sharing and feedback.

More recent research has applied automated facial expression analysis to video-based social media. (Holiday, Hayes, Park, Lyu, & Zhou, 2023) conducted a multimodal study of 402 Instagram "InstaMom" influencer videos, using face analysis software to quantify the creators' expressions in each video. They found that both the amount of emotion an influencer displayed on their face and the particular type of emotion (discrete expressions like happiness, surprise, etc.) had a meaningful impact on viewer engagement measured by likes, comments,

and views. In other words, an influencer's emotional expressiveness was directly related to how audiences responded to the video. This study also noted that contextual factors, e.g. the follower count, modulate the effect of emotions on engagement, suggesting that while emotion matters, its impact can interact with audience expectations and creator identity.

2.3 Contribution

Our work aims to build on existing literature by analyzing continuous frame-level data of face presence and emotion across full-length videos from a single YouTube channel while correlating these measures with engagement metrics. This holistic approach aims to show how sustained emotional expressiveness and on-screen presence can influence views, likes, and comments going further than the first impression a thumbnail or isolated snapshots give alone. Future research can test these insights across multiple channels and genres, integrate multimodal signals such as audio sentiment or comment analysis, and explore real-time emotion-driven recommendations to further refine social media engagement models.

3 Methodology

This study employs a multi-stage pipeline to investigate the relationship between video-level characteristics and audience engagement on YouTube. The overall data processing flow is illustrated in Figure 1, which outlines the sequence of operations from raw video input to video-level feature aggregation. As shown, the procedure consists of three main streams: (1) frame-level extraction and facial emotion analysis based on detected faces, (2) computation of visual features across all frames, and (3) title-based sentiment scoring using a language model. All intermediate outputs are ultimately consolidated into a unified dataset for regression analysis.

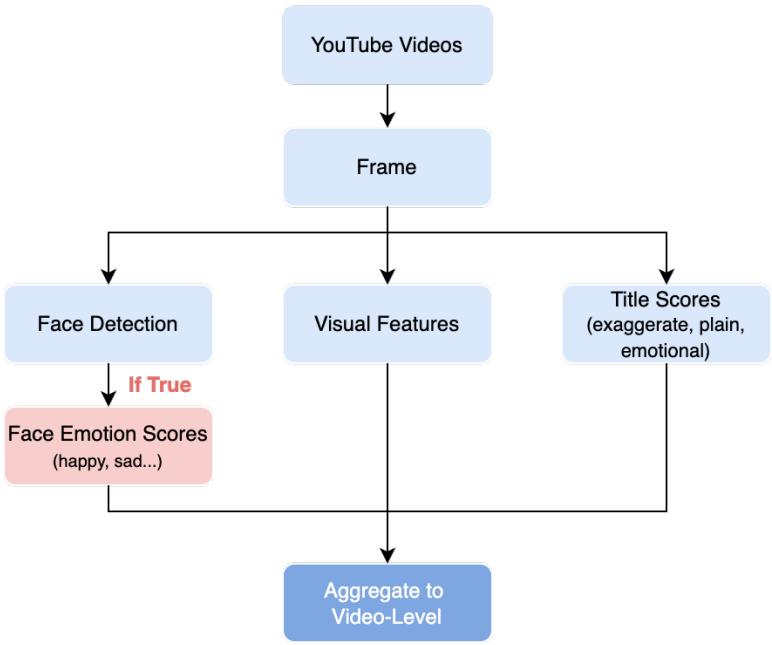

Figure 1: Overview of the Data Processing Pipeline

3.1 Data Collection

For the purposes of this experiment, the YouTube channel "Casey Neistat" was selected as a case study due to its consistent content release schedule, substantial audience size, and discernible stylistic identity. Using YouTube Data API v3 (Google Developers, 2023), a total of 1,115 videos were accessed programmatically from the channel's designated "Uploads" playlist. Metadata pertaining to each video was extracted, including unique video identifiers, titles, publication dates, aggregate view counts, like counts, comment counts, and video duration. To ensure the temporal relevance of the analysis and to account for the potential evolution of the platform, the study was limited to videos published from 2020 onward. The video content was downloaded in MP4 format through yt-dlp and all

metadata and associated media files were systematically archived for subsequent processing and analysis.

