Popularity Analysis for Youtube Makeup Videos

Shiyi Cheng
University Of California, Riverside
schen470@email.ucr.edu

Kexin Wang
University Of California, Riverside
kwang164@ucr.edu

Ruchen Zhang
University Of California, Riverside
rzhan120@ucr.edu

Zhuocheng Shang
University of California, Riverside
zshan011@ucr.edu

Chiyuan ma
University of California, Riverside
cma062@ucr.edu

ABSTRACT

Makeup vlogs take a significant place among YouTube videos nowadays. Although YouTube trending analysis has already be generated thoroughly, the study of popularity in specific category is still needed to explore. Our group is getting to wonder what factors will affect the trending about these makeup videos on YouTube and why the audience prefers specific videos than others. In this project, two main processes are included. For the data preprocessing part, we fetched large data from Youtube and clean data by using NLTK, bag of word and Tf-idf Vectorizer. After we have a dataset we want, we compute four machine learning methods to study the correlation between word vectors and ratio of like and dislike, which are Neural Network, Linear Regression, K Nearest Neighbors and Logistic Regression including confusion matrix. It had been found the results as follows: Neural network has the highest accuracy, KNN's accuracy is the same as LR and Linear Regression is the worst. The public focus more on the keywords on "makeup tutorial" or "eye makeup" and pays less attention on makeup videos in recent years.

Keywords

Youtube, data cleaning, text analysis, videos, machine learning.

1. INTRODUCTION

Since 2007, YouTube has launched local pages in 75 countries and regions, supporting 61 languages - not even 50 languages with more than 100000 entries on Wikipedia. At the same time, about 60% of the average number of hits per video uploader on YouTube comes from abroad. What's more, in the past decade, everything on YouTube has influenced countless users and industries around the world. As YouTube's video sharing policy allows users to communicate freely around the world, YouTube has become one of the preferred online platforms for the digital community of makeup lovers. In fact, YouTube seems particularly well suited to launch new cosmetics collections, product reviews, and video tutorials. In recent years, make-up videos have sprung up on YouTube. In 2016, according to Statista, there were only 55 billion videos of make-up. But by 2017 / 2018, it had grown to 880 / 169 billion. Today, through online digital video stations like youtube, we have built a digital community full of passion and beauty related to beauty and fashion. In the past, they were more willing to share stories, inspiration and good skills. We have noticed that the bloggers who published the beauty videos on YouTube not only have the beauty masters with more than one million fans, but also the beauty lovers of plain people. Ignoring the influence factors of the number of video bloggers' fans, we focus on the relevant characteristics of the video itself, and study how they affect the popularity of video.

Therefore, the motivation of the project has two aspects: first, from the video description and label, we can roughly understand the reasons for the success of the makeup videos on YouTube; second, we can understand the relationship between the popular video and the amount of playback. The data

is mainly obtained by capturing the video information under the popular YouTube makeup channel through the network, and the complex data cleaning and processing are carried out, which will be further explained in Section 3. In addition, four data mining algorithms are used to model and train the data. The models are explained in Section 4. Finally, the evaluation results are presented in Section 5, and the limitations and subsequent steps are discussed as well.

2. RELATED WORK

Gloria Chatzopoulou, Cheng Sheng and Michalis Faloutsos presents a study of commenting and comment rating behaviors and relationships between other features such as views, topic categories, and comment rating[1]. The paper explores the dataset from two aspects. Firstly, the article applies a classification model for deciding comments acceptance based on attitudes of certain words. Support vector machine and Naïve Bayes are implemented during this process. Secondly, this article performs both qualitative and quantitative studies of each term to explore the correlation between sentimental values and comment rating using SentiWordNet thesaurus. However, the article should consider cleaning up the data first in order to get a more accurate dataset. Further, the report may expend the search area to studying different language-based comment ratings.

