How to write answers with stronger traffic-driving capability on quora?

Jie Meng, Yue Zhu, YanPeng Hu, YiRen Lu, Xiaowen You
SIST, Shanghaitech University

ABSTRACT

Concerning getting more upvotes in Quora, length of sentence, Mtld lexical diversity ,sentiment polarity, readability, number of words and subjectivity are most important features.

We built a model to estimate the future upvotes for new posts. The final cross-validated accuracy of our model is 89%

Acknowledgements

First of all, we would like to extend our sincere gratitude to Prof. Yizhou Lu for her useful advices on our project so that we finally increased our upvote predicting model's accuracy from 51% to nearly 90%. Her profound knowledge of web text mining and instructive comments triggered us to delve even deeper into the project and possibly further improve the model's performance in the future.

High tribute shall be paid to our teaching assistants, Huifeng Dong and Xiaohe He for their insightful responds to our questions and earnest preparations for the labs.

Problem Statement & Business Goal

In this era of internet, page view is money. Though originally designed as platforms to share knowledge and opinions, websites like Quora and Zhihu had gradually became important traffic driving machines for a great many who make fortune by increasing the clicks on their personal urls.

Our project aims at extracting the very features that contributes to a Quora answer's traffic-driving capability. We chose the number of upvotes as the measurement of traffic-driving capability instead of views since people who view the answer do not necessarily see the outlinks contained in the post.

Our project provides guidelines to authors who want to write better answers and our upvote predicting model is a useful tool to estimate future upvotes for new posts.

Web crawling & Data cleaning

The structure of Quora and Zhihu are similar: each answer corresponds to a single question while each question belongs to one or several topics. The number of answer under each questions is not limited.

We crawled two topics: Republican Party and Democratic Party. We get the raw text and other non-text features for each question and answers under the question. Around 1.8 thousand samples were finally retrieved.

We exempted answers which have not existed for more than a month from our dataset since the number of upvotes for those answers are unstable.

Feature Description

✓ **5 non-text features**

Tags for each question, number of answer's author's followers and question's followers as well as number of outlinks and pictures in each answer.

✓ **20+ text features**

- **Writing style features:** total word count, average sentence length, total sentence count, paragraph count, word ratio of different part of speech categories et al.
- **Dall-Chall readability score:**
$$0.1579\left(\frac{\text{difficult words}}{\text{words}}\times 100\right)+0.0496\left(\frac{\text{words}}{\text{sentence}}\right)$$
- **MLTD lexical diversity:** The richness of different word stems in a given text

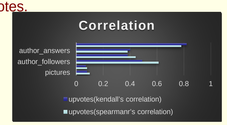
✓ **Content Topics**

✓ **Sentiment features**

- **Sentiment Polarity:** The range is between -1 to 1. The higher score is, the more positive the emotion is.
- **Subjectivity**: The range is between 0 to 1. The higher score is, the more subjective the answer is.

Non-text feature analysis

Using the Spearman Order Correlation Coefficient and the Kendall Order Correlation Coefficient to measure the correlation between each feature and the number of upvotes, we found that the author's number of followers and views of questions are highly correlated with his answers' upvotes.



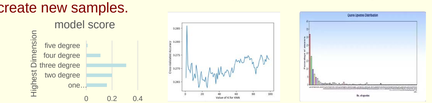
So we ignore these two features.

Next step, we will use these three features to build the model:

(1) number of author answers on quora
(2) number of links in the answer
(3) number of pictures in the answer

Then we use two types of model, KNN and polynomial regression, to predict the accuracy. As shown in the right, the distribution of upvotes are a long-tailed.

So we use cost-sensitive learning model and Smote algorithm to create new samples.



Although we tried other models like logistic regression, the accuracy rate hardly achieved 30%. This means other text feature need to be considered in the model.

