Predicting the Popularity Rating of Thai TV Drama by Text Mining of Social Network

Pornpimol Chaiwuttisak

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand
Corresponding author. E-mail address: pornpimol.ch@kmitl.ac.th
Received: 12 October 2020; Revised: 10 February 2021; Accepted: 15 February 2021; Available online: 6 May 2021

Abstract

The objectives of this study were to predict the popularity ratings of Thai TV drama programs with a prediction model, based on found and synthesized factors affecting them, and to check the accuracy of the model in terms of Root Mean Square Error (RMSE) of the predicted outcomes. The analyzed data were both structured and unstructured data. The structured data included the TV channels airing the programs, type of drama, on-air time, number of episodes, average time per episode, number of viewers watching already aired programs, number of viewers watching the highlight of already aired programs, and number of viewers listening to program soundtracks. The unstructured data included messages posted on Twitter. The messages were processed by sentiment analysis, and the sentiments found were statistically analyzed together with the structured data by multiple regression, yielding predicted popularity ratings. The results show that comments on Thai TV drama programs in social media significantly affected the predicted popularity ratings of those programs. A factor affecting the predicted ratings was 'message with positive sentiment'. A factor, the number of viewers watching the highlight of already aired programs, positively affected the popularity ratings when other factors were kept fixed. Another factor, number of viewers watching already aired programs, negatively and significantly affected the popularity ratings (< 0.05). Finally, the RMSE of the prediction model was 0.717 on the training data set containing data from 430,256 people, and the RSME of the prediction model was 0.41 on the test data set containing data from 246,133 people. Our findings may directly benefit Thai TV drama program producers and TV channel administrators in their effort to provide programs that will fully satisfy most viewers.

Keywords: Text Mining, Sentiment Analysis, Multiple Regression, Twitter

Introduction

A broadcasting media plays an important role in social and economic development in Thailand especially terrestrial television or free TV. There were major six analog terrestrial television channels: Channel 3, Channel 5, Channel 7, Channel 9 (or Modern Nine), NBT and Thai PBS. In April 2014 there was a transition to digital broadcasting which was a disruption of Thai television business. Consequently, there are increasing digital terrestrial television channels. The less popular channel cannot survival in the fierce competition situation. Now the remaining 15 digital TV channels still exist (Jada, 2017). From the report of the department of business development, the ministry of commerce, the revenue, profit and overall rating of each TV channel station in 2019 (TV Digital Watch, 2020) are presented in Table 1.

Table 1 revenue, profit and rating of some TV channel stations

TV Channel Stations	Revenue (million baht)	Profit (million baht)	Rating
Channel 7 HD	4,832.28	-1,400.00	1.862
Channel 33 HD	4,092.30	-397.20	1.166

Table 1 (Cont.)

TV Channel Stations	Revenue (million baht)	Profit (million baht)	Rating
Channel One	2,683.43	219.73	0.595
Channel 34 Amarin	3,268.46	167.72	0.355
GMM channel	823.13	-350.30	0.143

Thai TV drama is a kind of entertainment programs in Thailand to make the considerable revenue for the channel station. One of success measures of TV drama including the channel is a rating system. Thus, each TV channel station tries to increase the rating by improving a variety of drama contents and producing good dramas for the public. Besides watching the TV drama, most Thai people talk about Thai TV dramas through tweets and hashtags on twitter (Sodprasert, 2015). As a result, a wave on social media is the increase in the popularity or rating of the drama and drive to the spread of Thai drama across the Asia.

Social media refers to online applications and platforms which people of social networking services (SNS) such as Twitter, Instagram and Facebook build individual profile and social relationships with other people to share information, ideas, and comments, broadening their relational networks (Boyd & Ellison, 2007). Twitter is a micro-blog and social network service on which users post and interact with messages known as tweets. It was developed in the United States of Jack Dorsey, Evan Williams, BizStone such as jointly in 2006. San Francisco venture company. The maximum character that can be written at one time is 140 characters.

