Weibo - Getting the best out of your post

John Mohoang, Liu Liang, James Mochizuki-Freeman

MACALESTER COLLEGE

1. Reading the files

The dataset was obtained ned from a research team at the Journalism and Media Studies Centre, The University of Hong Kong (JMSC).[1] For this project, we only sampled week 1, which had over 4 million tweets in it.

To read the files we wrote the functions **UnicodeDictreader()**, **reader()** and **remessage()**. The **UnicodeDictreader()** decodes the data we have because it is encoded in unicode. the **reader()** function yields rows of the data we have. Each row yielded by the **reader()** function corresponds to a single tweets. The **remessager()** function yields only tweets that have been retweeted.

2. Project Idea

Research Question - What are the characteristics that influence popularity of reblogs and the reblogging rate of bloggers?

Weibo is one of the most famous microblogging services in China. Since most of the papers we studied in class deal with USA centric websites, we believe it is worthwhile to see if what we learned could also apply in other parts of the world, hence Weibo.

3. Methodology and Analysis

To try and answer our project question, these are some of the characteristics we thought would be useful to study:

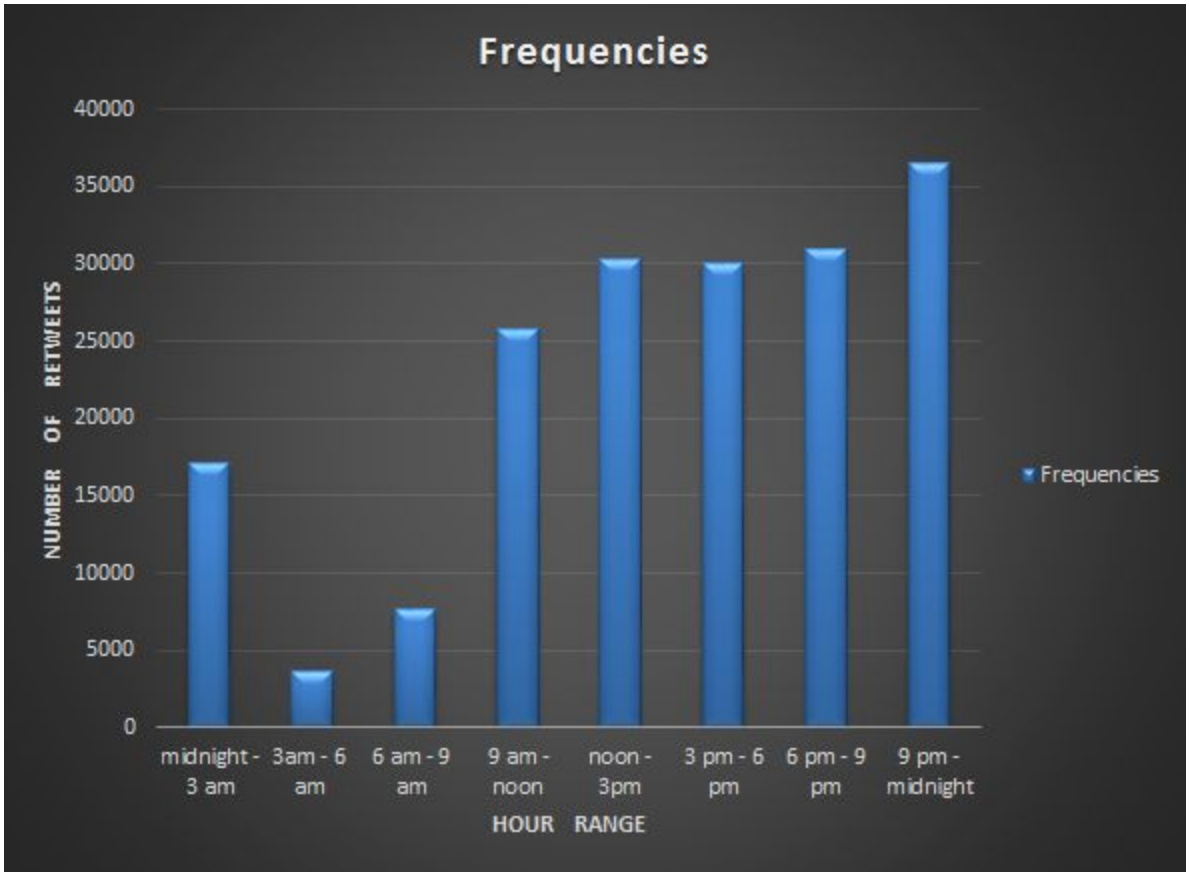
- The time when tweets are made
- Length of the tweets
- How active microbloggers on Weibo are
- Whether or not the tweets have images
- A combination of all of the above

