

Weibo – Getting the Best From Your Post

John Mohoang, Liu Liang, James Mochizuki-Freeman



Macalester College – St. Paul, MN

Research Question

What are the characteristics that influence the popularity of a blog post on Weibo?

Background

Weibo is one of the most famous microblogs 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 to other parts of the world.

Approach

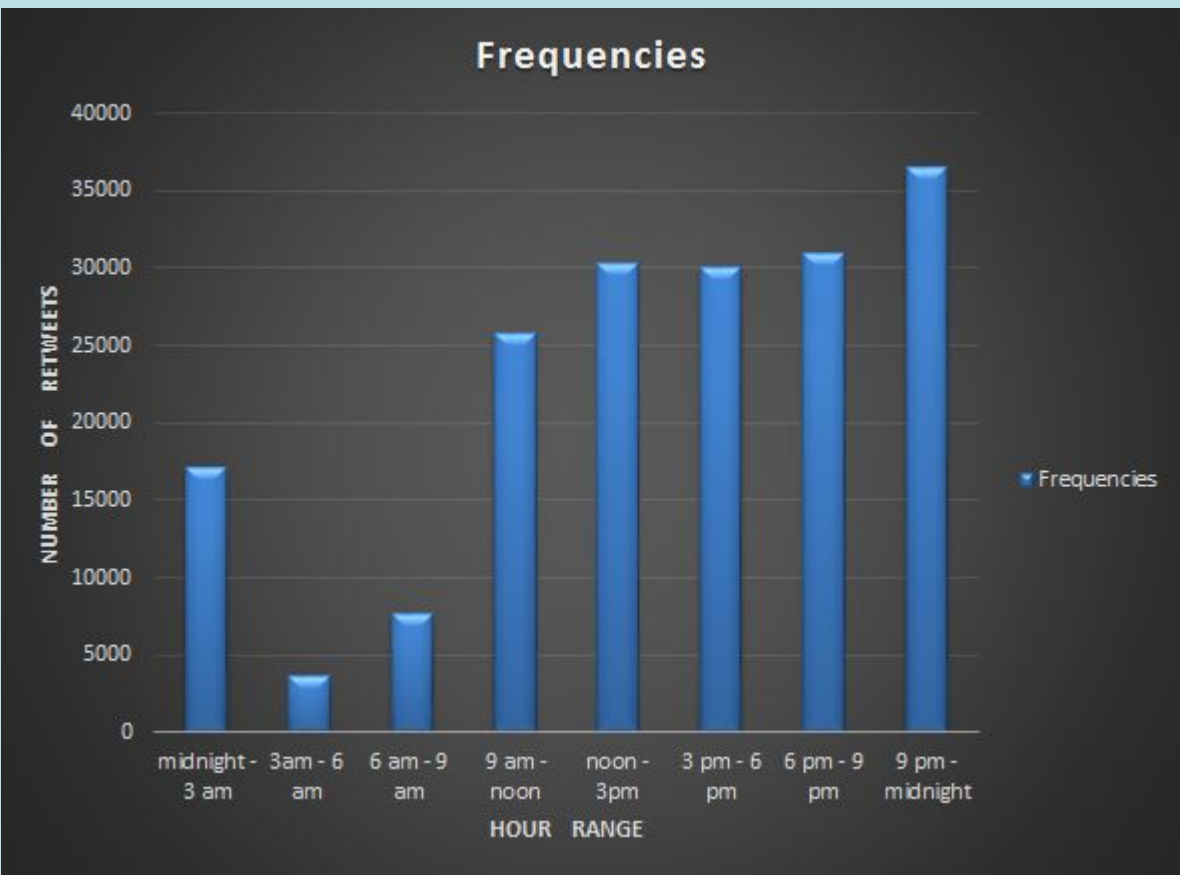
To try and answer our project question, these are some of the characteristics we thought it would be useful to study:

