Can reddit tell us what makes a question stupid?

Winston DeGraw, Data Scientist

Background

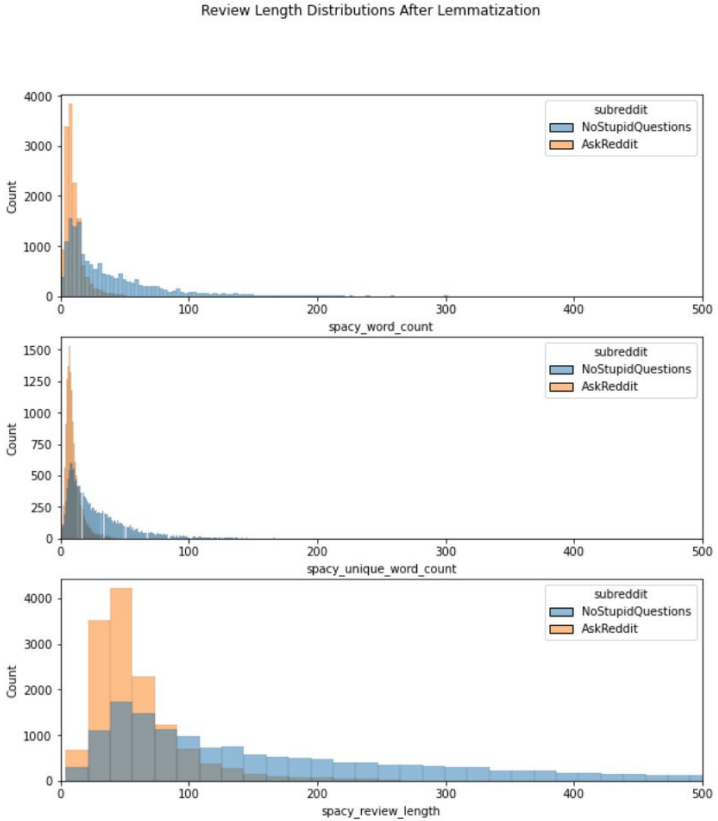
- Scraped submissions from r/AskReddit and r/NoStupidQuestions to build a model to distinguish between the two
- Why else ask a question in r/NoStupidQuestions unless you fear it is a dumb question
- Goal is to determine the features that make up a ‘stupid question’

Unique Posts per subreddit

NoStupidQuestions	15798
AskReddit	13798

Methodology

- Combined text and titles to increase data population
- Lemmatized the text and used TFIDF vectorization for creating features
- Limited the document frequency to ensure features smaller than sample size
- Compared the output of several models to select best performing
- Conducted a hyperparameter grid search for each model option
- Included review length (post lemmatization) for best model performance



Model Selection

- Selected from four models
- Struggled with high variance in each case
- In the end **Random Forest** proved to be the best performing model

model	train_scores	test_scores
kNN	0.994	0.531
Naive Bayes	0.843	0.741
Logistic Regression	0.888	0.762
Random Forest	0.784	0.767

Common Features of a Stupid Question

- Too long; keep it concise
- Self-referential
- Positivity/negativity
- Uncertainty
- Opinion-based

Words most likely to appear in a stupid question

Some Successes and Failures(?)

- Correctly predicted stupid question:

  "Why do people watch Mukbangs? I seriously get nauseous just looking at the thumbnails for these videos. And from what I know, the videos themselves are super messy. I just don't see what type of entertainment people get out of these videos."

- Correctly predicted smart question:

  "What's the funniest way to die?"

- Incorrectly not labeled as a stupid question:

  “When cheating happens, is it the fault of the married man or the mistress?”

- Incorrectly labeled as a stupid question:

  “What is your view on the male social hierarchy? (e.g, Sigma, beta and alpha male).”

If you find yourself asking...

Possibly stupid question:

Is this a stupid question?

No

Summary

- With ~77% accuracy a Random Forest classifier could distinguish between posts from r/NoStupidQuestions and r/AskReddit
- Stupid questions are most strongly predicted by their length, followed by referring to reddit itself (it's stupid to ask about reddit)
- Classification is somewhat successful based on text alone in this case, though engineering other features aided the model in reducing its variance and increasing its accuracy.
- A future model would include the presence of 'selftext' as a feature, as I expect this would help as a predictor