Linear Statistical Models

SYS 6021

Final Project

Analysis of "Buzz" in Twitter

Sriram Raju Dandu
sd9aj@virginia.edu

Summary

There are a couple of recommendations to start a 'Buzz' in Twitter.Firstly,attract more authors to talk about the topic and also should make the authors interact among themselves. Secondly, make sure that the burstiness level is increasing along with length of the discussion.The evidence I drew from the linear models I designed demonstrated that the variables of number of authors, author interaction, burstiness level and average discussion length had significant impact on the metrics to predict Buzz. There are a couple of metrics , namely, mean of gradient and cross correlation with number of active discussions. The 95% confidence intervals of the parameters for both the metrics are [1.05×10^{-2} , 1.15×10^{-2}] and [1.41×10^{-2} , 2.65×10^{-2}] respectively. This shows that the predictors and the 'Buzz' period are directly proportional.

Honor Pledge: On my honor, I pledge that I am the sole author of this paper and I have accurately cited all help and references used in its completion.

1 Problem Description

1.1 Situation

There are 7.2 billion people on the planet out of which there are approximately 3 billion active Internet users (i.e. 45% of the worlds population).There are nearly 2.1 billion people have social media accounts and close to 1.7 billion people have active social media accounts. Social networking platforms include Twitter, Facebook, Instagram, Google + and so on. Twitter has 284 million active users and on an average 500 million tweets are posted per day.

In [5], authors claim that in recent years, social media has become ubiquitous and important for social networking and content sharing. In this paper, authors demonstrate how social media content can be used to predict real-world outcomes. They show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. They further demonstrate how sentiments extracted from Twitter can be utilized to improve the forecasting power of social media.

In [4] authors say that platforms dedicated to social networks are the scene of new social phenomenon. Certain keywords experience a steep increase in their popularity, characterized by a large number of threads over a short period of time. These activity peaks are frequently qualified to "buzz".

It is characterized by: a support theme (i.e. a set of keywords), a period, and finally a measure of attention. The buzz period begins when the media theme captures a sufficient share of attention available on the target social network. This ends when the attention by theme picked up by the media is no longer sufficient. Buzz can occur gradually or abruptly.

Figure 1 shows the pattern of number of active discussion (NAD) over time. We can see that the cluster tends to move up as look from time 0 to 6. The table 1 also shows that the NAD values tend to increase over time. If the NAD value is higher then it means that there is a lot of activity happening (or Buzz). Thus, the figure 1 and the table 1 show that the NAD values tend to increase with time in the data set I have. This is the reason I will be using NAD as my response variable to analyze 'Buzz'.

NAD wrt time	Mean
0	168.3002
1	164.7896
2	191.437
3	219.4985
4	244.3768
5	266.7924
6	265.4408

Table 1: The mean values of NAD over time

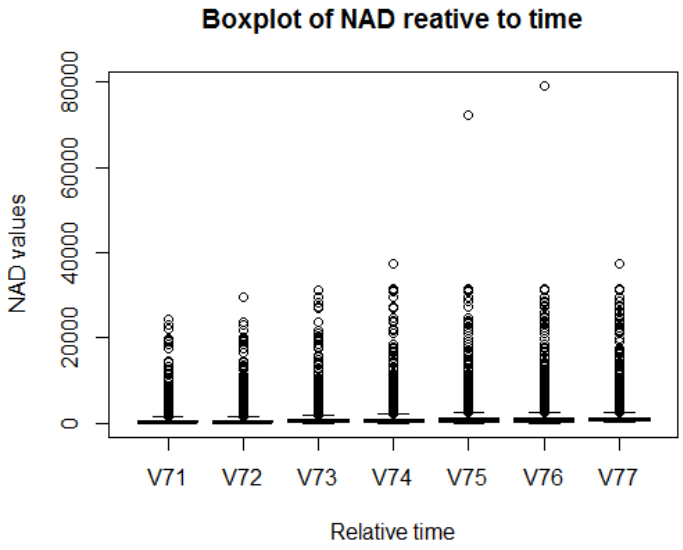

Figure 1: Boxplot of NAD with respect to time

Figure 2 shows the correlation between the features ADL,CS,Aint,BL and AtL with NAD. The abbreviations of the features can be found in table 2. We can see that there is high correlation between NAD and AtL. There is also high correlation between ADL and Aint which shows that discussion length and author interaction are closely related. Also burstiness level (BL) and contribution sparseness (CS) are highly correlated.
Figure 3 shoes that there is almost one-to-one correspondence between NAD and number of authors (NAu).This might be a good features to look out for to meet my goal.
This report will be written using a template provided by the instructors of the SYS6021 course at the University of Virginia [1] and according to the assignment description also provided [2].