3.2 Frame-Level Processing

Each video was decomposed into discrete frames at uniform 1-second intervals utilizing OpenCV (Bradski, 2000). To isolate human-centered content, Haar Cascade classifiers were implemented to detect facial presences within each frame. Upon identification of a face, the frame was analyzed through the DeepFace library, facilitating the extraction of emotion scores in a spectrum of seven categories: anger, disgust, fear, happiness, sadness, surprise, and neutrality (Serengil & Ozpinar, 2021). Currently, low-level visual attributes were computed for each frame, including brightness, mean RGB values, and blurriness. These attributes were subsequently aggregated at the video level through the computation of means across all sampled frames. Furthermore, the proportion of frames containing faces was calculated for each video, thereby quantifying the prevalence of direct human presence.

3.3 Title-Level Processing

To assess the rhetorical features present in the video titles, a zero shot classification model was used, sourced from the Hugging Face Transformers library. Zero-shot classification refers to the ability of a model to assign labels to inputs without having been explicitly trained on those labels, leveraging its understanding of natural language semantics (Puri & Catanzaro, 2019). In this study, the model evaluated each title based on its alignment with three predefined rhetorical categories: exaggerated, plain, and emotionally touching. Returns a probability distribution across these categories reflecting the degree to which each title matches the specified tones. These probability scores were then integrated into the dataset to serve as predictive variables in the regression analysis.

3.4 Final Dataset Integration

The curated metadata, which encompasses emotional and visual characteristics, facial detection frequencies, and title classification metrics, was synthesized into a consolidated data set. This integrated dataset provides the foundation for subsequent regression analyses, utilizing view counts, like counts, and comment counts as outcome variables. The synthesized features facilitate the modeling of audience engagement as a multivariate function of emotional valence, visual properties, and title rhetoric.

4 Analysis and Results

This section presents the analytical findings of the study. It is structured around three components: descriptive and exploratory statistics, feature importance analysis via tree-based ensemble models and SHapley Additive exPlanations (SHAP)(Lundberg, Erion, & Lee, 2019), and validation through linear regression.

4.1 Descriptive Statistics and Exploratory Data Analysis

Feature aggregations were also explored for interpretability and dimensionality reduction. Positive emotion are defined as the sum of happy and surprise scores, negative emotion as the sum of angry, disgust, fear, and sad, and emotional intensity was computed as the complement of the neutral score. These engineered features may capture broader affective trends.

Engagement metrics, specifically, views, likes, and comments have strong right skewness (8.56, 5.22, and 2.69 respectively) and extreme kurtosis (82.99, 35.18, and 10.49), indicating a heavy-tailed distribution dominated by a minority of viral videos. Among emotion-related variables, disgust and fear show moderate to high skewness (2.87 and 2.06) and kurtosis (13.00 and 9.26), suggesting infrequent but extreme expressions. In contrast, visual features such as brightness and mean RGB values are approximately symmetric with near-normal kurtosis, indicating well-balanced distributions. These findings suggest logarithmic transformation of video engagement metrics (views, likes, and comments) for subsequent analyzes. The full Table 1 can be found in the appendix.

Figure 2 shows correlation between all features. Although likes (r = 0.82) and comments (r = 0.67) show the strongest correlations with log_views, they are not used as outcome variables for two reasons. First, these metrics are endogenously driven by the total number of views a video receives; they are behaviors of viewers who have already chosen to engage with the video. As such, they are not independent measures of engagement, but rather by-products of exposure. Including them as dependent variables could introduce post-treatment bias, as view count acts as a confounder that simultaneously affects both the likelihood of being liked/commented on and the features under study (e.g., video aesthetics, emotion, title tone). Second, while video titles, emotions, and visuals can be strategically crafted by creators to influence viewership, they have little to no control over whether an individual viewer chooses to like or comment. Therefore, likes and comments are not ideal measures of creator-driven engagement. As an alternative, one could also construct a composite engagement score, which may offer a more holistic yet less interpretable proxy for user interaction.

