Proceedings of the International Conference on Global Politics and Socio-Humanities
DOI: 10.54254/2753-7064/29/20230835

The Influence of Titles on YouTube Trending Videos

Yihong Wu[1,a,*], Mingli Lin[2,b], Wenlong Yao[3,c]

[1]Minjiang University, Fuzhou, 350100, China
[2]Senior High School of the High School Attached to Xi'an University of Technological, Xi'an, 710000, China
[3]School of AI, University of Shenyang Technology, Shenyang, 110870, China
a. titnhs84@gmail.com, b. titnhs84@gmail.com, c. yaowenlong125@gmail.com
**corresponding author*

Abstract: The global video platform market has been growing in a remarkable way in recent years. As a part of a video, title can compel people to view. However, few scholars have studied the relationship between video trendiness and title at present. This work studies the influence of sentiment polarity of videos using Valence Aware Dictionary Sentiment Reasoner (VADER) and investigated the feasibility of the application of video titles text on YouTube trending videos research using Doc2Vec. It is found that the text in YouTube trend video titles possesses predictive value for video trendiness, but it requires advanced techniques such as deep learning for full exploitation. The sentiment polawrity in titles impacts the video views and this impact varies across video categories.

Keywords: YouTube, Trending Video, Sentiment Analysis, Text Vectorization, Logistic Regression

1. Introduction

In recent years, with the popularization and development of information technology and intelligent equipment, many video platforms have rapidly expanded into the public eye with their fragmented dissemination methods. The size of China's short video market will reach 376.52 billion yuan in 2022, an increase of 83.6% year-on-year. It is expected that the market size of China's short video industry will reach 1,066.08 billion yuan in 2025 [1]. This can reflect the video platforms market is growing in a violent way.

As a new video mode, mobile video is much shorter and easier to shoot for TV series and movies, and at the same time can meet people's needs for personalized editing and beautification. Analysis of the operation of a short video platform from the perspective of communication, the content of short videos is short and easy to understand and the topic involves all aspects of people's work and life, which can be spread quickly by using fragmented time. Also push according to the different likes of each person, meeting the diverse needs of people of different ages [2]. At present, releasing and watching videos on video platforms has become a new way of entertainment. With the large-scale development of new video models and video platforms, video traffic is a significant point for people. Because many people take advantage of its rapid spread and wide audience characteristics to increase visibility so as to achieve advertising and goods to increase their income. The explosion of mobile video and its market development has also attracted the attention of many industry professionals and experts.

There have many video platforms and many factors affect video traffic. However, at present, only a few scholars have studied the relationship between titles and video traffic. This research will focus on the relationship between video titles and YouTube video traffic. A report shows users are spending more and more time watching YouTube on connected TVs in America, with connected devices expected to account for 36.4% of YouTube viewing time in 2022, rising to 39.2% by 2024 [3]. YouTube is one of the largest and most widely recognized video platforms in the world. Although it has a long history, it has been constantly innovating with the development of the era. The title is one of the most important carriers of information dissemination. It carries the critical function of conveying information and opinions and summarizing the main content of the video.

In view of the current development trend of video platforms and mobile video, its development potential is huge, so it is necessary to study the relationship between titles and traffic. This research will focus on the relationship between the title and video traffic, and what factors affect video traffic, such as the length of the title, the language of the title, and whether to hide dislike hints. According to the above background, this paper aims to explore the effect of sentiment of video titles and the feasibility of the application of video titles on YouTube trending videos research.

2. Data Preprocessing

The major part of the dataset, including video titles, video tags, view counts, ‘likes/dislikes’ counts, video categories etc., is obtained from kaggle.com, an online data science community providing credible data resources. In order to further explore the effects of video title and diminish the influence of other factors such as the popularity of the channel and category, additional data including subscriber count, video counts and total video views of channels are added through YouTube Data Application Programming Interface (an API provided in Google Cloud Platform, which allows developers to integrate functions on YouTube website to their own applications or websites), by requesting the data with channel ID in the original dataset, which are codes referring to the channels. Besides other basic data cleaning steps, data tested out to have neutral titles in the Valence Aware Dictionary and Sentiment Reasoner (VADER) test are removed since they are less informative to the target. This action does not mean to neglect the existence of neutral titles. However, it helps the research to focus on the most valuable data and analyze the effect of sentiment polarity and the reaction of viewers.