1.2 Goal

To find/analyze the parameters that trigger a Buzz in twitter

1.3 Metrics

For measuring the "Buzz" in twitter, I have chosen two metrics, namely,

- Mean of the gradient:
  Measures the average change in number of active discussions(NAD) with respect to time

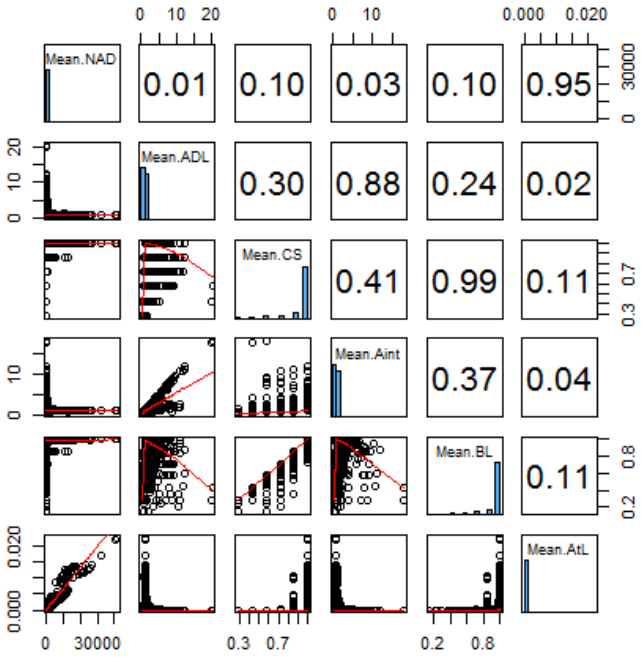

Figure 2: Correlation among various features

- Cross correlation with NAD:
  Measures the relation with the number of active discussion(NAD) with respect to time

1.4 Hypothesis(es)

Below are the hypothesis I inferred,
For Mean of the gradient :

- Increased number of authors doesn't increase the number of active discussions

For Cross correlation with NAD :

- Change in Burstiness level (BL) over time is not correlated with the change of NAD with respect to time.

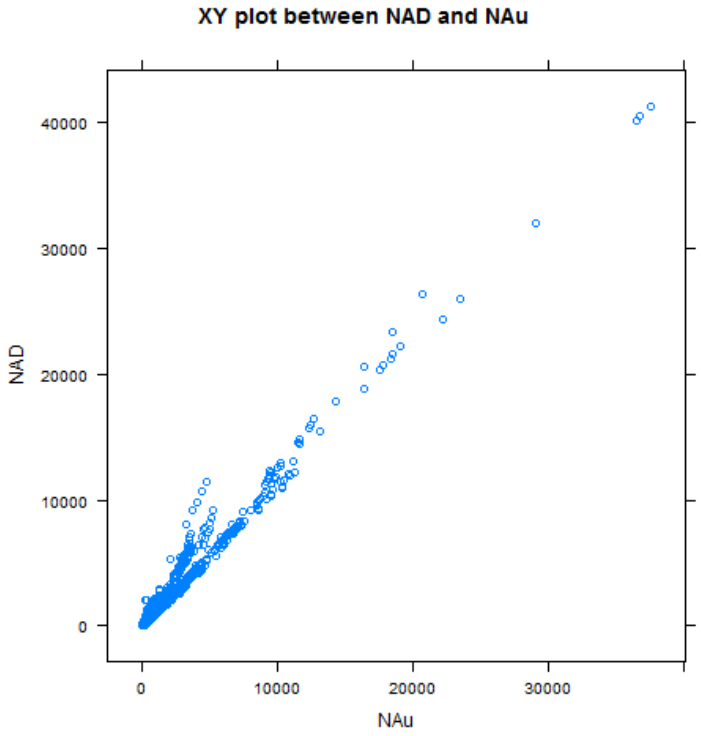

Figure 3: XY plot between NAD and NAu

2 Approach

2.1 Data

The data-set contains examples of buzz events from the social networking platform Twitter. I will be using the data set from [3]. The data set has 38393 observations with 77 features.Each instance is described by 77 features, those describe the evolution of 77 'primary features' through time. Hence each feature name is post-fixed with the relative time of observation. For instance, the value of the feature 'Nb Active Discussion' at time t is given in 'Nb Active Discussion t'.There are 11 different parameters and each of them have 7 time-relative recordings. So every time-relative recording of the parameter is considered as an attribute, making it 77 attributes. The 11 different parameters are mentioned in table 2.
All the features are numeric and there are no missing entities. The description for each of the parameter in mentioned below :

