What Makes a TikTok Video Become Viral?

Christy Natalia Jusman[1], Rainamira Azzahra[2]

Binus International University
Jakarta, Indonesia
[1]christy.jusman@binus.ac.id
[2]rainamira.azzahra@binus.ac.id

***Abstract*-- This paper is a research for the final project of Data Science course. In TikTok itself, some videos can be viral for some reasons and it is still unknown which means no one knows about it. Many people wondering the algorithms that is used by TikTok and causing some content became popular all of the sudden. Everyone want to get more know about it because they want to go viral too. By using the data by several popular content creators, this research is aimed to provide the most reasonable factors that affects some videos in TikTok went viral. Is it because of the hashtag that they used? or music? or maybe the number of followers?**

I. INTRODUCTION

TikTok has grown to become one of the most popular social media platforms in the world. Many people of all kinds of backgrounds enjoy TikTok. People spend their time scrolling through videos on TikTok every day. Creators use this platform to share short videos with a wide range of contents, including songs, dances, tutorials, education and numerous creative filters. Now a days, by using TikTok, people can be more easier to become famous. Users often can discover very interesting videos varying from the editing skills to the humour inside of the video.

According to Statista, Indonesia is TikTok's second largest market in the world in 2020 [2]. Indonesia is stated to have 22.2 million monthly active users, behind the United States, which has 65.9 million monthly active users. Moreover, since the Covid-19 outbreak, TikTok's popularity has skyrocketed. There are 92.2 million TikTok users in Indonesia by July 2021 [1]. TikTok users were only 37 million at the beginning of the epidemic, in April 2020 to be exact according to Internal Data, ID Audience.

Of course, this can make TikTok a relevant segment for digital marketing. As we all know, marketing is crucial to a business's success and influences whether or not its products will be accepted and acknowledged by people. A digital marketer must be able to generate distinctive and creative marketing to make a product or brand recognized by a lot of people.

Moreover, the content of the video that is shared has a major impact upon it. Unfortunately, creating promotional content takes skill and knowledge. That is where the assistance of content creators, or influencers, comes in handy. In this project, we will explore which factors correlate with the audience engagement and predict the amount of engagement a content creator's video will have. Engagement in social media is the interaction between the content with the audiences. If the audiences are interested in our contents, they will give us feedback and it can be likes, comments, and shares it to others [3].

II. DATASET

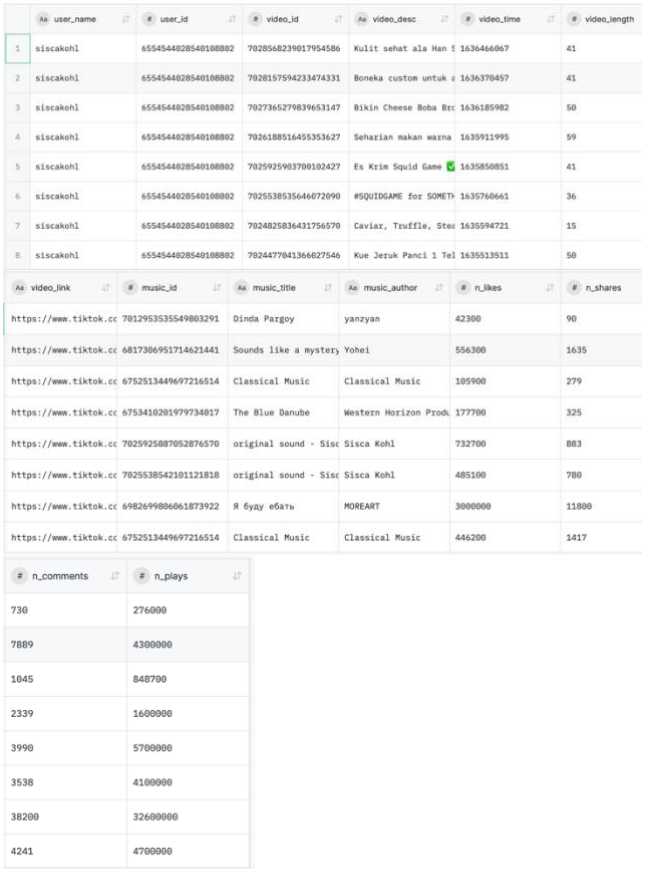
First, we conduct a survey by using google forms and ask 20 respondents, who are our friends that use TikTok frequently to write down their favorite TikTok accounts from different categories such as: fitness, culinary, beauty, fashion, comedy, travel, and technology. Then from each category, we will take 5 content creators based on the most written in the survey. Next, we will scrape the data from 50 videos from each chosen content creator.

For the dataset, we are going to scrape it by ourselves. To scrape the data, we are going to use the unofficial API for TikTok and it is made by David Teather on GitHub (https://github.com/davidteather/TikTok-Api). By using Python and pandas library, we are able to scrape the data and export it into csv file format. Therefore, inside the csv file it will have the number of likes, shares, comments, and plays from each video by the selected user.

For the datasets, there will be 2 files. The first one is for the videos. Here, we will have 14 rows and 50 columns. Fourteen rows including:

1. **user_name** - The content creator's readable id and it can be changed by the owner.
2. **user_id** - The unique number about the user id and it cannot be changed.