3. Exploratory Analysis

The preliminary exploratory data analysis mainly involves correlation test and chi-square test. Table 1 shows the description of the columns involved in this study.

Table 1: Columns of YouTube trending videos dataset

Column	Description
categoryId	The ID representing the category to which the video belongs.
video_id	The unique identifier for each video.
title	The title of the video.
channelId	The unique identifier for the channel that uploaded the video.
channelTitle	The name of the channel that uploaded the video.
view_count	The number of views the video has received.
likes	The number of likes the video has received.

Table 1: (continued).

dislikes	The number of dislikes the video has received.
comment_count	The number of comments the video has received.
ChannelVideoCount	The total number of videos that the channel has uploaded.
subscriberCount	The number of subscribers to the channel.
title_length	The length of the video's title in terms of characters.

The dataset consists of data of trending videos from 2020 to 2023. Table 1 shows the columns of original dataset. Correlation test is conducted for the numerical data, which consists of various metrics of trending videos including view count, likes and dislikes count and comment count. These data (Figure 1) are found to be correlated. In order to further enrich the dataset, data from the original channels of the trending videos, namely the total views and the subscriber counts, are added to find out the characteristics of the dataset. As a result, similar pattern of correlation is observed. It suggests that potential interdependencies exist among these data across both individual trending videos and the channels they belong to, which align with our hypothesis of the dataset. As an official document of YouTube presents, videos marked as ‘trending’ are based on these considerations: view count, the rate at which a video gains views, the source of view etc., showing that the ‘trending’ feature is basically a system built around view count [4]. Based on this and the analysis of correlation test, our preliminary hypothesis is that since ‘comment’, ‘click the like/dislike button’ are all behaviors after the viewers watch the videos, multicollinearity may exist among view count and data related to these behaviors. Therefore, this paper can establish an indicator called ‘trending score’ incorporating view count and other numerical data with multicollinearity to estimate the scale of trend. The detail of trending score and its calculation will be discussed in later chapters.

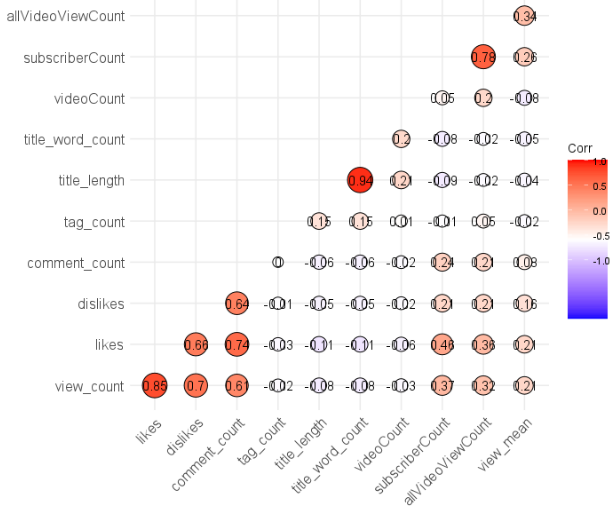

Figure 1: Correlation matrix of numerical variables in YouTube trending videos dataset

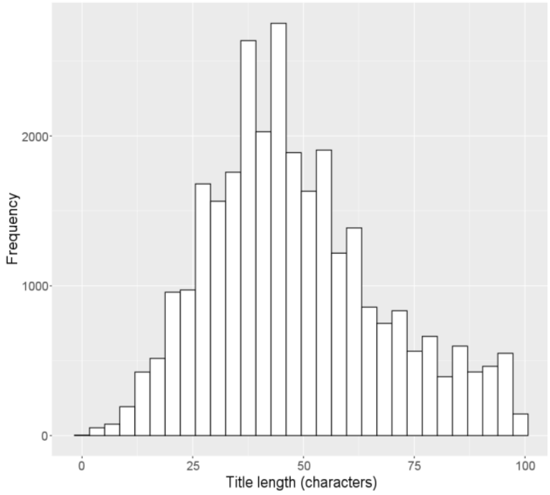

Figure 2: Distribution of title length

After the analysis above, variable ‘dislikes count’ is abandoned due to YouTube’s policy of ‘hiding dislikes’ since 2021 in later analysis and ‘title length’ is also abandoned because of the lack of correlation with other variables. ‘Dislike’ is a button from which viewers can provide feedback of disagreement or dissatisfaction for the content creators on YouTube. In the late 2021, YouTube decide to keep the count of ‘dislikes’ private to viewers. According to the official blog, this action is based on the consideration of protecting creators and rendering a diverse and creative environment

[5]. Henceforth, the original dataset cannot request the dislikes count in a normal way from then on. After 10, 2021, all dislikes counts are replaced with 0 in the dataset, which caused the declination of correlation (correlation coefficient decreased from 0.70 to 0.43 after the update of ‘dislike’ button). To maintain the accuracy and timeliness of our dataset, we decide to forego the use of dislikes count as a component of trending score.

