UNIVERSITY OF CALIFORNIA
Los Angeles

Application of Text Mining and BERT on YouTube Fashion Videos

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Yunwen Kai

2021

ABSTRACT OF THE THESIS

Application of Text Mining and BERT
on YouTube Fashion Videos

by

Yunwen Kai
Master of Applied Statistics
University of California, Los Angeles, 2021
Professor Yingnian Wu, Chair

This thesis aims to discover the distinction between high view counts YouTube fashion videos and low view counts YouTube fashion videos. Fashion YouTubers are divided into two groups, Popular Group and Super Popular Group. By Text Mining, sentiment analysis is performed, most mentioned topics for two groups are analyzed, and two common topics "Haul" and "Vlog" are compared within two groups.

There is no clear differences found on sentiment analysis between two groups, and it is beneficial for Popular Group to make more trendy market- oriented videos, while advantageous for Super Popular Group to create content- oriented videos.

Several Bidirectional Encoder Representations from Transformers (BERT) based models are applied and compared on classification tasks' results, including BERT uncased, BERT cased, ALBERT, DistilBERT uncased, DistilBERT cased and RoBERTa. The DistilBERT uncased Multi-modal is preferred on classifying the YouTube video view ranks (high view counts, low view counts) with text and tabular data including title, video length, title length, uppercase ratio in title, topic and sentiment intensity.

The thesis of Yunwen Kai is approved.

Hongquan Xu

Michael Tsiang

Akram M. Almohalwas

Yingnian Wu, Committee Chair

University of California, Los Angeles

2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

YouTube is a User-generated Content (UGC) website, and it is dominating the video market. In the third quarter of 2020, 77 percent of U.S. internet users aged 15 to 25 years and 26-35 years accessed YouTube, and over 70 percent of US internet users aged 36-45 yeas and 46-55 years accessed YouTube [Cec21]. As YouTube's impact expands, it attracts advertising and promotion to the video market. The commercialization includes advertisements played before each video, and the collaboration between company and YouTubers [Hou19]. Because of the rich advertisement and promotion approach, YouTuber becomes a profitable job now with a competitive environment.

The view count is the key metric to evaluate a YouTuber, and title is an important element to attract audience. An appealing title becomes a key for the YouTuber to draw advertisement opportunities. NoxInfluencer gives recommendation on how to name a video to attract views, "Key words should be placed as far as possible to the left of the title; Headings should be completely relevant to the video content; Appropriate inclusion of numbers in titles such as the most popular YouTube video of 2020; If it's a series of videos, mark the episode or chapter; Add your channel name at the end of title" [Nox21]. However, this thesis tries to find the objective relationship between the title and the view counts.

This study focuses on fashion- and beauty- related YouTube videos to explore the relationship between title and video view counts. Beyond one's experience on how to name a video attractively, this thesis demonstrates whether there are common patterns of word choice and sentiment between well performed videos (high view counts) and poorly performed

videos (low view counts), and also, whether the topics are similar within field. Besides the word itself, whether capital words have correlation with video view counts. Also, this study predicts the videos' view rank (high view counts or low view counts) with the deep learning model, BERT, by videos' title and other quantifiable characters, including title length, video length, published dates, sentiment intensity and unsupervised learned topics by Latent Dirichlet Allocation.

The remaining paper is organized as follows. Section 2 introduces how the meta data is extracted from YouTube and processed. Section 3 presents the methodology and models used in the experiment. Section 4 explains the detailed experiment design. Section 5 demonstrates the results from text mining and the classification models. Finally, section 6 draws conclusion from the project.

CHAPTER 2

Data

Python is used for requesting data from YouTube Data API v3, and the code is from Python-Engineer on GitHub [PE19], and then the downloaded dataset is transformed from .json to .csv file. After generating .csv files from python, R language is used for later text mining, and Python is applied for BERT classification tasks. The interests of the thesis are to analyze YouTubers from the fashion area. Because YouTube does not categorize YouTubers, the analyzed YouTubers are selected from the article "100 Fashion YouTubers for Fashion Lovers." by Feedspot Blog. According to Feedspot Blog, they have a research team to rank the influencers by Relevancy, Blog post frequency(freshness), Social media follower counts and engagements, Domain authority, Age of a blog, Traffic Rank and many other parameters [Blo21]. By selecting YouTubers from this article, the category of the content is limited to fashion and beauty areas. However, the influencers on the lists are from different countries, speaking various languages. To control the similarity for both the YouTubers and their audience, only female social media celebrities who are located in the US are selected. People who only discuss single topics like DIY are excluded. The remaining eighteen celebrities' videos are about fashion, beauty and lifestyle. They are divided into two groups, Popular Group and Super Popular Group, which have subscriber counts below or above around one million on the data extraction date. The detailed channel name and subscriber count are shown in Table 2.1. The total video counts for two groups are 3469 and 3520.

Total seven features extracted from YouTube Data are included and analyzed, which are channel title, published at, video title, description, view count, duration and subscriber

Popular Group		Super Popular Group	
channelTitle	subscriberCount	channelTitle	subscriberCount
Naomi Boyer	440000	Jenn Im	2920000
Song Of Style	361000	Amber Scholl	3470000
Ashley Brooke	471000	Summer Mckeen	2250000
AMY LEE	533000	Olivia Jade	1860000
Vanessa Ziletti	349000	Tess Christine	2360000
Fiorella Zelaya	823000	Mel Joy	1260000
Chriselle Lim	747000	Kelsey Simone	1670000
Alex Centomo	638000	Sarah Rae Vargas	1410000
Audrey Coyne	572000	FashionByAlly	973000

Table 2.1: Popular Group and Super Popular Group's lists.

count. The features' descriptions and examples are provided in Table 2.2 below. According to the modeling needs, some metadata is transformed in the later data engineering step.

Variable	Description	Example
channelTitle	The channel name	Naomi Boyer
publishedAt	The video publish date and time	2018-08-23T15:11:37Z
title	Video's Title	Simple Things You Can Do To Look Better — Enhance Your Look
description	Video's description including the information from the content (Example shortened)	In this video, I give you a couple simple tips on the things you can do to look better and enhance your look. Items Mentioned: Teeth Whitening- http://shopstyle.it/l/QWpb http://shopstyle.it/l/QWpf VISIT MY BLOG: www.stylestaycation.com FOLLOW ME.... INSTAGRAM — http://instagram.com/naomiboyer BUSINESS INQUIRIES — naomi@stylestaycation.com
viewCount	The number of views	1000
duration	The length of the video	PT10M4S
subscriberCount	The number of subscribers	440000

Table 2.2: Scraped dataset from YouTube API.

CHAPTER 3

Methodology

3.1 Text Mining Preparation

The basic text mining preparation process includes the corpus preparation, text cleaning, transformation to tibble and unnest tokens, which are stated in Figure 3.1. The cleaning step intends to remove unnecessary characters from the corpus, including transforming the corpus into lower words, removing numbers, punctuation and stop words, stripping white space, and stemming. The common stop words like “for”, “very”, “and”, “of”, “are” etc. do not provide meaningful information, and specific stop words can be removed based on requirements. After the text preparation, further analysis, such as finding frequent words, comparing the wordcloud, NRC sentiment analysis, are performed.

With the cleaned corpus, a tibble is generated. Later, tokens, the separated words, are created based on document tibble. The example corpus, tibble and token are shown in Figure 3.2.