- NCD : The number of discussions created at time step t and involving the instance's topic.
- AI : Number of new authors interacting on the instance's topic at time t

Parameter	Acronym
Number of Created Discussions	NCD
Author Increase	AI
Attention Level (measured with number of authors)	AL
Burstiness Level	BL
Number of Atomic Containers	NAC
Attention Level (measured with number of contributions)	AtL
Contribution Sparseness	CS
Author Interaction	Aint
Number of Authors	NAu
Average Discussions Length	ADL
Number of Active Discussions	NAD

Table 2: The R^2,Adjusted R^2 ,AIC, BIC values for all the linear models formulated for hypothesis of Mean of the gradient

- AL : The attention payed to a the instance's topic on a social media
- BL : The ratio of NCD and NAD
- NAC : The total number of atomic containers generated through the whole social media on the instance's topic until time t
- AtL : The attention payed to a the instance's topic on a social media
- CS : Spreading of contributions over discussion for the instance's topic at time t
- Aint : The average number of authors interacting on the instance's topic within a discussion.
- NAu : The number of authors interacting on the instance's topic at time t
- ADL : The average length of a discussion belonging to the instance's topic
- NAD : The number of discussions involving the instance's topic until time t

2.2 Analysis

The analysis section is divided into two sub sections, where one deals with the mean of the gradient metric and the other deals with cross correlation with NAD.

2.2.1 Mean of the Gradient

The mean of the gradient for all the features over time is calculated using the equation (1),

$$Y_i = \frac{\sum_{t=0}^{6} X_i(t)}{7} \tag{1}$$

where $X_i(t)$ refers to a feature 'i' at time t.

The strong interest periods are when the slope/gradient is increasing rather than stable or decreasing. If the activity of NAD tends increase over time, then it can be considered as a potential period of high interest.

So this metric is computed for all the available features. Then the instances whose mean of the gradient of NAD is higher than 20, are considered only as instances of interest and the rest are discarded. There are 5228 instances whose mean of the gradient is higher than 20. But there were a couple of outliers i.e. instances 3077,3801 which had really high mean of the gradient values as seen in Figure4. These two instances have been discarded for now, assuming there are abnormal instances. Now, using this refined dataset I will be testing the hypothesis I mentioned in 1.4.

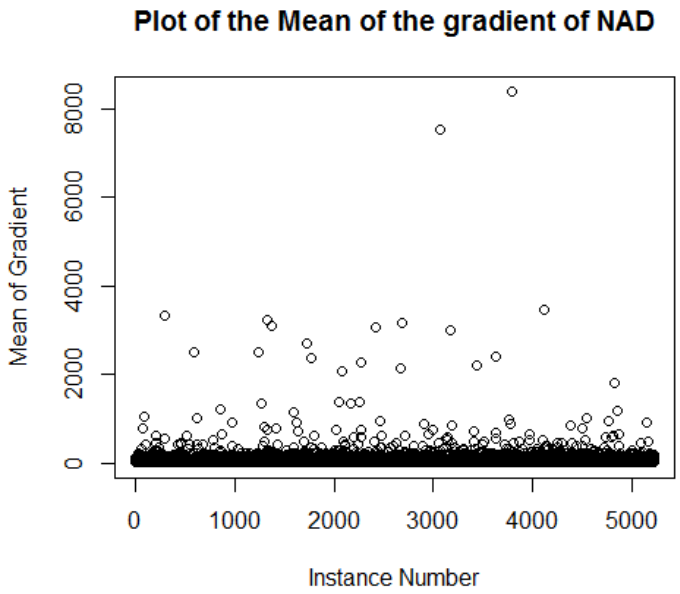

Figure 4: Mean of the Gradient of NAD

The hypothesis is tested using linear regression models because the mean of the gradient of NAD (which is the response variable) are numeric.
The following linear models are constructed to test the hypothesis:

- Linear Model 1 : $Y_{NAD} = \beta_0$
- Linear Model 2 : $Y_{NAD} = \beta_0 + \beta_1.Y_{NAu}$
- Linear Model 3 : $Y_{NAD} = \beta_0 + \beta_1.Y_{NAu} + \beta_2.Y_{Aint} + \beta_3.Y_{AI}$
- Linear Model 4 : $Y_{NAD} = \beta_0 + \beta_1.Y_{NAu} + \beta_2.Y_{Aint} + \beta_3.Y_{AI} + \beta_4.Y^2_{NAu} + \beta_5.Y^2_{Aint} + \beta_6.Y^2_{AI}$
- Linear Model 5 : $Y_{NAD} = \beta_0 + \beta_1.Y_{NAu} + \beta_2.Y_{Aint} + \beta_3.Y_{AI} + \beta_4.Y^2_{NAu} + \beta_5.Y^2_{Aint} + \beta_6.Y^2_{AI} + \beta_7.Y_{AI}.Y_{Aint} + \beta_8.Y_{Aint}.Y_{NAu} + \beta_9.Y_{AI}.Y_{NAu}$
- Linear Model 6 : Step regression of Linear Model 5

In the models mentioned above, Y_i is the mean of the gradient of feature i.

The models constructed above use the available features which correspond to "Authors". This way I can test my hypothesis which is to see if there is proportional relationship between authors and the active discussions. The linear model 2 is formulated based on just the number of authors(NAu) as predictors but the rest of the models (i.e. 3,4,5) are formulated based on all the features related to authors. The linear model 6 is a step regression model 5. Step-wise regression is an approach to selecting a subset of effects for a regression model. Using these models one can decide what all features associated with authors effect (or don't effect) the number of active discussions.

Linear Model #	R^2	Adjusted R^2	AIC	BIC
1	172478020	0	69208.03	69221.15
2	8645999	0.949862	53567.71	53587.39
3	7443432	0.956819	52789.04	52821.84
4	7430327	0.956871	52785.83	52838.32
5	7414938	0.956935	52780.99	52853.17
6	7415165	0.956942	52779.15	52844.77

Table 3: The R^2,Adjusted R^2 ,AIC, BIC values for all the linear models formulated for hypothesis of Mean of the gradient

From Table 3, The step-wise regression model has the highest adjusted R^2 and the lowest AIC and BIC values among all the models. The step wise regression model is :

$$Y_{NAD} = \beta_0+\beta_1.Y_{NAu}+\beta_2.Y_{Aint}+\beta_3.Y_{AI}+\beta_4.Y^2_{NAu}+\beta_5.Y^2_{Aint}+\beta_6.Y^2_{AI}+\beta_7.Y_{Aint}.Y_{NAu}+\beta_8.Y_{AI}.Y_{NAu} \quad (2)$$

Now that the linear model 6, (i.e. the regression model) is the best model among all the linear models. I looked into the diagnostic plots, show in Figure 5 ,to make sure that the model is the best.

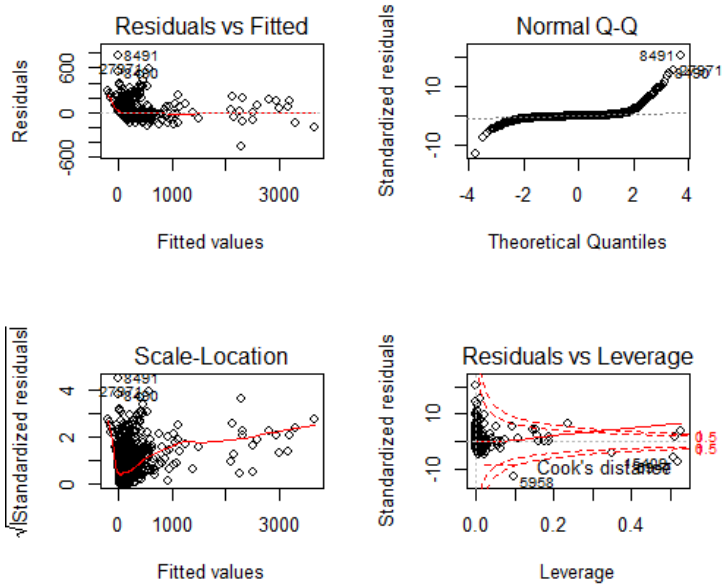

Figure 5: Diagnostic Plot of the linear model 6

From the figure 5, we can see that the residual vs fitted plot looks correlated and skewed, which is not desirable. Also in the QQ plot the points significantly deviate from the qq-norm line. It's the same case with Scale vs location plot and Residual vs Leverage. This shows that the response variable are not Gaussian distributed and have cluster with varying variance. Thus, I decided to transform the response variable (i.e. Y_{NAD}) using log transform and Box-cox transform to rectify this issue.

