DAT4 Project: What content features lead to more shares on DoS Facebook Properties?

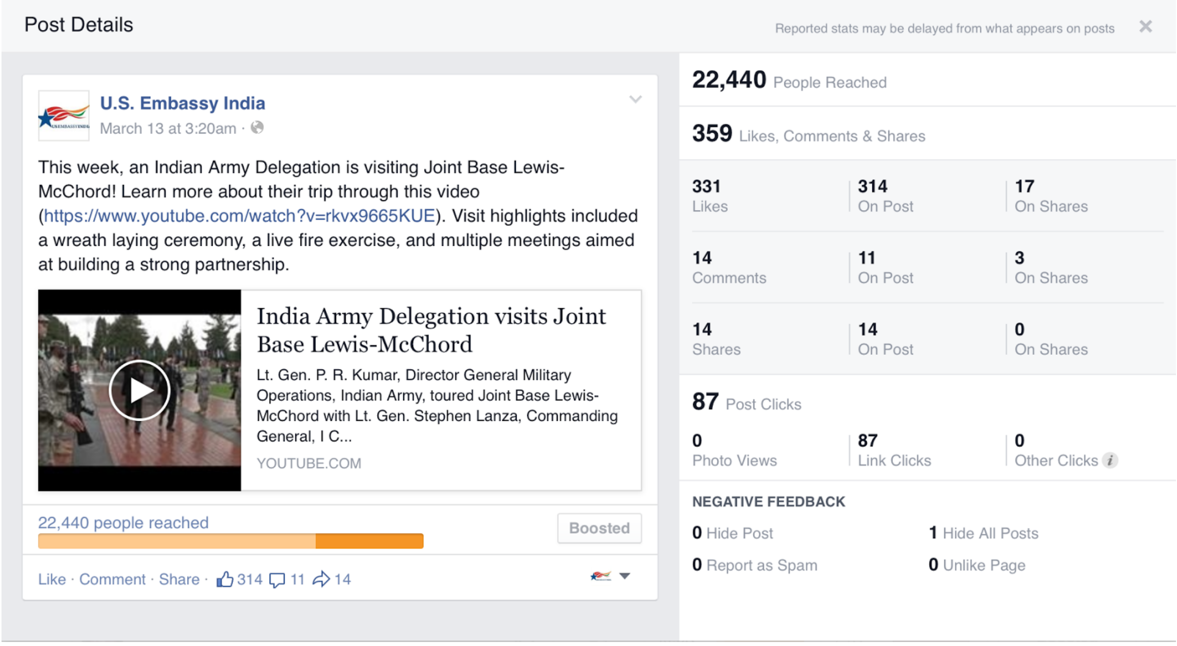
Jennifer Lambert 16 March 2015

Data:

The data set includes Facebook post level data from 12 DoS (Department of State) Facebook pages; two properties from each of DoS's six regions were included. The data also includes thematic tags from the IIP's (Bureau of International Information Programs) new digital platform, ShareAmerica (share.america.gov) when the Facebook post contains a link to an article on ShareAmerica. IIP's Office of Analytics then hand-coded all the other Facebook posts with similar thematic tags so that the content can be analyzed according to theme. The data set also includes a range of other metrics available from Facebook insights; important to this project will be the number of people reached by a particular post and the number of shares the post received. Below is an example of a Facebook post and some of the metrics available via Facebook Insights.

State Department Regions

AF = Sub-Saharan Africa
EAP = East Asia Pacific
EUR = Europe
NEA = Near East, Northern Africa
SCA = South Central Asia
WHA = Western Hemisphere



Post Details

Reported stats may be delayed from what appears on posts

U.S. Embassy India
March 13 at 3:20am ·

This week, an Indian Army Delegation is visiting Joint Base Lewis-McChord! Learn more about their trip through this video (https://www.youtube.com/watch?v=rkvx9665KUE). Visit highlights included a wreath laying ceremony, a live fire exercise, and multiple meetings aimed at building a strong partnership.

India Army Delegation visits Joint Base Lewis-McChord

Lt. Gen. P. R. Kumar, Director General Military Operations, Indian Army, toured Joint Base Lewis-McChord with Lt. Gen. Stephen Lanza, Commanding General, I C...

YOUTUBE.COM

22,440 people reached

Boosted

Like · Comment · Share · 314 11 14

22,440 People Reached

359 Likes, Comments & Shares

331 Likes	**314** On Post	**17** On Shares
14 Comments	**11** On Post	**3** On Shares
14 Shares	**14** On Post	**0** On Shares

87 Post Clicks

0 Photo Views	**87** Link Clicks	**0** Other Clicks

NEGATIVE FEEDBACK

0 Hide Post
1 Hide All Posts
0 Report as Spam
0 Unlike Page

Methodological Approach:

The file was converted into a .csv file and read into Python. The first step involved cleaning up the column names and the thematic tag column. Our team had several different spellings of entrepreneurship and some interesting ways of denoting English language learning and freedom of expression. Each of the thematic tags were turned into a separate column and a dummy variable was used to account for each post that was tagged with that particular theme. Dummy variables were also created for post type (link, status update, photo, or video) and whether the post contained a link to IIP's new digital platform, ShareAmerica. Since I decided to use the Stats Models version of linear regression, I also had to get rid of spaces and different characters in the column names.

Next, I decided to use the linear regression model from sklearn. Since my response variable is shares per user reached, a continuous variable, and this is an example of supervised learning, regressions are an appropriate model choice. I used test-train-split to segment my data into a training set and a test set. I ran the model on the training data and then got an R squared value based on the test data. I then computed the root mean squared error (RMSE) for the model's predictions for the test data set compared to the actual shares per user reached in y_test. The RMSE was 0.16. I then used a 10-fold cross validation and the RMSE was 0.18072.

I then decided to also try a decision tree regressor from sklearn so I could compare two different models. Using the training set, I fit the model and then compared the predicted values for X_test to the actual values in y_test. The RMSE for this model was 0.166. Using 10-fold cross validation the RMSE was 0.18767.

Findings:

The RMSE when using cross validation for both models if fairly close (0.18072 vs. 0.18767). When looking at the mean squared error for all 10 models used in cross validation for the decision tree regressor, the variance was pretty high. Thus, the difference between the two models' RMSE suggests that it's not really meaningful. Essentially, the two models performed about equally well.

Next, so that I could see which feature variables had low p-values, I also ran a linear regression with Stats Models. Several features had p-values lower than 0.05, suggesting that the variables are statistically significant. A couple variables were quite close to this mark, with p-values around 0.08.

Feature Variable	p-value	coefficient (pos or neg?)
About America	0.005	positive
Sports	0.001	positive

Science & Tech	0.046	positive
Development	0.043	negative
ShareAmerica	0.016	negative
Mission Affairs	0.001	negative
Study in the USA	0.082	positive
Photo (post type)	0.079	positive

The high variance of the 10 different models used in cross validation of the decision tree regressor, however, suggests that I might have a high variance problem.