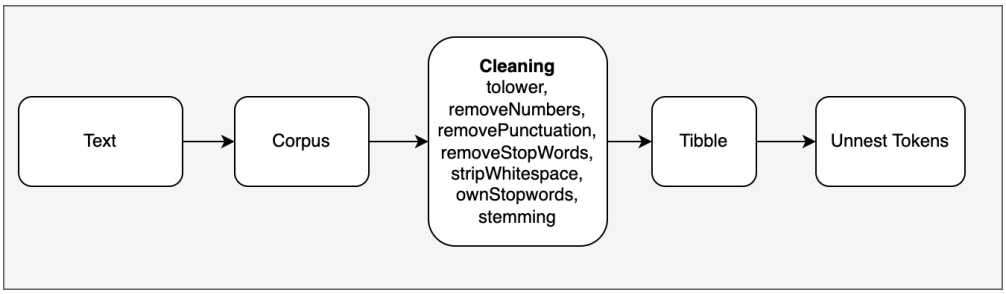
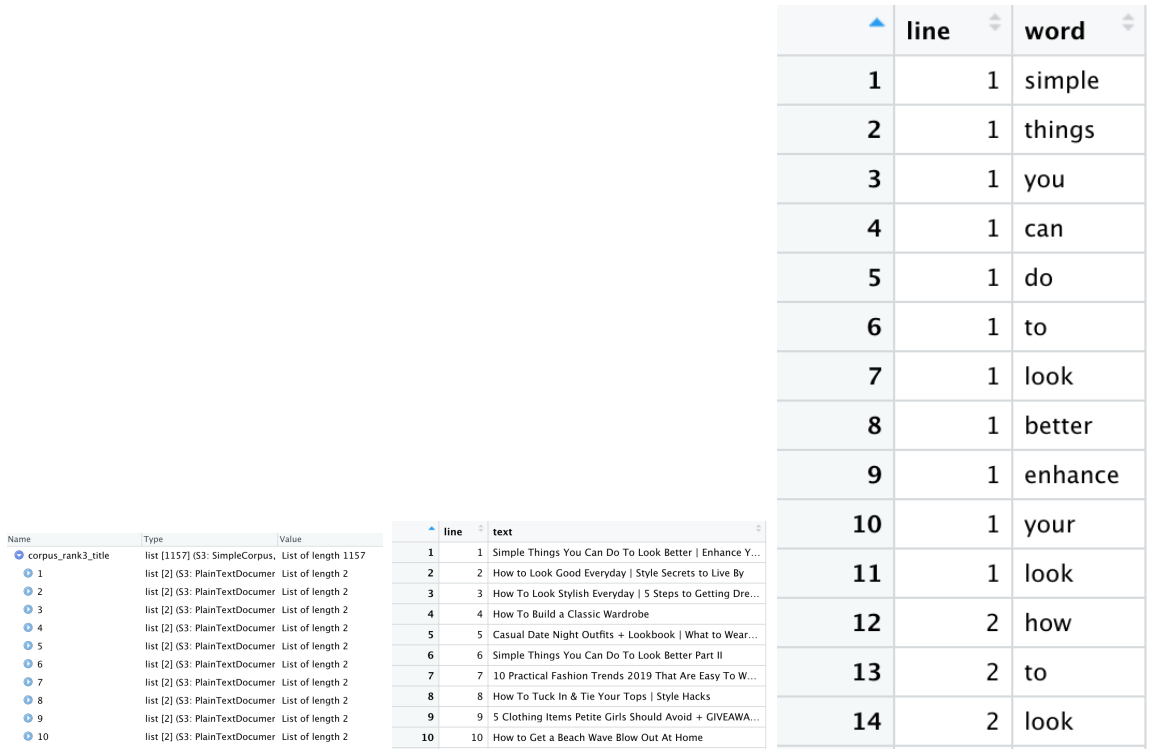


Figure 3.1: Text Mining process.



Name	Type	Value
corpus_rank3_title	list [1157] (S3: SimpleCorpus,	List of length 1157
1	list [2] (S3: PlainTextDocumer	List of length 2
2	list [2] (S3: PlainTextDocumer	List of length 2
3	list [2] (S3: PlainTextDocumer	List of length 2
4	list [2] (S3: PlainTextDocumer	List of length 2
5	list [2] (S3: PlainTextDocumer	List of length 2
6	list [2] (S3: PlainTextDocumer	List of length 2
7	list [2] (S3: PlainTextDocumer	List of length 2
8	list [2] (S3: PlainTextDocumer	List of length 2
9	list [2] (S3: PlainTextDocumer	List of length 2
10	list [2] (S3: PlainTextDocumer	List of length 2

	line	text
1	1	Simple Things You Can Do To Look Better | Enhance Y...
2	2	How to Look Good Everyday | Style Secrets to Live By
3	3	How To Look Stylish Everyday | 5 Steps to Getting Dre...
4	4	How To Build a Classic Wardrobe
5	5	Casual Date Night Outfits + Lookbook | What to Wear...
6	6	Simple Things You Can Do To Look Better Part II
7	7	10 Practical Fashion Trends 2019 That Are Easy To W...
8	8	How To Tuck In & Tie Your Tops | Style Hacks
9	9	5 Clothing Items Petite Girls Should Avoid + GIVEAWA...
10	10	How to Get a Beach Wave Blow Out At Home

	line	word
1	1	simple
2	1	things
3	1	you
4	1	can
5	1	do
6	1	to
7	1	look
8	1	better
9	1	enhance
10	1	your
11	1	look
12	2	how
13	2	to
14	2	look

Figure 3.2: Example corpus(left), tibble(middle) and token(right).

3.2 Sentiment Analysis/Intensity

Sentiment Analysis is a hot Natural Language Processing field that analyzes texts' emotions and inherent opinions. Sentiment Intensity can be calculated by sentiment classification on positivity and negativity [HG14]. Fitting text or document into a trained/labeled dictionary is a common way to score sentiment. Valence Aware Dictionary for Sentiment Reasoning (VADER), which incorporates classic benchmarks like LIWC and ANEW, is a rule-based sentiment dictionary trained on social media text like Twitter and Facebook [HG14]. VADER generates polarity from text including positive, neutral and negative scores, and then normalizes the three scores as a compound score ranging from -1 to 1 as the overall sentiment intensity of a text. It is also reported in the original paper that VADER performs better than human raters with higher F1 classification accuracy [HG14]. Later in the experiment, the compound score would be calculated from the VADER library for each title, which stands

for the title's sentiment intensity.

3.3 LDA - Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is an unsupervised topic modeling machine learning method. LDA is a Bayesian probabilistic model [HBB10]. In order to generate topics from many documents, the "K" (number of topics) needs to be set manually first. Second, the LDA analyzes and extracts "K" topics from all the documents. Each topic would be a mixture of words. According to the distribution over topics, topic weights over each document and posterior distribution, the LDA can match each document to synthesized topics with probabilities [HBB10]. For example, when the "K" is set to 3, document one can be 70% topic one, 20% topic two and 10% topic three. The thesis choose the highest probability topic as the title's assigned topic. To discover whether the topic would impact on a video's view counts, LDA is performed to generate topic class for each video, and the topic class is one of the features to predict video's view rank (high view count videos or low view count videos).

3.4 Transformer

Transformer is a popular architecture now, which was first introduced by Google. The Transformer is a breakthrough because it only depends on self-attention mechanisms, which connect the encoder and decoder, and it abstains from recurrence [VSP+17]. The attention is calculated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.1}$$

where matrix Q is a set of packed queries, matrices K and V are packed keys and values, and d_k is the dimension of queries and keys [VSP+17].

With the multi-head attention, the Transformer does not consider the distance in the input or output sequences, but the global dependencies between input and output [VSP+17]. The reduced sequential computation allows more parallelized computation, which significantly increases the training speed and saves time.

$$\begin{aligned} MultiHead(Q, K, V) &= Concat(head_1, \cdots, head_h)W^O \\ where\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3.2}$$

where $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^K \in R^{d_{model} \times d_k}$, $W_i^V \in R^{d_{model} \times d_v}$, $W_i^O \in R^{hd_v \times d_{model}}$. There is $h = 8$ parallel attention layers, and $d_k = d_v = d_{model}/h = 64$, where d_v is the dimension of values [VSP+17].

As the model structure suggested in Figure 3.3, the Transformer incorporates encoder and decoder. Both the encoder and decoder include six identical layers. The encoder has two sub-layers, multi-head self-attention mechanism and position-wise fully connected feed-forward network, while the decoder's one sub-layer only executes multi-head attention mechanism.

During the training, the Transformer applies Adam optimizer and regularizations like residual dropout and label smoothing [VSP+17]. Most importantly, Transformer shows good results, which outperform traditional ensembles in several datasets.