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

Results and Analysis

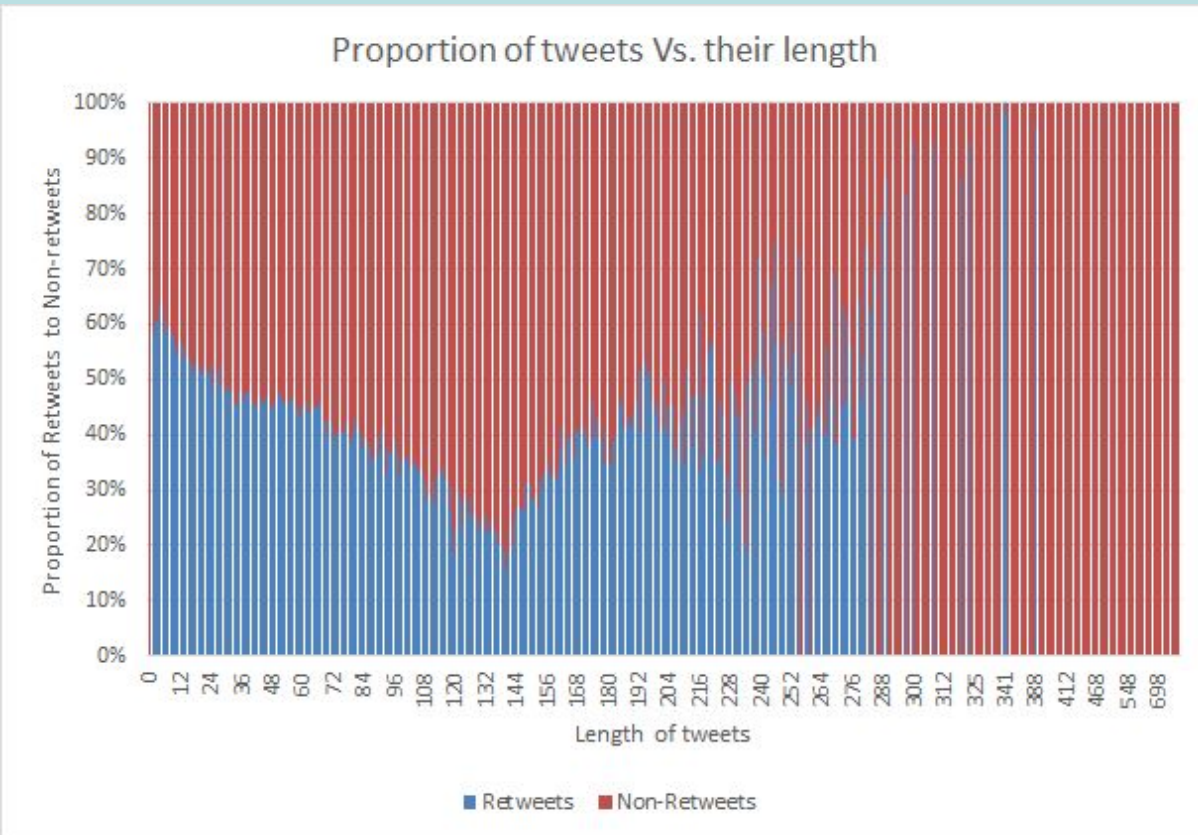
Time

There are many more retweets in the afternoon than in the morning, so that suggests that a tweet that is retweeted right away will be more likely to be retweeted in the afternoon or evening compared to the morning.



Length of tweets

Short posts of less than 50 characters and posts that are between 180 and 280 characters are the most likely to be retweeted, with a sharp drop in retweet rate after 280 characters.



Users' Activity

A users is considered to be an active retweeter if they have retweeted at least one tweet. Most users are retweeters, according to our data, so it follows that the more content a user produces, the more likely it is that someone else will retweet it.