3. **video_id** - The random id to specify the video and it must be unique.
4. **video_desc** - The short description about the video including the hashtags.
5. **video_time** - It shows when the video is released by the content creators.
6. **video_length** - The duration of the video in seconds.
7. **video_link** - The URL link for the video
8. **music_id** - The id of the song that is used for the video.
9. **music_author** - The name of the music owner used.
10. **music_title** - The title of the music used for the video.
11. **n_likes** - The number of users liked this video.
12. **n_shares** - The number of people sharing the current video to others.
13. **n_comments** - The number of comments for the corresponding video.
14. **n_plays** - It shows how many times the video has been played.



	user_name	user_id	video_id	video_desc	video_time	video_length
1	siscakohl	6554544028540108802	7028568239017954586	Kulit sehat ala Han S	1636466067	41
2	siscakohl	6554544028540108802	7028157594233474331	Boneka custom untuk a	1636370457	41
3	siscakohl	6554544028540108802	7027365279839653147	Bikin Cheese Boba Bro	1636185982	50
4	siscakohl	6554544028540108802	7026188516455353627	Seharian makan warna	1635911995	59
5	siscakohl	6554544028540108802	7025925903700102427	Es Krim Squid Game	1635850851	41
6	siscakohl	6554544028540108802	7025538535646072090	#SQUIDGAME for SOMETH	1635760661	36
7	siscakohl	6554544028540108802	7024825836431756570	Caviar, Truffle, Stea	1635594721	15
8	siscakohl	6554544028540108802	7024477041366027546	Kue Jeruk Panci 1 Tel	1635513511	50

	video_link	music_id	music_title	music_author	n_likes	n_shares
	https://www.tiktok.cc	7012953535549803291	Dinda Pargoy	yanzyan	42300	90
	https://www.tiktok.cc	6817306951714621441	Sounds like a mystery	Yohei	556300	1635
	https://www.tiktok.cc	6752513449697216514	Classical Music	Classical Music	105900	279
	https://www.tiktok.cc	6753410201979734017	The Blue Danube	Western Horizon Produ	177700	325
	https://www.tiktok.cc	7025925887052876570	original sound - Sisc	Sisca Kohl	732700	883
	https://www.tiktok.cc	7025538542101121818	original sound - Sisc	Sisca Kohl	485100	780
	https://www.tiktok.cc	6982699806061873922	Я буду ебать	MOREART	3000000	11800
	https://www.tiktok.cc	6752513449697216514	Classical Music	Classical Music	446200	1417

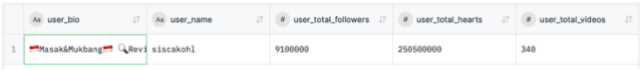
n_comments	n_plays
730	276000
7889	4300000
1045	848700
2339	1600000
3990	5700000
3538	4100000
38200	32600000
4241	4700000

Fig. 1. Results of scraping data from user siscakohl and it only shows the latest 8 data. There are 14 rows in here and there is no null value.

Each .csv file consists of 50 columns and it indicates the metadata for the latest 50 videos created by the user. The date of scraping this data is 10 November 2021. Therefore, it will show the videos from 10 November till the oldest 50 videos.

The second one is for the user profile. In this dataset, there will be 4 rows and a column. Here, each user has different csv files. The rows consist of:

1. **user_name** - The content creator's readable id and it can be changed by the owner.
2. **user_bio** - The description of the user's profile.
3. **user_total_followers** - The total followers that the user has.
4. **user_total_hearts** - The total of hearts from all public videos by that user.
5. **user_total_videos** - The number of videos that are created by the selected user.



	user_bio	user_name	user_total_followers	user_total_hearts	user_total_videos
1	Masak&Mukbang Revi	siscakohl	9100000	250500000	340

Fig. 2. The results of scraping the data from the user profile who is siscakohl.

Here, we tried to visualize the data by using Tableau Public. Under the technology categories, we show the average engagement that consist of total followers, shares, comments, and likes between five content creators compared to their total followers.

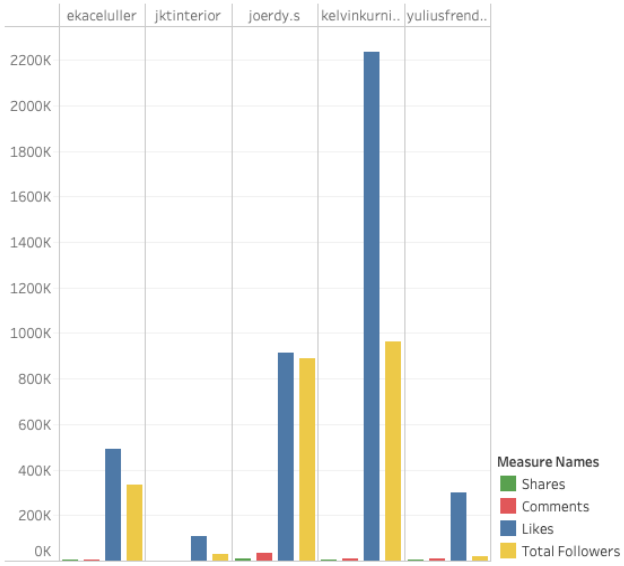


Fig. 3. The visualization to compare the number of total followers with shares, comments, and likes.