3.5 BERT

Bidirectional Encoder Representations from Transformers (BERT) is a Transformer based Machine Learning model, and it only includes encoder from the Transformer model, so it only uses attention mechanisms and feed-forward layers [DCLT18]. It considers two sentence segments, sequences and tokens, as a single input. The BERT framework includes two steps, pre-training and fine-tuning. It is pre-trained on large uncompressed text, including BOOKCORPUS and English WIKIPEDIA [DCLT18]. The pre-train BERT used maksed LM and next sentence prediction (NSP), and then the pre-trained model parameters are

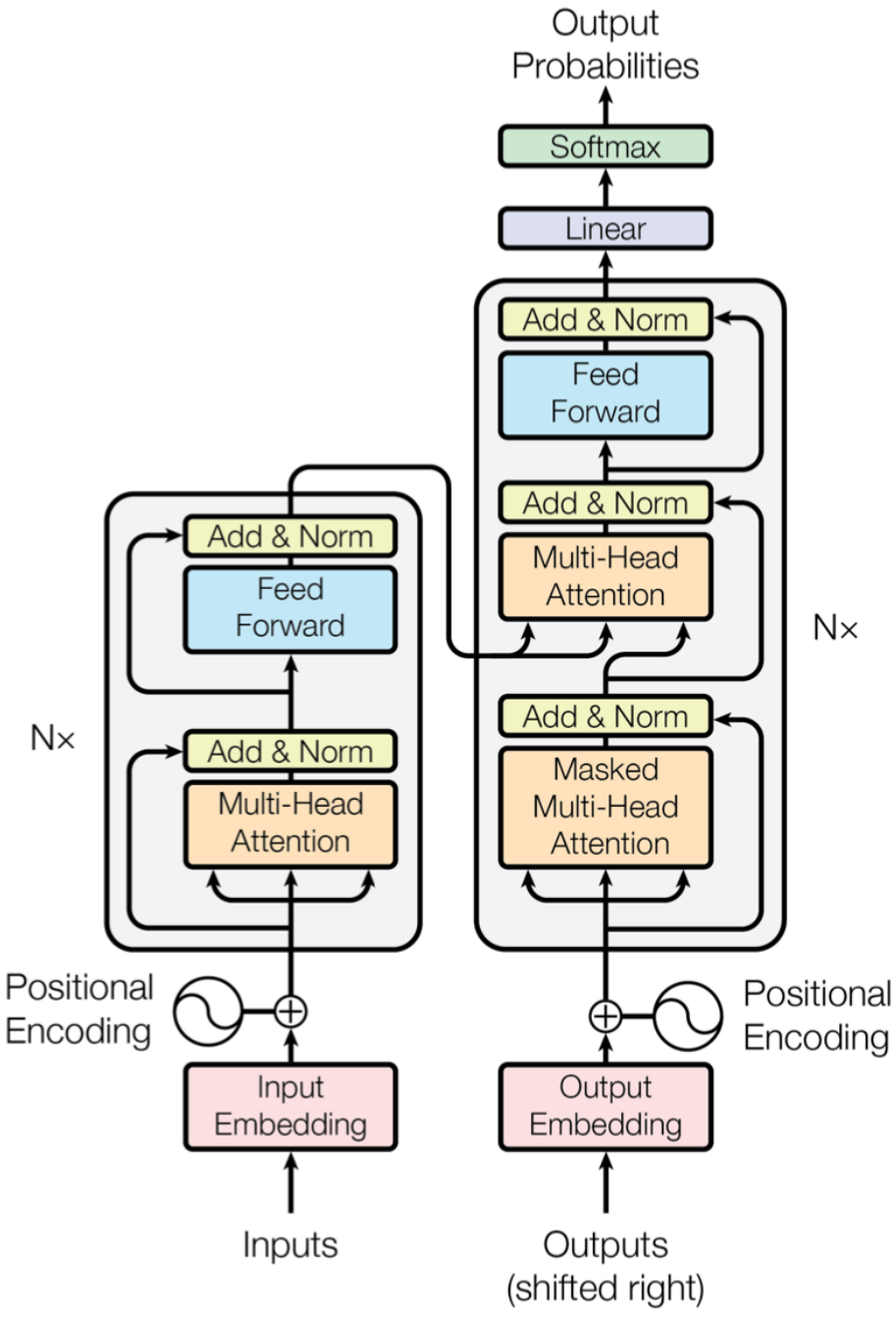


Figure 3.3: The Transformer - model architecture[VSP+17]

fine-tuned in the fine-tuning step, which is inexpensive [DCLT18].

$$h = W \times x \tag{3.3}$$

where h is the thought vector, W is the encoding matrix and x is the input vector [DCLT18].

$$\Delta h_m = \sum_{k=1}^{M} attention_{m \to k} \times value_k \tag{3.4}$$

3.5.1 RoBERTa

Robustly optimized BERT approach (RoBERTa) is a pre-trained BERT model with longer and bigger batches and longer sequences, which removes the next sentence prediction objectives and changes the masking pattern [LOG+19]. According to Liu's experiment, RoBERTa with improved pretraining procedure lead to a better downstream task performance on GLUE, RACE, SQuAD and CC News than the traditional BERT.

3.5.2 DistilBERT

Although larger models usually result in better results, the environmental cost and real-time computation become challenging [SDCW19]. Sanh and his team proposed a lighter DistilBERT model, which adopts BERT architecture. DistilBERT design comprises student architecture, student initialization and distillation [SDCW19]. The three techniques reduce the number of layers, spot the simple initialization and exclude next sentence prediction. The training loss L_{ce} is set as follows:

$$L_{ce} = \sum_i t_i * log(s_i) \tag{3.5}$$

where $t_i (resp. s_i)$ is a estimated probability [SDCW19].

The softmax-temperature p_i is set as follows:

$$softmax\ temperature: \ p_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{3.6}$$

where T is the smoothness and z_i is the model score for the class i [SDCW19].

The simple architecture design makes DistilBERT 40% smaller and 60% faster than the original BERT, while keeps 97% language understanding abilities [SDCW19].

3.5.3 ALBERT

A Lite BERT (ALBERT) is an enhanced BERT model by reducing parameters [LCG+19]. ALBERT incorporates basic BERT architecture, but innovates in three aspects, 1) factorized embedding parameterization, 2) cross-layer parameter sharing and 3) Inter-sentence coherence loss [LCG+19]. As stated by Lan, the above transformations decompose embedding parameters into smaller matrices, improve parameter efficiency and increase language understanding by applying next-sentence prediction (NSP) loss and sentence-order prediction (SOP) loss [LCG+19]. As a result, although the model size is increased, the ALBERT architecture consumes lower memory and leads to a better result on downstream tasks.

3.5.4 Multimodal Learning

Since the text data is usually not solely exists in real life, Multimodal Learning is designed for regression or classification tasks on text and tabular data with Transformer architecture [GB21].

As shown in Figure 3.4, Multi-modal takes text input first. After processing with the Transformer model, the text output is reprocessed with categorical and numerical variables, which then generates prediction on the combining module [GB21]. According to Gu's experiments on three datasets, the Transformer prediction results on only text are always lower than the Multi-modal prediction results, which takes text and structured data as inputs.

Since the YouTube meata data consists of both text and tabular data, Multi-modal from Gu is employed for the classification tasks. Multi-modal incorporates Hugging Face's [WDS+19] existing API. In total, six BERT based models are tested and analyzed, includ-

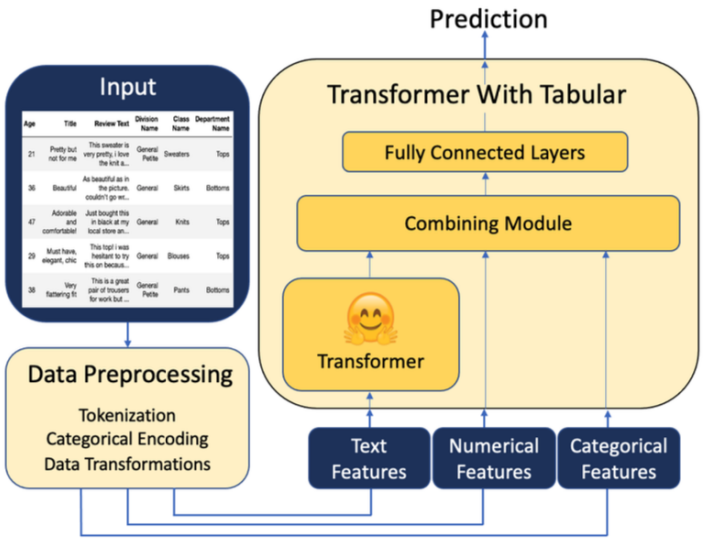


Figure 3.4: The framework of Multimodel-Toolkit [GB21]

ing BERT uncased, BERT cased, ALBERT, DistilBERT uncased, DistilBERT cased and RoBERTa.

CHAPTER 4

Experiment

The experiment design comprises the text mining process and classification modeling, which consists of feature engineering and modeling.

4.1 Text Mining

4.1.1 Analysis on Video Description

The descriptions for beauty- and lifestyle- videos serve as an information box, and YouTubers usually provide links for products they mentioned in the videos. The analysis for video descriptions is straightforward, shown in Figure 4.1. First, transform the text data from .csv files into corpus for both groups, Popular Group and Super Popular Group. Second, clean the corpus by converting all letters to lower letters, remove numbers, remove punctuation, remove English stop words, strip white-spaces, remove own stop words and stem. The self-defined stop words include YouTubers' names, since they put their own name in descriptions for better searchability. Third, transform the cleaned corpus into tibble and unnest tokens. When the text is converted to cleaned tokens, the most frequent terms can be found in two groups and see what the similarities and differences are.