As for the length of video title, figure 2 suggests that the length of title is distributed normally while the other numerical data of trending videos follow a skewed distribution, which means the length of title varies among the trending videos. A reason may be that content creators have found a propriate title length that satisfies viewers’ reading habit. Hence, such preference leads to the peak in the distribution of title length around median length. Furthermore, previous correlation test suggests a lack of significant correlation among title length and other numerical variables. Combing this with the observed distribution of title length, this variable is excluded in this study.

Shifting focus to the video category, it is found to be an important factor of trending videos to consider in later analyses, which can significantly influence the scale of trend. Figure 3 and Figure 4 show that Entertainment, Gaming and Music are the most trending categories on YouTube, taking over more than half of the proportion of trending videos. In order to verify the significance of category, a distribution comparison is performed. Obvious difference in distribution emerged upon setting a threshold of view count. To be more specific, view count was utilized as a proxy measure for the degree of trendiness of a video and a threshold was set at the 75th percentile of the view count to categorize videos into "highly trending" and "less trending" groups. As Figure 3 and Figure 4 shows, the Music category rose to the first place, surpassing Entertainment and Gaming. And Film & Animation surpasses Comedy as well.

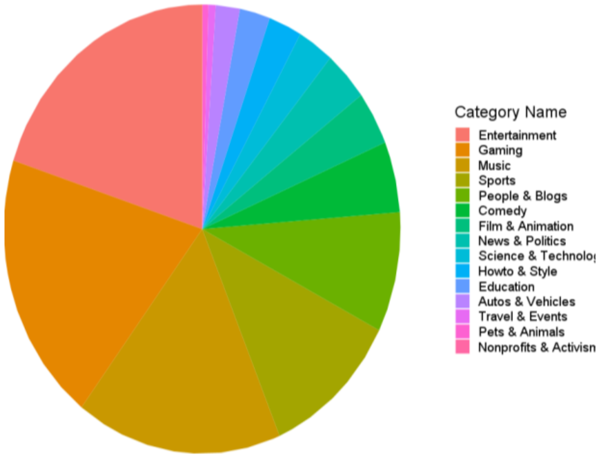

Figure 3: Pie chart of frequency of video categories

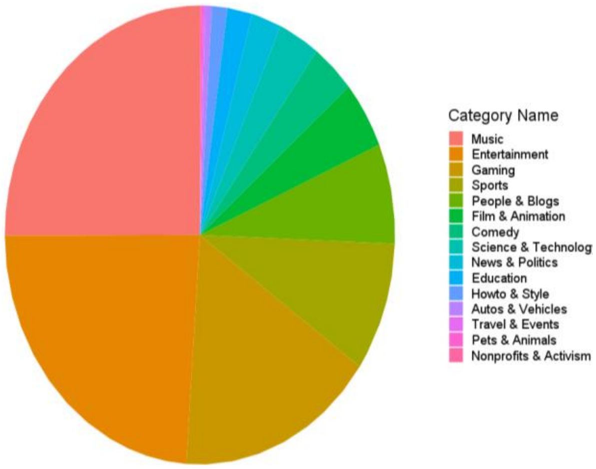

Figure 4: Frequency of video categories with threshold of view counts > 75th percentile

To further substantiate these findings, a Chi-Square test was conducted, which is a statistical test to determine the significance of difference among distributions. As a result, the test gives out a p-value < 2.2e-16, rejecting the null hypothesis that there is no association between category and the degree of video trendiness. Table 2 shows the median value of numerical data in the dataset by different categories. It basically tallies with the previous observation as Music and Entertainment remains the most trending categories with highest view count (its calculation will be discussed in later chapters), indicating that these categories are more likely to become trending.