Among predictive features, the strongest positive correlation with log_views

was found for video duration (r = 0.28), followed by the exaggerated tone of video titles (r = 0.25) and the proportion of facial presence (r = 0.22). Notably, positive emotions such as *happy* showed a negative correlation with engagement (r = -0.21).

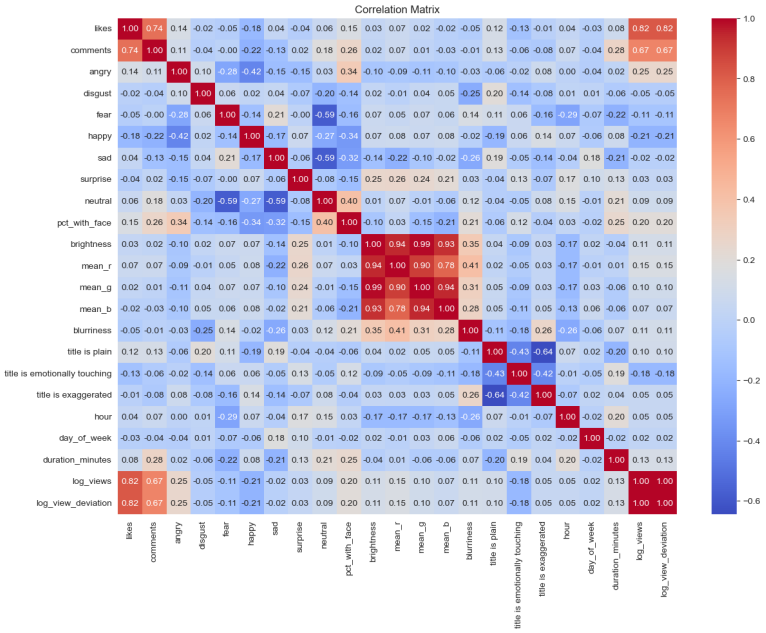

Figure 2: Correlation Matrix of Features and Engagement Metrics

Video durations ranged from under one minute to over 25 minutes, with a concentration between 5 and 12 minutes (Figure 5).

4.2 Feature Importance Extraction

To understand which factors most strongly influenced video engagement, two tree-based ensemble models for regression are trained: a Random Forest Regressor and an XGBoost Regressor. The target variable for both models was the log-transformed deviation of each video's view count from the channel mean, which captures the relative performance of a video independent of channel-wide trends. These models are selected for their robustness to nonlinear interactions and their built-in mechanisms for feature importance ranking.

Feature importance first assessment is done using each model's native impurity-based metric (i.e., Gini importance for Random Forest(Breiman, 2001), gain-based importance for XGBoost(Chen & Guestrin, 2016)). As shown in Figure 3, both models reveal similar trends, in which video duration, happiness, and anger. XGBoost also gave a high importance to the mean blue value and if a title is exaggerated.

To validate and give a direction to these findings, SHAP values are computed for both models. SHAP assigns each feature a contribution score per predic-