[2] mainly focuses on analyzing factors correlation with YouTube videos' popularity. Three basic popularity concepts are argued in the paper: time factors, multiple popularity metrics, and various categories and labels. The main parameter in this paper is the number of times a video is watched, and four features are taken into consideration while generating potential influence values: number of comments, number of ratings, number of favorites and the average rating. The strength of this paper is that it uses a large dataset, which increases the accuracy of the model. However, the project only implements a linear regression to explore the correlation between attributes. The paper could do more research on causality analysis.

[3] is writing about the Support Vector Machine using in sentiment analysis model. The sentiment analysis is used to find a pattern or a certain character of the one we need to analyze. It is a process of picking up and favorability of a natural language. Support Vector Machine is used to classify the opinion into positive, neutral and negative class and Lexicon Based method and Confusion Matrix is used, which are used to know the result of weighting percentage of analysis to SVM. SVM is used in the process of data retrieval, analysis, until the conclusion. Sentiment analysis can be used to find out how far the performance based on the comments on Youtube. However, data recording should be high in numbers to achieve the accuracy of the results and conclusions in further research.

According to [4],detection of sentiment polarity is a very challenging task due to the limitations in current sentiment dictionaries. In the present ones, there are no proper sentiments of terms created by the community. Thus, analysis of user comments is a source through which useful data may be achieved for many applications. Different techniques are adopted for sentiment analysis for user comments and for this purpose sentiment lexicon called SentiWordNet is used in this paper. The basic survey framework covers three main issues, which includes event classification, detection of sentiment polarity and predicting comments. Precision or positive predictive value is a part of relevant retrieved instances, while recall or sensitivity is the fraction of retrieved relevant instances. Precision and recall are based on two facts.

In [5],they have presented a multimodal sentiment analysis frame-work, in particular, the textual sentiment analysis module has been enriched by sentic-computing-based features, which have offered a significant improvement in the performance of the textual sentiment analysis system. Visual features also play a key role to their perform the state-of-the-art. And ELMs offer significant advantages such as faster learning speed, ease of implementation, and minimal human intervention. They have analyzed the importance of each feature used in the classification task. The best accuracy was obtained when all features were used together. However, they still have some shortcomings. For example, the performance of audio information

analysis in the model needs to be improved, and the time complexity of the algorithm can be reduced.

Word cloud is a visual representation of a collection of text documents. A new method of placing multiple time word labels on a series of specific shapes under different constraints by using rigid body dynamics method was proposed in [6]. With the help of geometric constraints, aesthetic constraints and time consistency constraints, this method can not only align the text labels with their respective shapes, but also temporarily deform, smoothly change the text cloud over time, and generate the text cloud. Therefore, through the proposed frame by frame deformable text cloud, the shape will produce various interesting time visualization effects. From the change of the shape, we can observe the time-varying text data in the whole story, and from the frame of the text cloud, we can observe the detailed experimental results.

Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, Katja Filippova want to do Option Polarity on Youtube comments[7]. The authors finally define a systematic approach by modeling classifiers for predicting the opinion polarity and the type of comment and proposing robust shallow syntactic structures for improving model adaptability. They don't choose bag-of-words method but choosing tree kernel technology to automatically extract and learn features which with better generation power. A large number of empirical evaluations have been carried out on their manually annotated YouTube comment corpus. The results show that the classification accuracy is high and the advantage of structural model in cross domain setting is highlighted.The authors make a systematic study OM targeting Youtube comments which has three contributes.

[8]'s goal is to find the relationships between content features, video attributes and parasocial attributes(the characteristics that could lead to the creation of parasocial relationships) among the top most subscribed Youtube channels. They use several useful methods containing MANOVA to solve the problem and use graphs to show their results. They use appropriate sampling method. However, the overall sample size of 234 videos is relatively small and findings are only generalizable to the population of top YouTube channels. Furthermore, while this allowed for comparisons among top channels, conclusions could not be drawn about the potential efficacy of parasocial appeals employed by these YouTube channels compared to less successful channels. And this study did not analyze channels containing non-English content.