Kim, Kim, and Choi (2016) explored the relationship between Average Minute Rating (AMR) and Share Rating (SHR) based on online Twitter data. The result showed that tweets have implications for AMR and SHR in practice. Acharya, Gupta, and Shankar (2019) presented a predictive model to predict the popularity of TV shows based on user comments from social media by using three different data mining techniques: Decision Tree, Naïve Bayes, and XGBoost. Cheng et al. (2013) proposed an anticipating model that is created in the television program fan pages by watchers and Artificial Neural Network to perform conjectures on program evaluations. Kim, Jeon, Kim, Park, and Yu (2012) used Core-Topic-based Clustering to analyze tweets of particular dramas and extracted significant topics.

The objective of the research is to analyze whether the drama's popularity was based on sentiment analysis of tweets on twitter and other factors consisting of number of episodes, duration of the first teaser trailer's release, average time per episode, number of viewers watching already aired programs, number of viewers watching the highlight of already aired programs, number of viewers listening to program soundtracks, TV channels airing the programs, type of drama, and on-air, and to develop a model for predicting the popularity of the Thai TV drama using Multiple Regression Analysis.

Methods and Materials

Data used in the study are 141,283 tweets with hashtag according to the titles of Thai TV dramas in year 2019 by using GetOldTweets tools created by Jefferson-Henrique (Henrique, 2017). They can circumvent

the Twitter API restrictions which are unable to retrieve historical data over 7 days. Moreover, the GetOldTweets tools can work on both Python 2.x and Python 3.

Because of the complexity of Thai language, there is no punctuation like any other languages. Also, there are various and specific word formats, such as royal words. This is a result to makes a difficult for word segmentation. However, there are various tools for the natural language processing on Thai language. In this case, the module "word_tokenize" with engine "newmm" in PyThaiNLP which is Python library is used for tokenization.

Messages on social media express the opinion and feeling. In additional to the text, it still includes emoji symbols or URLs. It is not involved in this research. Therefore, it has to be eliminated in order to obtain research related content by using the Regular Expression (Regex) library. Moreover, the duplicated messages are removed from the data set as well.

Later the text corpora for sentiment analysis are created as a text file with two polarity categories: positive and negative. Corpus is used to consider the number of positive and negative words that appear in a given message. If the number of positive word appearances is greater than the number of negative word appearances, it returns a positive sentiment, and vice versa. If the numbers are even, it will return a neutral sentiment.

For next stage, a model for predicting the popularity of a TV drama is built. The independent variables consist of number of positive sentiment messages, number of negative sentiment messages, and including the other variables that have impact on the rating of Thai TV drama: number of episodes, duration of the first teaser trailer's release, average time per episode, number of viewers watching already aired programs, number of viewers watching the highlight of already aired programs, number of viewers listening to program soundtracks, TV channels airing the programs, type of drama, and on-air time.

The root mean square error (RMSE) is used to evaluate the efficiency of the model because RMSE is the most popular evaluation metric used in regression problems. The research procedure is presented in Figure 1.

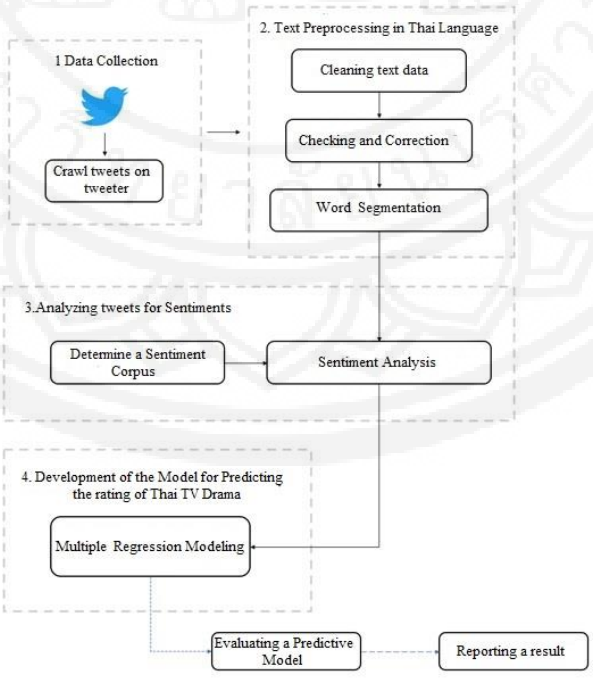

Figure 1 A diagram of the research procedure

The independent variable and the dependent variable must be quantitative variable or have a measurement level as interval scale or ratio scale. In this case, some independent variables have a measurement level as nominal scale or ordinal scale. Thus, they must be converted into dummy variables. Both independent variables and dependent variables are described in the table 2.