- Linear Model 7 : Box-cox transform (L=0.3 ; computed in R-studio)
  $(Y_{NAD})^{0.3} = \beta_0 + \beta_1.Y_{NAu} + \beta_2.Y_{Aint} + \beta_3.Y_{AI} + \beta_4.Y_{NAu}^2 + \beta_5.Y_{Aint}^2 + \beta_6.Y_{AI}^2 + \beta_7.Y_{Aint}.Y_{NAu} + \beta_8.Y_{AI}.Y_{NAu}$
- Linear Model 8 : Log transform
  $\log(Y_{NAD}+1) = \beta_0 + \beta_1.Y_{NAu} + \beta_2.Y_{Aint} + \beta_3.Y_{AI} + \beta_4.Y_{NAu}^2 + \beta_5.Y_{Aint}^2 + \beta_6.Y_{AI}^2 + \beta_7.Y_{Aint}.Y_{NAu} + \beta_8.Y_{AI}.Y_{NAu}$

The diagnostic plots of the linear models 7 and 8 are shown in figures 6 and 7 respectively. The diagnostic plots of these models improved compared to the linear model 6, but there is still some pattern in residual vs fitted and also the QQ plot deviates from the norm line. The influential points (i.e. the instances with high leverage) are 4123,1323 and 3964. Due to the lack of annotations, we can't neglect these instances.

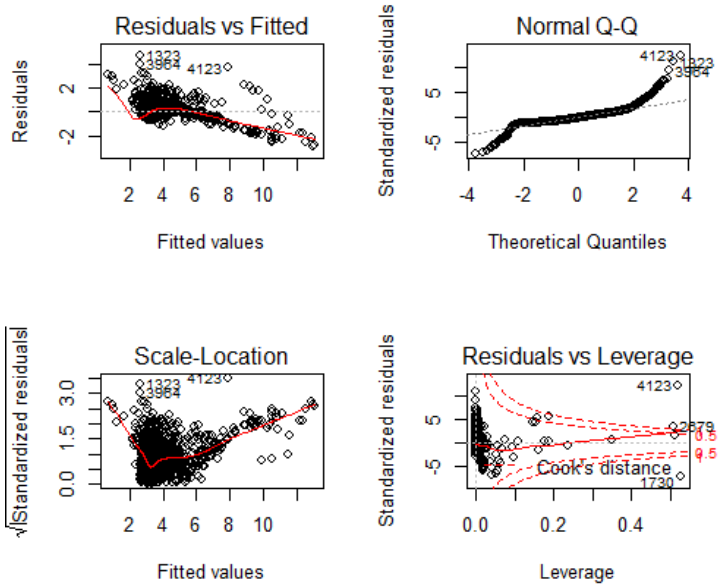

Figure 6: Diagnostic Plot of the linear model 7

2.2.2 Cross correlation with NAD

cross-correlation is a measure of similarity of two series as a function of the lag of one relative to the other. For my case, I will be computing the cross correlation of the available keyword features with NAD relative to time. To explain it further, I will be doing the a cross correlation between a feature and the NAD for every instance.This way we can identify there is any time-dependent correlation between the feature and NAD. To analyze this correlation, I will be using the maximum value of the correlation.

$$C(X_i, X_{NAD})(n) = \sum_m X_i(m).X_{NAD}(m+n) \tag{3}$$

where n is the lag value, $X_i(m)$ is the feature 'i' at time m and $C(X_i, X_{NAD})(n)$ is the correlation between the feature and NAD for different lags.

$$Z_i = max(C(X_i, X_{NAD})(n)) \tag{4}$$

where is Z_i is the max cross correlation value of feature 'i' with NAD for a particular instance.

