A Data Driven YouTube Title Ranking System For Content Creators

Rohit Raju, Prerana Hadadi, Jagadeesh Gurram

Abstract

In the current advanced landscape of content creation with YouTube leading other platforms, Youtube demands content creators to bring out innovation and optimization in their content to help maximize the reach of the video. This study primarily focuses on building a model capable of informing the content creators regarding the strength of the title and its potential in gaining views. Taking help of advanced large language model's capabilities of classification and advanced ML model's capabilities of regression, the system provides the understanding of the strength of any given youtube title. This information will help content creators to optimize their title based on the requirement, helping them build a stronger title across various genres. This powerful tool can help content creators to enhance the overall success on youtube.

1. Introduction

In the ever-evolving landscape of digital content creation, YouTube stands as a dynamic platform where visibility is often synonymous with success. Acknowledging the critical role of captivating titles in driving audience engagement, the study presents a solution – the Data-Driven YouTube Title Ranking System for Content Creators.

The system aims to provide content creators with a powerful tool for analyzing the strength of any given title across various genres. When presented with a title, the system classifies it into one of four classes: Weak Title, Could Be Better, Strong Title, Very Strong Title. This classification assists creators in adjusting their titles for better viewership.

The classification of titles into classes is accomplished by running them through powerful large language architectures in an attempt to learn patterns in the title that may correlate with views. The models used in this study include [1] BERT (Bidirectional Encoder Representations from Transformers), a powerful large language architecture (as shown in Fig. 1) designed to learn patterns in text documents. BERT is trained to perform various tasks such as sentence completion, text insertion, text deletion, sentence correction, and classification. This report primarily focuses on utilizing BERT for classification, although it is also partially used for regression. In all the experiments, BERT consistently assumes a crucial role in encoding text data before it is loaded into the systems. The experiments focused on regression employ an [2] Elastic Net Regression model, aiming to predict views based on titles. This initial experiment marked the beginning of the study, which later evolved to delve into classification, as discussed in subsequent sections of the report.

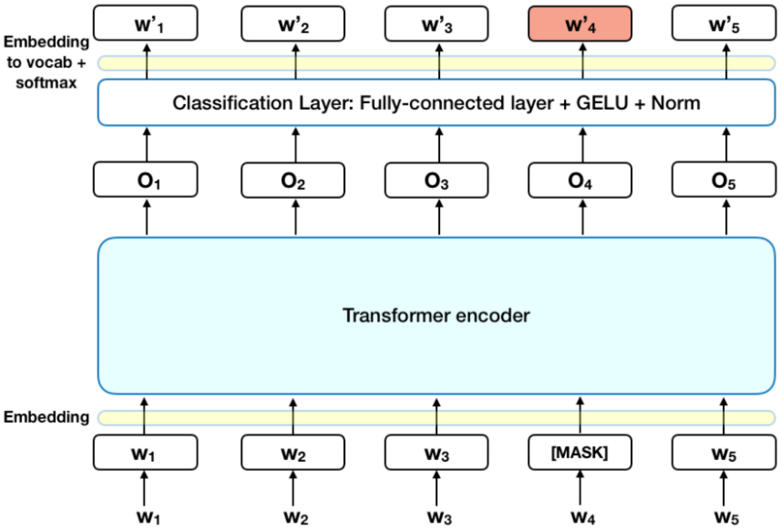

Fig. 1 BERT Architecture

1.1 Data description

The Data collection process was conducted by employing advanced [3] web scraping techniques, to carefully gather data on title names, publishing dates, views, comments, and month and year of publication. The data was meticulously collected using [2] Google API across genres, painting a vivid picture of viewer preferences within distinct content categories. Diving into the vast pool of YouTube content, the team has meticulously compiled data across diverse genres, including News & Current Events, Gaming, Comedy, Film, Entertainment & Music, Education, and Family/Lifestyle. The focus is on unraveling view count patterns, strategically grouped into five categories: 0-500, 500-1000, 1000-10k, 10k-50k, and 50k-300k. A total of 50,000 videos were collected. The data collection spanned the past two years, with the exclusion of the last three months to maintain the integrity of the dataset. Analyzing the influence of titles on views without allowing for a minimum time would undermine the validity of the analysis. This data-driven process promises content creators an understanding of their viewer's behavioral pattern of converting impressions to views, enabling them to tailor their strategies on building a robust title.