4.1.2 Analysis on Video Title

To find out hot videos' common characteristics on titles, the videos with high view counts is focused. Since every YouTuber has a different number of followers, categorizing hot videos

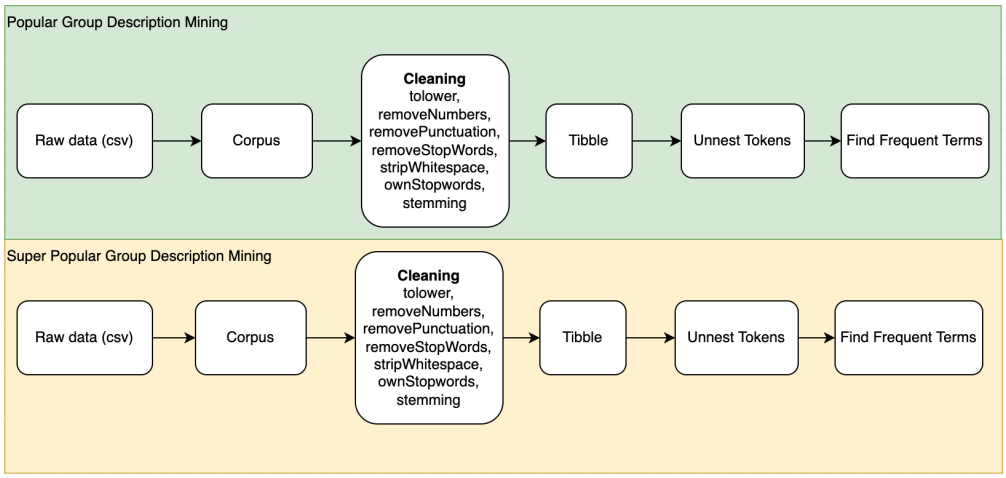


Figure 4.1: Processing steps for descriptions

directly by the view counts is unreasonable. As a result, the videos are divided into three groups by the rank of view counts for each YouTuber. The best performed videos (with the highest view count) and worst performed videos (with the lowest view count) are compared and analyzed.

There are four groups after splitting Popular Group and Super Popular Groups into two video groups, the most viewed videos group and the least viewed videos group. The same process will be applied on the four groups, shown in Figure 4.2. First, create corpus, clean the corpus, create tibble and unnest tokens just like the processing steps for descriptions. With the cleaned tokens, the NRC sentiment analysis is applied to check whether there are significant differences in titles' mood for the four groups. Then, the Wordcloud is plotted with the minimum number of words being 20 for all four groups. "Vlog" and "haul", the two big topics for beauty and lifestyle videos, are analyzed. With the same applications for the four groups, this study compares the results and understands what common traits hot videos' titles share.

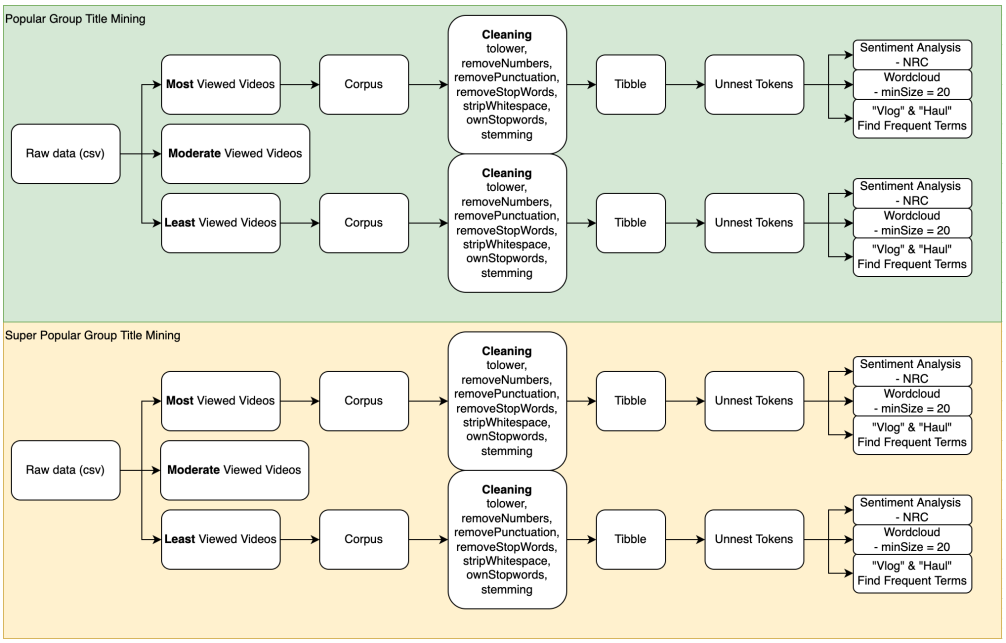


Figure 4.2: Processing steps for titles

4.2 Modeling

4.2.1 Feature Engineering

Features are transformed for prediction models. Since the subscriber counts are varied for YouTubers, instead of predicting the exact view counts, the goal is to classify whether a video is a well performed video (with high view counts) or a poorly performed video (with low view counts). The response variable is the view rank, which quantile (equally) cuts each YouTuber's videos into two groups, 0 for low view counts videos and 1 for high view counts videos. The published time is converted to "days passe" by subtracting the publish date from the data extraction date, "2021-03-06". Video length only keeps the minute. For example, "PT10M4S" is changed to "10", which stands for a 10 minute long video. Title length is the word count of a title. Capital letters ratio is also calculated by dividing the capital letter counts to the title length (word counts). The capital letter ratio can be greater than 1, because usually the first letter of a word is capitalized, and a higher number means

more capital letters, which may be caused by capital words. For example, the uppercase ratios are 1 and 3.5 for "Simple Outfit" and "SIMPLE Outfit".

Besides basic feature engineering, the Latent Dirichlet Allocation (LDA) from Scikit-Learn is modeled [PVG+11], and Sentiment Intensity Score from nltk is calculated [HG14]. Among all the titles, LDA is performed for topic modeling to extract ten topics in total, and the topics would be assigned to video titles with probabilities. Each video title would be categorized as the topic with the highest probability. The vader lexicon from nltk library helps calculate a video title's sentiment intensity by normalizing positive, neutral and negative mood. The final input dataset example is provided in Table 4.1.

Feature	Description	Type	Example
text	cleaned text	string	How To Look Stylish Everyday Steps to Getting Dressed
vid-length	video length	integer	10
title-len	title word count	integer	11
uppercase-ratio	capital letter ratio	float	1.1
max	topic allocation from LDA (1 to 10)	integer-category	1
compound	sentiment intensity from ntlk	float	0.4404
label-response	view count rank (0 or 1)	integer-category	0

Table 4.1: Input dataset.

4.2.2 Training

The data is separated into training and validation data by the ratio 0.8 and 0.2. The modeling is a binary classification task, which is executed in Multi-modal [GB21] with bert-base-uncased, bert-based-cased, albert, distilbert-based-uncased, distilbert-based-cased and roberta-base [WDS$^+$19]. After comparing the classification results on the validation dataset, the best model is picked. For every model, the training data is trained with a mini-batch size of 32 and 5 epochs. Based on the Multi-modal setting, the features are divided into text, categorical data and numerical data, and then trained. According to the Multimodal default setting, the Relu activation function is used, dropout ratio is set as 0.1.

To compare and evaluate the models, F1 [S$^+$07] and ROC AUC (Area Under the Receiver Operator Characteristic Curve) [Faw06] are plotted with steps. Precision, Recall and False Positive Rate are insightful criteria. Both F1 and ROC AUC are common metrics to evaluate classification problems since they combine Precision, Recall and False Positive Rate. Precision measures the fraction of true positive labeled samples among all positive labeled samples, while the recall measures the fraction of the true positive labeled samples among all correctly labeled samples. F1 score, $F1 = 2 * \frac{precision * recall}{precision + recall}$, balance between Precision and Recall, which is high when both Precision and Recall are high. ROC AUC compares the True Positive Rate (Recall) and False Positive Rate. The higher ROC AUC value stands for a greater distinction between true positives and true negatives [Faw06]. Usually the imbalance data has a larger impact on F1 than ROC AUC. Since the dataset has an equal number of 0's (low view counts) and 1's (high view counts), both F1 and ROC AUC can evaluate the results well [YP17].

CHAPTER 5

Result and Analysis

5.1 Text Mining Data Analysis

5.1.1 Similarities and Differences in Description

The most frequent words showing in the description are listed in Table 5.1. There are similarities and differences. Two groups share 4 types of words.

Popular Group	Super Popular Group
Type 1: instagram snapchat twitter facebook Type 2: link Type 3: subscribe channel Type 4: sponsor	
business inquiries dress outfit look vlog blog	thank watch hope love enjoy like

Table 5.1: Similarities and difference in descriptions for two groups.