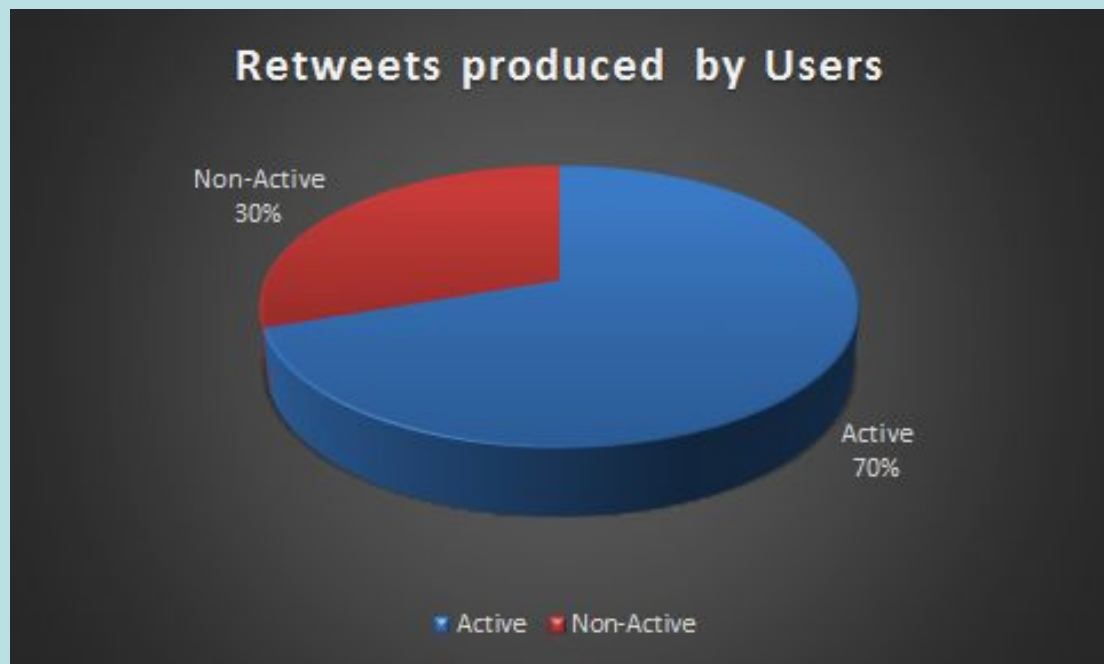


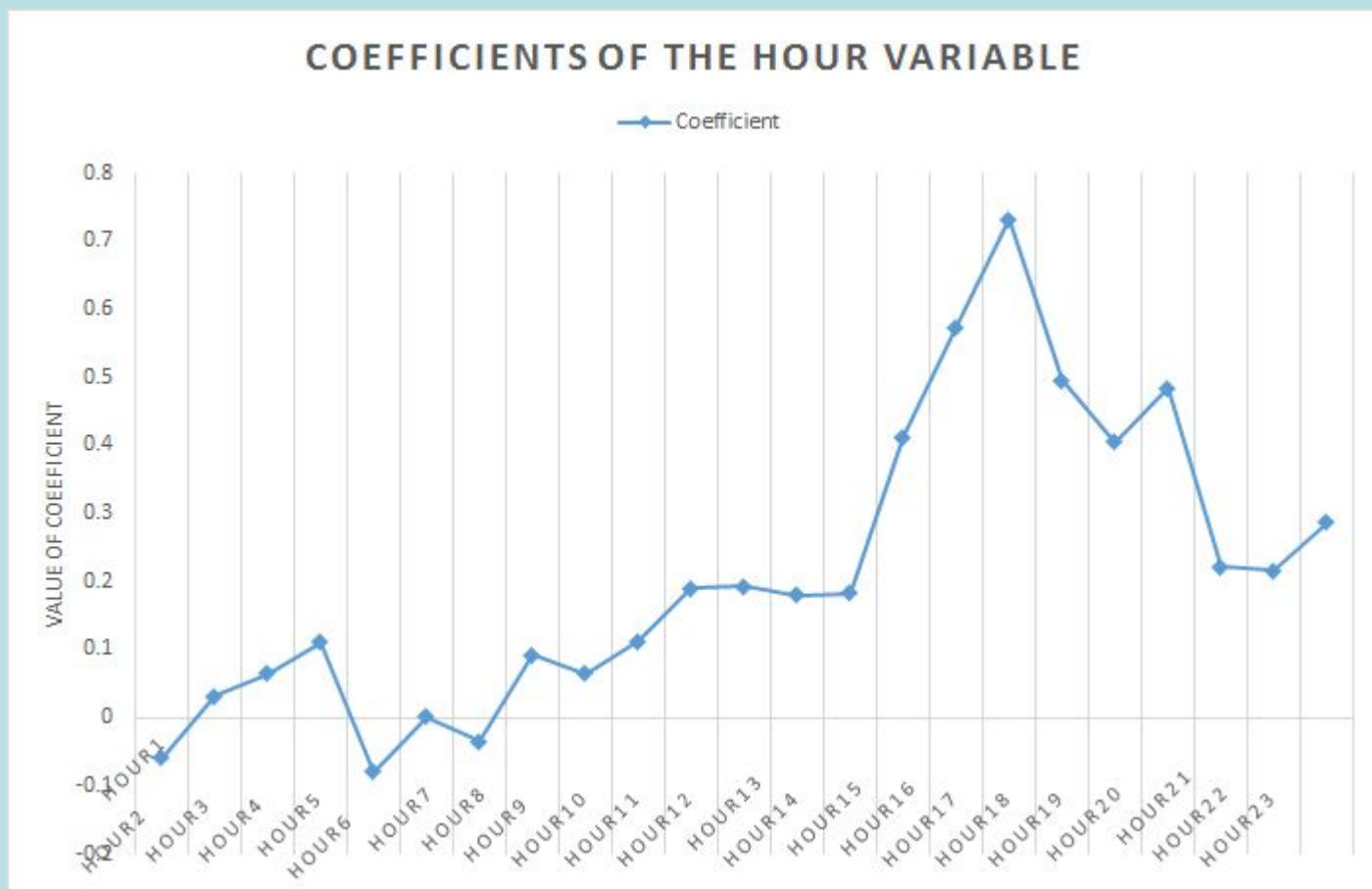
Image Presence

Almost all of the tweets in our dataset did not have images, so we concluded that we did not have enough data to be able to reliably determine whether the presence or absence of an image had any effect on the likelihood of a tweet being retweeted.