The data consists of features that include title channel ID, Video ID, title name, publishing data, views, video category. The count of videos per video category is shown in Fig 2. which shows how carefully the data is collected from various video categories. The purpose of collecting videos from various categories and across various view brackets was to capture the population data.

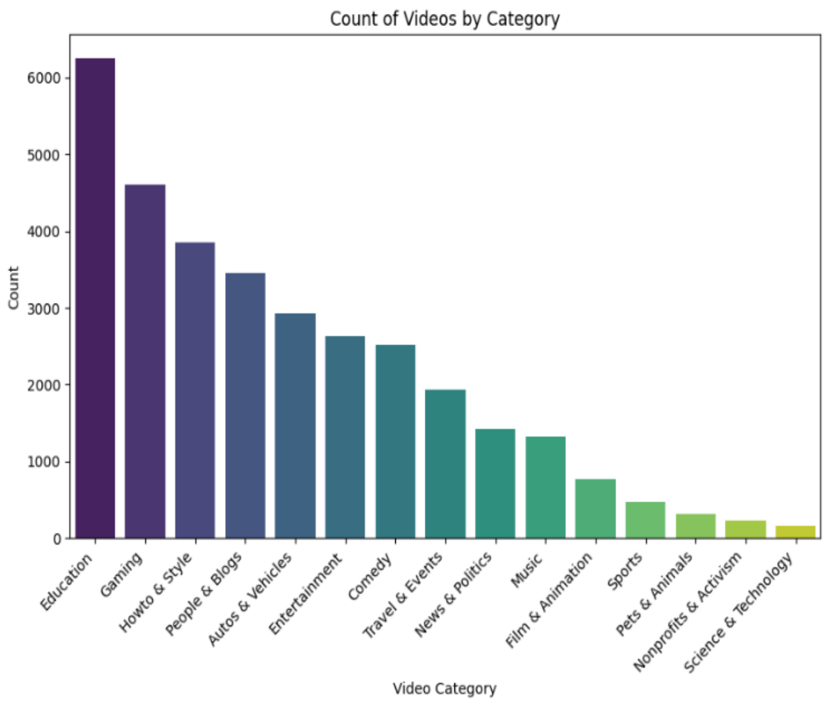

Fig. 2 Count of videos by Category

1.2 Research Questions

The research question posed in this study solely surrounds the influence of title on views. The question could be framed as follows,

Research Question: Does video title significantly influence the number of views? If it does, then could a Large Language model be trained on the data to rank the titles?

- Null Hypothesis (H0): Mean views of videos with title length less than 50 characters is equal to the mean views of videos with title length greater than 50 characters.
- Alternative Hypothesis (H1): The mean views of videos with title lengths less than 50 characters are not equal to the mean views of videos with title lengths greater than 50 characters.

Conducting hypothesis tests on these factors aims to enhance the understanding of how titles contribute to increased views. This can help us to build confidence in the subsequent model-building phase

2. Methods

The project begins with exploratory data analysis to enhance the understanding of the research questions. This is followed by data preprocessing for hypothesis testing and finally, model building.

2.1 Data Analysis

The data collected from [5] YouTube has been cleaned in the collection process and hence further cleaning of data was not required. To understand the impact of title length and video category on views, a heatmap is plotted as shown in Fig. 3

