1 Data Exploring

The "Trending YouTube Video Statistics" is selected as the dataset [1]. In general, the dataset focus on daily records of trending viedos on Youtube. It consists of informations of trending viedos in 10 countries (USA, Great Britain, Germany, Canada, France, Russia, Mexico, South Korea, Japan and India). Entries from each country are stored in a CSV file. For each entry, the video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count are recorded, where video title, channel title, tags and description are text and others are digits. Among them, the title, tags, description, views, likes, dislikes and comment count are selected for further tasks.

1.1 Distributions of numerical data

The distributions of all numerical data in the dataset are obtained and those distributions of different countries shares similarity.

Table 1. Descriptive Statistics of Video Metrics (CA)

Statistic	Views	Likes	Dislikes	Comment Count
Count	39,585	39,585	39,585	39,585
Mean	1,169,234.01	40,596.94	2,058.69	5,159.72
Std. Dev.	3,437,842.10	134,596.73	19,312.58	21,899.59
Min	733	0	0	0
25%	149,715	2,395	104	442
Median	383,120	9,244	314	1,357
75%	983,139	29,670	976	3,821
Max	137,843,120	5,053,338	1,602,383	1,114,800

According to Table 1, the basic statistics information of numerical data in Canada is provided. Both the mean and median values of views are much larger than those of likes, dislikes and comment count. Additionally, the standard deviations of all metrics are larger than their mean but smaller than ten times of the mean, which means their distributions are dispersive, such as the exponential distribution.

For each numerical metrics, their distributions are generated and the distributions in Canada are provided. And the probability distribution functions (PDF) obtained fromstimation (KDE) kernel density and exponential fitting are used. According to Figure 1, Figure 2, Figure 3, Figure 4, their distributions are exponential distributions, meaning that most of the data are small and close with few extremely high data. Therefore, all numerical data are preprocessed into log-scale in further tasks to ensure the models to converge.

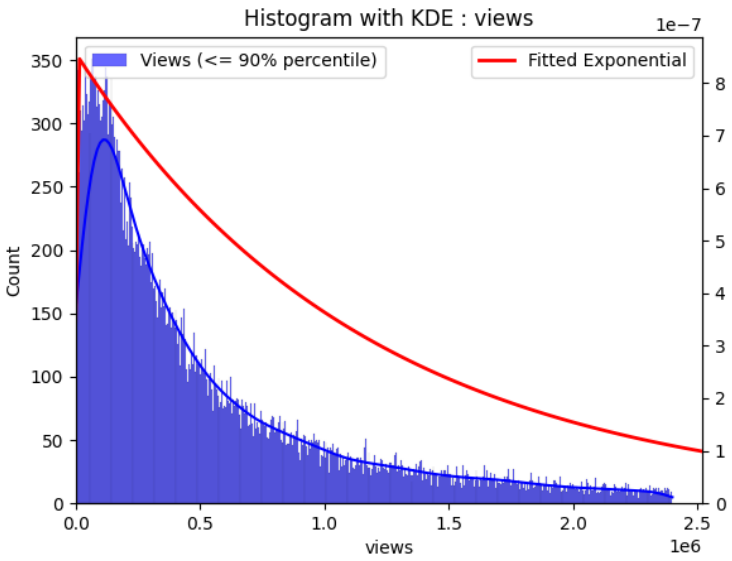

Figure 1. The distribution of views in Canada. The PDFs generated form KDE and exponential fitting are used. The max 10% data are ignored.

