IDS 702 Final Project

YouTube Comments Sentiment Analysis

Vishaal Venkatesh

1. Summary:

Becoming successful on YouTube is an incredibly creative and difficult task and is far from something that could be accurately modelled. However, in this analysis, we attempt a novel approach and trying to model the Likes to Dislike ratio (L/D) on a video using the sentiment of the comments. The sentiments were calculated using a lexicon and rule-based sentiment analysis tool called VADER. It was observed the compound score was indeed a significant, strong predictor of L/D ratio. Moreover, the same model applied to a collection of popular but controversial YouTube channels performed comparably well. Finally, correlations between the sentiments of the top comments under a video, the sentiment of the title and the average sentiments of all other comments under a video was studied. A top comment is likely to be positively correlated with the title of the video and is likely to be uncorrelated with the average sentiments of all other comments.

2. Introduction:

YouTube's algorithm to monetarily compensate content creators is vastly complex. It depends on the views on the video, the number of subscribers of the content creator, the advertisement-appropriateness of the content of the video, the likes to dislike (L/D) ratio etc. Some sources even suggest that that a content creator's monetary success depends on how much interest a video is able to instill in the viewer. For example, if a video on black holes consistently makes viewers watch more recommended videos on black holes, then the creator of the first video is proportionally compensated as their video instilled in the viewer an interest to watch more videos on blackholes. As suggested earlier, this can become incredible intricate, however, the L/D ratio serves as a good standalone metric to predict and/or gauge a YouTuber's success. Maintaining a high like to dislike (L/D) ratio is critical to a YouTube content creator's long-term success. While there are exceptions to this, most consistent, top YouTuber's have very high L/D ratios. Viewers may voice their opinions either by liking/disliking the video, commenting on the video or by doing both. The goal of this study is to answer the following questions.

1. Develop a model to study the association between a YouTube video's L/D ratio and the sentiments of the comments.
2. See if the same model can be used to predict the L/D of a popular collection of controversial YouTube channels (Daily Show with Trevor Noah, CNN, ABC News, Cardi B, MILO, Logan Paul Vlogs, WatchMojo.com, Legally Armed America, ProPublica, Courageous Conservatives PAC.)
3. Analyze the correlation between the sentiment of the most popular comment under a video and its general sentiment relative to both the sentiment of the video's title and all other comments under that same video.

The sentiment of the comments and titles were quantified into a numerical score using the text sentiment analysis tool called Valence Aware Dictionary for sEntiment Reasoning (VADER). VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. VADER uses a lexical approach to sentiment analysis, i.e., it maps words, phrases and sentences to a dictionary of sentiments. Following this matching, the sentence is provided with four different scores – a positive, neutral, negative and compound score. The first three scores are on a scale of [0, 1] while the final compound score is the standardized sum of the first three scores on a scale of [-1, 1]. One reason for VADER's success and popularity lies in its ability to function without having to train a model. Everything we need to analyze a sentence is contained in the so-called dictionary of emotions. It is also noteworthy that

VADER is very capable of dealing with emoticons, slang and a variety of contextual punctuation marks – all of which are very useful when dealing with social media text analysis.

It is important to understand what defines a popular controversial YouTube channel, and in what way are the above-mentioned channels are controversial. A popular controversial channel, as defined for this analysis, is any channel that receives on average above 10,000 views per video, has a L/D ratio greater than 5 and has an average negative compound score for all its comments. In other words, the video is popular because it is generally well-received as indicated by the views and relatively high L/D ratio but is controversial due to the negative nature of the comments. News channels like CNN, ABC News etc. and investigative journalism outlets like ProPublica may have controversial news stories or may report on stories with an inherent political bias. This might garner negative responses from the audience. Channels like Courageous Conservatives PAC and legally Armed America may be targeted towards a specific type of audience and may be negatively received by other due to the controversial topics they might potentially base their content on. Other channels like Logan Paul Vlogs may be controversial due the cult-like following they might have and the different factions their fans might get into (one popular incident of fans taking sides was during the boxing event between Logan Paul and KSI in 2019 – not reflected in our data). It has to be noted that the data we have is from late 2017 and this was before the infamous and controversial incident when Logan Paul filmed an unfortunate hanging man's corpse in Japan and made a mockery of the whole situation. The situation drew huge backlash from the online community, but it happened in early 2018 and is not reflected in our data. It was surprising to discover that entertainment channels like Cardi B's and Trevor Noah's were controversial at the time. Cardi B, understandably, was just getting into the music scene in 2017. This may have resulted in some mixed opinion from the public. It was however difficult to uncover to why Noah's channel emerged controversial.