A. Time when retweets are made

To test the effect that time has on the number of tweets that have been retweeted, we created the the function **plotRetweetVsHour()** which read through the retweeted messages and classified each in bins, each bin being a three hour period. The frequency of each bin is diagramed below;

[1] The link to the dataet : http://t.co/zTZYVvmxCb

Retweet Vs. Hour

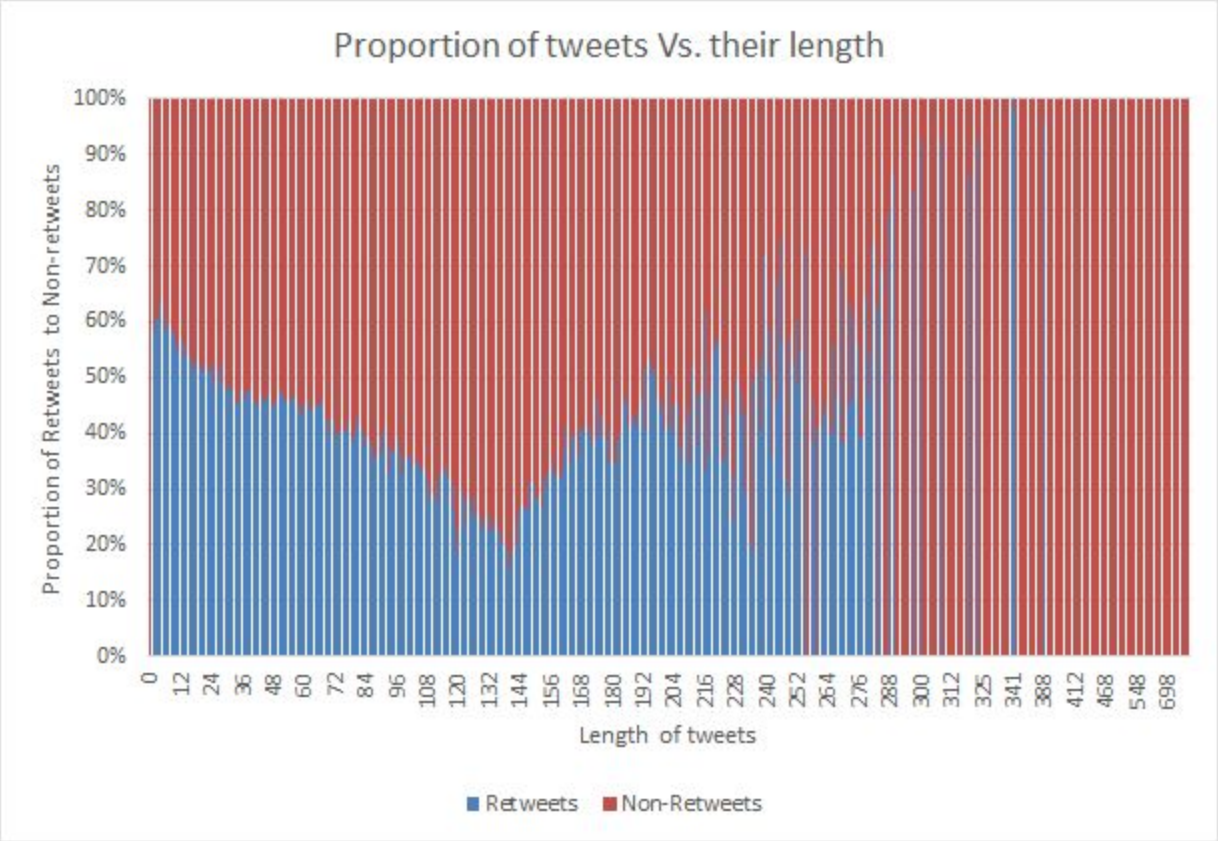


As can be seen in the above diagram, significant increase in the number of tweets being retweeted is between noon and 3 pm. it is also, worth noticing that the largest number of retweets are of posts done within the last three hours of the day. From this results may be because people are generally asleep after midnight and that people generally have binding activities to do in the late mornings and afternoons, hence the peak in weibo activity at night.

From this analysis, it is clear that to increase the chances of a tweet being retweeted, one should post it at night, especially close to midnight.

B. Length of the tweets

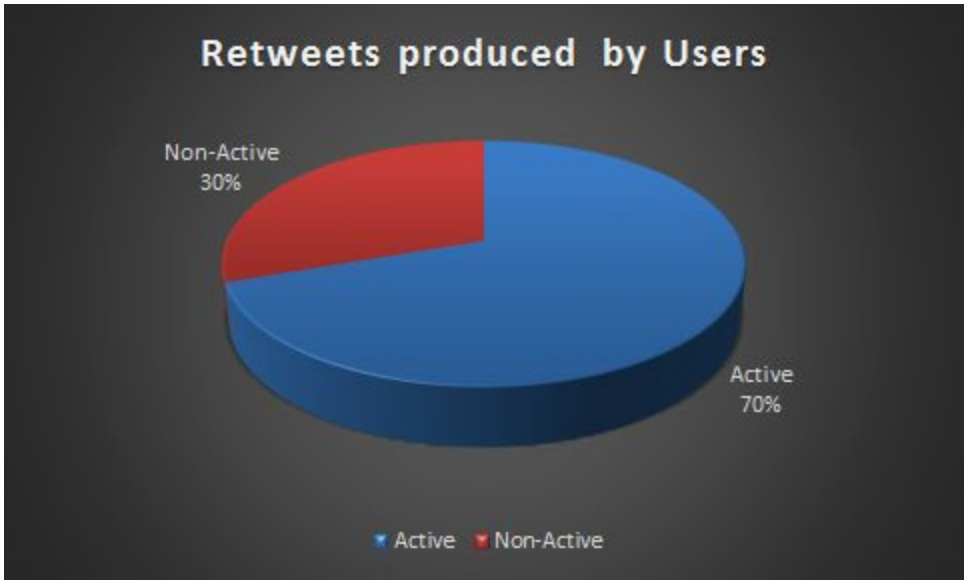
To test the effect of the tweet length, we created the function **plottweetLengthVsHour()** which essentially wrote in an excel file the lengths of each tweet and whether or not it was retweeted. The graph produced by information from that file is shown below;



We had earlier predicted that the longer the tweet, the more likely it is to be positive, in terms of sentiment. This may be true for only a short portion of the tweets with lengths between 160 and 210. Even so, the proportion of retweeted posts is still not as high as that of shorter posts, meaning that, a conclusion can be made that the longer the post becomes, the less likely it is to be reposted. This may be because when there are many people producing content on Weibo, people become lazy too read instead they would prefer getting more varied information that one in-depth piece of information.

C. Activity of microbloggers

To test whether or not the activity of Weibo users has an effect on the number of retweets produced, we used our function **activeUser()** and plotted the information as shown in the diagram below;



From the above graph, it is clear that the more content a user produces, the more likely that some of that content will be reposted by some other user. This does not, however, mean that the proportion of tweets that are retweeted will be proportional to content produced.

D. Images

To test whether or not having an image in a tweet would increase the chances of a tweet being retweeted, we made used the function **imageCheck()** and found out that almost all the the tweets don't have images. This means that whether or not a tweet has an image has no effect on its chances of being retweeted.

4. Overall Predictors (All the above predictors combined)

At this point, we were trying to find out which of the predictors are more significant in predicting whether or not a tweet will be retweeted.