In [9], the author has two goals. First, get a general idea of what makes YouTube's make-up videos successful in terms of the number of times you watch them, and second, understand how sponsorship particularly affects viewers. Mainly collect data through the network to grab and use the Youtube API.Add other variables to the dataset by reducing the title, description box, transcript, and markup to a bag of words. Besides, some word frequencies are combined with other word frequencies that represent the same concept. In addition, two prediction models have been created to estimate the number of Views: model A (which only includes the variables that exist when the video is published) and model B (which also includes dynamic variables, such as like, dislike and comment mood). Lasso and ridge regression are used to fit both models. Finally, the results of these regressions are explained [9].

Phakhawat Sarakit, Thanaruk Theeramunkong, Choochart Haruechaiyasak, Manabu Okumura propose a method to classify the emotions of Thai media clips on YouTube by using the comments given to the clips The six basic emotions considered are anger, disgust, fear, happiness, sadness and surprise[10] The performance of three alternative machine learning algorithms (MNB, DT and SVM) is compared. They use the open source machine learning tool Weka to classify emotions All algorithms are set using default settings. A framework is proposed to classify the emotions in Thai texts in YouTube comments.As a result, SVM achieves the highest accuracy in ad category, while MNB achieves the best result in MV category [10].

3. DATA PREPROCESSING

This section describes the dataset examined in the project. It contains four parts involving data collection, attributes description, data cleaning, and data analysis.

3.1 Data Collection

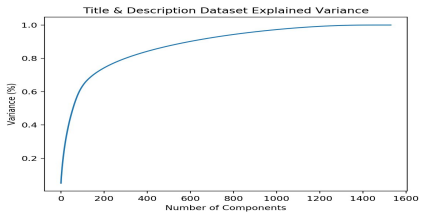
The dataset is collected through crawling videos information under popularity Youtube makeup channels. We basically collect “video id”, “title”,”publish date”, ”like count”,”dislike count, “view count”, and “descrption” from each website. In this project, our team decided to set the combination of title and description as independent variable, and to set the ratio of like and dislike to be the dependent variable. In order to void some null value in the dataset, we normalize the ration with the formula

$$Y = like\ count / (dislike\ count + like\ count).$$

For the cluster category, we firstly calculate the mean value of Y discussed above. If the $Y_i < Y(mean)$ we said that the video is “unpopular", otherwise the video labeled as “popular”. The vectorization of the bag of words (title & description) will be discussed in the data cleaning section.

3.2 Data Cleaning

The most important data needed to clean is text data in our dataset. We have two mainly text data: title and description. Before numeric the text data, we need to clean the dataset. In the project, we use the package Beautifulsoup and NLTK to remove HTML tags, stop words, digits, punctuation and null variables. We apply Tf-idf to count the importance of a word for each combination of title and description. We use PCA to reduce dimensionality. The figure 3.1 shows the process of how to pick the most accuracy component and the curve shows that 200 is the best component for this dataset.

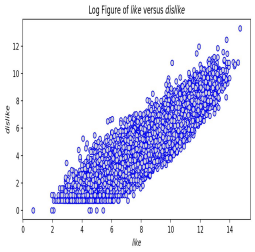


[figure 3.1] Component and variance

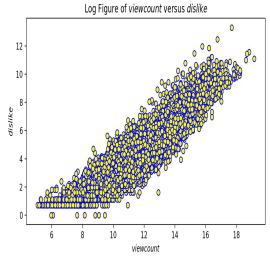
3.3 Data Analysis

In this project, we focus on colarations between four attributes: like count, dislike count, view count, and word vectors (calculated from the combination of title and description).