3. Data:

Two different datasets were used for this analysis. The first dataset contained data on the videos themselves, while the second dataset contained information on the comments under each video listed in the first dataset. The data is about videos featured in the trending category on YouTube between the dates of 13th September and 22nd October 2017 in the United States. Both datasets were linked by a unique video ID. The datasets were obtained from the popular data science website by the name of Kaggle. Following is a brief schema of the simplified videos dataset with all the relevant columns (Table 1).

Table 1. *Schema of Videos Dataset*

video_id	date	title	channel_title	category_id	views	likes	dislikes	L/D

The video dataset contained data on the name of the video, date when the video was uploaded, name of the channel, category (one of fifteen categories like news, entertainment etc.), views, likes and dislikes. The L/D ratio was calculated to simply be the ratio of the number of likes to the number of dislikes. It has to be noted that there were in all 73 video that had no dislikes whatsoever. These videos had to be removed from the dataset as a they would result in a L/D of infinity. There were no instances of missing data and no imputations or deletions had to be performed. Furthermore, about 2/3 of the videos in the dataset were repeated with different dates associated with them. These were simply a single video trending for multiple days and accumulating different number of views, likes and dislikes over these days. For the sake of this analysis, the likes, dislikes and views were summed up by video_id and each video were treated as a single entity despite trending on multiple days. Following the videos dataset, the schema for the comments video set has been included below (Table 2).

Table 2. *Schema of Comments Dataset*

video_id	comment_text	likes	replies

As evident from the above schema (Table 2), the comments dataset contained information of the content of the comment, the number of likes each comment received and the number of replies. It has to be noted that there were precisely 9 occurrences of incorrect textual data in the reply's column. These columns were removed from the dataset and were not used in the analysis. Furthermore, for future analysis, the most popular comment under a video was the comment that received the greatest number of likes for a given video_id. Finally, a one-to-many, inner join on the video_id was performed on the two datasets before using VADER to score the comments.

A few plots and insights from the Exploratory Data Analysis (EDA) have been summarized below to provide better segue into the model. From the correlation heat map given below (Figure 1), it was observed that there were some obvious positive correlations such as the number views and the likes on the video. An obvious negative correlation was between the positive score of a comment and the negative score of the same comment – a comment cannot be highly positive and negative at the same time. More interesting was the correlations between the dependent variable (L/D) and the various was possible predictors. L/D was slightly positively correlated with likes, positive score and compound score and was mildly negatively correlated with views, dislikes neutral and negative scores. It remains to be seen which of these predictors will emerge significant. A few potential interaction effects were also observed between the compound score and the category id when predicting L/D ratio (Figure 2). This is depicted by the intersection lines and although the interactions do not seem to be major (the lines don't intersect perpendicularly), it still remains to be seen if any of the interaction terms significant. Note than only a handful of categories were included in the interaction plot to aid readability.