III. Data Preparation and Processing

Next, we combine each user profile and video data into a new dataset based on the categories. We would like to see the results on which variable has the biggest impact for each category. As an example below, we merge the user profile data from siscakohl, anakkuliner, disthemeatguy, hndriaditya, and separuhakulemakk under the culinary category. Every file has 15 columns and 250 rows.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 15 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Unnamed: 0    250 non-null    int64
 1   user_name     250 non-null    object
 2   user_id       250 non-null    int64
 3   video_id      250 non-null    int64
 4   video_desc    250 non-null    object
 5   video_time    250 non-null    int64
 6   video_length  250 non-null    int64
 7   video_link    250 non-null    object
 8   music_id      250 non-null    int64
 9   music_title   250 non-null    object
 10  music_author  250 non-null    object
 11  n_likes       250 non-null    int64
 12  n_shares      250 non-null    int64
 13  n_comments    250 non-null    int64
 14  n_plays       250 non-null    int64
dtypes: int64(10), object(5)
memory usage: 29.4+ KB
```

Fig. 4. The information regarding the data of combined videos based on the category. It shows the column name, the number of non-null columns, and the data type of each column.

For the next step, we combine the user profile with other users in the same category by using Pandas library and convert it again into a new csv file. In the picture below, it will have 5 rows and 6 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5 entries, 0 to 4
Data columns (total 6 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0            5 non-null      int64
 1   user_bio              5 non-null      object
 2   user_name             5 non-null      object
 3   user_total_followers  5 non-null      int64
 4   user_total_hearts     5 non-null      int64
 5   user_total_videos     5 non-null      int64
dtypes: int64(4), object(2)
memory usage: 368.0+ bytes
```

Figure 5. The details about files from the combined user data based on the category. It has 6 columns and 5 rows. It also shows the column name, non-null count, and the data type for each column.

All datasets will be combined into one big new dataset per category that will be processed later to see the results that are not specific to a single Tik Tok influencer. We merge video and profile datasets together based on the user_name column using the Pandas library. The datasets will show the video factors along with the ownership. Then, we convert the new dataset into a new csv file. As shown below, the dataset contains 250 rows and 21 columns in total. We repeat the procedure until each category has their own datasets. Later, n_plays will be used as a label that we want to predict. The variables that will be used as predictors are n_shares, n_comments, and n_likes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 21 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Unnamed: 0            250 non-null    int64
 1   Unnamed: 0_x          250 non-null    int64
 2   user_bio              250 non-null    object
 3   user_name             250 non-null    object
 4   user_total_followers  250 non-null    int64
 5   user_total_hearts     250 non-null    int64
 6   user_total_videos     250 non-null    int64
 7   Unnamed: 0_y          250 non-null    int64
 8   user_id               250 non-null    int64
 9   video_id              250 non-null    int64
 10  video_desc            250 non-null    object
 11  video_time            250 non-null    int64
 12  video_length          250 non-null    int64
 13  video_link            250 non-null    object
 14  music_id              250 non-null    int64
 15  music_title           250 non-null    object
 16  music_author          250 non-null    object
 17  n_likes               250 non-null    int64
 18  n_shares              250 non-null    int64
 19  n_comments            250 non-null    int64
 20  n_plays               250 non-null    int64
dtypes: int64(15), object(6)
memory usage: 41.1+ KB
```

Figure 6. Both users' videos and users' profiles are combined into one file based on their categories. Here is the information regarding the columns of the combined file.

Before we begin data processing, data cleaning is needed to make sure that the data we are working on is precise and reliable. In this step, we must look for duplicate values, missing values, unused values, and other variables that might lead to data interpretation mistakes. Pandas library will be used again to load the dataset. Here, we drop 10 columns which are Unnamed: 0, Unnamed: 0_x, Unnamed: 0_y, video_time, user_bio, user_total_hearts, user_id, music_id, video_id and video_link. For 3 unnamed columns, they are all just the index of the data and we will not need it for further. Video_time and video_link will not be used because they do not have any relations with the goal of this project.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 250 entries, 0 to 249
Data columns (total 16 columns):
user_bio                250 non-null object
user_name               250 non-null object
user_total_followers    250 non-null int64
user_total_hearts       250 non-null int64
user_total_videos       250 non-null int64
user_id                 250 non-null int64
video_id                250 non-null int64
video_desc              243 non-null object
video_length            250 non-null int64
music_id                250 non-null int64
music_title             250 non-null object
music_author            250 non-null object
n_likes                 250 non-null int64
n_shares                250 non-null int64
n_comments              250 non-null int64
n_plays                 250 non-null int64
dtypes: int64(11), object(5)
memory usage: 31.4+ KB
```

Figure 7. The results after we cleaned our datasets. We deleted several columns that are not useful for our researches.

At this point, we check if there are any rows that contain a null value. Since there is some parameters that have a null value present in the dataset namely video_desc, music_title, and music_author. We begin to fill the null value. For video_desc, we fill the null value with no description. For music_title, we also gonna fill it with similar value which is no music. And last, music_author will be filled with the same value as music_title that can be seen in the figure below.

```
#INSERT THE NULL VALUES
combined_profile['video_desc'].fillna("no description", inplace = True)
combined_profile['music_title'].fillna("no music", inplace = True)
combined_profile['music_author'].fillna("no music", inplace = True)
```

Figure 8. Fill the null value inside the dataset

Then, we would like to categorized the video_length column to make calculations easier when the model created later. We will categorized the video_length to three categories which are short, medium, and long. The short category has a range between 0 until 15 seconds. The medium has a range between 15 until 60 seconds. Last the long category has the range of 60 seconds to 180 minutes.

```
#CLASSIFY VIDEO LENGTH
video_length_category = pd.cut(combined_profile.video_length,bins=[0,15,60,180],labels=['Short','Medium','Long'])
combined_profile.insert(6,'video_length_group',video_length_category)
```

Figure 9. Classify the video length to 3 category (short, medium, long)

Next, we will convert the video_desc column into categorical values. In this process we would like to find out whether or not hashtags are used in a video description. If no, it will be indicated by 0. Otherwise, it will be 1.