Table 2: Median values of numerical by top 10 video category, sorted by view count

Category Name	View Count	Likes	Comments	Channel Video Count	Subscriber Count	Channel Total View
1 Nonprofits & Activism	2758540	67342	6640	846	2570000	422118795
2 Music	1377312	83128	4761	159	2870000	1953884150
3 Entertainment	1106646	48692	2913	801	4170000	1341920896
4 Science & Technology	1099054	46892	3015	517	4280000	696811141
5 Film & Animation	1094386	56555	3862	425	3170000	952286336
6 Comedy	968310	67394	3412	279	4140000	967221195
7 Education	945864	52470	3296	370	4220000	866503088
8 Gaming	851120	39843	2542	619	2220000	568177604
9 Sports	849671	13966	1998	6354	2060000	999170198
10 Pets & Animals	817272	38544	2383	274	4010000	1407413484

4. Methodology

4.1. Principle Component Analysis

Principle Component Analysis (PCA) is a common strategy in the dimensionality reduction of dataset with high dimensions invented by Pearson. In this study, the effect of PCA is to turn multivariate data into univariate data. In brief, PCA first establish a new coordinate on the data center. This coordinate is then rotated to find out the angle where the data points' projections on the axes, namely the principal component, have the maximum variance. This angle can be obtained by calculating the eigenvectors of the covariance matrix of the dataset, which is a process of linear transformation.

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{bmatrix} \tag{1}$$

Formula 1 shows a typical covariance matrix involves data with 2 dimensions, x and y. σ_x^2 and σ_y^2 are the variance of x and y respectively, and σ_{xy} is the covariance of x and y.

$$\sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{2}$$

$$\sigma_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3}$$

The formula above shows the calculation of covariance. It is a statistical measure that quantifies the relationship between two variables. If two variables tend to vary similarly, i.e., one increase and the other increase as well, then the covariance is positive. The covariance is negative when the other variable decrease.

The calculation of eigenvectors is associated with the concept of eigenvalue. Assume that the covariance matrix Σ have an eigenvector denoted v, this is a non-zero vector satisfy that:

$$\Sigma v = \lambda v \tag{4}$$

In the equation above, scalar λ is the eigenvalue of v. By sorting the eigenvalues in descending order, we can form the principal components, e.g., the component with the maximum eigenvalue is considered as the first principal component. Specifically, assume that X is the matrix of original dataset, Y is the principal component matrix and E^T is the transposed matrix of eigenvectors. The original dataset can be represented as the following equation:

$$X = Y * E^T \tag{5}$$

Every column, namely the features, of the original dataset X, can be represented as a linear combination of the principal component scores and the weights. The weight of every component is the quotient of the corresponding eigenvalue and the sum of all eigenvalues. And the component score is coordinate of the data point in the new axis.

$$x_i = y_1 * e_{i1} + y_2 * e_{i2} + \cdots + y_p * e_{ip} \tag{6}$$

As the equation above shows, x_i represents the i-th column in the original data matrix X (i.e., an original feature). y_j represents the j-th column in the principal component matrix Y, which is the score of the j-th principal component. e_{ij} is an element in the loading matrix E, indicating the correlation between the i-th original feature and the j-th principal component.

In this study, PCA is performed to simplify multidimensional data into a single dimension which we termed “trend score”.

4.2. Valence Aware Dictionary and Sentiment Reasoner

Developed in 2014 by Gilbert and Hutto, Valence Aware Dictionary and Sentiment Reasoner (VADER) is a simple but effective sentiment analysis model particularly for social media text. In social media, quantitative abbreviations, internet slangs and emoticons etc. exist, therefore it’s a difficult task for traditional sentiment analysis like Linguistic Inquiry and Word Count (LIWC), which does not consider these features on social media and sentiment intensity [6]. VADER has an sentiment dictionary containing common English words, emoticons, common internet slangs and punctuations, and every entry in the dictionary has a sentiment score to represent the sentiment tendency. VADER has 3 methods to measure the sentiment score, which considers negative sentiment, positive sentiment, and compound sentiment. For instance, after evaluation based on its rule (considering the context, adverbs etc.), compound method gives out a score ranging from -1 (negative) to 1 (positive), and 0 stand for neutral sentiment. Due to its ability of context understanding, the input of VADER does not need complex text preprocessing like tokenization, lemmatization, removal of stop words etc.