Figure 1. *Heat Map of Correlations*

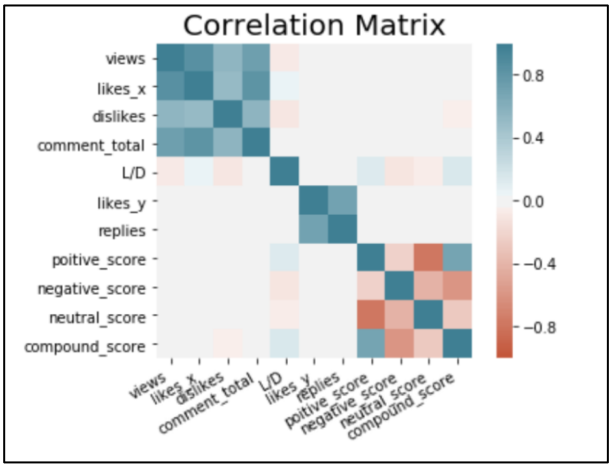


Figure 2. *Interaction b/w categories and compound score*

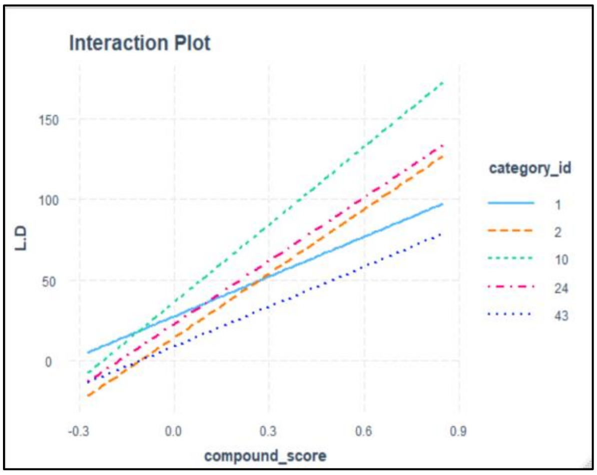
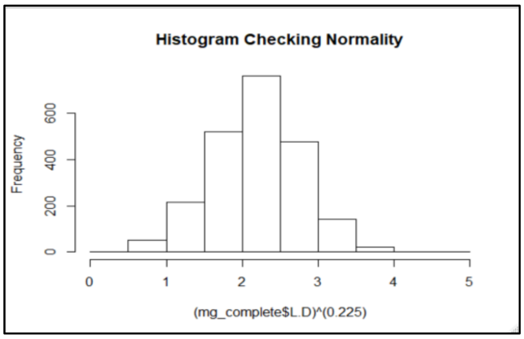


Figure 3. *Normality post transformation*



Finally, a note on the normality of the dependent variable – L/D ratio. The distribution, as given was not remotely normal. Achieving normality using conventional transformations such as logarithmic and simple power transformations did not help achieve skewness and kurtosis similar to that of a normal distribution. Therefore, a method known as the Tukey's Ladder of Power was used. This method identifies a power to which the dependent variable may be raised while preserving the order and proximities between observations i.e., using smooth functions. The power is optimized to make the resulting transformation as similar as possible to a normal distribution. Using this method, the optimal power was determined to be 0.225. The histogram above (Figure 3) depicts the transformed plot.

4. Model:

The following section has been broken down into three different categories to answer the three questions posed in the introduction section of this report.

(4.1) A multiple-regression model including interactions effects was used to predict the L/D ratio using the average compound score of the comments and other predictors. Data on the controversial channels studied in the second part of this section was not included when building the model. The model developed in this section is the model trained on all other channels – that shall then be used on the popular controversial channels to study how well it can predict L/D ratio. The base model was developed using the variables views, positive, negative, neutral and compound score – as revealed by the correlation heat map from above. The variables like and dislikes were not included as these were obviously correlated to the L/D ratio. The variables view and compound score emerged as significant at the 0.05 significance level. These variables were retained, and the others were discarded. It was imperative to then check for multicollinearity as some strong correlations were evident from the correlation heat map above (Figure 1). This was done by calculating the Variance Inflation Factors (VIF). All VIFs were less than 5, suggesting acceptable levels of multicollinearity and hence lesser chances for overfitting. Next, it was important to analyze the interaction effects between the compound score and category id when predicting L/D ratio. The category id variable was thrown into this base model as a predictor (sans the interacting term). Some categories did emerge significant, and therefore the variable category id was included in the model. Following this addition, a nested F-test was performed to between this model and another model with the added interaction term between category id and the compound score. The p-value resulting from the F-test was less than 0.05 and hence significant. The interaction term was now included into the base model. Following this development, a step-wise regression using backward model selection was performed to corroborate our model thus far. The model result from the step-wise regression was identical to the model that was manually developed. The regression summary has been included in the appendix along with the 95% CI for the estimates. The category ids have been replaced with the actual category names to aid in interpretation.