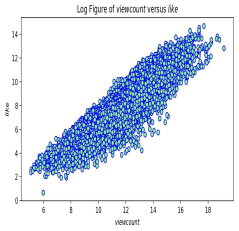
3.3.1 coloration between like count, dislike count and view count.



[figure 3.2] like vs dislike count



[figure 3.3] view count vs dislike count

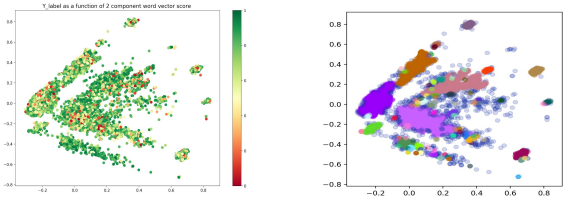


[figure 3.4] log of view count vs like

[figure 3.5] view count vs like & dislike

The four figures all represent a linear coloration. Especially in the figure 3.5, it shows if view counts as a function of like count and dislike count. If a video has a high view count, it also will be liked or disliked by larger amount audience.

3.3.2 coloration between word vector and Y



[figure 3.6 & 3.7] Y as a function of word vectors and the clustered plot after applying DBSCAN

Firstly, we use PCA to reduce 200 dimensional word vectorizational dataset to a 2 dimensional dataset. The X-axis and Y-axis are the 2 components in word vector dataset. This figure 3.6, presents a nonlinear relationship between word vector value and Y label which is the ratio of like and dislike count. Hence, we apply DBSCAN to cluster the data, and there are four most dense clusters. Our group analysis these four clusters and extract top five words that appear

most frequently from each of them. The results are shown in the table below.

Table shows top 4 clusters' frequency words

Cluster	word 1	word 2	word 3	word 4	word 5
1	video	tutorial	use code	somkey eye	eye makeup
2	use code	affiliate link	code lauralee	tutorial	video
3	affiliate link	use code	coupon	sponsored	tutorial
4	use code	code laurale	affiliate link	code james	coupon

4. MODELS

In this section, we explain four models applied in the project to predict the popularity of a makeup video based on its given title and description.

4.1 Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values. After a regression line has been computed for a group of data, a point which lies far from the line (and thus has a large residual value) is known as an outlier. Such points may represent erroneous data, or may indicate a poorly fitting regression line. If a point lies far from the other data in the horizontal direction, it is known as an influential observation. The reason for this distinction is that these points have may have a significant impact on the slope of the regression line. Once a regression model has been fit to a group of data, examination of the residuals (deviations from the fitted line to the observed values) allows the modeler to investigate the validity of his or her assumption that a linear relationship exists. Plotting the residuals on the y-axis against the explanatory variable on the x-axis reveals any possible non-linear relationship among the variables, or might alert the modeler to investigate lurking variables.

Here we set one variable as X of word vectors and the other one as Y of like/(like + dislike). Then define the functions to calculate the residuals standard error, r-squared and log likelihood. According to these values, especially the log likelihood, the collation between these two variables will be showed. These two variables show a linear increasing relationship. Linear Regression is also used in analyzing the keywords for tag result. First, we define a function to read all the words into one dictionary and skip the empty line. Then we use the key-value pairs to count the number of one word and get the final key-value pairs. By generating word clouds, it is much easier to figure out which tag is the most popular.

4.2 Logistic Regression

Logistic regression is a generalized linear model, so it has many similarities with multiple linear regression analysis. Their model form is basically the same, they both have w'x + b, where w and b are the parameters to be sought, and logistic regression uses the function L to correspond w'x + b to a hidden state p, p = L (w ' x + b), and then determine the value of the dependent variable according to the magnitude of p and 1-p. If L is a logistic function, it is logistic regression. Logistic regression measures the relationship between the dependent variable (the label we want to predict) and one or more independent variables (features) by using its inherent logistic function to estimate the probability. These probabilities must then be binarized to truly predict. This is the task of the logistic function, also known as the Sigmoid function. The Sigmoid function is a sigmoid curve. It can map any real value to a value between 0 and 1, but it cannot take 0 or 1. Then use a threshold classifier to convert values between 0 and 1 to 0 or 1. We want to maximize the probability

of random data points being correctly classified, which is the maximum likelihood estimation. Maximum likelihood estimation is a general method for estimating parameters in statistical models.