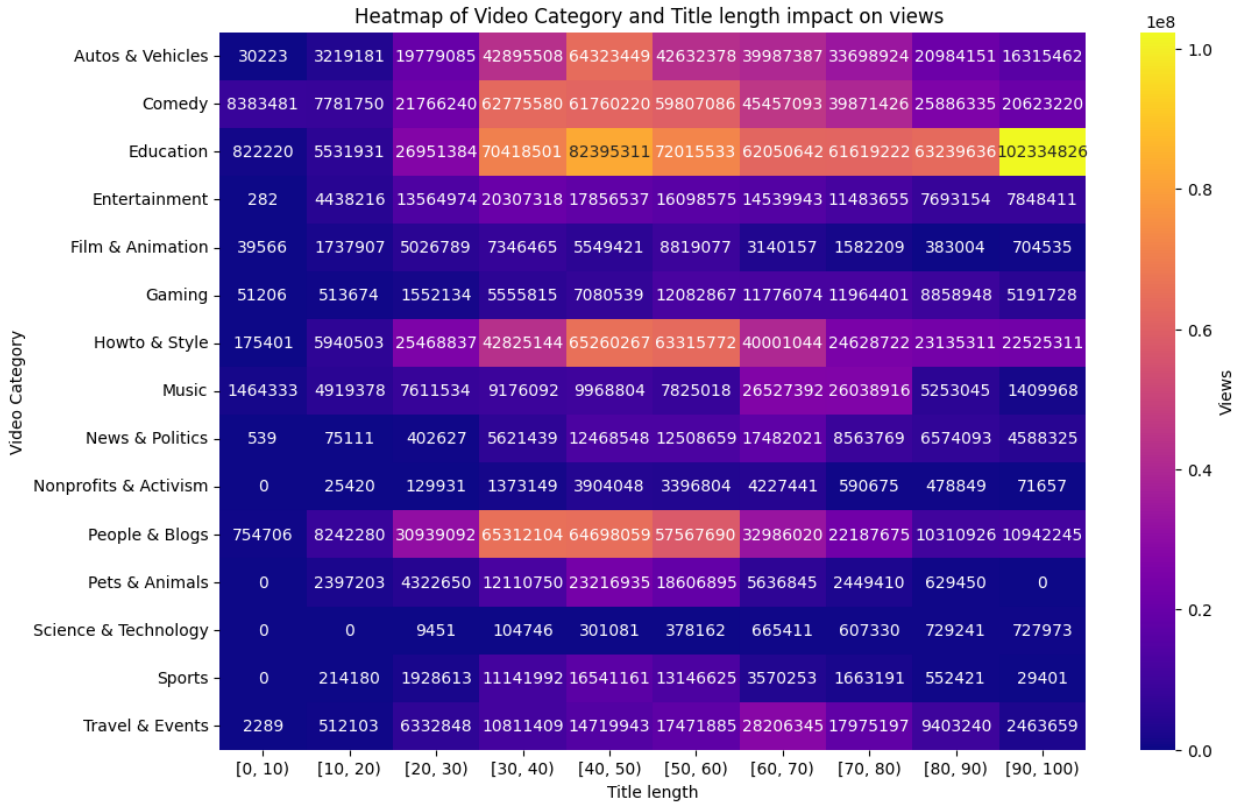

Fig. 3 Influence of title length and video category on views

The heatmap suggests that as the title length increases, there appears to be a corresponding increase in the number of views across various categories of videos in the sample data. Additionally, it is noticeable that certain video categories outperformed others. Specifically, videos categorized as 'Entertainment' and 'Comedy' seem to receive significantly more views than videos in other categories.

Based on this analysis, it can be inferred that titles may have an influence on views. To validate and confirm this prediction, hypothesis testing should be conducted on the population, moving beyond the sample data.

2.2 Data preprocessing

Data preprocessing was conducted in two stages,

2.2.1 Data Preprocessing for hypothesis testing

To conduct hypothesis testing, it is essential for the data to be normalized, and the variance between classes must be approximately equal. The dataset utilized in this study exhibits a highly right-skewed distribution of views, which can impact the variance between classes. To address this issue and achieve normality and equal variance, the 'views' column of the data has been log-transformed. The data before log transformation is depicted in Fig. 4, while the data after log transformation is presented in Fig. 5.