Social media account Staying connected with subscribers can bond the community. Videos are updated only once or twice a week on YouTube, while YouTubers are eager to connect with followers more often, so they listed their other social media accounts in videos' descriptions. If the audience likes a YouTuber, they would follow the YouTuber on different platforms. Also, guiding followers to other social media, like Instagram can also

increase YouTubers' advertisement revenue on different social medias. Beauty gurus can not only monetize their videos [Hou19], but also monetize their social media posts if they want.

Products' link Introduced or used products' links are mostly provided for two reasons, first, viewers would wonder about the products' information; second, YouTubers receive rebates when viewers click the link or purchase through the link. The rebate links are not a secret for content creators or audiences, and both YouTubers and video viewers need it.

Ask for support The metadata, view count, subscriber count, comment count, like count and dislike count influence the evaluation of a YouTuber's business value. Beauty gurus usually ask for support at the end of a video, and they also mention a lot in the description. When there is a new viewer, they would want to turn them into subscribers.

Sponsorship Sponsorship is sensitive in beauty and lifestyle videos. Since there are too many soft advertisements embedded in the videos. To bond with the community, YouTuber usually specify whether the video is sponsored or not.

5.1.2 Popular Group's Uniqueness in Description

Emojis YouTubers from Popular Group use emojis or symbols in description. Four out of nine people from Popular Groups use emojis in their description, while none of YouTubers form Super Popular Group uses.

Business Cooperation Because of the relatively low number of subscribers, Popular Group provides contact information after business inquiries to attract business cooperation.

Searchability Since the words in descriptions also can be searched, the Popular Group amplifies video topics in description again to increase video exposure when audiences search

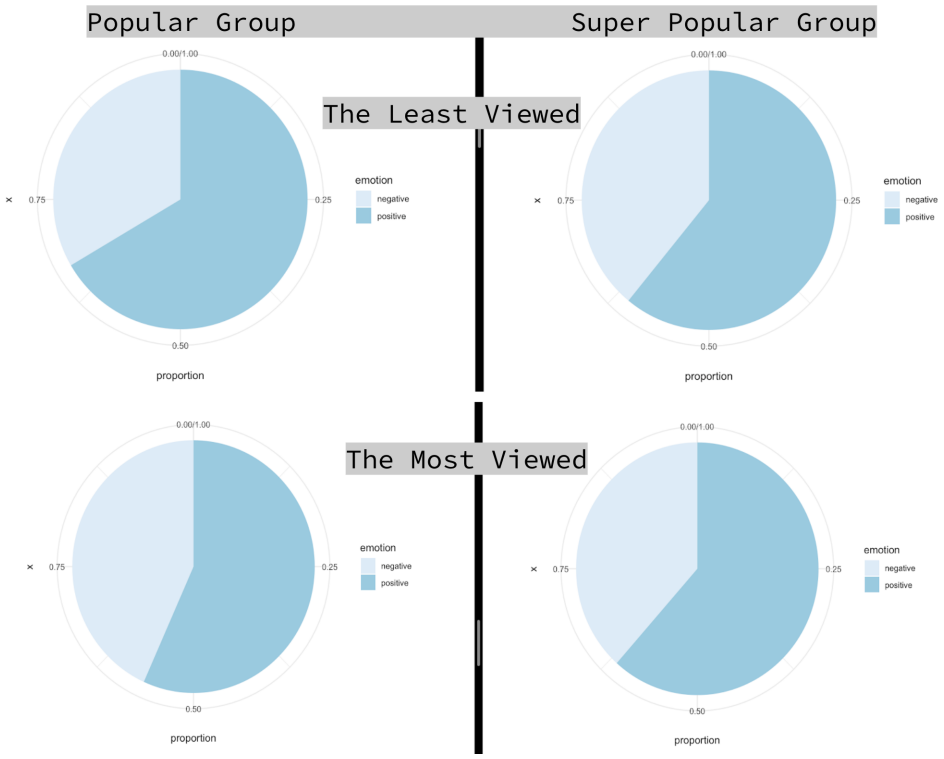


Figure 5.1: Comparison of sentiment among four groups, the least viewed videos in Popular Group, the most viewed videos in Popular Group, the least viewed videos in Super Popular Group and the most viewed videos in Super Popular Group.

related topics.

5.1.3 Super Popular Group's uniqueness in Description

Connection Compared to Popular Group, Super Popular Group focus more on connecting with followers, by putting sentences like "Thank you for watching", "Hope you love/enjoy/like it".

5.1.4 No Sentiment Preference

The NRC-Sentiment Analysis suggests that there is no fixed pattern for audience' sentiment preference between the four groups: Popular Group's most viewed videos, Popular Group's least viewed videos, Super Popular Group's most viewed videos, Super Popular Group's least viewed videos. As Figure 5.1 shows, for Popular Group, the most viewed videos' titles have fewer positive words than the least viewed videos' titles have. However, the opposite happens for Super Popular Group, the most viewed videos' titles have more positive words than the least viewed videos have. Also, there is no clear connection between the positive and negative words used in Popular Group's titles and in Super Popular Group's titles. The sentiment and emotions in the video titles usually propose the video tone and users' preference to the video's mood depends on their feelings. When designing the recommendation algorithm, it is better to recommend the videos that are more related to users' current emotions [CCY17]. No single certain mood wins all, but the appropriate mood delivered to the audience at the right timing increases the view counts. As a result, sentiment analysis or sentiment classification matters more in recommendation algorithms, not content creation.

5.1.5 Normality of Beauty Contents

The frequent words showing in the Wordcloud (Figure 5.2) gives similar words in titles for all four groups, Popular Group's most viewed videos, Popular Group's least viewed videos, Super Popular Group's most viewed videos, Super Popular Group's least viewed videos. Since titles are usually short, the video's topics are mostly included in the video titles. The main topics in the sample include "vlog", "haul", "make up", "favorite", "fashion" etc., and other frequent words are related to the topics, such as "look", "fall", "plus size" etc. Hou [Hou19] indicates the normality in beauty- and lifestyle- related videos, which many beauty gurus catch up with several trendy or viral topics to increase searchability.

Even though the topics are similar, the word choices are distinct. The same conditions are

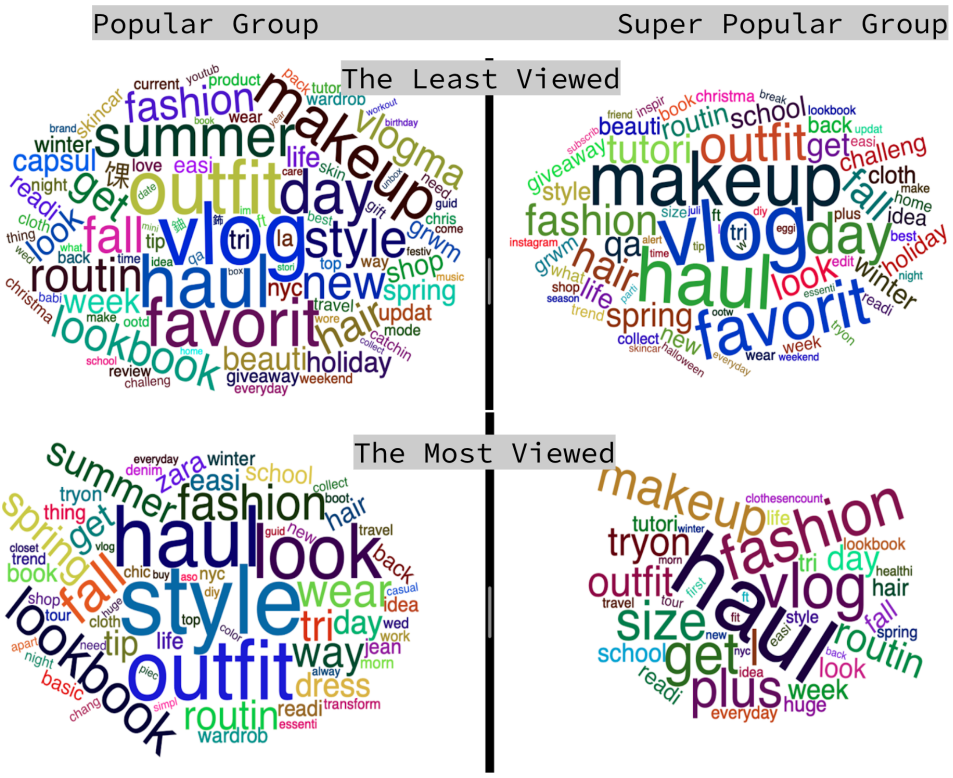


Figure 5.2: Comparison of Wordcloud among four groups, the least viewed videos in Popular Group, the most viewed videos in Popular Group, the least viewed videos in Super Popular Group and the most viewed videos in Super Popular Group.