Logistic regression is the preferred method for binary classification tasks. In Section 3, we introduced the quantification of popularity of videos in the dataset according to $Y = like\ count / (dislike\ count + like\ count)$. Here the dependent variable has two values of 0 or 1, which are suitable for logistic regression models to make classification predictions about whether videos are popular or not.

4.3 K-NN Model

KNN is classified by measuring the distance between different eigenvalues. The idea is that if most of the k most similar (ie, the nearest neighbors in the feature space) samples of a sample belongs to a certain category, the sample also belongs to this category, where K is usually not greater than an integer of 20. In the KNN algorithm, the selected neighbors are all objects that have been correctly classified. This method only determines the category to which the sample to be classified belongs based on the category of the nearest sample or samples.

The kNN method is mainly based on the limited neighboring samples, rather than the method of discriminating the class domain to determine the category. Therefore, the kNN method is more effective than other methods for the set of samples to be divided that have more overlapping or overlapping class domains. Also KNN has high accuracy, is insensitive to outliers, and assumes no data input.

To summarize the idea of the KNN algorithm: that is, when the data and labels in the training set are known, input the test data, compare the characteristics of the test data with the corresponding features in the training set, and find the most similar K Data, the category corresponding to the test data is the one that appears most frequently among the K data, and the algorithm description is:

1) Calculate the distance between the test data and each training data;

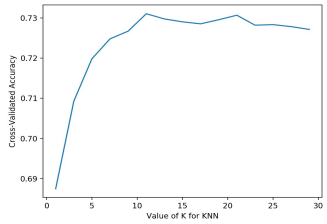
2) Sort by increasing distance;

3) Select K points with the smallest distance;

4) Determine the frequency of occurrence of the category of the first K points;

5) The category with the highest frequency among the first K points is returned as the prediction classification of the test data.

In our project, we consider the word vector formed by the title and the introduction as the point coordinates in the multidimensional space, and calculate the Euclidean distance between the points. The classification label is like / (like + dislike).We use the average of the like rate of all videos as the cut-off point of the label for KNN model training. And we tried different k to run the model and analyzed the relationship between k and accuracy. KNN model get best accuracy 0.73 when K 's value is around 11.



[figure 4] K vs model accuracy

4.4 Neural Network Model

A neural network is a neural network that simulates the human brain in order to implement artificial intelligence-like machine learning techniques. Now I'll explain the process of neural network. Firstly we should create M hidden layers, establishing the connection between the input layer and the hidden layer in order, and finally establish the connection between the hidden layer and the output layer. Select an activation function for each node of each hidden layer. Solve the weight of each connection and the bias value of each node. The so-called activation function is a function that is further enhanced after summing the inputs of various paths. The training nature of the so-called neural network problem is to know y1, y2 .... yn and x1, x2 .... xm, and to solve the weight of each connection and the deviation

value of each neuron. For a neural network with a single-layer activation function of RELU, y = max (sum (w * x) + b, 0). Knowing y and x, solve w and b. [11]

Now I'll explain methods of training. For the above solutions to the values of w and b, scientists have found that they can be solved by a combination of backpropagation and gradient descent. It is to initialize the weight of each connection with random numbers at first, and then compare the y value calculated by the neural network with the real y value. If the value differs greatly, modify the weight of the connection of the current layer. When this value is found to be not much different, the weight of the lower layer is modified. This step is repeated all the time, passing the weights to the first layer step by step.