Table 2 Dependent variable and independent variables and their respective dummy variables

Name of variable	Descriptive	Type of variable	Number of categories	Number of dummy variables
Independent Variable (X)				
Episodesnum	Number of episodes	Ratio		
Showtitle	Duration of the first teaser trailer's release	Ratio		
Aver.Minutes	Average time per episode	Ratio		
Onlineviewers	Number of viewers watching already aired programs	Ratio		
Highlights	Number of viewers watching the highlight of already aired programs	Ratio		
Music	Number of viewers listening to program soundtracks	Ratio		
Positive	Number of positive messages	Ratio		
Negative	Number of negative messages	Ratio		
Network	TV channels airing the programs	Nominal	3 (TV Channel7, TV Channel 3, TV Channel ONE)	Two = TV Channel 3 & TV Channel ONE "TV Channel 7" is the reference category
Type	Type of drama	Nominal	6 (Tragedy, Romantic, Melodrama, Comedy, Action, Horror)	5 = Romantic, Melodrama, Comedy, Action & Horror "Tragedy" is the reference category
Day	On-air time	Nominal	2 (Monday-Thursday, Friday-Sunday)	1 = Friday-Sunday "Monday-Thursday" is the reference category
Dependent Variable (Y)				
Rating		Ratio		

Results

For Natural language processing to analyze the rating of TV drama through comments on Twitter, it was found that the channel that has the number of messages with the most hashtags of TV drama titles is the TV Channel 7. There were 57,717 messages in the total, representing 40.85 percent of all messages. There are 14,909 positive messages, 15,129 negative messages and 27,679 neutral messages. Following the second popularity TV Channel is TV Channel 3 with 43,641 messages, accounting for 30.89 percent of all messages. The number of positive, negative and neutral-polarized message is 12,080, 10,064, and 21,497 respectively. Finally, TV Channel ONE has 39,925 messages, representing 28.26 percent of all messages. There are 8,360 positive messages, 9,973 negative messages and 21,592 neutral messages.

The frequencies of words have appeared in the tweets about TV drama of Thai TV Channel. The most frequent word for Thai TV channel is negative, that is "ไม่" or "no", followed by positive polarity, that is "ชอบ" or "like" (see in Figure 2).

Figure 2 Word cloud of tweets

A result of the main assumption of Multiple Regression Analysis (MRA) is the following as:

1) Test of Linearity: The investigation of each independent variable is linearly correlated with the dependent variable using Pearson's correlation and Spearman correlation. Correlation coefficient between independent variables and the dependent variable is shown in Table 3.

Table 3 Correlation coefficient

Hypothesis	Correlation Coefficient	p-value	Interpretation
H_0: There is no linear relationship between number of episodes and rating H_1: There is the linear relationship between number of episodes and rating	-0.254	0.266	Accept H_0
H_0: There is no linear relationship between the duration of the first teaser trailer's release and rating H_1: There is the linear relationship between the duration of the first teaser trailer's release and rating	-0.427	0.054	Accept H_0

Table 3 (Cont.)

Hypothesis	Correlation Coefficient	p-value	Interpretation
H_0: There is no linear relationship between average time per episode and rating H_1: There is the linear relationship between average time per episode and rating	0.533	0.013	Reject H_0
H_0: There is no linear relationship between number of viewers watching already aired programs and rating H_1: There is the linear relationship between number of viewers watching already aired programs and rating	-0.599	0.004	Reject H_0
H_0: There is no linear relationship between number of viewers watching the highlight of already aired programs and rating H_1: There is the linear relationship between number of viewers watching the highlight of already aired programs and rating	0.098	0.673	Accept H_0
H_0: There is no linear relationship between number of viewers listening to program soundtracks and rating H_1: There is the linear relationship between number of viewers listening to program soundtracks and rating	0.361	0.108	Accept H_0
H_0: There is no linear relationship between number of positive messages and rating H_1: There is the linear relationship between number of positive messages and rating	0.635	0.002	Reject H_0
H_0: There is no linear relationship between number of negative messages and rating H_1: There is the linear relationship between number of negative messages and rating	0.421	0.058	Accept H_0
H_0: There is no linear relationship between TV channels airing the programs and rating H_1: There is the linear relationship between TV channels airing the programs and rating	-0.876	0.00	Reject H_0
H_0: There is no linear relationship between type of drama and rating H_1: There is the linear relationship between type of drama and rating	0.318	0.160	Accept H_0
H_0: There is no linear relationship between on-air time and rating H_1: There is the linear relationship between on-air time and rating	-0.050	0.829	Accept H_0