VADER is estimated as a robust method. According to the developer of VADER, its performance is comparable to human raters [6]. (Hutto & Gilbert, 2014). In a study of Twitter texts, VADER shows good accuracy in sentiment classification [7]. In this paper, VADER is called using Python code, by importing NLTK (Natural Language Toolkit) package, and the sentiment score is obtained using compound method.

In this study, VADER is implemented using Python codes. It is used for sentiment analysis of video titles. Using the compound method, we get the sentiment scores ranging from -1 to 1, which can indicate whether the video title is negative or positive.

4.3. Doc2Vec Vectorization

Doc2Vec is a document vectorization method based on Word2Vec, which can process the entire document, not just isolated words. Compared with average word embedding, a simple document vectorization method, Doc2Vec can capture the order of words and the semantics of the sentence, since average word embedding only simply represent sentence vectors by adding word vectors. Released in 2013 by Mikolov, Word2Vec calculate the word vector using continuous bag of words (CBOW) and skip-gram. In short, CBOW predicts the target word from the context while skip-gram predicts context words from the target word. These methodologies enable Word2Vec generate the word vectors.

This paper performs Doc2Vec on titles of YouTube trending videos and the dimensionality of video titles is reduced to 1 to keep simplicity so that quantitative analysis can be reached.

4.4. Logistic Regression Model

Logistic regression is a methodology for binary classification. It can be simply regarded as an extension of linear regression where a Sigmoid function is applied to the outcome of linear equation. Sigmoid function constrains the predicted value within the range of 0 and 1, making it suitable for binary classification problems.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{7}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{8}$$

Formula 7 shows the calculation of linear regression. And formula 8 shows the Sigmoid function, which can map any real number to the interval (0, 1), making it applicable for binary classification problems.

$$p(Y = 1|X) = \sigma(z) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \tag{9}$$

We can see from Formula 9, logistic regression first calculates the result of linear regression, z, and then use this result as the input to the Sigmoid function.

This paper applies logistic regression to investigate the influence of sentiment polarity and the vectorized title on the logarithmically transformed video views, which are categorized into high and non-high views.

5. Result Analysis

5.1. Trend Score of Videos

Principle Component Analysis (PCA) is performed in this study, simplifying multidimensional data into a single dimension which we termed “trend score”. Trend score is the weighted sum of selected features of a trending video, including view count of video, number of ‘likes’, comment count of video, total view count of the channel and subscriber number of the channel. The calculation is shown as below, λ_i is the weight of the corresponding variable.

$$TrendScore = \lambda_1 * ViewCount + \lambda_2 * LikesNumber + \lambda_3 * ChannelTotalView + \lambda_4 * ommentCount + \lambda_5 * SubscriberCount \tag{10}$$

Table 3: omponents of trend score

	View Count	Likes Number	Channel Total View	Comment Count	Subscriber Count
Coefficient	0.495736	0.530708	0.338024	0.449152	0.395723

The table 3 above shows the result of PCA. The highest weights are attributed to video view count and likes number, indicating that they play significant roles in trend score. The number of likes not only reflect the view count of a video, but also viewers' positive reactions, which might explain the reason why likes number have higher wight than view count.

In conclusion, the calculation of trend score provides a unified measurement to quantify the trendiness of YouTube videos. It encapsulates great amount of information about a video's trendiness into a single score, thereby simplifying the process of trend analysis.

5.2. Cross Analysis of Trend Score and Sentiment Score

VADER is implemented using NLTK library of Python. About 27% of video titles are found neutral. Neutral titles are removed since they do not contribute significantly to the polarity of sentiment, which is the focus of our analysis. The remaining video titles are subsequently divided into positive and negative categories.