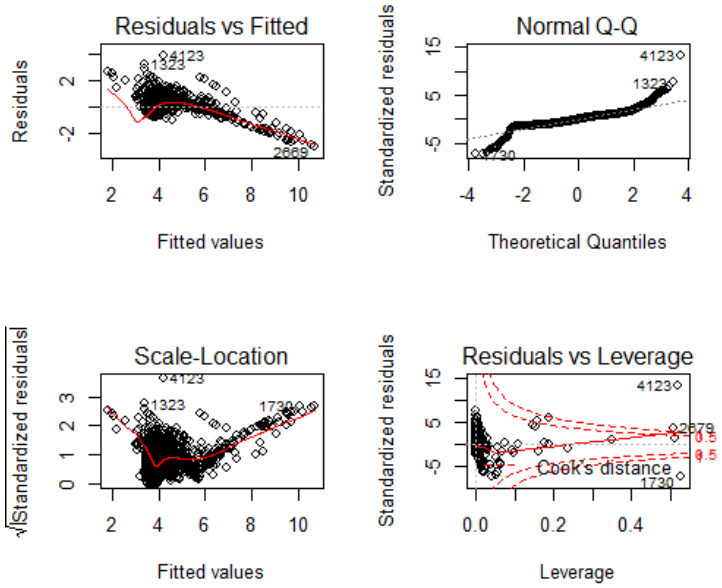

Figure 7: Diagnostic Plot of the linear model 8

I computed this feature for all the keyword features for all the instances. But there is an issue which I realized after computing the values i.e. there were a lot redundancies in the cross correlation values. It is because the NAD values or the other feature is almost constant in time. These instances are also not potential buzz periods because its not attracting buzz activity. To reduce this redundancy, I computed the standard deviation of NAD over time for all the instances and discarded the instances whose standard deviation is less than 20.Different values can be chosen as a threshold, but for my case I chose 20. There are 16400 instances whose standard deviation of the NAD values is higher than 20. Then I computed the cross-correlation for all these features and used just these instances for further analysis. There were a few outliers which had exceptionally high NAD auto correlation which were discarded. These discarded instances are 10012,15072. These can clearly spotted in Figure 8.
Now to test our hypothesis mentioned in1.4, I will be using this processed data. I have constructed a few linear regression models to test my hypothesis. I employed linear regression model because my response (i.e. auto correlation of NAD) are numerical values. The linear regression models I will be using for my analysis are mentioned below :

- Linear Model 1 : $Z_{NAD} = \beta_0$
- Linear Model 2 : $Z_{NAD} = \beta_0 + \beta_1.Z_{BL}$
- Linear Model 3 : $Z_{NAD} = \beta_0 + \beta_1.Z_{BL} + \beta_2.Z_{NCD}$

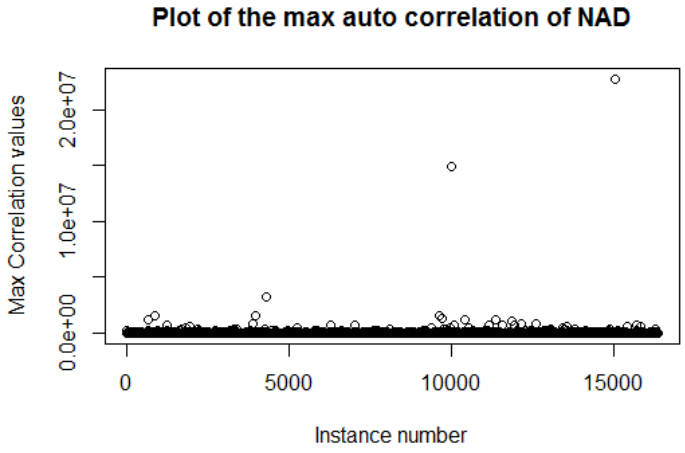

Figure 8: Plot of the max auto-correlation of NAD

- Linear Model 4 : $Z_{NAD} = \beta_0 + \beta_1.Z_{BL} + \beta_2.Z_{NCD} + \beta_3.Z_{ADL}$
- Linear Model 5 : $Z_{NAD} = \beta_0 + \beta_1.Z_{BL} + \beta_2.Z_{NCD} + \beta_3.Z_{ADL} + \beta_4.Z_{BL}.Z_{ADL} + \beta_5.Z_{BL}.Z_{NCD} + \beta_6.Z_{NCD}.Z_{ADL}$
- Linear Model 6 : $Z_{NAD} = \beta_0 + \beta_1.Z_{BL} + \beta_2.Z_{NCD} + \beta_3.Z_{ADL} + \beta_4.Z_{BL}.Z_{ADL} + \beta_5.Z_{BL}.Z_{NCD} + \beta_6.Z_{NCD}.Z_{ADL} + \beta_7.Z^2_{NCD} + \beta_8.Z^2_{BL} + \beta_9.Z^2_{ADL}$
- Linear Model 7 : Step regression of Linear Model 6

I have used the predictors burstiness level (BL), number of created discussions (NCD) and average discussion length (ADL) because all these features are number of tweets and keywords. The Table 4 shows the R^2,adjusted R^2, AIC and BIC values for all these models.