The independent variables that have a linear relationship with the dependent variable are an average time per episode and rating, number of viewers watching already aired programs, number of positive messages, TV channels airing the programs

2) Tests of Normality: The statistical test that the independent variable is drawn from a normal population by using Shapiro-Wilk test which requires the sample size to be between 3 to 50 (Shapiro & Wilk, 1965). Moreover, Shapiro and Wilk did not extend their test beyond samples size of 50 (D'Agostino, 1971). The hypotheses used are:

H_0: The sample rating is not significantly different than a normal population.

H_1: The sample rating is significantly different than a normal population

Table 4 Shapiro-Wilk test

	Shapiro-Wilk		
Rating	Statistic	df	p-value.
	0.967	21	0.668

From Table 4, the test statistical value = 0.967, p-value 0.668, which was greater than the significance level = 0.05, indicated that the rating variable has a statistically significant normal distribution.

3) Test to detect Multicollinearity: The state occurs when independent variables are correlated. It indicates that changes in one variable are associated with changes in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently.

Table 5 Collinearity among the variables in a regression model

Independent variable	VIF
Episodesnum	5.837
Showtitle	2.317
Aver.minutes	22.786
OnlineViewers	1.253
Highlights	1.222
Music	2.054
Positive	1.433
Negative	1.475
Day	1.431
CH_3	3.944
CH_one	1.369
Type_2	1.103

Table 5 (Cont.)

Independent variable	VIF
Type_3	1.068
Type_4	1.073
Type_5	1.050
Type_6	1.045

The Variance Inflation Factor (VIF) measures the impact of collinearity among the variables in a regression model. A value of VIF that exceed 10 is often regarded as indicating multicollinearity. From Table 5, an average time per episode, which is the independent variable, is related to each other with VIF value of 22.786, therefore it is necessary to eliminate the highly correlated variable from multiple regression analysis.

Table 6 Model summary

R	R Square	Adjusted R Square	Durbin-Watson
0.905	0.820	0.775	1.811

As Table 6, the R column represents the value of R, the multiple correlation coefficient. R can be considered to be one measure of the quality of the prediction of the dependent variable. That is "Rating". A value of 0.905 indicates a good level of prediction. The R Square column represents the R^2 value (also called the coefficient of determination), which is the proportion of variance in the dependent variable that can be explained by the independent variables (technically, it is the proportion of variation accounted for by the regression model above and beyond the mean model). The value of 0.820 means that the independent variables explain 82.0% of the variability of "Rating" dependent variable.

Table 7 ANOVA

Model	Sum of Squares	df	Mean Square	F	p-value
Regression	17739704932957.080	4	4434926233239.270	18.177	.000
Residual	3903874849900.068	16	243992178118.754		
Total	21643579782857.150	20			

The F-ratio in the ANOVA table (see Table 7) tests whether the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable, $F(4, 16) = 18.177$, $p < .005$.

The general form of the equation to predict the popularity rating from number of viewers watching the highlight of already aired programs, number of viewers watching already aired programs, TV channels airing the programs, number of positive messages is:

Predicted popularity rating = 5,831,034.395 + (443.959 x Positive) +(- 1244179.653 x OnlineViewers) + (- 835183.907 x CH_one) + (549174.104 x Highlights)

This is obtained from the Coefficients table, as shown in Table 8.

Table 8 Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	p-value
	B	Std. Error	Beta		
(Constant)	5831034.395	1688583.392		3.453	.003
Positive	443.959	192.486	.293	2.306	.035
Online Viewers	-1244179.653	246377.557	-.600	-5.050	.000
CH_one	-835183.907	267519.738	-.388	-3.122	.007
Highlight	549174.104	221082.148	.292	2.484	.024

Testing the assumptions for residuals

1. Test for Normality of Residuals: The test statistical value = 0.954, p-value 0.404 which is greater than 0.05 indicates that the values of the residuals are normally distributed.