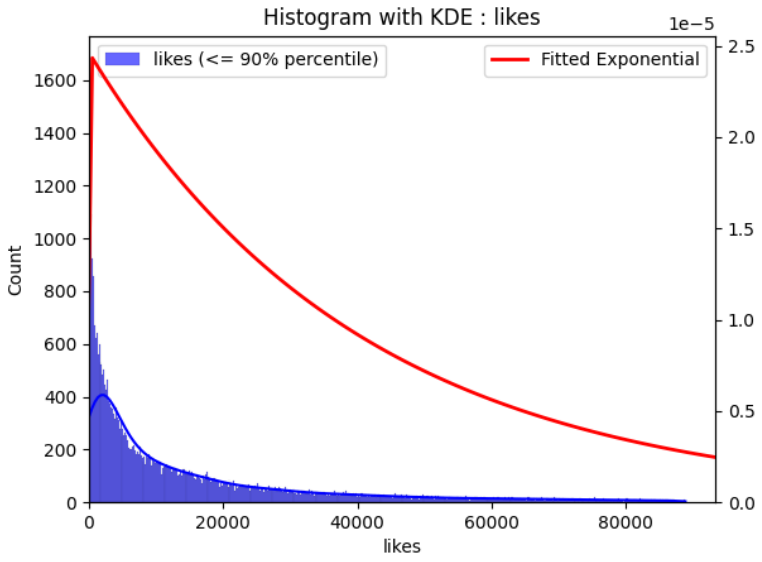

Figure 2. The distribution of likes in Canada. The PDFs generated form KDE and exponential fitting are used. The max 10% data are ignored.

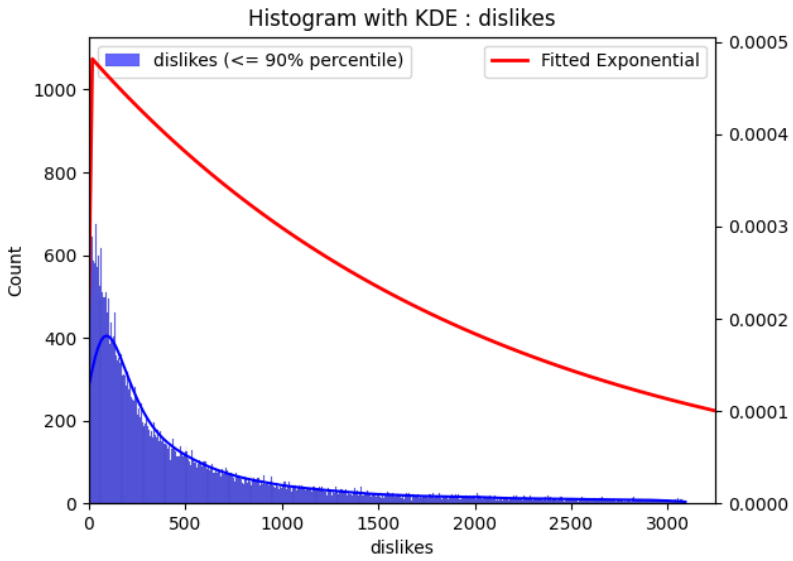

Figure 3. The distribution of dislikes in Canada. The PDFs generated form KDE and exponential fitting are used. The max 10% data are ignored.

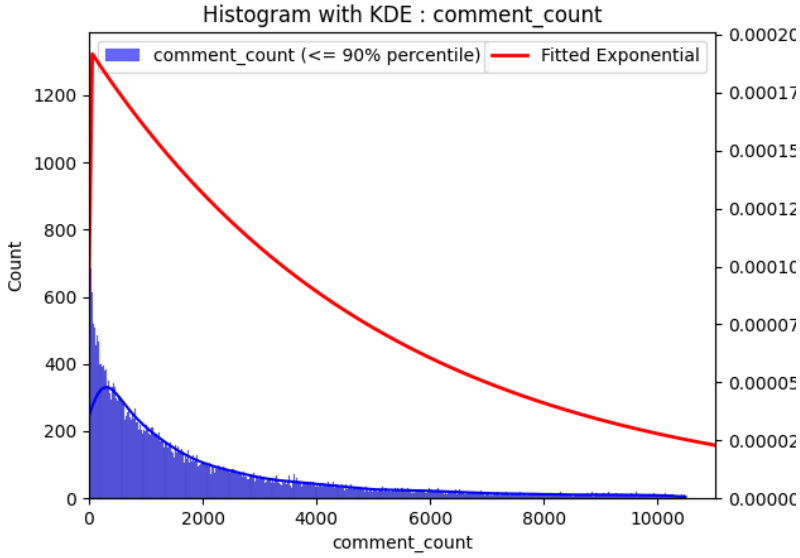

Figure 4. The distribution of comments count in Canada. The PDFs generated form KDE and exponential fitting are used. The max 10% data are ignored.