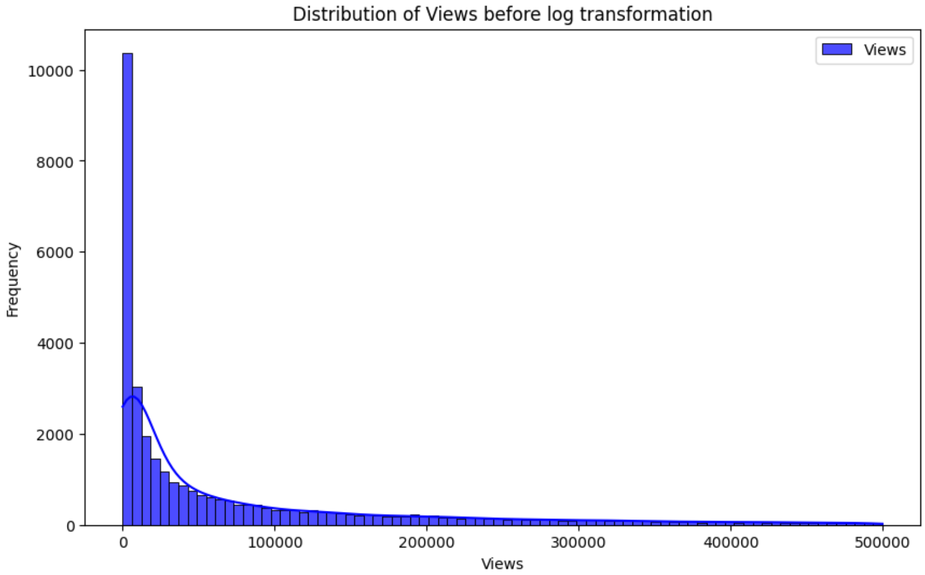

Fig. 4 Distribution of Views before log transformation

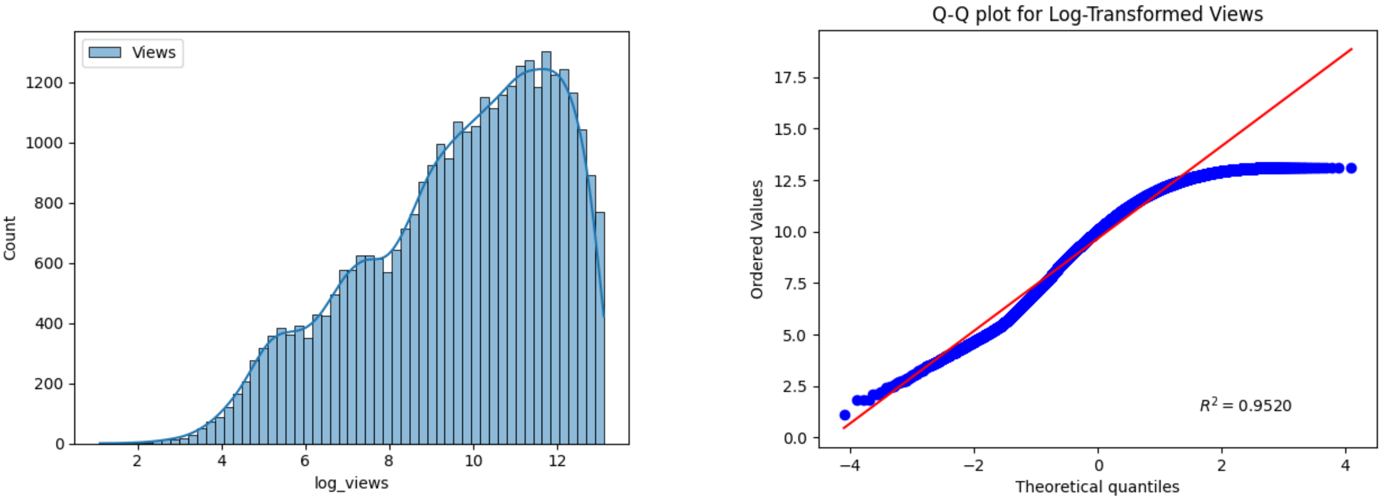

Fig 5. Distribution of views after log Transformation along with Q-Q plot

From Fig. 5 it is observed that the R^2 value is 0.95 which suggests that the sample is an optimal representation of the population and a fair normality of distribution is also achieved. The data is now ready for hypothesis testing.

