Humor Detection in New Yorker Cartoon Captions

Hari Anbarasu

Department of Computer Science, Yale University

LILY

Introduction

Humor is one of the most interesting sub-fields of natural language processing as computers are fairly incapable of detecting humor. Humor detection by a computer is especially difficult, given that there is no definition for humor amongst humans. What one person finds funny may fall flat in front of a different audience. Furthermore, there are many types of humor such as puns, sarcasm, etc. thus making it nearly impossible to create a taxonomy of humor. Finally, since most humor requires cultural context a computer would have to be trained for cultural relevancy in addition to the subtleties of humor itself. For these reasons, humor has been dubbed the "final frontier" of artificial intelligence. In this work, I attempt to formalize humor detection as a prediction task in which I assign each of existing New Yorker captions a score based on their received votes, and I attempt to predict the score of an existing caption by using linear regression with 10-fold cross validation. To generate the linear regression model, I explored the semantic structure underlying humor from the perspective of incongruity, ambiguity, polarity and size.

Materials and Methods

The data set was provided by Bob Mankoff of The New Yorker. I cleaned the data set by removing all captions with fewer than 50 votes, and then I assigned a score to all the captions using the lower bound of the Wilson score confidence interval for a Bernoulli parameter. After removing any other incorrectly formatted captions, I tried to find the underlying semantics of each sentence. I looked for the incongruity, ambiguity, polarity and size of a sentence. For incongruity, I used Google's Word2Vec, with Google News' trained model, to calculate the semantic meaning distance between words. From this, I calculated the minimum and maximum similarity between words. For ambiguity, I used WordNet and calculated the number of possible meanings a sentence could have, and the farthest/closest away two senses in a sentence were. For polarity I used SentiWordNet to calculate the polarity of each word, and then summed that for the sentence. For size, I took the number of words in the sentence.

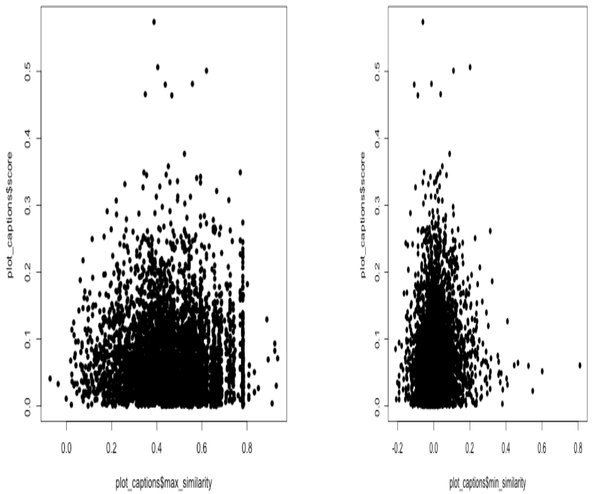

Figure 1. Two plots of min and max similarity vs. score.

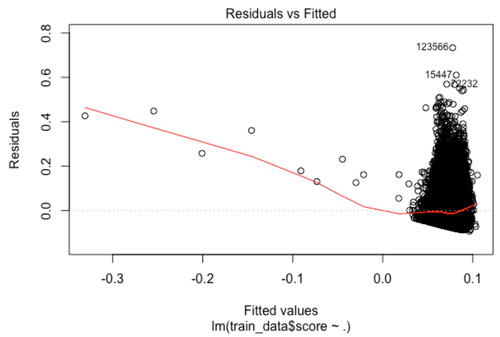

Figure 3. Fit of Linear Regression

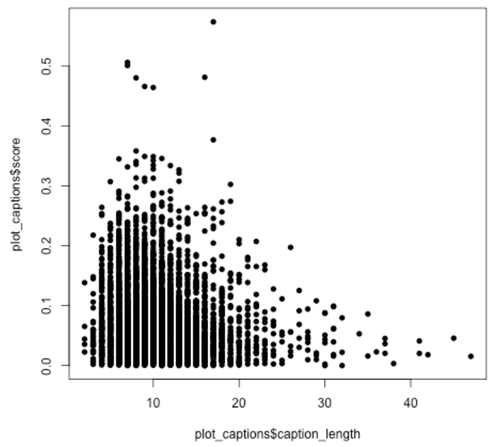

Figure 2. Relationship between caption length and score

```
Call:
lm(formula = train_data$score ~ ., data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09376 -0.04019 -0.01316  0.02490  0.73314

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.750e-01  9.928e-03  47.850  < 2e-16 ***
cartoon          -7.428e-04  1.825e-05 -40.699  < 2e-16 ***
min_similarity    1.327e-02  2.659e-03   4.990 6.04e-07 ***
max_similarity    4.919e-03  1.279e-03   3.847  0.00012 ***
caption_length   -1.136e-03  6.987e-05 -16.251  < 2e-16 ***
sense_combination 8.746e-04  1.430e-04   6.118 9.53e-10 ***
sense_farmost     4.166e-03  2.111e-03   1.973  0.04849 *
sense_closemost   6.200e-02  1.120e-02   5.538 3.06e-08 ***
positive          1.604e-03  1.841e-03   0.871  0.38372
negative         -1.788e-03  1.651e-03  -1.083  0.27891
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05665 on 108861 degrees of freedom
Multiple R-squared:   0.02,     Adjusted R-squared:  0.01992
F-statistic: 246.9 on 9 and 108861 DF,  p-value: < 2.2e-16
```

Figure 4. Summary of Linear Regression

Results

Initial results seemed to indicate that I had received good results. For background knowledge, the average score of a caption, when looking at all captions, was 0.079 and the average score of one of the top 20 captions, from each cartoon, was 0.361. Both the linear regression and the support vector machine approach resulted in similar results as they both produced an average error rate around 0.04. Specifically, the linear regression had an average error 0.042 and the SVM approach resulted in an average error rate of 0.0404. This meant that given a caption, on average, the actual score of a caption was within plus or minus 0.04 of the predicted score. Looking more closely however, it was evident that these models did not fit the data well at all. The maximum score predicted by the linear regression model was 0.102 and the maximum score predicted by the SVM was 0.11. Since out of the 130,000 captions only a handful of them had a relatively high score, both the linear regression and SVM model predicted that almost every caption was bad. While, overwhelmingly, the majority of captions did have a poor score, my models did not predict any caption to have a high score so it is clear that the model needs tweaking.

Conclusion

Overall, I would say that my prediction model did not work, but it did highlight many of the relationships between the semantic structures and its relationship to humor as it pertains to the caption contest. For example, sense combination, max similarity, min similarity and caption length all seemed to have a quadratic relationship with the score of the caption. Both positive and negative sentiments seemed to be negatively correlated with the score of a word. While the direct linear regression or SVM method might not be able to be used to predict the score of a caption, these underlying trends might.

Acknowledgement

I want to thank Professor Radev for his help and support as well as Bob Mankoff for supplying the data!