1.2 Word Frequency

The frequency of words in the text data provided signification information about the video. After removing the irrelevant informations such as stopping words and urls, the word frequency of title, tags and descriptions are generated as the word clouds.

Figure 5. The word cloud of titles in Canada.

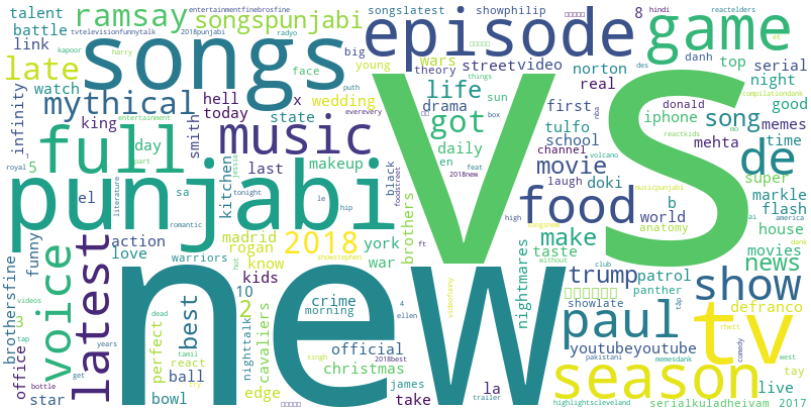

Figure 6. The word cloud of tags in Canada.

Figure 7. The word cloud of descriptions in Canada.

2 Task

Since all features selected are expected to share strong releations, each feature can be potentialy predicted by other features. Taking consideration of the real world application, the prediction of views of a viedo is selected as the task. Since views of a viedo can significantly indicate its popularity, the effective prediction can direct the creators while improve the recommendation methods of platforms.

As the prediction target is views, a regression model is expected. Initially, the simple logistic regression models with other numerical features such as likes, dislikes and comment count is selected as baselines. Further, the Term Frequency-Inverse Document Frequency (TF-IDF) approach is expected to perform well with text features, such as titles, tags and descriptions. The MSE, MAE and R^2 are selected for evaluation.

All numerical features are converted into log-scale and all text features are preprocessed by removing irrelevant informations.

3 Model

To predict the views of a viedo, all other metrics in the entry can be considered as features. Since the views, likes, dislikes and comment count share the same distributions, the regression models using likes, dislikes and comment count as inputs are expected to perfomr well. However, the prediction task can not be actually completed with these models since the views data is available if likes, dislikes and comment count are available. And using features to predict an existed data is meaningless.

The title, tags and descriptions of a viedo are available once it is uploaded. Therefore, these informations can be applied as features for prediction. Then the TF-IDF approach can be used, which provides the contribution of a word to text. The word with higher frequency in a sample text and lower frequency in the whole text is assigned with higher weight, and vice versa. The method focus on the releation between each signle word and the sample text. However, the context information is aborted and all words are considered as independent.

Introduced by [5], the self-attention based approach : Transformer, is proven to perform well in processing texts. It weights the importance of releations among different words, enabling it to extract the context and global information. In general, the model develops the encoder-decoder structure. In the encoder, inputs are converted into tokens, which are the minimum units of information. Then, linear layers are

applied to map the input tokens into three matrixes: Query, Key and Value. And the attention weights are generated by the matrixes.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

In the decoder, a prediction head is used to generate the prediction result from the weights.

To apply Transformer in the views prediction task, the features (title, tags and descriptions) are converted into tokens, which are the basic unit of the text. After that, a pre-trained Transformer model with signle head is adopted. In each epoch of training, the model's performance on the validation set is evaulated and early-stopping is adopted to avoid overfitting.

4 Literature

Similar prediction tasks are completed with the dataset, such as likes prediction, Category Prediciton [2, 3]

Although Transformer based approach archives the best performance, it is limited by huge time and space consumption since each token has interactions with all tokens. Although its self-attention mechanism can obtain the context information, the computation cost increases rapidly as the text seqence increasing. The Receptance-Weighted Key-Value (RWKV) model combines the RNN and self-attention, archives high performance on large scale tasks with acceptable costs [4]. Its RNN structure reduces the time and space consumption and its time decay mechanism performs well on processing long scale information.