2.2.1 Data Preprocessing for Model building

Data was preprocessed for regression and classification problems,

1. Z scale normalization: Views were normalized using the Z scale normalization and a new column of Z scaled views were introduced
   Formula:
   $$z = \frac{x - \mu}{\sigma}$$
   $\mu =$ Mean
   $\sigma =$ Standard Deviation

2. Min-Max Normalization: Views were normalized using the Min-Max normalization formula and a new column of Min-max views were introduced
   Formula:
   $$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

3. View-category column introduction: The 'Views' feature of the data is categorized into various classes for training classification models. These classes include:

- Weak title (0-1,000 views): Titles scoring views in the range of 0-1,000.
- Could be better (1,000-10,000): Titles scoring views in the range of 1,000-10,000.
- Strong title (10,000-100,000): Titles scoring views in the range of 10,000-100,000.
- Very strong title (100,000-300,000): Titles scoring views in the range of 100,000-500,000.

2.3 Hypothesis testing

A two-tailed [6] t-test has been conducted for the presented hypothesis to determine the t-value and p-value of the hypothesis test. Based on the obtained p-value and t-value, the null hypothesis is either rejected or accepted. Detailed result of the hypothesis test is provided in the subsequent ‘result’ section of the report.

2.4 Model Building

The study initially aims to construct a regression-based large language model for predicting the number of views a video may gain based on its title. Various models, including the Elastic net regression model (BERT encoding, Z scaled normalization), Elastic net regression model (BERT encoding, Min-Max normalization) were employed in this pursuit. However, given the scarcity of literature on regression using Large Language Models, the focus shifts toward a classification problem. Views are categorized into classes, namely 'Weak title (0-1000 views),' 'Could be better (1000-10000 views),' 'Good title (10000-100000 views),' and 'Very strong title (100000-500000 views).' Subsequently, the BERT Large Language Model is employed to classify titles into these classes based on inherent patterns identified in the titles.

3. Results

The results of the hypothesis test and the models constructed on the dataset are presented in the following subsection of the report.

3.1 Hypothesis:

Null Hypothesis (H0): The mean views of videos with the title length less than 50 characters is equal to the mean views of videos with the title length more than 50 characters.

Alternative Hypothesis (H1): The mean views of videos with the title length less than 50 characters is not equal to the mean views of videos with the title length more than 50 characters.

The results are as follows,

Hypothesis	**t-value**	**p-value**
Hypothesis 1	19.73	$6.28\times e^{-25}$

Table 1 Hypothesis test results

The results from Table 1 indicate that the hypothesis test has extremely low p-values, suggesting that the null hypotheses can be rejected. The positive t-value in the hypothesis provides evidence that the mean views of videos with a title length less than 50 characters are significantly greater than those with title length greater than 50 characters. These findings strongly support the idea that titles have a significant influence on views. Consequently, models built on titles could yield fruitful results.

3.2 Model Evaluation

The experiment was run on 3 models specifically,

1. Elastic net regression model (BERT encoding, Z scaled normalization)
2. Elastic net regression model (BERT encoding, Min-Max normalization)
3. BERT classification model with bert-large-cased

Models	Mean Squared Error (MSE)
Elastic net regression model (BERT encoding, Z scaled normalization)	0.925
Elastic net regression model (BERT encoding, Min-Max normalization)	0.684

Table 2 Model evaluation: Regression

Models	Accuracy	Precision	Recall	F1 score
BERT with bert-base-uncased (classification)	35%	36%	35%	35%
BERT with bert-large-cased (classification)	76%	75%	76%	75%

Table 3 Model evaluation: Classification

The results from Table 2 indicate that regression models did not perform well compared to the classification model. The Elastic Net regression model exhibited the poorest performance with a relatively high Mean Squared Error (MSE) score of 0.925. This high MSE is considered unfavorable, especially given Z-scaled normalization, which brings values within the range [-1, 1]. Similarly, the Elastic Net regression model with Min-Max normalization also demonstrated suboptimal performance, yielding an MSE of 0.684. Despite the values falling within the [0, 1] range, this score is still considered poor. The limitations of basic regression models to capture intricate patterns might contribute to these unsatisfactory results.

