Can NLP be used to distinguish puns from truths?

By Chris Ratigan

The Subreddits

- TechnicallyTheTruth (TTT)
  - Founded in 2017
  - Devoted to unexpected truths
  - Often reposts of images from other sites

- Dadjokes
  - Founded in 2011
  - When it becomes apparent
  - Primarily text-based posts.

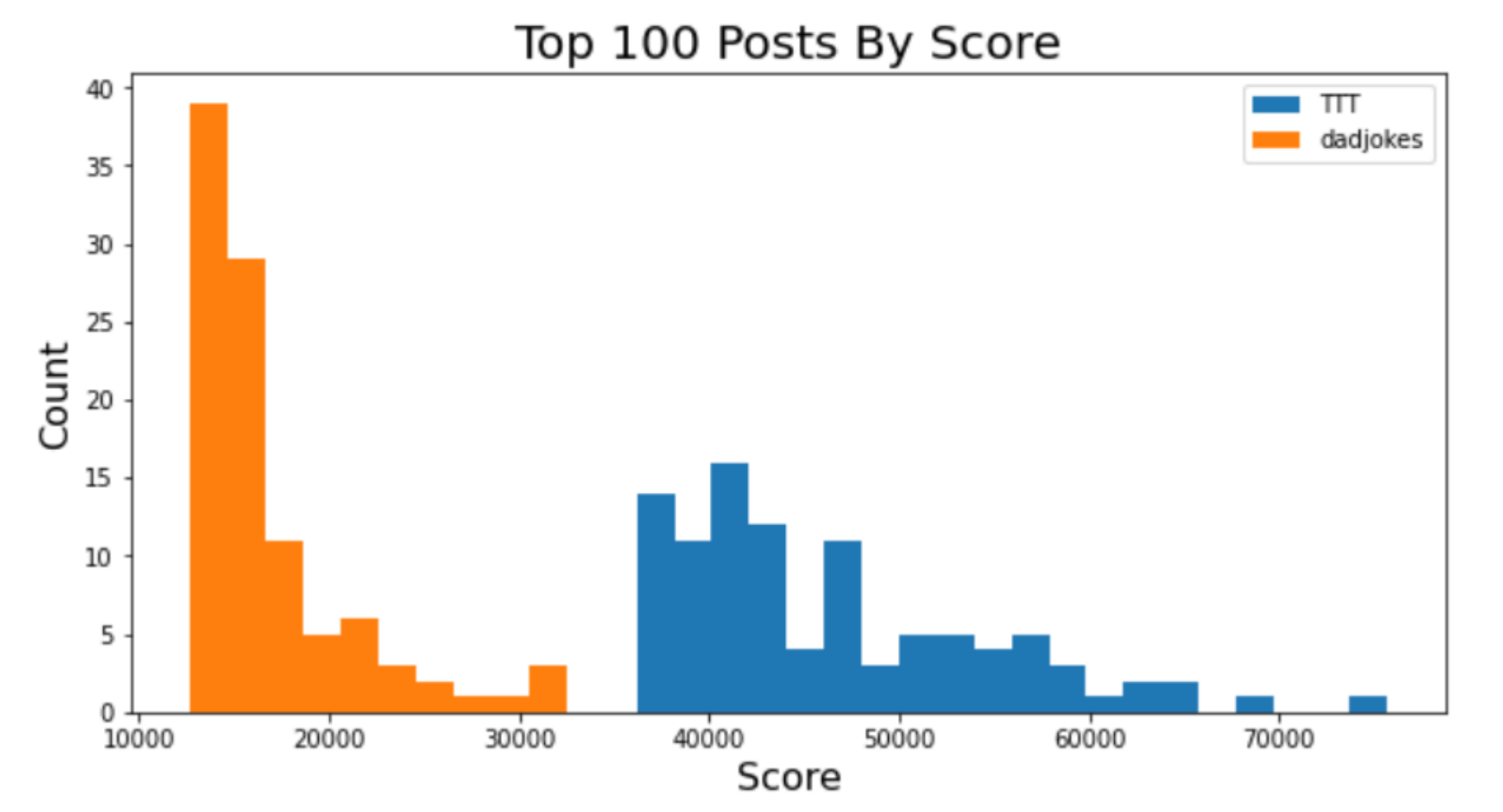
Why was 4 scared to ask out 5?