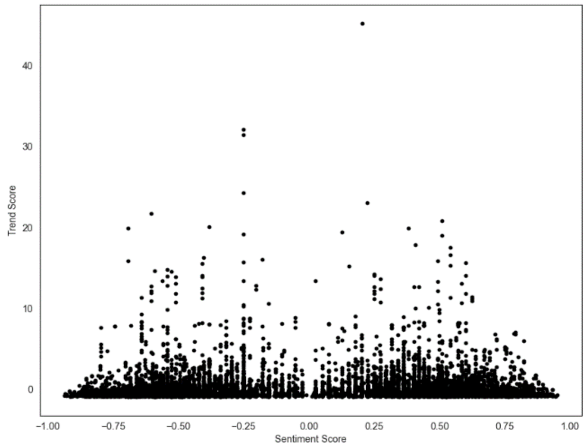

Figure 5: Sentiment score vs. trend score

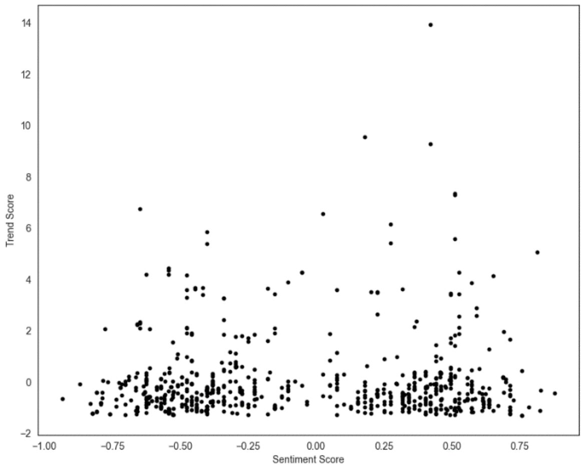

Figure 6: Sentiment score vs. trend score in comedy category

As figure 5 suggests, videos with titles that have slight negative sentiment and medium positive sentiment are more likely to obtain high trendiness.

Further study shows that the effect of sentiment is influenced by video category. In certain categories including "Howto & Style" and "Comedy", the effect of sentiment becomes more obvious, which is shown by figure 6.

5.3. Effect of Video Title Text

Vectorization using Doc2Vec is performed, from which we get quantified titles for further study. After the vectorization, the dimensionality of video titles is reduced to 1 to keep simplicity for the data analysis task and reduce computational complexity, and more meaningful results can be reached. The resultant vectors are then incorporated into our model to examine their predictive ability.

To further examine the effect of video titles, videos are categorized into outstanding videos and non-outstanding videos by setting a threshold on the 'trend_score' variable, a measure derived from other numerical data related to the trendiness of videos, then a density plot of distribution of them is generated. In figure 7 we can see that the feature on x-axis, which is derived from a dimensionality reduction of the vectorized video titles, presents an interesting pattern in the density plots. The distributions of outstanding and non-outstanding videos diverge along the x-axis, where not-outstanding videos precedes the other one. However, there is also a large portion where the trend and non-trend distributions overlap, suggesting a commonality in the video titles of both outstanding and non-outstanding videos for certain ranges of titles. Therefore, video titles hold predictive potential for determining whether a video will trend or not. However, given the overlap observed in the distributions, it is likely that the vectorized titles would need to be utilized in combination with other features to enhance the accuracy of predicting a video's trending status.

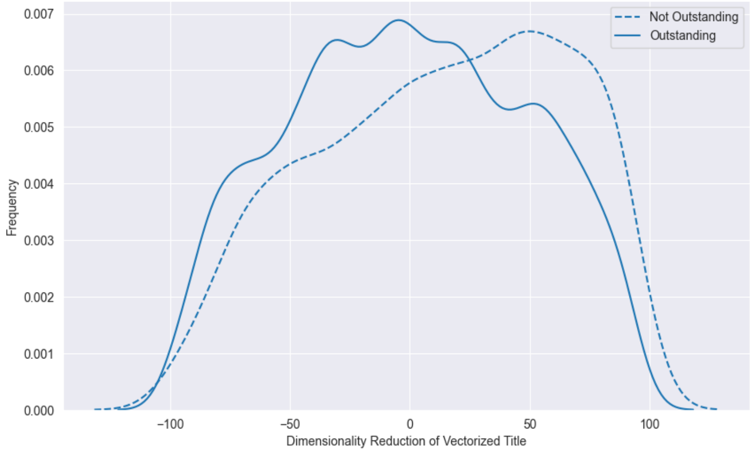

Figure 7: Frequency distribution of dimensionality reduction of vectorized title

5.4. Logistic Regression Model

A logistic regression model is established to examine the effect of sentiment polarity and the vectorized title. Sentiment score, likes count, comment count, subscriber count, video count and dimensionality reduction of vectorized titles are considered to be the independent variables. And the dependent variable is the log-transformed video views, which is categorized into high views compared to other videos in the ‘trend videos’ dataset and non-high views, based on a set threshold. The threshold is set as the first quantile of view count. Given that the dataset exclusively comprises trending videos, there is significant variability in view counts. High-view videos can vastly exceed the views of others, resulting in a substantial number of outliers distributed across various quantiles. These considerations result in the relatively low threshold.