5 Experiments and Results

Table 2. Performance Metrics of Models

Model	MSE	MAE	R^2
Single Feature (likes)	0.77	0.63	0.74
Single Feature (comments)	1.19	0.77	0.60
Single Feature (dislikes)	0.76	0.64	0.74
TF-IDF (tags)	0.43	0.33	0.86
TF-IDF-SVD (tags)	1.27	0.42	**0.58**
TF-IDF (description)	0.80	0.31	0.74
TF-IDF-SVD (description)	0.55	0.39	0.82
TF-IDF (title)	0.32	0.28	0.89
TF-IDF-SVD (title)	0.44	0.38	0.85
Transformer (tags)	0.44	0.38	0.85
Transformer (description)	0.30	0.35	0.90
Transformer (title)	**0.18**	**0.27**	0.93

Table 2 provides the performance of 12 different models. The dataset is likes, comments, dislikes, tags, description and titles in US, with size of 40739. After shuffle, 90% of the dataset is devided as the training set and 10% is the validation set. The metrics of MSE, MAE and R^2 are used.

The baselines are in three groups: regression models with signle features, TF-IDF models and TF-IDF-SVD models. For TF-IDF models, the max 20000 imprtant words are used. For TF-ID-SVD models, the Singular Value Decomposition approach is introduced for dimension reduction from 20000 to 5000.

As the result, the Transformer model with titles as features archives the best performance: 0.18 in MSE and 0.27 in MAE, proving the effectiveness of Transformer on texts prediction tasks. Furthermore, all models with titles perform better than other text features, indicating that the title of a video contributes more with its views than tags and descriptions.

Acknowledgments

To Robert, for the bagels and explaining CMYK and color spaces.

References

[1] [n. d.]. *Trending YouTube Video Statistics.* https://www.kaggle.com/datasets/datasnaek/youtube-new/data
[2] [n. d.]. *youtube likes prediction.* https://www.kaggle.com/code/hetulmehta/youtube-likes-prediction
[3] [n. d.]. *Youtube Videos Category Prediciton.* https://www.kaggle.com/code/emirfarukerman/98-accuracy-youtube-videos-category-prediciton
[4] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, Xuzheng He, Haowen Hou, Jiaju Lin, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartlomiej Koptyra, Hayden Lau, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Bolun Wang, Johan S. Wind, Stanislaw Wozniak, Ruichong Zhang, Zhenyuan Zhang, Qihang Zhao, Peng Zhou, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the Transformer Era. arXiv:2305.13048 [cs.CL] https://arxiv.org/abs/2305.13048

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

A Research Methods

A.1 Part One

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi malesuada, quam in pulvinar varius, metus nunc fermentum urna, id sollicitudin purus odio sit amet enim. Aliquam ullamcorper eu ipsum vel mollis. Curabitur quis dictum nisl. Phasellus vel semper risus, et lacinia dolor. Integer ultricies commodo sem nec semper.

A.2 Part Two

Etiam commodo feugiat nisl pulvinar pellentesque. Etiam auctor sodales ligula, non varius nibh pulvinar semper. Suspendisse nec lectus non ipsum convallis congue hendrerit vitae sapien. Donec at laoreet eros. Vivamus non purus placerat, scelerisque diam eu, cursus ante. Etiam aliquam tortor auctor efficitur mattis.

B Online Resources

Nam id fermentum dui. Suspendisse sagittis tortor a nulla mollis, in pulvinar ex pretium. Sed interdum orci quis metus euismod, et sagittis enim maximus. Vestibulum gravida massa ut felis suscipit congue. Quisque mattis elit a risus ultrices commodo venenatis eget dui. Etiam sagittis eleifend elementum.

Nam interdum magna at lectus dignissim, ac dignissim lorem rhoncus. Maecenas eu arcu ac neque placerat aliquam. Nunc pulvinar massa et mattis lacinia.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009