Modelling and Prediction

Logistic Regression

A logistic regression of the data gave a negative coefficient for both a higher length of tweet and a greater user activity. This leads to the conclusion that active users that post long posts are actually not as successful as those that post shorter, more infrequent posts. Other than that, as can be seen in the graph, the time of day can slightly improve the chances of a post being retweeted, but that is still insignificant.

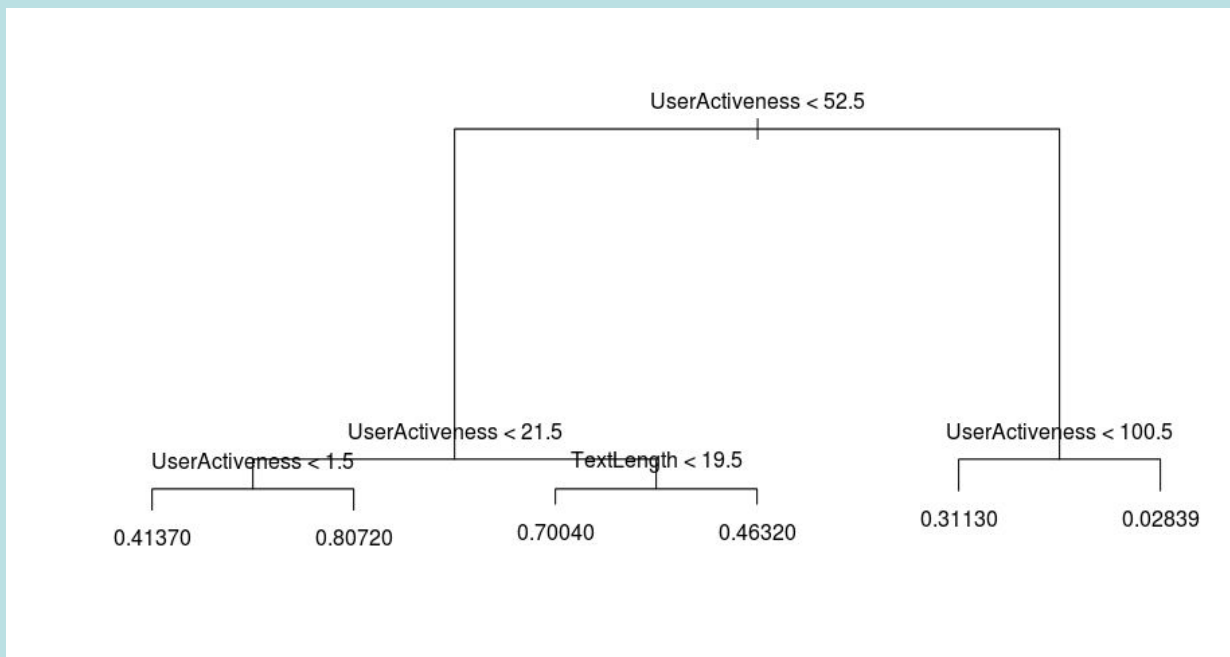


	0	1
0	34155	7011
1	15579	43255

From this confusion matrix, we can see that we correctly classified 77.41% of our test data, meaning that our test error rate is only 22.59%.

Decision Tree Model

Our decision tree classifier, on the other hand, gave a different result than logistic regression, although the error rate was roughly the same. According to the decision tree, a higher User Activeness (how many retweets the user has posted) leads to a higher likelihood of the user's own tweets being retweeted. A text length above 19 characters also performed lead to higher retweet likelihoods. So, it follows that if you want your own original tweets to be successful, you need to also retweet content from others and not make very short posts.



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 post should generally be short. It does not matter whether the post has images or not.

•UserActiveness and TextLength are the two most 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.

Future Work

•Add other interesting variables, such as the popularity of the poster or the time since last tweet, to the dataset.

•Use other analysis and classification methods to try to improve the success rate.

•Compare the results with similar published works about sites such as Twitter, to determine any similarities or differences between social networks in the USA and in China.