Video Count for "vlog" and "haul" Series

VLOG	"vlog" video	Least Viewed	Most Viewed
Popular Group	282	128	58
3469	8.13%	45.39%	20.57%
Super Popular Group	278	95	101
3520	7.90%	34.17%	36.33%

HAUL	"haul" video	Least Viewed	Most Viewed
Popular Group	296	72	110
3469	8.53%	24.32%	37.16%
Super Popular Group	430	101	158
3520	12.22%	23.49%	36.74%

Figure 5.3: Video count for “Vlog” and “Haul” series videos.

implemented for the four groups, with minimum word size 20. While the size of Wordcloud is different. Words are more diverse in less viewed or Popular Group video’s title. Popular Group use more words than Super Popular Group use, and the least viewed videos’ titles use more words than the most viewed videos titles use. Overall, the Popular Group’s least viewed videos (top left) has the most words in the Wordcloud, while the Super Popular Group’s most viewed videos (bottom right) has the least words in the Wordcloud. Popular Group’s videos and the least viewed videos have larger sizes of frequent words than Super Popular Group’s videos and the most viewed videos. The normality of topics signifies the homogeneous content, but the view counts of videos verify the strategies’ significance.

5.1.6 Further Analysis on “Vlog” and “Haul”

There are two general categories of fashion videos, market-oriented and content-oriented [GR17]. “Vlog” is a content-oriented video series and “haul” is a market-oriented topic. The word “vlog” comes from “video of blog”, which records personal routines through video.

The frequencies of two types of videos for Popular Group and Super Popular Group are shown in Figure 5.3. The percentages of both series (7.90% and 12.22%) for Super Popular Group are much higher than Popular Group’s two series videos (8.13% and 8.53%). With the fact that Super Popular Group has less frequent words, Super Popular Group makes

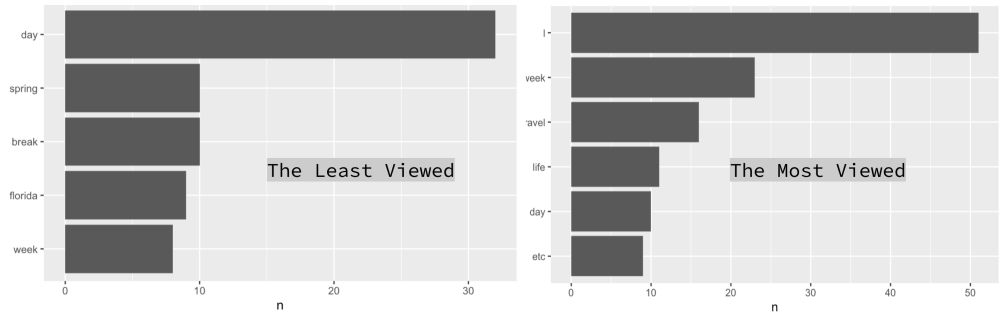


Figure 5.4: Comparison of "Vlog" series videos for Super Popular Group.

more series videos than Popular Group does.

For "vlog", by comparing the most viewed videos counts and the least viewed videos counts, it shows that Popular Groups' vlog are not favored. Popular group's vlog videos are more likely to be in the least viewed video list than in the most viewed video list. While the chance is similar for Super Popular Group. Things get better in "haul" series videos. For both Super Popular Group and Popular Group, more haul videos are categorized as the most viewed videos than in the least viewed videos.

Combining the statistics in the market- and content-oriented videos. When the follower is not enough, market-oriented videos can attract more volume. García-Rapp [GR17] suggests that know-how content tends to be searched by non-subscriber viewers, because the audience already knows the series of videos; when the audience turned to loyal subscribers, vlog videos maintain them. With an increasing number of subscribers, content-oriented videos can help bond the community, since beauty gurus communicate with audiences through vlogs [MOG+08].

5.1.7 Personal "Vlog" is Favored

Due to the low number of Popular Group's "vlog" video, only the Super Popular Group's "vlog" videos are analyzed. As Figure 5.4 shows, there is an overlap of words in two groups, the least viewed videos and the most viewed videos, such as "week" and "day". The topics

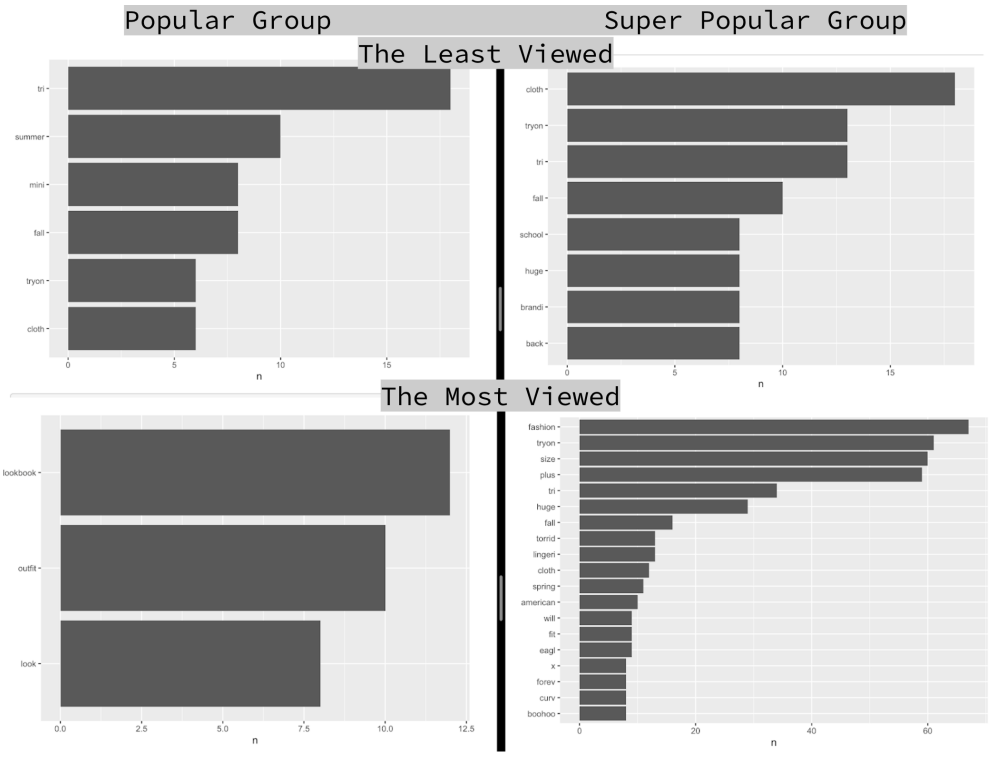


Figure 5.5: Comparison of "Haul" series videos among four groups, the least viewed videos in Popular Group, the most viewed videos in Popular Group, the least viewed videos in Super Popular Group and the most viewed videos in Super Popular Group.

under "vlog" series are limited since it's the film of daily life. However, comparing the most common words for the two groups, "day" and "I" appear more in the most viewed videos. Audiences prefer personally exciting stuff starting with "I", like "I moved to a new city". Since the main targeted audience for vlogs is loyal subscribers, popular community-oriented contents need some exposure of personal life to interest the audience.

5.1.8 "Haul" - Expand Series Videos

Comparing the four groups' frequent word numbers (Figure 5.5), the number of words in Super Popular Group is much more than the number of words in Popular Group, which contradicts the fact we see the general word counts of titles. Generally, the number of words

is more in Super Popular Group or the most viewed videos. Because the number of the most viewed videos is more than the number of the least viewed videos, it is comprehensible that the number of words is more in the most viewed video list than in the least viewed video list. Combining the fact that Popular group has more various words overall, YouTubers should make series videos, and expand the topic within the trendy topics.

5.2 BERT Model Classification Comparison

This section presents the results of Multimodal classification on view ranks (high view counts, low view counts) with the video titles and other tabular features with BERT models including BERT uncased, BERT cased, ALBERT, DistilBERT uncased, DistilBERT cased and RoBERTa.

For each model, F1 and ROC AUC (Area Under the Receiver Operator Characteristic Curve) are displayed with training steps as the x-axis. The training time and training loss are compared between all models at the end. The Figure 5.6 shows the corresponding relationship between epochs and steps. In total, there are 5 epochs and 875 steps, and each epoch consists of 175 steps.

5.2.1 BERT uncased and BERT cased

Both the F1 and AUC of the BERT uncased model reach their highest points around step 300, which is around epoch 1.5 (Figure 5.7). The F1 score attains 0.72. The ROC AUC rises to 0.71, which means there is a 71% chance that the BERT uncased model will be able to distinguish between positive and negative classes. The ROC AUC is relatively stable through the whole epochs, while the F1 score drops significantly after step 300, which may be because of Recall drops. Also, there is an overfitting problem, the evaluation loss increases dramatically after step 300.