Linear Model #	R^2	Adjusted R^2	AIC	BIC
1	4.1×10^{13}	0	401391.7	401407.1
2	4.07×10^{13}	0.005224368	401306.8	401329.9
3	269779529	0.9999934	205738.6	205769.4
4	264134723	0.9999936	205393.8	205432.4
5	232804080	0.9999943	203329.4	203391
6	227950334	0.9999944	202989.9	203074.7
7	227969839	0.9999944	202989.3	203066.3

Table 4: The R^2,Adjusted R^2 ,AIC, BIC values for all the linear models formulated for hypothesis of cross correlation with NAD

From Table 4, the model 7 has the highest adjusted R^2 and the least AIC and BIC. Linear model 7 is the step regression model of linear model 6. The step wise regression of the linear model 6 is given below :

- Linear Model 7 :
  $Z_{NAD} = \beta_0 + \beta_1.Z_{BL} + \beta_2.Z_{NCD} + \beta_3.Z_{ADL} + \beta_4.Z_{BL}.Z_{ADL} + + \beta_5.Z_{NCD}.Z_{ADL} + \beta_6.Z^2_{NCD} + \beta_7.Z^2_{BL} + \beta_8.Z^2_{ADL}$

Now let's look into the diagnostic plots of the linear model 7. The figure 9 shows the diagnostic plot of linear model 7. The residual vs fitted plot looks clustered and skewed. The QQ plot also deviates from the norm line. Similarly for scale-location plot and the residual-leverage plot. So I employed the Boxcox transformation and the log transform to tackle this issue. The boxcox transform did not improve the diagnostic plots as it did a transform the response variable to the power of 1. But the log transform did improve the diagnostic plots, shown in figure 10. The response variable almost follows Gaussian distribution but the residual vs fitted plot is still skewed. The influential points (i.e. the instances with high leverage) are 9819,4287 and 4259. Due to the lack of annotations, we can't neglect these instances. The model equation for the log transformed linear model 7 is given below :

- Linear Model 8 :
  $\log(Z_{NAD}+1) = \beta_0 + \beta_1.Z_{BL} + \beta_2.Z_{NCD} + \beta_3.Z_{ADL} + \beta_4.Z_{BL}.Z_{ADL} + + \beta_5.Z_{NCD}.Z_{ADL} + \beta_6.Z^2_{NCD} + \beta_7.Z^2_{BL} + \beta_8.Z^2_{ADL}$

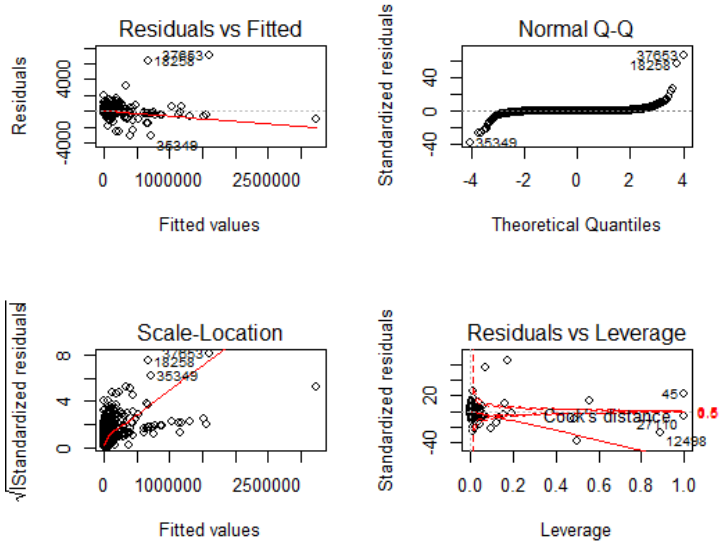

Figure 9: Diagnostic Plot of the linear model 7

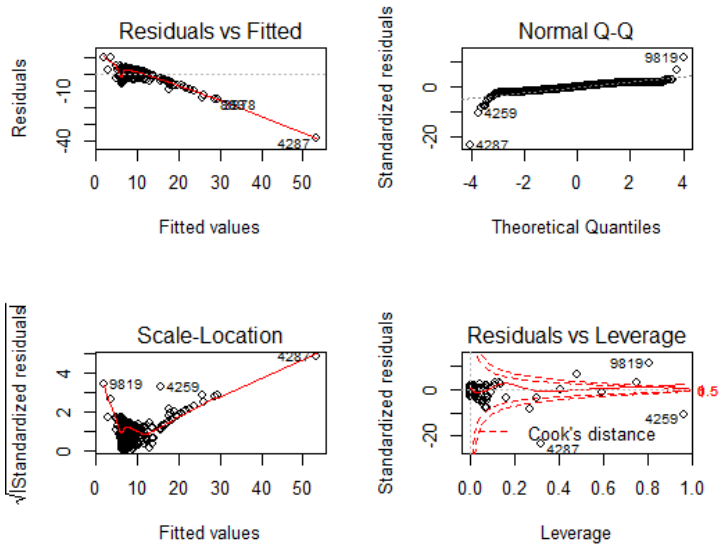

Figure 10: Diagnostic Plot of the linear model 8

3 Evidence

In this section, I will describe the credibility of my models and also evaluate our hypothesis for the each of the measures, namely, Mean of the Gradient and Cross correlation with NAD.