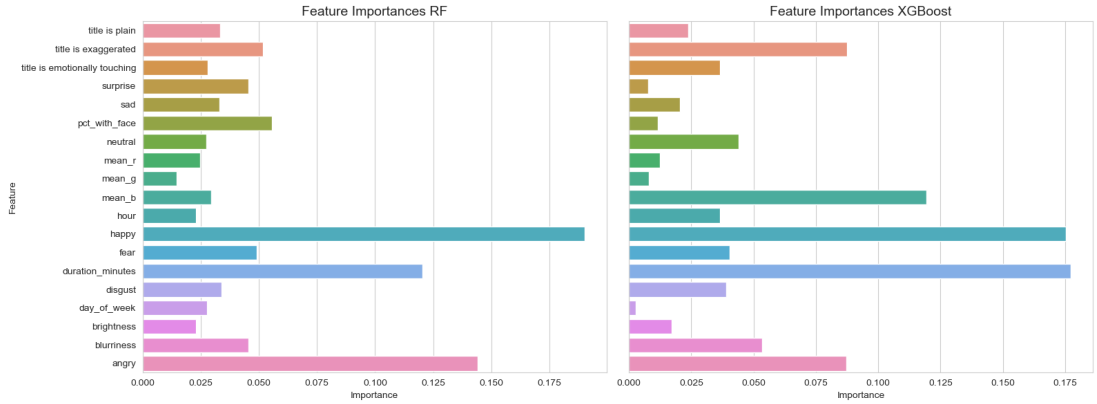

Figure 3: Feature Importance Ranking for RF and XGBoost

tion, revealing both size and direction of influence. While plots are generated for XGBoost and Random Forest, focus lies on the latter due to slight distribution differences. The XGBoost plot is included in the appendix (Figure 6) for reference.

The SHAP summary plot for the RF (Figure 4) reinforces patterns observed earlier in the correlation and feature importance analyses. Notably, longer video durations consistently contributed positively to predicted engagement deviations, indicating that extended content may enhance viewer retention or algorithmic visibility. Similarly, higher values for facial presence and anger expression were also associated with increased view count.

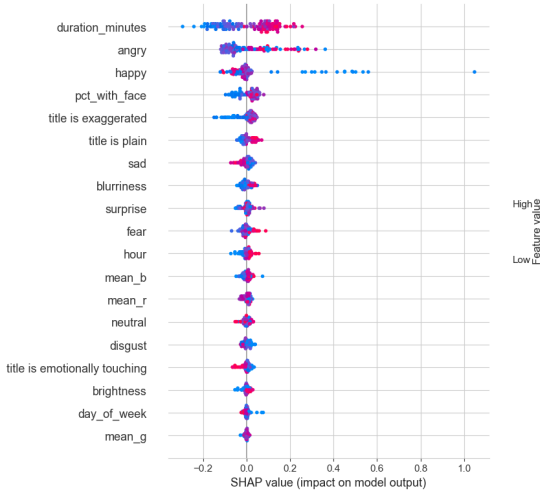

Figure 4: SHAP Value Summary – Random Forest

In contrast, some emotional features, particularly sadness and neutrality, tended to exhibit slightly negative SHAP values, implying a marginal suppressive effect on engagement performance. Interestingly, the presence of a plain title style occasionally pushed predictions upward, despite its overall lower importance. This may reflect noise in the rhetorical classification model or unobserved inter-

actions not captured by other features. Such findings highlight the complexity of modeling human attention and platform dynamics, where both highly stylized and unexpectedly minimalistic presentation strategies may trigger engagement under different contexts.

Taken together, these results underscore that engagement performance is shaped not by a single dominant factor, but by the interplay between structural (e.g., duration), affective (e.g., happy, angry), and rhetorical (e.g., exaggerated title) cues. These insights have practical implications for content creators seeking to optimize viewer engagement on algorithmically driven platforms like YouTube.

4.3 Linear Regression Analysis

In addition to tree-based models, a series of linear regression analyses estimate the marginal associations between video-level features and audience engagement. The goal of this is to assess whether interpretable, low-complexity models could capture consistent directional effects of key features. A composite engagement score based on a normalized linear combination of views, likes, and comments serves as a dependent variable for linear modelling. The first model is specified as follows:

$$\begin{aligned}\text{engagement\_score}_i = \beta_0 + \beta_1\,\text{pos\_emotion}_i + \beta_2\,\text{pct\_with\_face}_i \\ + \beta_3\,\text{published\_at}_i + \beta_4\,\text{duration\_minutes}_i + \epsilon_i \qquad (1)\end{aligned}$$

The regression results reinforce that longer videos and more facial presence are positively associated with the engagement score, controlling for time and emotional intensity. Positive emotion also exhibits a modest positive coefficient, though its statistical significance varies with model specification.