The F1 and ROC AUC for BERT cased peak at around step 500, which is around epoch

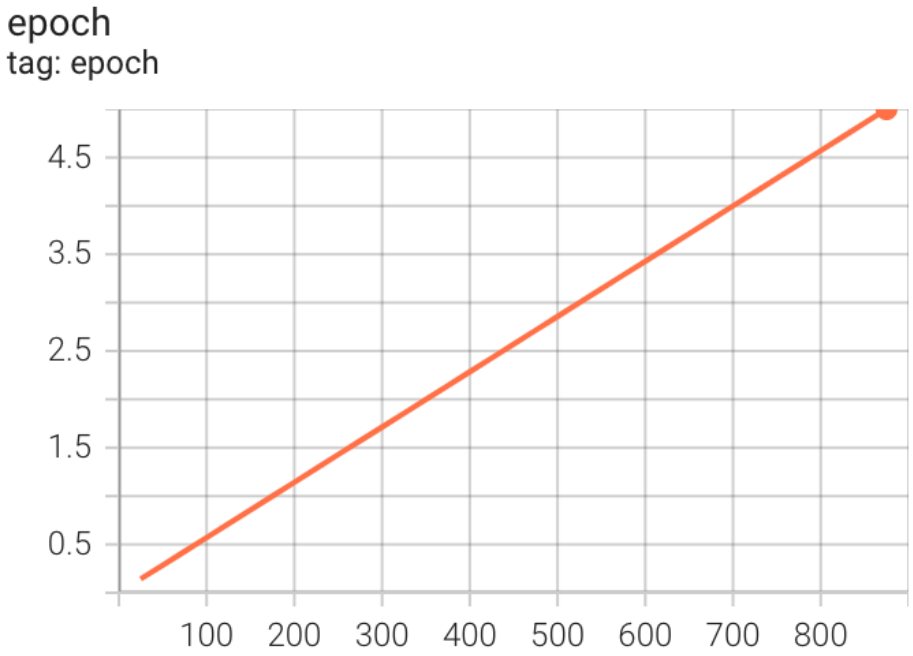


Figure 5.6: Corresponding relationship between epoch and steps.

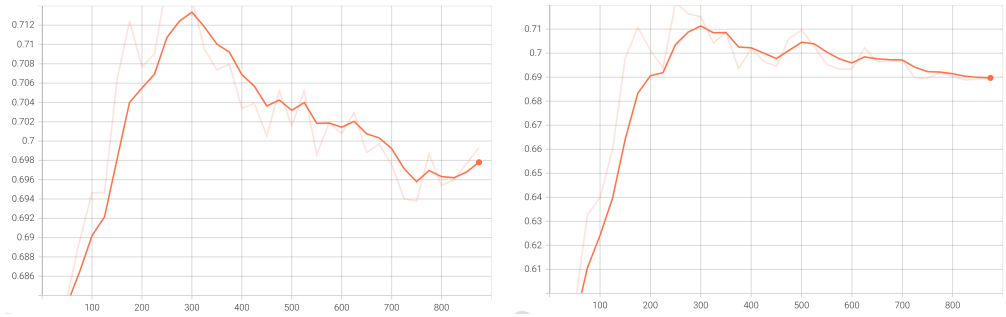


Figure 5.7: F1(left) and ROC AUC(right) results for BERT uncased model.

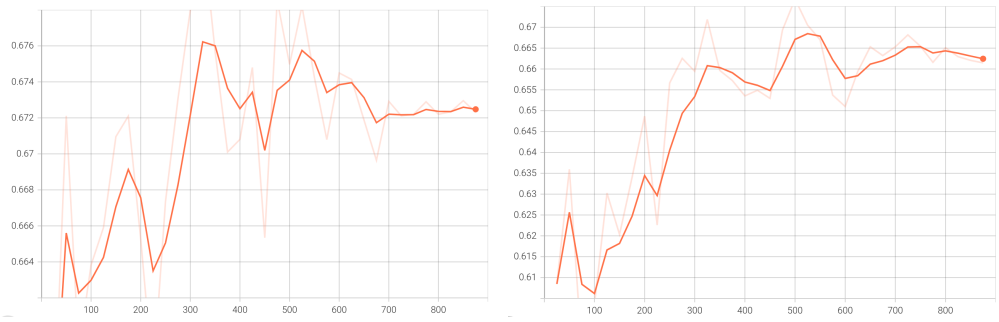


Figure 5.8: F1(left) and ROC AUC(right) results for BERT cased model.

3 (Figure 5.8). The F1 is 0.67 and the ROC AUC is 0.67 as well, which means there is a 67% chance that the BERT cased model will be able to distinguish between positive and negative classes. There are no considerable drops within the 5 epochs. Both the F1 and ROC AUC are lower for the BERT cased model compared to BERT uncased model, which contradicts usual assumptions that the upper or lower letters matter in classification of view counts rank. Also, with the fact that uppercase ratio is included as one feature, further cased analysis might be an extra.

5.2.2 RoBERTa

The F1 for RoBERTa reach their peaks at around step 350, which is around epoch 2, while ROC AUC reaches the highest point later at step 550, which is around epoch 3 (Figure 5.9). The highest F1 score is 0.71 and the ROC AUC is 0.69, which means there is a 69% chance that the RoBERTa model will be able to distinguish between positive and negative classes. The F1 score drops quickly after step 300, and the ROC AUC increases stably for the whole 5 epochs. Checking the details, the true negative and false negatives increase along the training process, while the true positive and false positive decrease significantly after step 300.

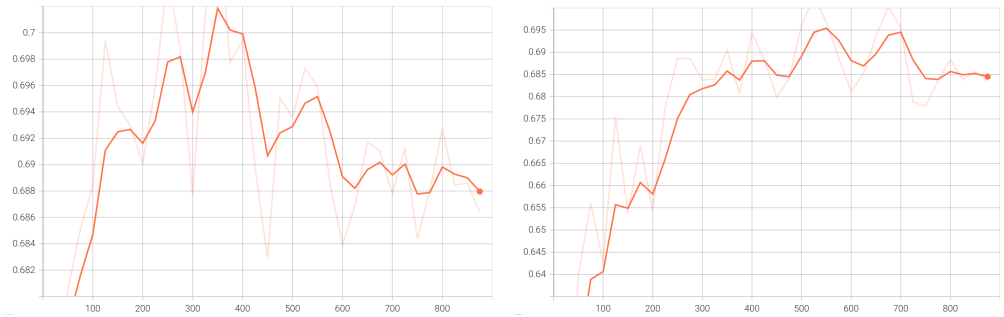


Figure 5.9: F1(left) and ROC AUC(right) results for RoBERTa model.

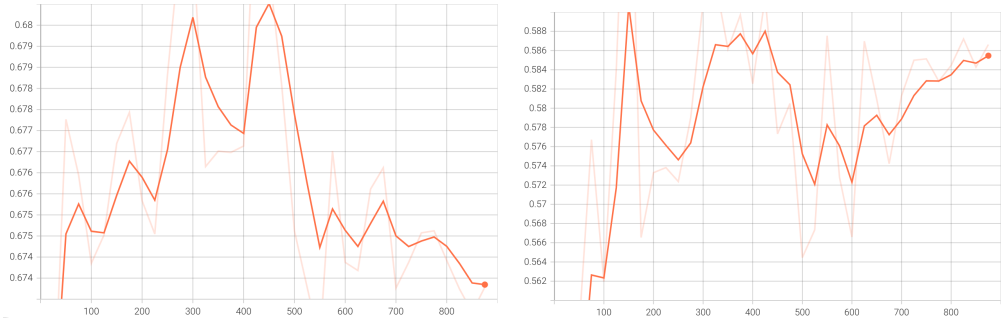


Figure 5.10: F1(left) and ROC AUC(right) results for ALBERT model.

5.2.3 ALBERT

The F1 and ROC AUC for ALBERT reach their high points at around step 450, which is around epoch 2.5 (Figure 5.10). The F1 is 0.685 and the ROC AUC is 0.59, which means there is a 59% chance that the ALBERT model will be able to distinguish between positive and negative classes. Compared to the balanced F1 and ROC AUC, the ROC AUC for ALBERT is relatively lower than F1.