Understanding the limited capability of models to regress and find patterns in the titles, the study progressed to experiments on classification algorithms. The BERT model trained on bert-base-uncased performed better than the regression models with a low accuracy of 35% while BERT model trained on bert-large-cased demonstrated a relatively high accuracy of 76% (as seen in Table 3), followed by high recall, precision, and F1 score.

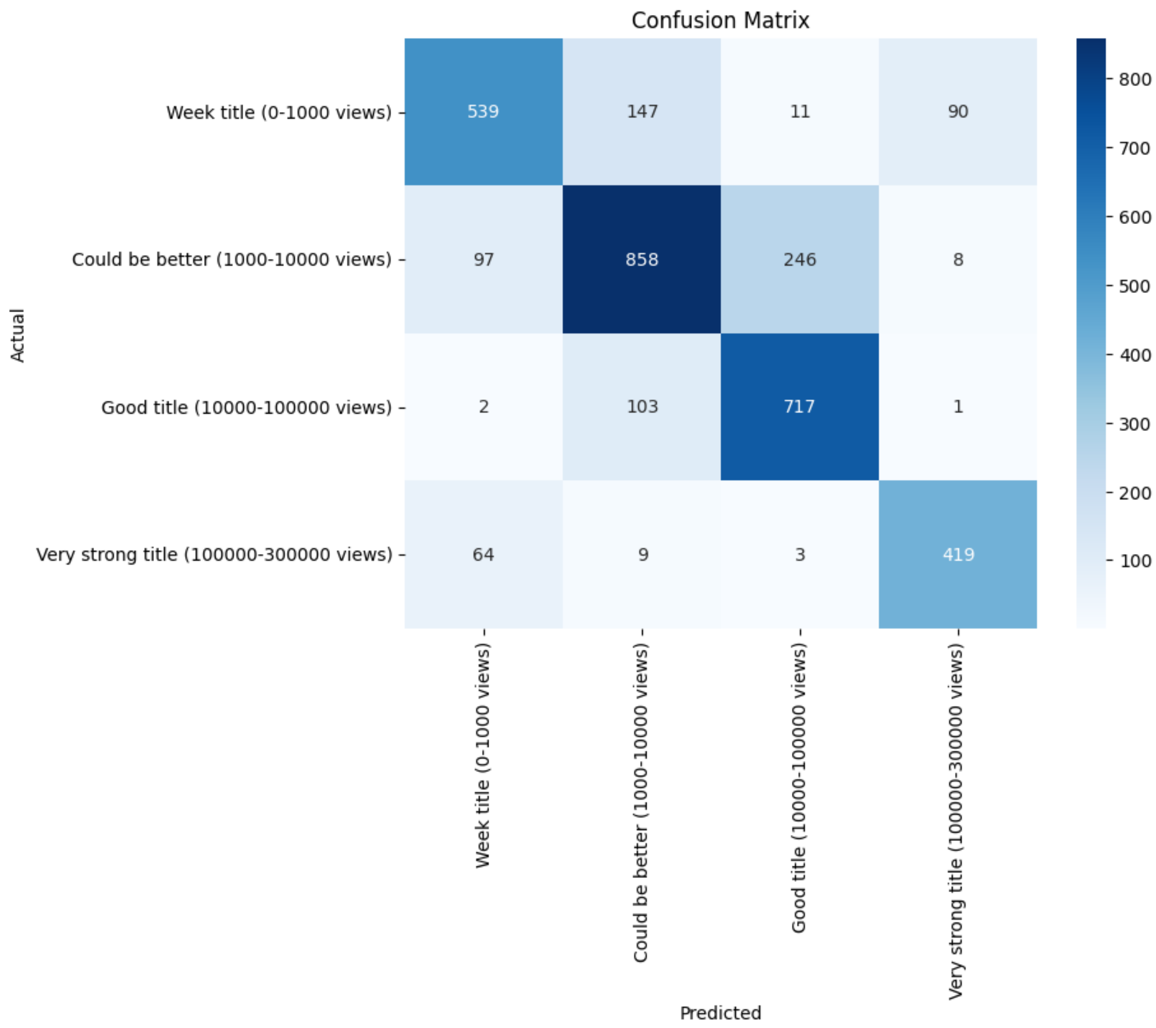

Fig. 6 Confusion matrix for the BERT classification results.