To address multicollinearity, Principal Component Analysis (PCA) is applied separately to the emotion scores and title tone features. The first principal component of each is retained, capturing the majority of variance in each domain. These features are used in a second linear model and while it achieves similar explanatory power, interpretability is limited due to the abstract nature of PCA-transformed features. Although coefficients are statistically significant, causal interpretations should be avoided due to the lack of temporal or experimental control.

Even though the linear models offer statistically significant coefficients, they are not intended to imply causal relationships. The limited sample size (N = 128) and potential confounding variables limit the generalizability of these findings. Future work with larger datasets may enable more rigorous causal inference.

5 Discussion

This study investigated how facial emotional expressions, video structure, and title rhetoric relate to viewer engagement on YouTube. By analyzing 128 videos from a single creator (Casey Neistat) using facial emotion detection, visual attributes, and natural language title analysis, key patterns that contribute to viewership were identified.

5.1 Facial Emotion as a Predictive Signal

One of the core findings of this study is the relevance of emotional expressiveness, particularly anger and overall facial presence, in predicting video engagement. SHAP values and feature importance analyses both suggest that anger, despite its negative connotation, is associated with increased engagement. This aligns with prior research on emotional arousal, where high-arousal emotions (regardless of valence) are more likely to attract attention and provoke action (Berger & Milkman, 2012). Notably, happiness, a traditionally "safe" emotion often favored in marketing, had weaker or even negative associations with engagement. This suggests that audiences may be more drawn to emotionally charged or provocative content rather than uniformly positive affect.

5.2 Title Rhetoric and Platform Optimization

In line with expectations, exaggerated video titles correlated positively with engagement, according to SHAP values. The presence of such language likely contributes to improved click-through rates and initial viewer interest. Interestingly, a plain rhetorical style also occasionally showed positive SHAP contributions, suggesting that minimalist or understated titles can sometimes perform well, potentially due to novelty effects or audience fatigue with hyperbole. These nuances reinforce the need for a more context-aware understanding of what constitutes effective rhetoric in title construction.

Zero-shot classification offered a scalable way to annotate titles with semantic tones, but its use also highlights limitations. While the classifier provided useful distinctions among broad tone categories, it is agnostic to cultural, temporal, or audience-specific nuances. Further work using fine-tuned language models or incorporating comment sentiment may yield a more accurate picture of rhetorical impact.

5.3 Video Duration and Structural Elements

Longer video duration consistently ranked among the top predictors of higher engagement. While longer videos may benefit from algorithmic favorability on platforms like YouTube, where total watch time influences recommendation algorithms, they also allow for more narrative development and emotional pacing. However, this finding should not be interpreted to mean that longer is always better. Excessive length without meaningful content may harm retention, and optimal video length likely varies by genre and audience expectations.

Low-level visual features like brightness and color composition were less influential, suggesting that raw aesthetic properties may play a more minor role than facial cues and structural factors. Nonetheless, future research might examine how changes in visual mood (e.g., contrast, saturation) modulate engagement over time within a single video.

5.4 Limitations and Future Directions

While the analysis provides valuable insights, several limitations should be acknowledged. First, the study is confined to a single YouTube creator, Casey Neistat, whose content and audience style may not generalize to other creators or genres. Second, the dataset is limited in size ($N = 128$), which restricts model complexity and statistical power. Third, the use of cross-sectional regression models constrains causal inference; longitudinal or experimental designs would be needed to confirm directional effects.

Moreover, while facial emotion models like DeepFace offer high accuracy, they are not without error—particularly under varied lighting, occlusion, or facial angles. Title tone classification, too, is subject to ambiguity and potential bias from pre-trained language models. Therefore, the findings should be interpreted as correlational, not definitive.

Future research could scale this framework across multiple creators and incorporate additional modalities, such as audio sentiment, eye gaze tracking, or viewer comment sentiment. Real-time emotion detection and engagement feedback loops could further enable adaptive content strategies. Additionally, testing these features in experimental or A/B-tested environments could validate their causal impact on audience behavior.