5.2.4 DiltilBERT cased and DiltilBERT uncased

The F1 and ROC AUC for DistilBERT cased reach their high points at around step 300, which is around epoch 1.5 (Figure 5.11). The F1 is 0.684 and the ROC AUC is 0.64, which means there is a 64% chance that the DistilBERT cased model will be able to distinguish

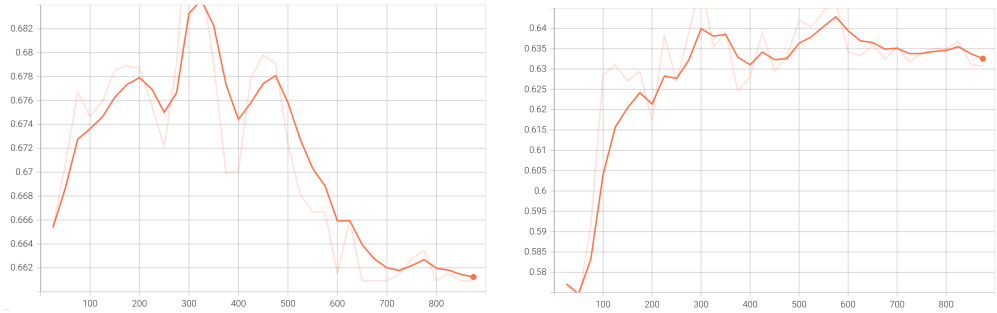


Figure 5.11: F1(left) and ROC AUC(right) results for DistilBERT cased model.

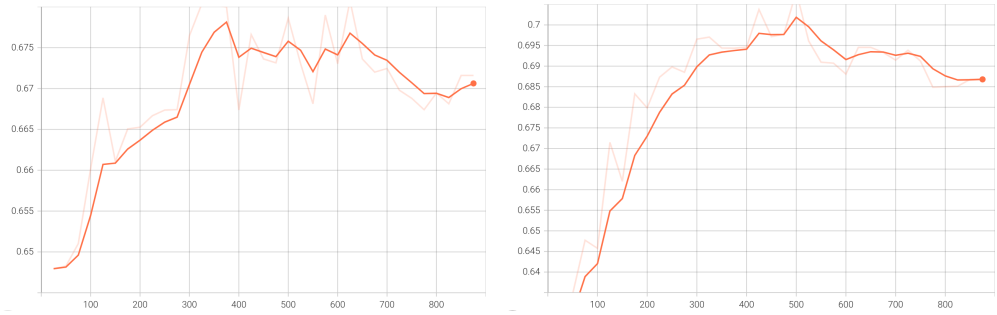


Figure 5.12: F1(left) and ROC AUC(right) results for DistilBERT uncased model.

between positive and negative classes.

The F1 and ROC AUC for DistilBERT uncased reach their peaks at around step 500, which is around epoch 3 (Figure 5.12). The F1 is 0.7 and the ROC AUC is 0.675, which means there is a 67.5% chance that the DistilBERT uncased model will be able to distinguish between positive and negative classes.

Similar to BERT cased and uncased results, DistilBERT uncased model has a better and stable performance in both F1 and ROC AUC than DistilBERT cased model. The scores are higher and there is no noticeable drops for DistilBERT uncased model.

5.2.5 Summary

Comparing all six BERT models' results (Table 5.2) on evaluation datasets' F1 and ROC AUC, 1) the uncased model has better performance than on the cased model; 2) BERT uncased, RoBERTa and DistilBERT uncased are the top three performed models; 3) the training times are largely depend on model structure. DilstilBERT, the lightest model, acquires the least training time and the RoBERTa requires almost triple time compared to DistilBERT uncased model. Considering all the factors, DistilBERT uncased model is the ideal model for the classification task on videos' view counts rank with both title and tabular data including video length, title length, uppercase ratio in title, topic, sentiment intensity, because it has good results, little training time and low training loss. Although training time does not seem to be important in this project, since this small dataset with around 8000 records does not take long time to train, the training time does matter in real life application.

What's more, the pattern, existing in all six classification models, Recall is much higher than the Precision value. The average Recall is about 0.95, while the average Precision is about 0.55, which signifies that the false positive samples are much higher than the false negative samples. The BERT model tends to classify samples to positive(high view counts) over negative (low view counts).

Model	bert-base-uncased	bert-base-cased	albert	distilbert-base-uncased	distilbert-base-cased	roberta
Training Time	2m30s	3m23s	2m50s	1m20s	1m52s	3m20s
Training Loss	0.4260	0.4773	0.6904	0.4881	0.4223	0.5427
Eval ROC AUC	0.72	0.67	0.69	0.70	0.68	0.71
Eval F1	0.71	0.67	0.59	0.68	0.64	0.69

Table 5.2: Model results summary.

CHAPTER 6

Conclusion and Discussion

Fashion/Beauty videos are populating YouTube. Apart from the subjective analysis of beauty- and lifestyle- related video contents and titles, this thesis examines them in an objective way by text mining and classification modeling. Beauty guru's ambition can be found in description. Each group shares social media accounts to connect with followers more, they share products' links, ask viewers to subscribe to channels, and clarify sponsorship. Popular Group focuses more on attracting business collaboration and increasing video searchability, while Super Popular Group focuses more on communication with the audience in the descriptions.

The videos for Popular Group and Super Popular Group are ranked by the view counts. By comparing the four groups' video titles, there is no clear pattern in sentiment analysis, and it may help with recommendation algorithms more than content creation. By analyzing the general frequent words and series videos frequent words, it is found that beauty social celebrities benefits from making more trendy market- oriented videos since it increases searchability and attracts new audiences. While within a certain topic, YouTubers should expand small topics related to it. When a YouTuber has a large number of loyal audiences, creating content-oriented videos can bond the community.

Among six BERT based models, DistilBERT uncased model is the ideal model for the classification task on videos' view counts rank with both title and tabular data including video length, title length, uppercase ratio in title, topic, sentiment intensity, because it produces decent results and requires little time. The classification results on videos' view

rank (high view counts or low view counts) are not high compared to other tasks using BERT. It is reasonable for this experiment design. Because besides basic objective information for videos, view counts are related to multiple factors, such as quality of the video and YouTube's recommendation algorithm. YouTube's recommendation algorithm would also be highly related to the quality of the video and audience interests, which are subjective and hard to quantify. Ideally, with audiences' behavioral data, the recommendation and classification would have smaller errors. However, in this experiment, the goal is to predict whether a video would be a popular video or fair video before it reaches the audiences. In other words, the BERT model gives insights to fashion YouTubers whether the video metrics would have a high view counts.

The limitations of the study are as follows: 1) the sample size is not large enough; 2) the subset of two groups, the subscriber counts for Popular Group and Super Popular Group do not vary a lot, and it may cause the overlap of traits. In the future, with increased sample size, the data can be analyzed according to year and see the trend changes. Also, other metadata like comment count, like count and dislike count can be included to expand the study topics.

The code we used to train and evaluate our models is available at https://github.com/yunwen-kai/MAS-thesis-YouTube-title/tree/main

REFERENCES

[Blo21] Feedspot Blog. 100 fashion youtubers for fashion lovers you must follow in 2021, Nov 2021. URL: https://blog.feedspot.com/fashion$_Y ouTubers/$.

[CCY17] Yen-Liang Chen, Chia-Ling Chang, and Chin-Sheng Yeh. Emotion classification of youtube videos. *Decision Support Systems*, 101:40–50, 2017.

[Cec21] L. Ceci. U.s. youtube reach by age group 2020, Oct 2021. URL: http://www.statista.com/statistics/296227/us-YouTube-reach-age-gender/.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Faw06] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[GB21] Ken Gu and Akshay Budhkar. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, 2021.

[GR17] Florencia García-Rapp. Popularity markers on youtube's attention economy: the case of bubzbeauty. *Celebrity Studies*, 8(2):228–245, 2017.

[HBB10] Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23:856–864, 2010.

[HG14] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

[Hou19] Mingyi Hou. Social media celebrity and the institutionalization of youtube. *Convergence*, 25(3):534–553, 2019.

[LCG+19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[LOG+19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[MOG+08] Heather Molyneaux, Susan O'Donnell, Kerri Gibson, Janice Singer, et al. Exploring the gender divide on youtube: An analysis of the creation and reception of vlogs. *American Communication Journal*, 10(2):1–14, 2008.

[Nox21] NoxInfluencer. Youtube video title generator and tips, 2021. URL: http://www.noxinfluencer.com/YouTube/video-title.

[PE19] Python-Engineer. Python-engineer/youtube-analyzer: Extract statistics for a youtube channel with the youtube data api, 2019. URL: https://github.com/python-engineer/youtube-analyzer.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[S+07] Yutaka Sasaki et al. The truth of the f-measure. 2007. *URL: https://www.cs.odu.edu/˜ mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf*, 2007.

[SDCW19] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[WDS+19] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[YP17] Yahya and Paul. F1 score vs roc auc, Jul 2017. URL: https://stackoverflow.com/questions/44172162/f1-score-vs-roc-auc.