Fig. 6 depicts the confusion matrix obtained for the BERT (bert-large-cased) classification model. Observing the matrix reveals that 'Very Strong titles' achieved the highest accuracy, with almost 80% of predictions being accurate. Similarly, 'Weak titles' and 'Could be better' categories also demonstrated high accuracy, reaching 78%. However, the 'Good titles' category exhibited relatively lower accuracy at 72%, potentially contributing to the overall model accuracy of 76%. Providing additional focus on training models specifically for the 'Good titles' category may help address this issue.

4. Future Scope

Future studies must focus on feature selection. To build a strong model for classification one could pass features such as thumbnail and video duration along with the title to a more advanced model to increase the robustness of the model. Doing so, can help in the development of a very strong model, capable of classifying videos on a broader set of factors beyond titles. Additionally, stronger models can be developed by using advanced LLM models as the current study only experiments with a basic BERT classification model.

5. Conclusion

The research underscores the influence of titles on the improvement of views. The study begins with the collection of YouTube video details conducted via web scraping, followed by a brief data analysis indicating a potential influence of titles on views. The research progresses by stating hypotheses to strengthen understanding of the influence of titles on views. Furthermore, the hypothesis tests indicate a very strong relationship between titles and views. Finally, the research moves on to building a classification model with the capability to learn patterns in titles and classify them into brackets of views. This approach yields a significantly higher classification score and has substantial room for improvement. It can be concluded that titles play a major role in the increase/decrease of views. Hence, training powerful classification and regression models could help build an architecture that assists content creators in developing robust titles, ultimately improving views.

6. Acknowledgement

This study is a continuation of an existing statistics project by the same team, from which the hypothesis test results were derived to build confidence in the subsequent model-building phase. The primary emphasis of this study was on constructing a classification model, a goal that was successfully accomplished with a notable accuracy of 76%.

7. References

[1] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, May 24). *Bert: Pre-training of deep bidirectional Transformers for language understanding*. arXiv.org. https://arxiv.org/abs/1810.04805

[2] Al-Jawarneh, A. S., Ismail, M. T., Awajan, A. M., & Alsayed, A. R. (2020). Improving accuracy models using elastic net regression approach based on empirical mode decomposition. *Communications in Statistics - Simulation and Computation*, *51*(7), 4006–4025. https://doi.org/10.1080/03610918.2020.1728319

[3] V. Singrodia, A. Mitra and S. Paul, "A Review on Web Scrapping and its Applications," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2019, pp. 1-6, doi: 10.1109/ICCCI.2019.8821809.

[4] Google. (n.d.). Google. https://developers.google.com/

[5] YouTube. (n.d.). YouTube. https://www.youtube.com/

[6] Mishra, P., Singh, U., Pandey, C. M., Mishra, P., & Pandey, G. (2019). *Application of student's t-test, analysis of variance, and covariance*. Annals of cardiac anaesthesia. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6813708/
[7] X. Cheng, C. Dale and J. Liu, "Statistics and Social Network of YouTube Videos," 2008 16th Interntional Workshop on Quality of Service, Enschede, Netherlands, 2008, pp. 229-238, doi: 10.1109/IWQOS.2008.32.
[8] M. Brbić, E. Rožić and I. Podnar Žarko, "Recommendation of YouTube Videos," 2012 Proceedings of the 35th International Convention MIPRO, Opatija, Croatia, 2012, pp. 1775-1779.
[9] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (n.d.). *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. ACL Anthology. https://aclanthology.org/2020.acl-main.703/
[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to Sequence Learning with Neural Networks," arXiv.org, Sep. 10, 2014. https://arxiv.org/abs/1409.3215v3
[11] Transformers: simpletransformers.ai/docs/usage/ [attention] A. Sojasingarayar, "Seq2Seq AI Chatbot with Attention Mechanism," arXiv.org, Jun. 04, 2020. https://arxiv.org/abs/2006.02767v1

Links to presentation, code and data

1. Video Presentation Link: YouTube Video Presentation
2. Link to Code and Data: Github Code link