5.5 Conclusion

This study demonstrates that multimodal signals, including facial emotion, structural video properties, and title rhetoric, jointly contribute to the dynamics of

video engagement on YouTube. Unexpectatly, expressive but intense emotions such as anger, sustained on-screen facial presence, and rhetorically exaggerated titles are stronger predictors of engagement than positive affect or aesthetic beauty alone. These findings suggest that viewer attention on algorithmic platforms may be more responsive to emotional salience and narrative intensity than to traditional notions of likability or positivity.

For content creators and platform designers alike, the implication is clear: emotion matters, but not always in the ways we expect. Engagement is not simply driven by what looks good, but by what feels compelling, and that feeling is shaped by how humans appear, act, and communicate in the frame.

A Appendix

A.1 Kurtosis and Skewness

Variable	Skewness	Kurtosis
views	8.56	82.99
likes	5.22	35.18
comments	2.69	10.49
angry	0.85	1.06
disgust	2.87	13.00
fear	2.06	9.26
happy	0.59	0.49
sad	0.47	-0.17
surprise	1.63	3.27
neutral	0.61	1.91
pct_with_face	1.14	1.52
brightness	-0.32	0.78
mean_r	-0.22	0.75
mean_g	-0.26	0.77
mean_b	-0.05	0.83
blurriness	2.12	6.71
title is plain:	0.28	-0.43
title is emotionally touching	0.97	1.16
title is exaggerated	0.78	0.21
hour	-1.63	4.76
day_of_week	0.37	-0.81
duration_minutes	1.42	5.91