To do this, we decided to use R to build our models.
We used the function **overall(n)** in python to make a .csv file with headers being:
Retweeted - 0 if the tweet is not retweeted, 1 if it is.
TextLength - length of the text
UserActiveness - how active the user is in producing content in the given week.
OriginalTime - This is the time when the original tweet was made.

An important note about the **overall(n)** function is that, it builds an excel file with numbers of tweets that were retweeted being equal to the number of tweets that were not retweeted. This is

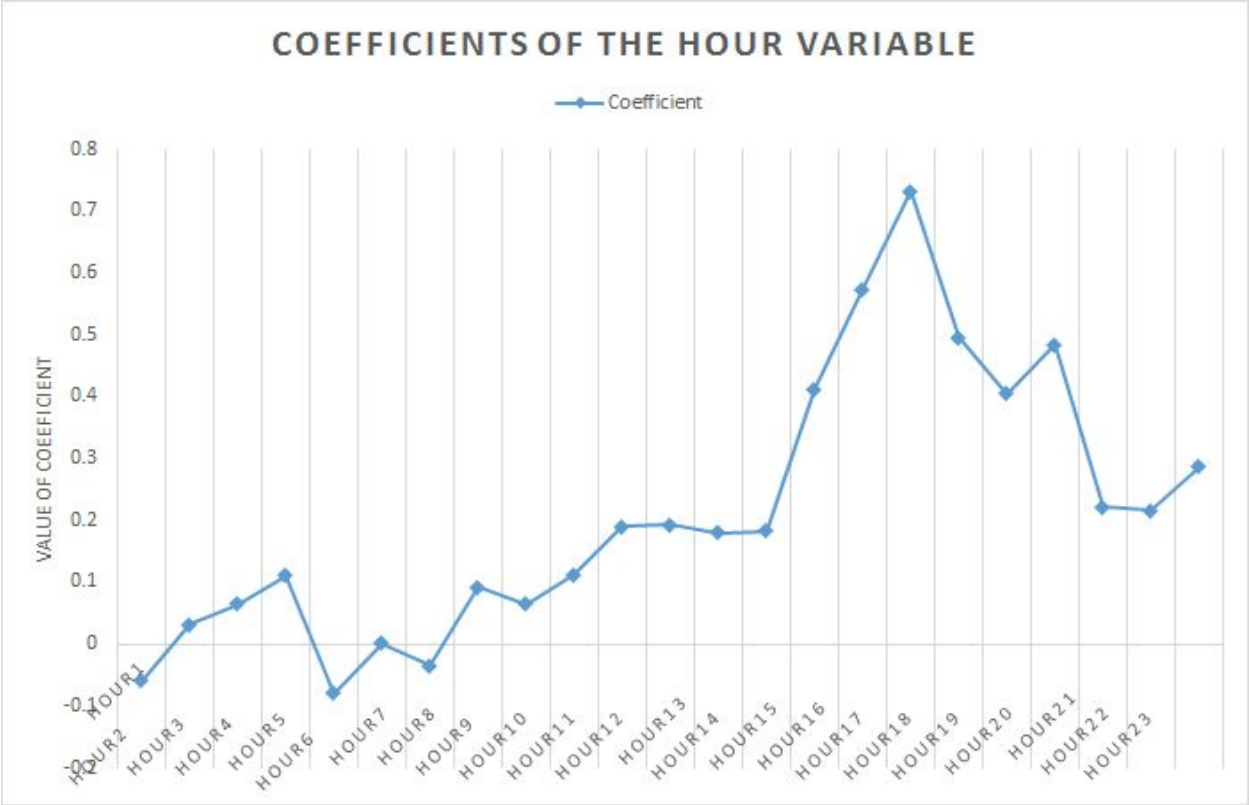
to make sure that our predicted models are accurate and not just predict that tweets will not be retweeted because the number of not-retweeted tweets is significantly larger that those that were retweeted.
The excel file made by **overall(n)** is named **overall.xls**. We then modified the format of the OriginalTime and created the **overall.csv** file which we exported to R.