I'll explain the neural network in our own project in the following paragraph. We build a fully connected neural network which has three layers. The first layer has 32 nodes, and the activation function we choose is ReLu. The second layer has 16 nodes, and we also choose Relu as activation function. The third layer has just one node. That's because our problem is actually a regression peroble. Finally we use adam optimizer to change the parameters of our model, and mean squared error metrics to evaluate our regression models and mean squared error loss to evaluate how accurate our predictions are. We use the word factor as our data and like/(like + dislike) as label to train the model. We choose like / dislike as our label initially but we find out that actually the ratio of like and dislike is unrelated as our word vector.

Finally, the accuracy of NN model is high, which means NN model works really well on our dataset.

5. EVALUATION

In this section, we evaluate all of our models using corresponding evaluation values.

5.1 Linear Regression

Residual Standard Error: 0.031073211796996715
r-Squared: 0.1860813189038045
Log-Likelihood: 62153.91636814299
Test score: 0.17740594576943547

5.2 Logistic Regression

Class 0: popular Class 1: unpopular

5.1 Confusion Matrix

	Pred_0	**Pred_1**
True_0	**18528**	**2082**
True_1	**5324**	**4348**

Train Accuracy: 0.7554322699953768
Test Accuracy: 0.7494551218545671

5.3 KNN

Labels: popular, unpopular for corresponding Y values discussed in the section 3.1.
K=11, the model has most accuracy around 0.73.

5.4 Neural Network

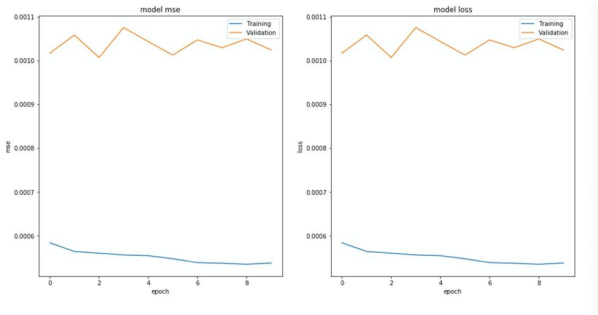


figure 5.2 epoches vs mse

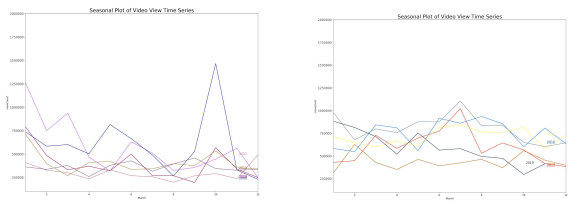
In this project, we use MSE as the value to determine the accuracy of our model. The figure 5.2 shows the MSE score changes based on different epochs. The figure presents a stable tendency of the MSE and it is always around 0.001.

5.3 Limitation Discussion

There is one limitation in our project is that the data we collected is not randomized enough. It is a case that since we collect data from each channel, we may collect too much data from one uplooder. It will lead to duplications of description in our dataset because it is likely that one uploader has similar description among his/hers videos. This would reduce certain accuracy of our model.

Also this project does not include the sentiment analysis toward video comments. It could be the future work of our project.

6. CONCLUSION & FUTURE WORK



[figure 6.1] mean viewcount monthly change from 2007 to 2019

In this paper, our major contribution is to build four models which can predict whether the new makeup videos will be popular. We build Linear Regression model, NN model, Logistic Regression model and KNN model. Then we compare the effect of the four model on our dataset. Also, we analyze the correlation between like count, dislike count and view count. And we analyze coloration between word vector and y. Y as a function of word vectors and the clustered plot after applying DBSCAN. We also analyze mean view count monthly change from 2007 to 2019.

Now I'll explain what we have done. We analyze the features that make the makeup videos popular. Specifically, we analyze the influence of title and description to popularity of videos. Firstly, we get the title and description of most popular 40377 makeup videos using web crawler. Then we clean the data, deleting wrong records (maybe some field of the records is empty). And we need to remove HTML tags, stop words, digits, punctuation. And then we need to transform title and description of a video into a word vector. We use Tf-idf to do this and then using PCA to reduce dimensionality. After data processing, we get word vectors of each videos. We use the number of like / (the number of like + the number of dislike count) as our labels.