Before interpreting the results of the model, model diagnostics were performed to make sure the model met the assumptions of linearity, independence, normality and homoskedasticity. The plots have been included in the Appendix. No obvious patterns were observed when the residuals were plotted against a continuous predictor suggesting the residuals were linear. Moreover, almost all the points on the qq-plot were along the 45^o line suggesting the normality assumption had been met. Furthermore, looking at the Residuals Vs Fitted values plot, it can be observed that the points were mostly uniformly distributed suggesting homoscedasticity (constant variance). Moreover, there were no obvious patterns in the Residuals Vs Fitted values plot suggesting independence. All assumptions of multiple regression have been met and the interpretations made thus forth can be considered valid.

Following are some of the main interpretations from the model. As hypothesized earlier, the average compound scores of all the comments under a particular video did emerge significant when predicting L/D ratios. It was not only significant but also a strong predictor (high magnitude of t-value). A complete turnaround of the compound scores from -1 (all very negative comments) to 1 (all very positive comments) results in an increase in the L/D ratio by 24.9 (95% CI [0.25, 190.9]). An increase of 0.5 in the compound score results in an increase of 0.05 (95% CI [0.0011, 0.402]) in the L/D ratio. Overall, this confirms our initial hypothesis that the compound score is positively correlated with the L/D ratio. It, however, has to be emphasized that this does not imply causality and only suggests association. Let us consider the interpretation of videos belonging to the sports category. If a video were to belong to the sports category, then a 0.5 increase in the compound score results in an increase of 0.381(95% CI [-0.21, 13.13]) in the L/D ratio – including the effect of both the categorical variable and the interaction term. Similar interpretations could be made for the other categories in the dataset.

(4.2) As mentioned earlier, the second question dealt with popular, controversial channels. We aim to see if the above developed model can be used to model the relation on the videos on the following channels - Daily Show with Trevor Noah, CNN, ABC News, Cardi B, MILO, Logan Paul Vlogs, WatchMojo.com, Legally Armed America, ProPublica and Courageous Conservatives PAC. Please note that data from these channels were not used when training the model. While ostensibly it might seem that the R^2 would be a good metric to compare fit on the new data, the R^2 is only clearly defined for a model for the training data set. Attempting to calculate the R^2 on new data may result in values of R^2 greater than 1. This is because, if the model is a very poor predictor of L/D on the new data, the magnitude of the residuals can be greater than the overall variability in the new data – resulting in an R^2 greater than 1. Instead, we decide to compare the in-model and out-of-model Mean Squared Error (MSE). The in-model MSE is simply the square of the RMSE – a measure of the error in the actual values and modelled values in the original data. The out-of-model MSE is a similar metric – but is calculated using the predicted values on the new data and the actual values in the new data. If these metrics were comparable, the model is performing comparably in both the datasets. In other words, the previously developed model does a good job in modelling the relationships between L/D ratio and compound score. The values of the MSEs have been included in the following table (Table 3). The out-of-model MSE was surprisingly lesser on the new data as compared the training data. While this is unusual, it has difficult to argue that the model performs poorly on the new data. The same model can therefore be used to model the controversial YouTube channels mentioned above.

Table 3. *Values of MSEs*

In-Model MSE	**Out-of-Model MSE**
0.2095	0.1848

(4.3) The final question that requires answering is how the sentiment of the top comment under each video correlates with both the sentiment of the video title and the average sentiment of all other comments in that particular video. To achieve this, the magnitude of the correlation between the sentiment of the top comment and the title sentiment was first calculated. This was mildly positively correlated at 0.13. This suggests that the top comment in each video is positively correlated with the sentiment of the title. A similar correlation was then calculated between the sentiment of the top comment and the average sentiment of all other comments in a particular video. This came out to be very mildly positively correlated at 0.0299. This is almost equal to zero and hence can be considered thus. This suggests that a popular comment is likely to be positively correlated with the sentiment of the title and likely to not be correlated with the average sentiment of all other comments.