Then, we will convert the columns that have values formatted in string into numerical values such as video_length, music title, music_author, and user_name. A Scikit-learn library named LabelEncoder will be used. This is needed because libraries that are used for prediction modeling require numerical data.

```
#ENCODING THE USERNAME
label_encoder = preprocessing.LabelEncoder()
combined_profile['username_encoded'] = label_encoder.fit_transform(combined_profile['user_name'])
combined_profile['music_title_encoded'] = label_encoder.fit_transform(combined_profile['music_title'])
combined_profile['music_author_encoded'] = label_encoder.fit_transform(combined_profile['music_author'])
combined_profile['hashtag_exist'] = label_encoder.fit_transform(combined_profile['hashtag_exist'])
combined_profile['video_length_encoded'] = label_encoder.fit_transform(combined_profile['video_length_group'])
```

Figure 10. The code to encode the username, music_title, misc_author, hashtag_exist, and video_length_encoded.

- music_title_encoded: There will be a chance that some videos use the same music title. Then, to make it more simpler and can be used for further research, the value is encoded into integer.
- username_encoded: It varies from 1 until 35 and it indicates the username of content's owner but in number.
- hashtag_exist: To check whether in the video description, there is a hashtag or no. If no, it will be indicated by 0. Otherwise, it will be 1.
- music_author_encoded: It indicates the singer of the music. Usually, one singer has more than one music. Therefore, the name of the singer is encoded.

The reason why the music author and music title are encoded is to know whether specific author or music title has impact to make a video become viral. We also added the encoded data to the datasets and arranged it into the order that we want. The order itself is based on our own needs and to make our researches become easier. As the total, we will have 1750 rows and 16 columns.

IV. Models and Techniques

We choose regression as our techniques because it can be used to show correlation between our data since our data are continuous. By using regression, we can predict how many views that we have to achieve to receive the numbers of likes, comments, and shares. There are several regression techniques that we use such as linear regression, ridge regression, lasso regression, and Decision tree.

Linear Regression is the simplest type of regression in machine learning. This type of regression is

used while the data is related linearly. Linear regression is built to understand the relationships between the input, output, and make the predictions of it [7]. Here, we are predicting how many views that we have to reach when we want to achieve several numbers of likes, comments, and shares. Because this project uses several inputs such as: likes, comments, and shares, it is called multiple linear regression [7]. This kind of regression consists of two variables: dependent variable (Y-axis) and predictor variable (X-axis) [8]. Linear regression will show the relationship between those two variables. Predictor variable is the variable that declares the cause to make changes of the dependent variable. However, the dependent variable is used to declare the effects of the predictor variable. For this project, the predictor variable is the number of likes, shares, and comments. For the dependent variable it is the number of viewers.

Ridge regression is used to analyze data where several independent variables in a model are correlated. This type of regression uses L2 regularization which penalizes the sum of the squared coefficients. It uses Lambda as the penalty term. The alpha argument will specify the value of lambda [7]. So, we can regulate the penalty term by adjusting the number of alpha. The greater the alpha value, the greater the penalty term and the coefficient size is lower [7]. Therefore, we choose this regression method to minimize our model complexity.

Lasso regression is similar to Ridge regression. It also uses the penalty term but unlike ridge regression which penalizes sum of squared coefficients, lasso penalizes the sum of their absolute values which is L1 regularization [8]. As a result, several coefficients are set exactly to zero under lasso for high values of lambda, which is never the case with ridge regression. So, we must carefully choose the value of lambda by iterating over a range of possibilities and selecting the one with the lowest error [13].

The Decision Tree algorithm solved machine learning problems by transforming data into a tree representation. This method is being used to reach an estimate by asking true/false questions across the tree starting from at the root node [5]. Then, distinct branches are constructed for each answer, and so on until the leaf node is obtained. The tree is built by recursive partitioning. Since a decision tree is a supervised machine learning model, it attempts to map data to the outputs during the model's training process [5]. This is achieved by feeding past data into the model that is relevant to the issue, as well as its real value, which the model should learn to forecast properly [5]. Decision tree regression examines an object's characteristics and trains a model in the shape of a tree to forecast future data and create meaningful continuous output. The output is not fixed, however it will predict the number from the input.

We chose Decision Tree for a number of reasons. For starters, Decision Tree does not require data pre-processing and data normalization [6]. So, less time is spent as opposed to other methods. Outliers and missing values have very little effect on decision trees [6]. Moreover, they can handle both numerical and categorical variables. The output of a decision tree may then be clearly comprehended. Overall, the decision tree is easy to understand, analyze, and visualize. We believe that benefits explained earlier enable us to work conveniently and efficiently.

Pearson correlation is a test to show the strength and direction in the term of linear relationships between two variables. To do the calculation, we need two or more variables and the research hypothesis predicts whether there is a linear relationship between these two variables. However, the variable must be continuous in order to use Pearson correlation. Hence, as the results, it will show the linear graph and the value range is from -1 until +1. When the value is near -1, it means there is no relationship between two variables and it means the variables are not related with each other. Otherwise, it shows the strong relationship between those two variables [7]. For this project, we use Pearson correlation to see which factors are related to each other.