Then we do some data analysis. We draw pictures of coloration between like count, dislike count and view count. Finally we find out that dislike count and like count, dislike count and view count, like and view count are all linear correlation. And then we draw pictures of coloration between word vector and y. Y as a function of word vectors and the clustered plot after applying DBSCAN. After the analysis, we find out that the relationship between word vector value and Y label which is the ratio of like and dislike count is nonlinear. Hence, we use DBSCAN to cluster the data, and there are four most dense clusters. And then we analyze mean view count monthly change from 2007 to 2019, and we get some results. After drawing the picture, we find out that The public pays less and less attention to makeup videos.

Then we use four methods to process the vectors and the corresponding labels. It's actually a regression problem. We use Linear Regression, Logistic Regression, Neural Network Model and KNN model. Initially, we use four hidden layers in our neural network, and the first hidden layer has 64 nodes and using Relu activation function. But after deleting the first hidden layer, we find out that the accuracy is the same as the four hidden layer neural network. So we delete the first hidden layer. After the four methods be used on our dataset, we can compare the four methods. Compared with NN, Linear Regression' accuracy is not high. Neural Network works well on our dataset. KNN also doesn't work very well for our dataset. Linear regression and KNN's effect are almost the same.

In the future, we will improve the accuracy of our KNN model, Linear Regression model and Logistic Regression model. Further, we could try to obtain more videos randomly not only just collect top channels. Moreover, we could expand our search area to sentimental analysis about video comments.

7. REFERENCES

[1] Gloria Chatzopoulou, Cheng Sheng and Michalis Faloutsos. (2010). A first step towards understanding popularity in YouTube. IEEE.

[2] Stefan Siersdorfer, Sergiu Chelaru, Wolfgang Nejdl and Jose San Pedro. (2010). How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings. Proceedings of the 19th international conference on World wide web (WWW '10).

[3] Fiktor Imanuel Tanesab1 , Irwan Sembiring2 and Hindriyanto Dwi Purnomo3. (2017). Sentiment Analysis Model Based On Youtube Comment Using Support Vector Machine. International Journal of Computer Science and Software Engineering (IJCSSE)

[4] Muhammad Zubair Asghar, Shakeel Ahmad, Afsana Marwat, Fazal Masud Kundi. Sentiment Analysis on YouTube: A Brief Survey.

[5] Soujanya Poriaa, Erik Cambriab, Newton Howardc, Guang-BinHuangd, Amir Hussain. (2015). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. NueroComputing Volume 174, Part A, 22 January 2016, Pages 50-59.

[6] Ming-Te Chi, Shih-Syun Lin, Shiang-Yi Chen, Chao-Hung Lin, Tong-Yee Lee. Morphable Word Clouds for Time-Varying Text Data Visualization. IEEE.

[7] Aliaksei Severyn, Alessandro Moschitti, Olga Uryupina, Barbara Plank, Katja Filippova. Opinion Mining on YouTube.

[8] Stephanie Orme, Jared LaGroue, Arienne Ferchaud, Jenna Grzeslo. (2018). Parasocial attributes and YouTube personalities: Exploring content trends across the most subscribed YouTube channels. Computers in Human Behavior

[9] Lavanya Sunder. (2016). Predictive Model for Views In YouTube Beauty Community.

[10] Phakhawat Sarakit, Thanaruk Theeramunkong, Choochart Haruechaiyasak, Manabu Okumura. (2015). Classifying emotion in Thai youtube comments. International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)978-1-4799-8565-4/15/$31.00 ©2015 IEEE

[11] Hansen L K, Salamon P. Neural network ensembles[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1990 (10): 993-1001.