2. Test for Mean of Residuals: t-test = 0.00, p-value = 1.00 is greater than the significant level at 0.05. It can state that the mean of residuals is also equal to zero.

3. Test for Independence of Residuals or No Autocorrelation of Residuals: The Durbin-Watson statistic is used to test the assumption that the residuals are independent or uncorrelated. This statistic can vary from 0 to 4. A value should be between 1.5 and 2.5. In this case, the value is 1.811, so it can be said that this assumption has been met.

4. Test for Homogeneity of Variance or Homoscedasticity: The last assumption of multiple linear regression is homoscedasticity. The Breusch-Pagan test (Breusch & Pagan, 1979) is used to determine if homoscedasticity is present. The null hypothesis of the test states that there is constant variance among the residuals. As Table 9, p-value is greater than 0.05. It cannot reject the null hypothesis and conclude that homoscedasticity is present in the data.

Table 9 Breusch-Pagan and Koenker test statistic

	LM	p-value
BP	5.837	0.120

Thus, the model can be applied to predict the rating of Thai TV drama. The performance of the above models can be measured in terms of the root mean squared error (RMSE). It indicates the spread of the residual errors. It is always positive, and a lower value indicates better performance. The RMSE of the model is equal to 0.717. When calculating the RMSE of the testing data set, the RMSE is equal to 0.410.

Discussion and Conclusions

The number of positive messages is account for approximately 25%, the same as the number of negative messages. The rests of them are the neutral messages. The most total of positive messages is the drama of TV channel 7 and the most total of negative messages is the dramas of TV channel ONE. The predictive model has a R^2 value of 0.820, indicating that the independent variables are selected into the multiple regression models. There are 5 variables as follows: number of positive messages, number of viewers watching already aired programs, TV channels ONE and number of viewers watching the highlight of already aired programs. They can describe the rating of Thai TV dramas correctly 82%. If the number of positive sentiment messages increased by 1 while the other variables were constant, the rating would increase by 0.7399.

For further research, we will study the other predictive techniques such as Multivariate Adaptive Regression spline (MARs) as a non-parameter regression technique to reduce the limitations of Multiple Regression Analysis, which has preliminary assumptions for data.

References

Acharya, S. S., Gupta, A., & Shankar, P. K. C. (2019). TV Show Popularity Analysis using Social Media, Data Mining. *International Journal of Innovative Technology and Exploring Engineering*, *8*(7), 23-26.

Boyd, D. M., & Ellison, N. B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, *13*(1), 210-230. https://dx.doi.org/10.1111/j.1083-6101.2007.00393.x

Breusch, T., & Pagan, A. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, *47,* 1287-1294. https://dx.doi.org/10.2307/1911963

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika*, *58*(2), 341-348. https://doi.org/10.1093/biomet/58.2.341

Henrique, J. (2018). *Get Old Tweets Programatically.* Retrieved from https://github.com/Jefferson-Henrique/GetOldTweets-python

Jada, P. (2017). Survey and Analysis of Studies on the Effects of Television Media's Transition from Analog to Digital. *NBTC Journal*, *4*(1), 160-173.

Kim, D., Kim, Y., & Choi, S. (2016). Predicting the popularity of TV-show through text mining of tweets: A Drama Case in South Korea. *Journal of Internet Computing and Services*, *17*(5), 131-139. https://doi.org/10.7472/jksii.2016.17.5.131

Kim, S., Jeon, S., Kim, J., Park, Y., & Yu, H. (2012). Finding Core Topics: Topic Extraction with Clustering on Tweet. *2012 International Conference on Cloud and Green Computing (CGC), 1-3 November 2012* (pp. 777-782). Chinese: Xiangtan.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika*, *52,* 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Sodprasert, S. (2015). Perception and Feedback Behavior of TV Series Viewers on Facebook. *Dhonburi Rajabhat University Journal*, *12*(1), 65-80.

TV Digital Watch. (2020). *Digital TV Revenue.* Retrieved from https://www.tvdigitalwatch.com/category/highlight/revenue/performance/