3.1 Mean of the Gradient

When I observe the diagnostic plots in figure 5 for the linear model 6, we can see that the model is not a good fit for the data, although it is significant. This can be seen in the QQ plot, as the model deviates signicantly from the normal line. I decided to log transform and also boxcox transform the response variable. For the log transform, I added 1 to the response variable before transforming to ensure that there are no zeros. This creates a better model. The diagnostic plots for these models are show in figures 6 and 7. Though the diagnostic plots don't look god enough, they improved from the diagnostic plot of linear model 6.

Now that we have the models 7 and 8, I looked into the significance level of the predictors and also the estimate. The significance level helps us understand whether the predictor is contributing or not. The estimate helps us understand whether the relationship between the response variable and the predictor is positively or negatively proportional. In both models the feature "Number of authors" features is significant (2×10^{-16})and also has a positive estimate ($+ 1.1 \times 10^{-2}$). This shows that we can reject our hypothesis in 1.4. This helps us say that the "With increase in number of authors there is an increase in number of active discussions". Also, there is one more interesting observation i.e. the combination of features "Number of authors" and "Author interaction" is also significant(3.2×10^{-6}) and has a positive estimate.

3.2 Cross correlation with NAD

Figure 9 shows the diagnostic plot of the Linear model 7 for Cross correlation with NAD. The residual vs fitted plot looks correlated and skewed, which is not dersirable. In the QQ plot, the tail of the plot deviates from the normal line which is not desirable either. It's the same case with Scale vs location plot and Residual vs Leverage. So I applied simple log transformation to rectify this issue. Though the diagnostic plots 10 didn't improve a lot, they looked better than the original diagnostic plot.
In order to evaluate my hypothesis, I looked into the significance level of the predictors and also the estimate. In model, the feature "Burstness Level (BL)" is shown to have significant value (i.e. 1.24×10^{-10}). This shows that we can reject our hypothesis, because the change in "NAD" is correlated with change in "BL". By using this evidence, we can reject the hypothesis mentioned in 1.4. Now, we can say that, the change in burstiness level is correlated with the change in NAD over time.
There is an interesting observation in this model as well i.e. the combination of features BL and "Average Discussion Length(ADL)" is also very significant (i.e. 5.86×10^{-8}).

4 Recommendation

4.1 Mean of the Gradient

Based on the findings, I can successfully state that increased number of authors increases the number of active discussions. The model we use to test our hypothesis shows that we have a high R^2 value and a relatively low AIC and BIC values compared to other models we tested, as shown in Table 3. Our 95% confidence interval for this model is [1.05×10^{-2} , 1.15×10^{-2}] which is a positive interval, indicating that an increase in authors to a positive increase in active discussions.
Our evidence also suggested that the number of active discussions tend to increase with increase of author interaction and number of authors together.

4.2 Cross correlation with NAD

On analysing our models,I can surely state that with change in the burstiness level over time, the number of active discussions also tend to change. Table 4 shows that our model (i.e. model 7) has the highest adjusted R^2 and also the least AIC and BIC values. The 95 % confidence interval of the BL parameter is [1.41 $\times$ 10^{-2} , 2.65 $\times$ 10^{-2}] which is positive indicating that the change in BL is directly proportional to the change in NAD.
One more evidence suggested that the number of active discussions tend to change over time with change of burstiness level along with average discussion length.

Finally based on our analysis and evidence, to create a "Buzz" in Twitter one should target at-least one of the following things:

- Attract authors to talk about the topic
- Make the authors interact among each other and attract more authors to get involved in the discussion
- Increase the Burstiness level i.e. make sure the number of created discussions is always higher than the number of authors
- Increase the discussion length along with increase in burstiness level.

5 References

[1] "Final Project template," 2015, class template in SYS 6021.

[2] A. A. F. Laura E. Barnes, "Analysis of Buzz periods in Twitter," 2015, class project in SYS 6021.

[3] Buzz in social media Data Set http://archive.ics.uci.edu/ml/datasets/Buzz+in+social+media+

[4] Prdictions dactivit dans les rseaux sociaux en ligne (F. Kawala, A. Douzal-Chouakria, E. Gaussier, E. Dimert), In Actes de la Confrence sur les Modles et lAnalyse des Rseaux : Approches Mathmatiques et Informatique (MARAMI), pp. 16, 2013.

[5] Asur, Sitaram, and Bernardo Huberman. "Predicting the future with social media." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.