In R, we extracted the hour from the OriginalTime and that was the variable to be used in our models. We then divided our data into training and testing data, and then explored two models; Logistic Regression and Decision Tree Model.

Logistic Regression
In Logistic Regression, when making the summary of the model it is important to note that it tells us that at midnight, we expect a blank message from a user who has never been active to actually be retweeted. This is clearly not going to be the case in practice.

The coefficients of our TextLength and UserActiveness are both negative, meaning that we expect that active users who post long posts will have lower chances of those posts being retweeted. This makes some sense in that if a post is too long, chances that it will be read may be low and if a user is hyper-active on weibo, then people are less likely to follow what that person is saying because they know what the user generally posts about.

Below we have graphed each hour with its coefficient as models in our Logistic Regression Model;



In the above model, it important to note that most tweets that are retweeted are posted between 4 pm and 8 pm.

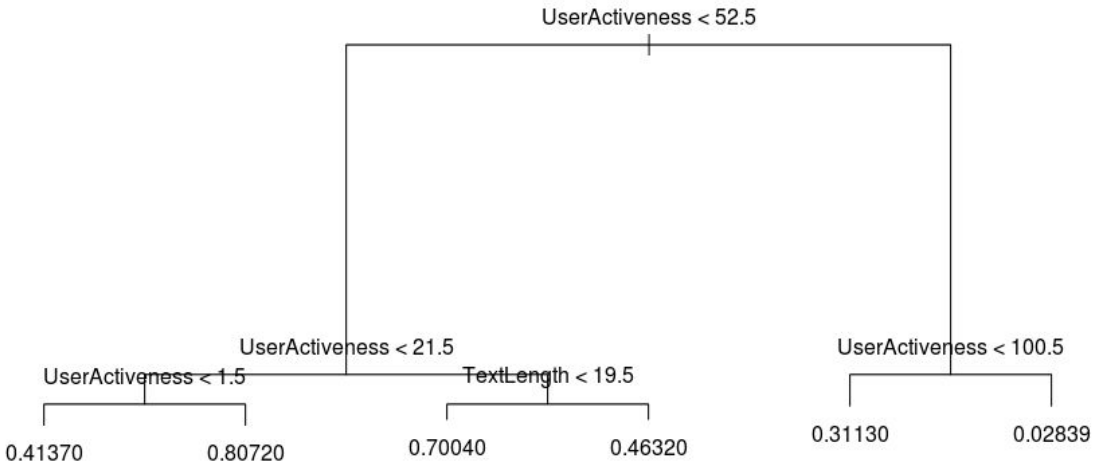
To predict our model on our test data, we built a Confusion matrix as follows;

	0	**1**
0	34155	7011
1	15579	43255

From this confusion matrix, we can see that we correctly classified 77.41% of our test data correctly, meaning that our test error rate is only 22.59%, which is not that bad,actually.

Tree Model

For the Tree model we made, it is worth noting that the summary of the model points out that UserActiveness and Textlength are most important variables used in the tree construction.
The tree looks as follows;



To check how accurately this tree model would work on the test data, we built a confusion matrix that looks as follows;

	0	**1**
0	39461	12420
1	10273	37846

From the confusion matrix above, we can see that we correctly classified 77.31% of the tweets and our error rate is only 22.69%

5. Conclusion

- To increase the chances of a tweet being retweeted, according to our models it should be tweeted at night when there is high traffic on Weibo, the posts should generally be short while each post does not have to have images.
- UserActiveness and TextLength are the two important variables in figuring out whether or not a tweet will be retweeted.
- Both the Linear Regression and the Tree model classify accurately most of the data though the tree model very slightly performs better, with an error rate that is only 0.1% better.

6. Future Work

- In the models done in R, it would have been nice to try other models like Boosted Tree models, Bagging, Random Forests etc and see if they can make accurate predictions. We did try, Boosting, but it did not work out, hence it is commented out in our final work.
- It would be cool to juxtapose our findings on this project on a similar project with another blogging site like Twitter.