We are going to use Python as the programming language. Python provides lots of packages that can be used to process the data and make the visualization. The packages that will be used for these projects are Pandas, NumPy, Scikit-learn, yellowbrick, scipy, os, glob, google.colab, pydot_graph and Matplotlib.

Since the raw data is stored on the google drive, we need google.colab library to access it. To change the directory between the files, os is used and they can handle operating system tasks. For glob, it is used to retrieve the file name based on the pattern that is provided.

Pandas is a library that can be used to analyze and process datasets. By using Pandas, we are able to import the datasets and work on the manipulation. We are able to delete the unwanted columns, and manipulate the null rows. NumPy is a powerful library and it will be used for calculations in arrays, and matrices. Scikit-learn is a powerful module for machine learning in Python. By using Scikit-learn, we do not have to create our own codes to perform linear regression, lasso regression, ridge regression, and decision tree regressor. It can also help to calculate the R^2 score and RMSE to check whether our training method is correct or not. Scikit-learn provides it all. Scipy is a library that provides functions to solve mathematical problems and things related to scientific

and technical computing [12]. For this research, scipy is used to create the pearson correlation. We also use yellowbrick as the library to help visualize the error from machine learning. It will show the actual and predicted values for it. Next, when we are done with all of the methods and techniques, it seems hard for us to read the data if it is in arrays. Hence, we need Matplotlib to make the visualization of the data so non-expert people are able to read the data and conclusion. To visualize the Pearson correlation table, therefore it can be read easily from the color provided by the library. Pydot_graph is an open source library and it is used to make graph objects that may be filled up with various nodes and edges. The input to make the graph itself is in DOT file format which can be obtained from the machine learning process. As an output, it will export the image into several formats such as png and it is more easier to read.

V. Evaluation Method

To analyse the performance of the model, we split the data into 70% train size and 30% test size and set the random_state into 365. By using the random_state, the accuracy of the test and the data will be constant. Usually the data proportions will be changed randomly and affect the accuracy. By using the scikit-learn library, we are able to split the data easily. Training data means the data will be used for machine learning to be trained. Later on, the result of learning will be used for testing data. As an example, in this dataset there are the number of viewers, likes, comments, and shares. It will learn how the predictors such as likes, comments, and shares affect the number of plays based on the data from training. Later on, the things that are learnt by using the training datasets, will be used to predict the data on testing part.

To evaluate the model's performance, we are going to use two evaluation metrics which are R-squared value and Root Mean Squared Error (RMSE). In general R-squared or coefficient determination values are in the range of 0 to 1 and presented as a percentage. It indicates the amount of variance caused by the independent variables which can be calculated using the formula below.

$$R^2 = 1 - \frac{sum\ squared\ regression\ (SSR)}{total\ sum\ of\ squares\ (SST)}$$

The total sum of squares is the distance between the data and the mean all squared. The closer the data to the mean all squared, the closer R-squared value to 1. Meanwhile, the sum squared regression is the sum of the residuals squared or difference between the predicted value and the actual value.

On the other hand, RMSE quantifies the mean magnitude of the residuals or errors to determine how good a regression line fits the data points [10]. Since it is scale dependent, it can only be used to evaluate prediction errors of different models or model configurations for a single variable, not between variables. RMSE can be calculated using the formula below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Usually a good model has a lower RMSE and a higher R-squared value. Both R-squared and RMSE are the most used regression evaluation methods. Also it is really easy to understand and compatible with most common statistical forecasts.

VI. Results and Discussion

To determine the labels in the dataset that have correlation with each other, Pearson correlation is used here.

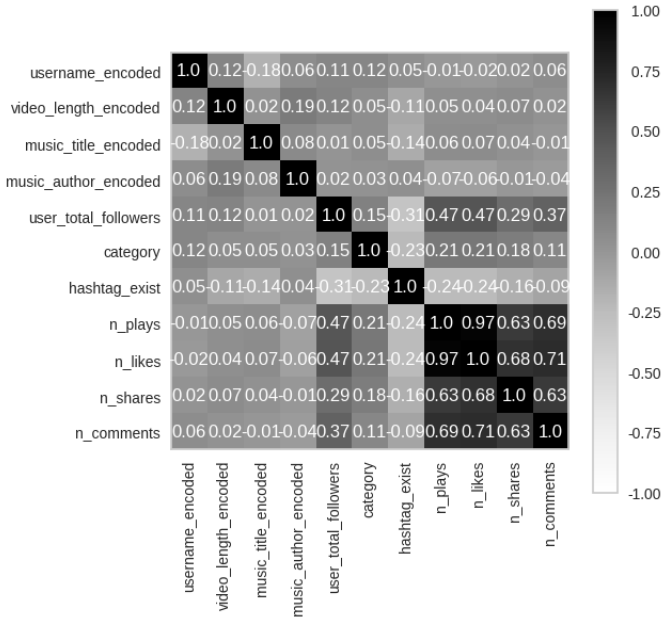


Figure 11. The table from Pearson Correlation and the data is from the columns on the csv files.