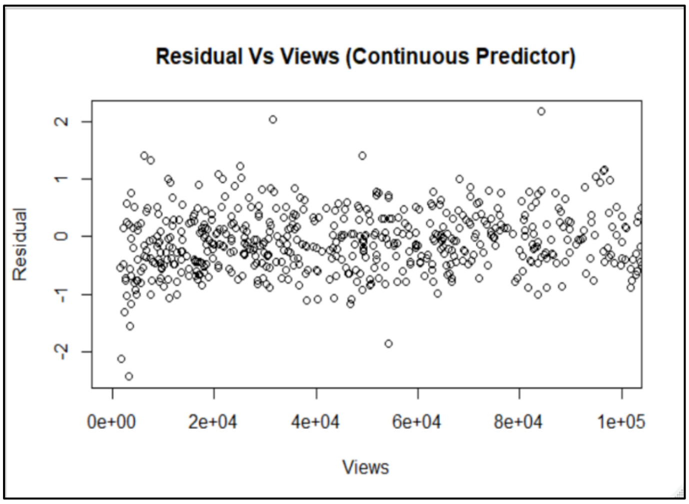
5. Conclusions:

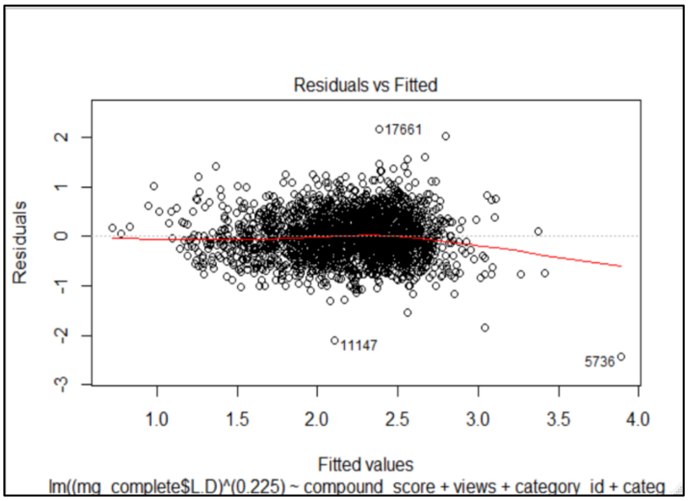
Overall, the analyses were successful, however, it is imperative to discuss the limitations of the analyses. Firstly, VADER is shockingly poor at detecting sarcasm. Some comments on YouTube may have been sarcastic but VADER may have misclassified the comment and miscalculated the sentiment score. That said, VADER is one of the best text sentiment analysis tools for social media, but the results have to be considered taking into account its limitations. Finally, the definition of popular controversial channels was very specific – any channel that averages above 10,000 views per video, has a L/D ratio greater than 5 and has an average negative compound score (less than zero) for all its comments. Moreover, the collection of controversial YouTube channels was limited to 11 channels in this analysis. Any changes to the definition or data may or may not affect the performance of our model on newer data. This has to be kept in mind when extending this data to newer models.

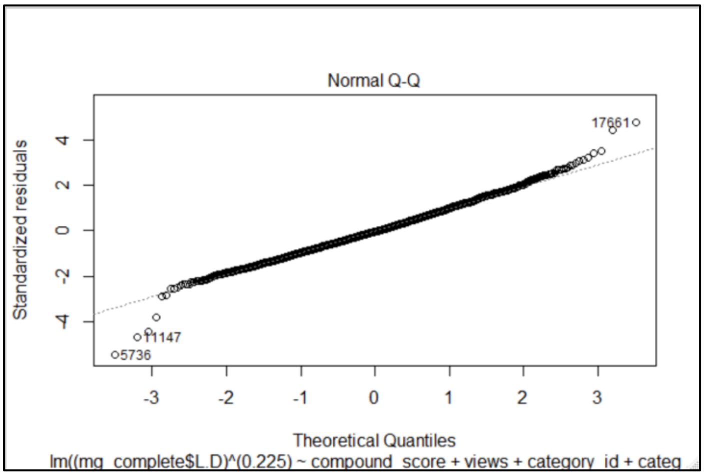
6. Appendix

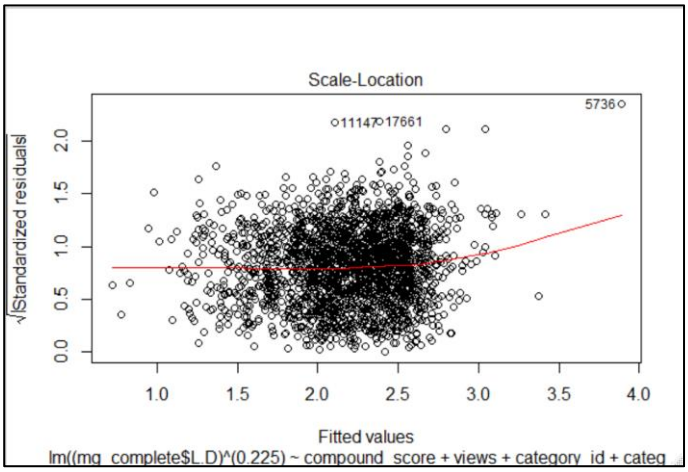
Model summary of the final model used.

