**Leonardo Ji**

**We looked at how the number of likes helped predict the Audience label.  We can turn that question around and ask how the content is related to the number of likes.   Can we predict how many likes something will get just by using the text?  What are some of the more "important" words in determining the level of likes? Anyone who turns in two paragraphs explaining how they answered the question and what the answer is (technical language allowed) by Wednesday before class can earn up to two bonus points.**

With just the text there is no good model to predict which important words is associated to level of likes.

First I drew a histogram, the number of likes range from 0 to close to 40,000. A lot of documents have 0 likes. A few documents have many likes. The “likes” data column is not a categorical variable, it is a continuous variable. First, we need to convert it to a categorical dummy variable. I used scale function to convert it to z values. Then I used z values to set it to three categories: low (lower than 40 percentile), medium (40 percentile to 90 percentile), high (high than 90 percentile). Next, I can split it to 70% training set and 30% validation set.

I used three models Decision Tree, Logistic Regression, Naïve Bayes to predict level of likes using count vectorizer feature spaces. The Decision Tree model’s overall precision ratio is only 45%, the Logistic Regression model’s overall precision ratio is only 46%, the Naïve Bayes model’s overall precision ratio is only 48%. All three models have even lower precision and recall ratio at predicting high number of “likes”. What this tells me is all three models fail to find words that causes high number of “likes” on the news. I also tried to switch to TF-IDF weight vectorizer instead of count vectorizer to get the feature spaces before running the models, and the prediction precision and recall ratios are also not good.

Fig. 1 Use only Text

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Overall Precision | Overall Recall | High number of “Likes” Precision | High number of “Likes” Recall |
| Decision Tree | .45 | .46 | .16 | .11 |
| Logistic Regression | .46 | .47 | .08 | .05 |
| Naïve Bayes | .48 | .45 | .16 | .28 |

After I added the “audience” variable to the logistic regression model the overall precision ratio improved to 92% and overall sensitivity ratio is 92%. What this tells me is that “audience” categorical variable is much better at predicting the number of “likes” than just the text.

Fig. 2 Use Text and Audience

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models | Overall Precision | Overall Recall | High number of “Likes” Precision | High number of “Likes” Recall |
| Logistic Regression | .93 | .93 | .97 | .92 |

The models and results are in the bottom sections of TM2017Fall\_5followup.html.