From the figure above, it can be seen that the strongest bond here is n_plays with n_likes since the value of it is 0.97. As the researcher expected, it also can be inferred that n_shares, n_plays, n_likes, and n_comments have strong correlation between each other as seen in the below right side of the matrix. Second strongest bond is n_plays with n_comments with the value of 0.69 and followed by n_shares with the value of 0.63. As for the negative bond, n_plays and hashtag_exist have the highest negative bond which is -0.24 and followed by music_author_encoded. As a result, we attempted to develop a model that predicts the number of views per video based on the number of likes, shares, and comments.

```
root mean squared error value of test using Decision Tree:  979763.6064949807
R2 score for testing using Decision Tree:  0.8849618985377403
root mean squared error value of train using Decision Tree:  0.0
r2 score for train using Decision Tree:  1.0
```

Figure 9. The result of R-squared and Root Mean Squared Error for decision tree regressor model.

At first, we did the training and testing by using Decision Tree Regressor. As can be seen from the result above, the R squared and Root Mean Squared Error score indicates the presence of overfitting. The first indicators is the model perform better on training set than the test set. Not only that, the R-squared score value for the training data is 1 and the Root Mean Squared Error score value is 0 which is suspicious because it is too perfect. Unfortunately, we found out that one of the way to solve the overfitting situation is by collecting more data and do training with more data. Most likely, we cannot do that since it's quite tricky to scrape more data from TikTok in such little time. So, we decided to search another way which we find, it's to use regularization. Regularization is a method to decrease a complexity of a model. The model that we choose are Ridge regression model and Lasso regression model that adopted lambda as the penalty term. In addition, we also test out other simple and basic model that is not utilizing regularization which is Linear regression model.

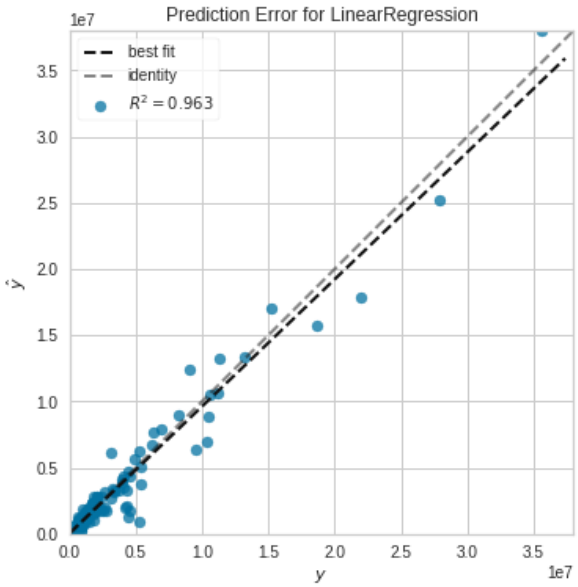
```
root mean squared error value of test using Linear Regression:  555439.6979448542
r2 score for test using Linear Regression:  0.9630280011459039
root mean squared error value of train using Linear Regression:  933364.3394834248
r2 score for train using Linear Regression:  0.9239829503578897

root mean squared error value of train using Ridge Regression:  933364.3394834248
r2 score for train using Ridge Rigression:  0.9239829503578897
root mean squared error value of test using Ridge Regression:  555439.6979448793
r2 score for test using Ridge Rigression:  0.9630280011459006

root mean squared error value of train using Lasso Regression:  933364.3394834247
r2 score for train using Lasso Regression:  0.9239829503578897
root mean squared error value of test using Lasso Regression:  555439.6979456046
r2 score for test using Lasso Regression:  0.9630280011458041
```

Figure 12. The result of R-squared and Root Mean Squared Error for linear regression, ridge regression, and lasso regression.

As we can see the model's ability is already improved and fixed. We can observed from the R-squared and RMSE values for each Linear, Ridge, and Lasso regression model, all the R-squared are above 0.90 or 90% which is very good. We print both the train and test data just to check if there is a big difference in value between the two set. Also we would like to make sure if the train accuracy is bigger than the test accuracy. If yes then maybe our models are not good enough and probably still indicates the present of overfitting, especially for RMSE. It turns out that the result score are not that far from each other. As seen in the figure, the value of RMSE in every model has a really big value reaching thousands in number compared with the R-squared result. It is because our range of data is also big and it can reach millions. We also generate some code to visualize the prediction error and residuals plot. The prediction error allows us to see how much variance is in the model as seen in the figure below. A 45-degree identity line is used to visually display the connection or pattern of the residuals, as well as to determine if the model is over- or under-fitting the provided values. Best fit attempted to build a line from the data that had previously been examined in order to determine the correlation between the expected and measured variables.