Variable	Estimate	CI 2.5%	CI 97.5%	Standard Error	t-value	p-value
Intercept	1.95	1.81	2.08	0.068	28.32	*<2E-16*
Compound score	1.03	0.434	1.63	0.30	3.39	*0.0007*
Views	1.03E-9	-7.58E-9	9.63E-9	4.39E-9	0.23	0.82
Cars & Vehicles	-0.19	-0.45	0.063	0.13	-1.47	0.14
Music	0.24	0.062	0.42	0.091	2.64	*0.08*
Pets & Animals	0.23	-0.13	0.59	0.18	1.23	0.22
Sports	-0.27	-0.43	-0.11	0.08	-3.23	*0.001*
Travel & Events	-0.19	-0.64	0.25	0.22	-0.874	0.38
Gaming	-0.16	-0.48	0.16	0.16	-0.99	0.32
People & Blogs	-0.24	-0.39	-0.081	0.08	-2.96	0.03
Comedy	-0.048	-0.21	0.12	0.23	-0.58	0.56
Entertainment	-0.12	-0.27	0.023	0.16	-1.65	0.098
News & politics	-0.41	-0.57	0.41	0.080	-5.29	*1.34E-7*
How-to & Style	0.24	-0.17	0.18	0.083	2.44	*0.014*
Education	0.23	-1.59	0.234	0.074	2.57	*0.01*
Science & Technology	0.0029	-1.13	0.68	0.078	0.03	0.97
Nonprofits & Activism	-0.69	-0.67	1.83	0.47	-1.49	0.14
Shows	-0.23	-0.58	0.85	0.46	-0.50	0.61
Compound Score: Cars & Vehicles	0.58	-1.13	1.55	0.63	0.912	0.36
Compound Score: Music	0.13	0.37	1.9	0.36	0.36	0.71
Compound Score: Pets & Animals	0.21	-0.082	3.33	0.68	0.31	0.75
Compound Score: Sports	1.12	-0.98	2.16	0.40	2.87	*0.004*
Compound Score: Travel & Events	1.62	0.92	2.29	0.87	1.87	0.062
Compound Score: Gaming	0.59	0.77	2.36	0.80	0.73	0.56
Compound Score: People & Blogs	1.61	0.24	1.57	0.35	4.6	*4.45E-6*
Compound Score: Comedy	1.56	-0.2	1.30	0.41	3.9	*0.00012*
Compound Score: Entertainment	0.91	-0.94	0.49	0.34	2.7	*0.0075*
Compound Score: News & politics	0.55	-0.27	1.21	0.38	1.43	0.152
Compound Score: How-to & Style	-0.23	-0.94	0.48	0.36	-0.62	0.532
Compound Score: Education	0.09	-0.74	0.92	4.24	0.2	0.841
Compound Score: Science & Technology	-0.02	-0.89	0.84	4.37	-0.05	0.96
Compound Score: Nonprofits & Activism	NA	NA	NA	NA	NA	NA
Compound Score: Shows	NA	NA	NA	NA	NA	NA
Adjusted R^2 = 0.3836 **Residual Standard Error = 0.4577**						


Residual Vs Views (Continuous Predictor)
Residual
-2
-1
0
1
2
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Views


Residuals vs Fitted
Residuals
-3
-2
-1
0
1
2
17661
11147
5736
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
lm((mg_complete$L.D)^(0.225) ~ compound_score + views + category_id + categ


Normal Q-Q
Standardized residuals
-4
-2
0
2
4
17661
11147
5736
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm((mg_complete$L.D)^(0.225) ~ compound_score + views + category_id + categ


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
5736
11147
17661
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Fitted values
lm((mg_complete$L.D)^(0.225) ~ compound_score + views + category_id + categ