In the exploratory analysis chapter, the dataset was initially examined for multicollinearity, which led to the creation of 'trending score'. However, as the research progressed, Variance Inflation Factor (VIF, a measure of how much the variance of the estimated regression coefficient is increased due to multicollinearity) was employed. The results from VIF analysis indicated that the multicollinearity among the parameters was not as significant as previously thought, and 'trending score' not compatible with the logistic regression model. Hence, to accommodate for this, the 'trending score' was deconstructed, with the processed 'view count' serving as the dependent variable. This revision is the respond to new insights gained during the further investigation of the dataset.

The vectorized video titles are removed in the final regression model because it is found to be statistically insignificant when included as a predictor in the logistic regression model. One plausible explanation for this could be that the contribution of vectorized video title on the trendiness of a video is indirectly captured by other variables. The example could be that an eye-catching title may lead to more likes and comments, which undermines the contribution of titles. However, it’s necessary to mention that this does not diminish the potential value of the video title in trending video prediction. It may be harnessed using other advanced techniques like deep learning methods.

The accuracy of this model is 75.10%, which is a reasonable level. And the pseudo R-squared for this model is 0.3910, suggesting that 39.10% of variance of log-transformed video views can be explained by the predictors. The formula of the logistic regression model is shown as below:

$$\begin{gathered} git\big(p(HighViewCount = 1)\big) = -4.2531761637770645e - 09 * SentimentScore \\ +3.126508712010787e - 05 * Likes + 5.201087286294733e - 05 * CommentCount \\ +2.9701416443121567e - 08 * SubscriberCount + 1.3404548902233257e - 06 \\ * VideoCount - 8.549142164225422e - 08 \end{gathered} \tag{11}$$

The model shows that all variables except sentiment score have a positive effect on high view count compared to other trending videos, which suggests that higher sentiment intensity may result in lower chance of obtaining high views. This result tallies with the previous observation that most video titles are neutral and only slight sentiment polarity is found in videos have high trend score.

6. Conclusion

This study has demonstrated that the text of trending video titles does indeed have a potential predictive value on the trendiness of the video. However, the exploitation of this potential requires advanced techniques such as deep learning.

We have also found an overall negative correlation between the sentiment polarity of the title and the rate of gaining high video views, but a slight degree of sentiment intensity in the title can benefit the trendiness of the vide. This finding suggests that content creators must strike a balance of sentiment intensity when crafting their video titles, as excessive intensity could potentially deter views. Interestingly, the influence of title sentiment polarity varies among different categories of YouTube videos. For instance, in categories like Comedy, a higher degree of sentiment intensity may foster trendiness.

In conclusion, though our understanding of what makes a video 'trending' on YouTube is far from complete, this study has glimpsed on some key factors and provides a stepping stone for future research in this field.

References

[1] iiMedia, "2023 China Short Video Industry Market Operation Monitoring Report," report.iimedia.cn, 2023. https://report.iimedia.cn/repo13-0/43328.html

[2] L. Ou, F. Zhang, and P. Chen, "Operation strategies of short video platforms of scientific journals from the perspective of communication studies: Taking Tik Tok, Bilibili, and WeChat Channel as examples," Chinese Journal of Scientific and Technical Periodicals, vol. 33, no. 58–66, Jul. 2021, doi: https://doi.org/10.11946/cjstp.202107050536.

[3] Insider Intelligence, "Most Trusted Social Media Platforms for Finding and Purchasing Products According to US Consumers, May 2022 (% of respondents)," Insider Intelligence, 2022. https://www.insiderintelligence.com/chart/257803/most-trusted-social-media-platforms-finding-purchasing-products-according-us-consumers-may-2022-of-respondents

[4] YouTube Help, "Trending on YouTube - YouTube Help," Google.com, 2019. https://support.google.com/youtube/answer/7239739?hl=en

[5] The YouTube Team, "An update to dislikes on YouTube," blog.youtube, Nov. 10, 2021. https://blog.youtube/news-and-events/update-to-youtube/

[6] C. Hutto and E. Gilbert, "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text," Proceedings of the International AAAI Conference on Web and Social Media, vol. 8, no. 1, pp. 216–225, May 2014, doi: https://doi.org/10.1609/icwsm.v8i1.14550.

[7] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and VADER sentiment," in Proceedings of the International MultiConference of Engineers and Computer Scientists 2019, 2019, p. 16.