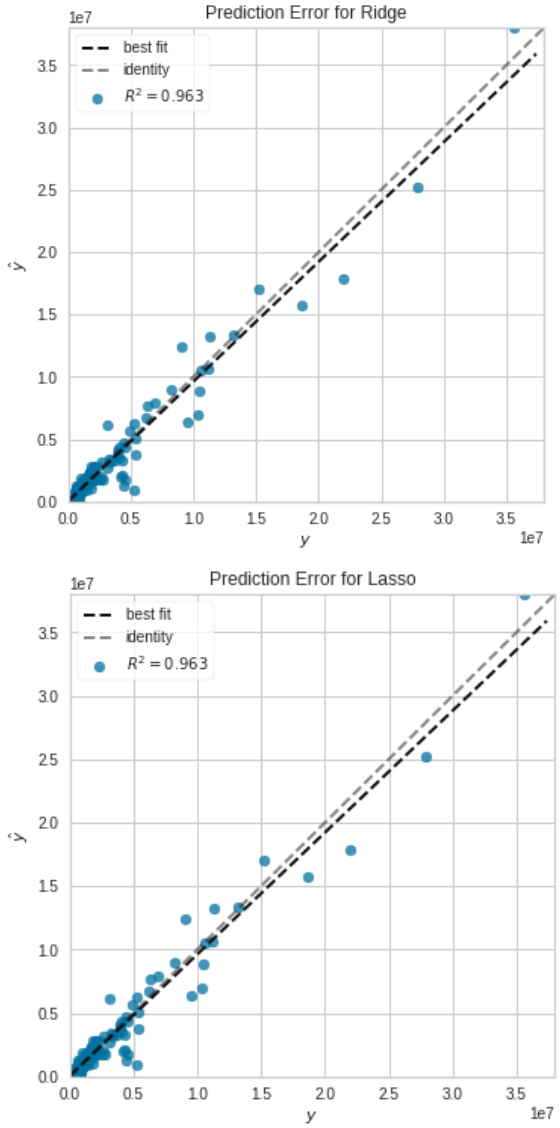


Figure 13. Visualization of prediction error for linear regression, ridge regression, and lasso regression model.

If we look at the figure, there really isn't a significant difference knowing the R-squared value has a slight difference between each other. Meanwhile the residuals plot allows us to see the difference between residuals on the vertical axis and the dependent variable on the horizontal axis. It also can be seen from the graph that the errors are distributed around zero which means the model is well fitted. Similar to prediction error, the residuals plot visualizations are also similar to each other which can be seen below.

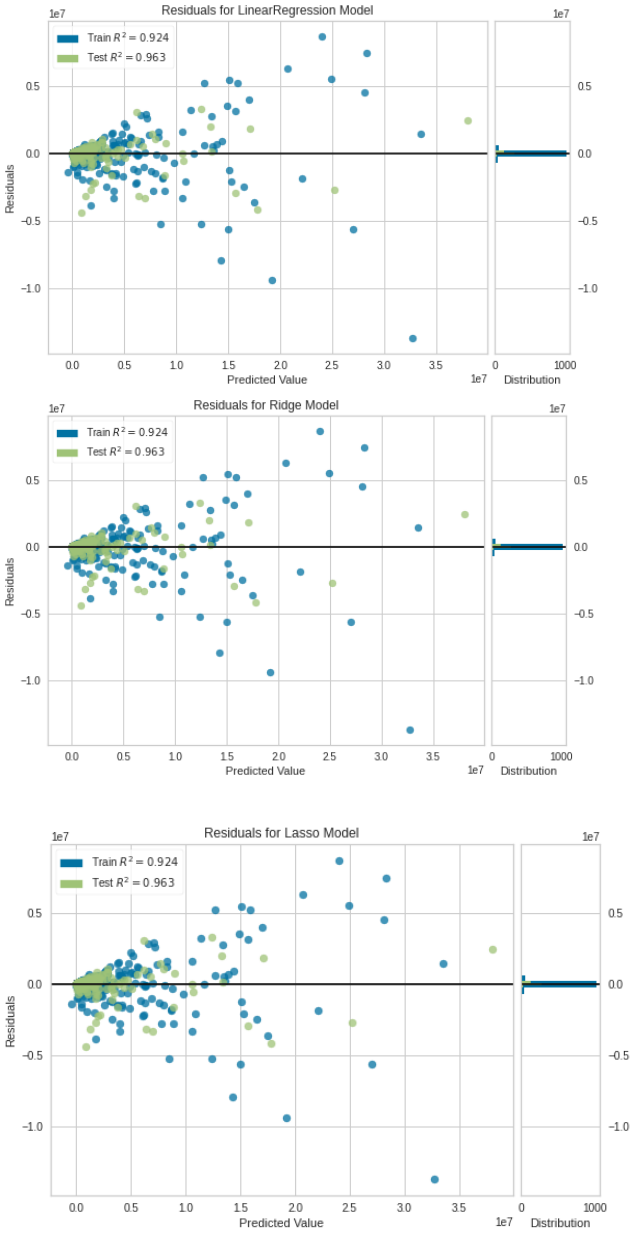


Figure 14. Visualization of residuals plot for linear regression, ridge regression, and lasso regression model.

Next, in this research we also provide the code if users want to know how many views that they can get for their video from the corresponding number of likes, shares, and comments. As it can be seen from the figure below, with 2500 likes, 1200 comments, and 530 shares, a predicted number of plays is around 76640 times. By using this method, people can predict how many numbers of plays that they should gather to receive those several numbers of engagements.

```
#THE ORDER OF THE INPUT IS LIKES, COMMENTS, SHARES
y_pred = LR.predict([[2500,1200,530]])
print("Predicted views: % d\n"% y_pred)

Predicted views:  76640
```

Figure 15. The example of predicted views using likes, comments, and shares as the argument.

VII. CONCLUSION AND RECOMMENDATION

Since TikTok users in Indonesia are quite high in number, it means there will be lots of users that can watch the content because we share the same interest and have similarities in our languages, cultures, and habits. The video will become viral if there are lots of people watching the creators content. That's why it's important to know how to get more viewers. As explained by several sources that Tiktok has a content-based algorithm where a preferred content can be boosted by its views by being disseminated again to other audiences [7]. With all of the results stated, we can draw a conclusion that we can achieve more views by receiving a certain amount of likes, comments, and times our video is shared to other people since the relationship is strong. However, it shows that in order to receive feedback from other users, the content must be entertaining and impactful to others. These predictors also include things that are easier for creators to control. For example, a creator can share his video with 10 friends and ask for likes or comments from each person. If their friends do as they told, they can get more engagement in the form of 10 likes and comments that can boost their video views.