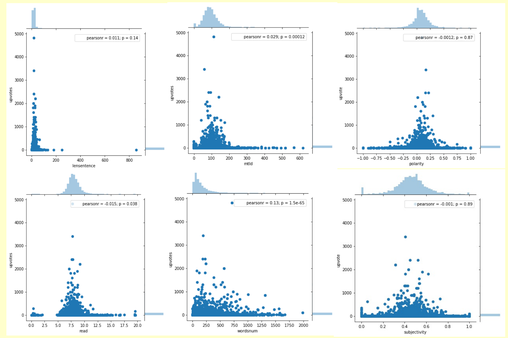
So in next step, we will dig text features and model with these features.

Ranking of text features

After applying PCA, five most important text features were found, and some observations were made:

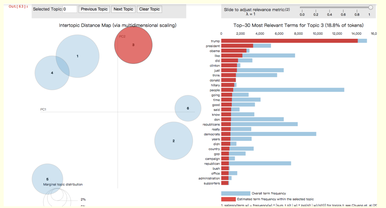
features	frequency	lower limit	upper limit
Length of sentence	0.888	1	30
Mtld	0.881	50	150
Sentiment polarity	0.923	-0.25	0.25
Readability	0.973	5.5	10
Number of words	0.720	2	500
Subjectivity	0.879	0.3	0.6

1. When average sentence length is within 30 words，it is more possible(88.8%) to get higher upvotes.
2. Emotion should be as rational as possible. When the polarity is between -0.25 and 0.25, 92.3% answers get higher upvotes.
3. The answer should be within 500 words.



LDA Topic Clustering

Using Latent Dirichlet allocation method, answers were clustered into six categories.



Next, remove the most common topic, because they're not representative. Also remove the unpopular topics. At last, 4 topics remained. Their ratios and upvotes were as following:

Number of upvotes for a topic is calculated by this foluma :
Topic Upvotes = Σ(t1,t2,t3,t4)*upvotes.
where t1 t2 t3 t4 is the the contribution to each topic.

topics	Ratio	upvotes
Parties and Elections	19.2%	39547.61
Presidents	18.8%	63620.85
Policy	15.6%	30486.41
Governments and Human Rights	12.6%	35534.63

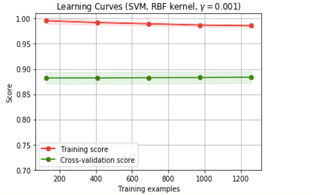
✓**Result:**
Topic about Parties and Elections are most popular, which accounts for 19.2%. But the topic, president, will attract more upvotes.

Upvote predicting model

✓ **Input data and pre-processing:**
Answers were divided into three categories labeled by 0,1 and 2, with their upvotes ranging from [0,10], [10,100] and [100,5000].

9 features were selected:
'author followers', 'author answers', 'number of words', 'length of word', ' long-short', 'number of sentence', 'readability', 'mtld', 'sentiment polarity'.

✓ **Classifier – SVM model**
Training set : test set = 7:3
Cross validation：each time with 30% data randomly selected as a validation set
Number of iterations: 60



✓ **Result analysis:**
There is almost no overfitting since training score and cross-validation score is close enough. The accuracy of the test set is 89%, which has proved that our model is effective enough to predict the upvotes. That is, the features we selected affect the model.

Conclusions & Future Expectations

In this project, we first use analytical method to inspect the most influential features that would increases an answer's upvote count on Quora. We found that the following five variables are important: length of sentence, Mtld lexical diversity, sentiment polarity, readability, total Words counts and subjectivity.

Then we used machine learning techniques to predict an answer's upvote count. Before the presentation, we mistakenly used only four most dominant features to train our model and the accuracy was at most 51%. After taking Prof. Zhou's advice, we added other seemingly ignorable features to the SVM classifier. Finally, the accuracy of our model became 89%, which is quite successful.

In the future, we may do more researches on deriving formula represented by upvotes and author's followers in order to better measure an answer's traffic-driving capability. We may also use Doc2Vec to measure the extent of relevance of an answer to a question and take it as another feature in our model.