Table 1: Skewness and Kurtosis for each variable

A.2 Video Duration

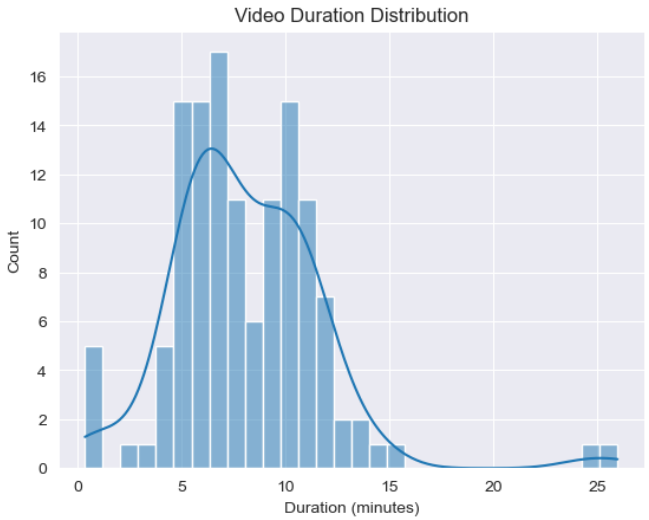

Figure 5: Distribution of Video Duration

A.3 SHAP Values XGBoost

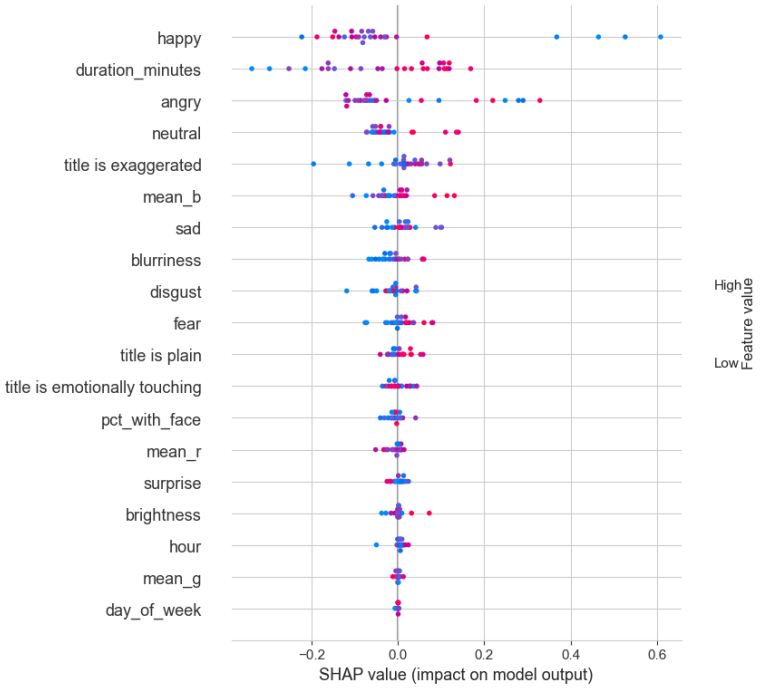

Figure 6: SHAP Value Summary – XGBoost

References

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, *49*(2), 192-205. Retrieved from https://doi.org/10.1509/jmr.10.0353 doi: 10.1509/jmr.10.0353

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Breiman, L. (2001, Oct 01). Random forests. *Machine Learning*, *45*(1), 5-32. Retrieved from https://doi.org/10.1023/A:1010933404324 doi: 10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). New York, NY, USA: Association for Computing Machinery. Retrieved from https://dl.acm.org/doi/10.1145/2939672.2939785 doi: 10.1145/2939672.2939785

Deng, J., Guo, J., Yang, J., Xue, N., Kotsia, I., & Zafeiriou, S. (2022, October). Arcface: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(10), 5962–5979. Retrieved from http://dx.doi.org/10.1109/TPAMI.2021.3087709 doi: 10.1109/tpami.2021.3087709

Google Developers. (2023). *Youtube data api v3.* Developer Guide. Retrieved from https://developers.google.com/youtube/v3

Hasan, M. (2020). What makes youtube videos go viral? emotional triggers and shopper's engagement in viral advertising. *SSRN Electronic Journal*. Retrieved from https://ssrn.com/abstract=3559034 doi: 10.2139/ssrn.3559034

Holiday, S., Hayes, J. L., Park, H., Lyu, Y., & Zhou, Y. (2023). A multimodal emotion perspective on social media influencer marketing: The effectiveness of influencer emotions, network size, and branding on consumer brand engagement using facial expression and linguistic analysis. *Journal of Interactive Marketing*, *58*(4), 414-439. Retrieved from https://doi.org/10.1177/10949968231171104 doi: 10.1177/10949968231171104

Jaakonmäki, R., Müller, O., & vom Brocke, J. (2017). The impact of content, context, and creator on user engagement in social media marketing. In *Hawaii international conference on system sciences.* doi: 10.24251/HICSS.2017.136

Lundberg, S. M., Erion, G. G., & Lee, S.-I. (2019). *Consistent individualized feature attribution for tree ensembles.* Retrieved from `https://arxiv.org/abs/1802.03888`

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *British machine vision conference.*

Puri, R., & Catanzaro, B. (2019). Zero-shot text classification with generative language models. *CoRR*, *abs/1912.10165*. Retrieved from `http://arxiv.org/abs/1912.10165`

Schroff, F., Kalenichenko, D., & Philbin, J. (2015, June). Facenet: A unified embedding for face recognition and clustering. In *2015 ieee conference on computer vision and pattern recognition (cvpr)* (p. 815–823). IEEE. Retrieved from `http://dx.doi.org/10.1109/CVPR.2015.7298682` doi: 10.1109/cvpr.2015.7298682

Serengil, S. I., & Ozpinar, A. (2021). Hyperextended lightface: A facial attribute analysis framework. In *2021 international conference on engineering and emerging technologies (iceet)* (p. 1-4). Retrieved from `https://ieeexplore.ieee.org/document/9659697/` doi: 10.1109/ICEET53442.2021.9659697

Shen, J. (2024, 11). The research of the factors that influence the popularity of youtube videos. *Theoretical and Natural Science*, *51*, 187-193. doi: 10.54254/2753-8818/51/2024CH0200

Stappen, L., Baird, A., Lienhart, M., Bätz, A., & Schuller, B. W. (2021). An estimation of online video user engagement from features of continuous emotions. *CoRR*, *abs/2105.01633*. Retrieved from `https://arxiv.org/abs/2105.01633`

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *2014 ieee conference on computer vision and pattern recognition* (p. 1701-1708). doi: 10.1109/CVPR.2014.220