This research also provides bright insights and new knowledge for us that can also be shared with creators who want their videos to go viral on TikTok or get more TikTok views. Besides that, we can also dismiss some theories about how Tiktok algorithm works to get higher views based on the result of our research. Some people claim starting from the use of hashtags and the selection of songs that are used as music can really affects the TikTok algorithm [11]. Even though if you look at the results of our research, there is no strong correlation between views and the presence of hashtags in the video description. Even the choice of songs used in the video is also has negative relationship with the views.

In the future, we can collect more data in order to make the research more valid and see other factors behind the algorithm of TikTok. Since there are some limitations in the data scraping that we did due to Tiktok's strict regulations, we can say that there are some internal data points that are still hidden and we cannot retrieve them. All of the data that we have taken is publicly available data. Therefore, to find out how the Tiktok algorithm actually works, deeper research is recommended and needed. Surely the actual algorithm is much more complex but with this research at least we can know a little about it.

REFERENCES

[1] A. Ahmad, "Pengguna TikTok di Indonesia Mengalami Peningkatan Tiga Kali Lipat Selama Satu Tahun," Bogor Suara, 23-Oktober-2021 [Online] Available: https://bogor.suara.com/read/2021/10/23/110736/pengguna-tiktok-di-indonesia-mengalami-peningkatan-tiga-kali-lipat-selama-satu-tahun. [Accessed 1 December 2021].

[2] Statista. 2021. *TikTok user base in selected countries 2020 | Statista*. [online] Available at: <https://www.statista.com/statistics/1202979/number-of-monthly-active-tiktok-users/> [Accessed 1 December 2021].

[3] S. Hidayatullah, "Pengertian engagement di media Sosial Dan Cara Mengukurnya," MarketingCraft, 26-May-2020. [Online]. Available: https://marketingcraft.getcraft.com/id-articles/pengertian-engagement-di-media-sosial-dan-cara-mengukurnya. [Accessed: 01-Dec-2021].

[4] K. Dhiraj, Top 5 advantages and disadvantages of Decision Tree Algorithm, 24-Dec-2020. [Online]. Available: https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a. [Accessed: 27-May-2019].

[5] Great Learning. 2020. What is ridge regression? [Online]. Available at: <https://www.mygreatlearning.com/blog/what-is-ridge-regression/>. [Accessed: 19-Jan-2022].

[6] Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. Procedia engineering, 201, 746-755.

[7] Brownlee, Linear regression for machine learning. Machine Learning Mastery, 16-Mar-2016. [Online]. Available: https://machinelearningmastery.com/linear-regression-for-machine-learning/. [Accessed: 19-Jan-2022].

[8] J, Aarshay, Complete Tutorial on Ridge and Lasso Regression in Python, 28-Jan-2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/01/ridge-lasso-regression-python-complete-tutorial/. [Accessed: 19-Jan-2021].

[9] S. Gawali, "Linear regression algorithm to make predictions easily," Analytics Vidhya, 09-Jun-2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/linear-regression-in-machine-learning/. [Accessed: 19-Jan-2022].

[10] S. Deepika, Linear, Lasso, and Ridge Regression with scikit-learn, 17-Jun-2019. [Online]. Available: https://www.pluralsight.com/guides/linear-lasso-ridge-regression-scikit-learn. [Accessed: 19-Jan-2022].

[11] B. Shukan, Studi kasus: mengetahui cara kerja algoritma Tiktok 2022, 28-Dec-2020. [Online]. Available: https://www.shukanbunshun.com/2020/05/cara-kerja-algoritma-tiktok.html. [Accessed: 19-Jan-2022].

[12] Great Learning Team, "SciPy tutorial for beginners: Overview of SciPy Library," GreatLearning Blog: Free Resources what Matters to shape your Career!, 13-Nov-2020. [Online]. Available: https://www.mygreatlearning.com/blog/scipy-tutorial/. [Accessed: 19-Jan-2022].

[13] Scikit-yb, Prediction Error Plot, Scikit-yb. [Online]. Available: https://www.scikit-yb.org/en/latest/api/regressor/peplot.html. [Accessed: 20-Jan-22]

Contact and Source Code

The researchers of this study can be contacted through the following channels:
Rainamira Azzahra

- Cell number: +62 8121 2058449
- Email: rainamira.azzahra@binus.ac.id

Christy Natalia Jusman

- Cell number: +62 8965 2171839
- Email: Christy.jusman@binus.ac.id

The source code of TikTok Data Scraping could be found on the link below:
https://github.com/rainamra/TikTok-Indonesia-Statistics.git

Datasets:
https://drive.google.com/drive/folders/1N-ZdDr8TVbBWOQi_6zWWxgy49QGgXE_n?usp=sharing

Questions and Answers

Q: In TikTok, which one is done first? Is it views, likes, comments, or shares?
A: The views first but our logic is when someone give engagements, the TikTok will boost the video to make it available on other people pages. In order to get more distributed, the person need to get the number of engagements

Q: I produce a video then do not have likes, comments, and shares. Then what should I do?
A: Share it to your friends to give comment likes. Hopefully TikTok will boost your videos to more people. TikTok also already declared that it is content-based. So, the engagement between audience and creator is really important so the algorithm will boost it.