Because 4 was 2^2

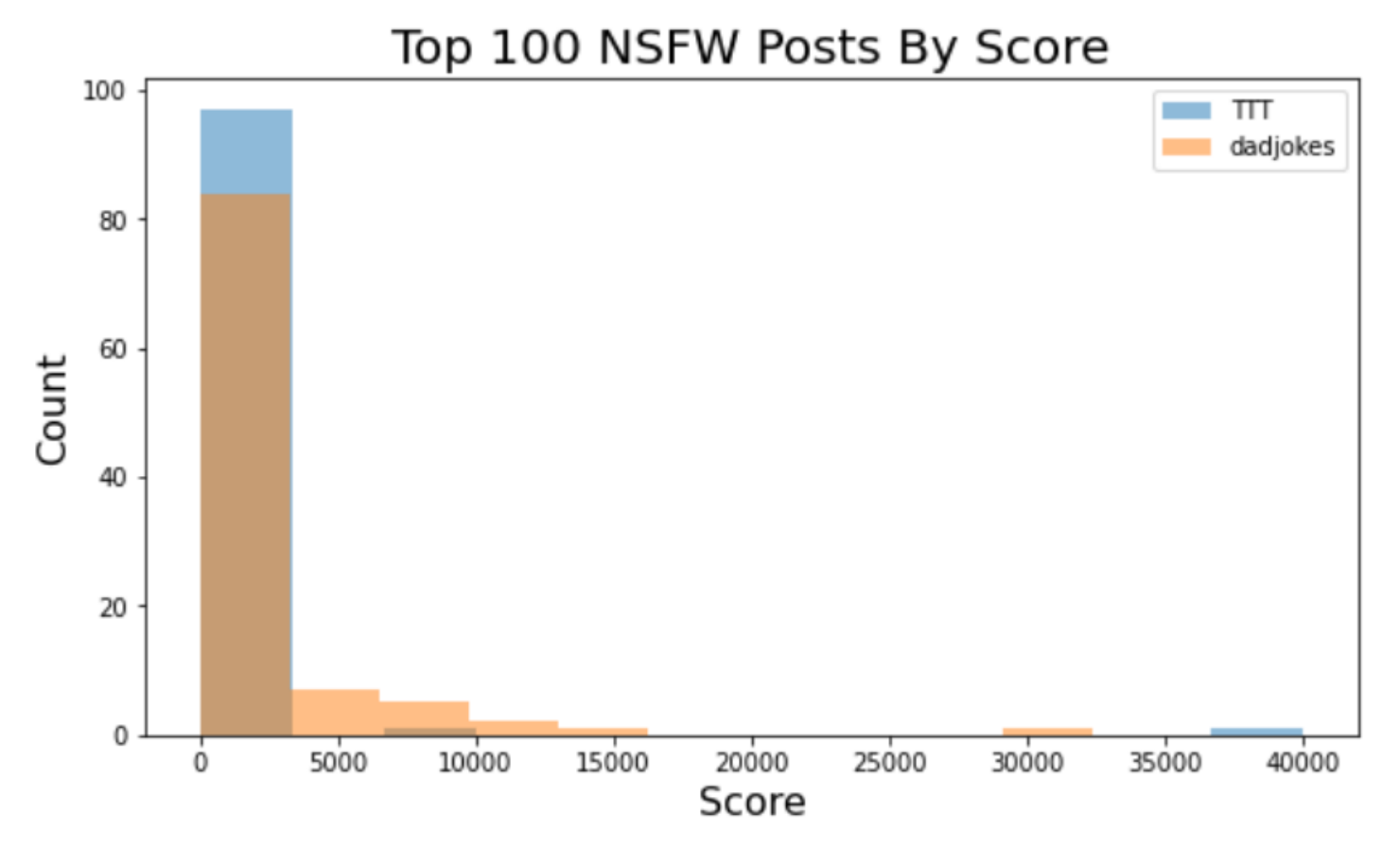
Getting the Data

- Pushshift API
- 300 posts from TechnicallyTheTruth and dadjokes:
  - Top 100 posts by score
  - Top 100 NSFW posts by score
  - Top 100 Spoiler posts by score

Some EDA


Top 100 Posts By Score
TTT
dadjokes
Count
Score
0
5
10
15
20
25
30
35
40
10000
20000
30000
40000
50000
60000
70000

NSFW

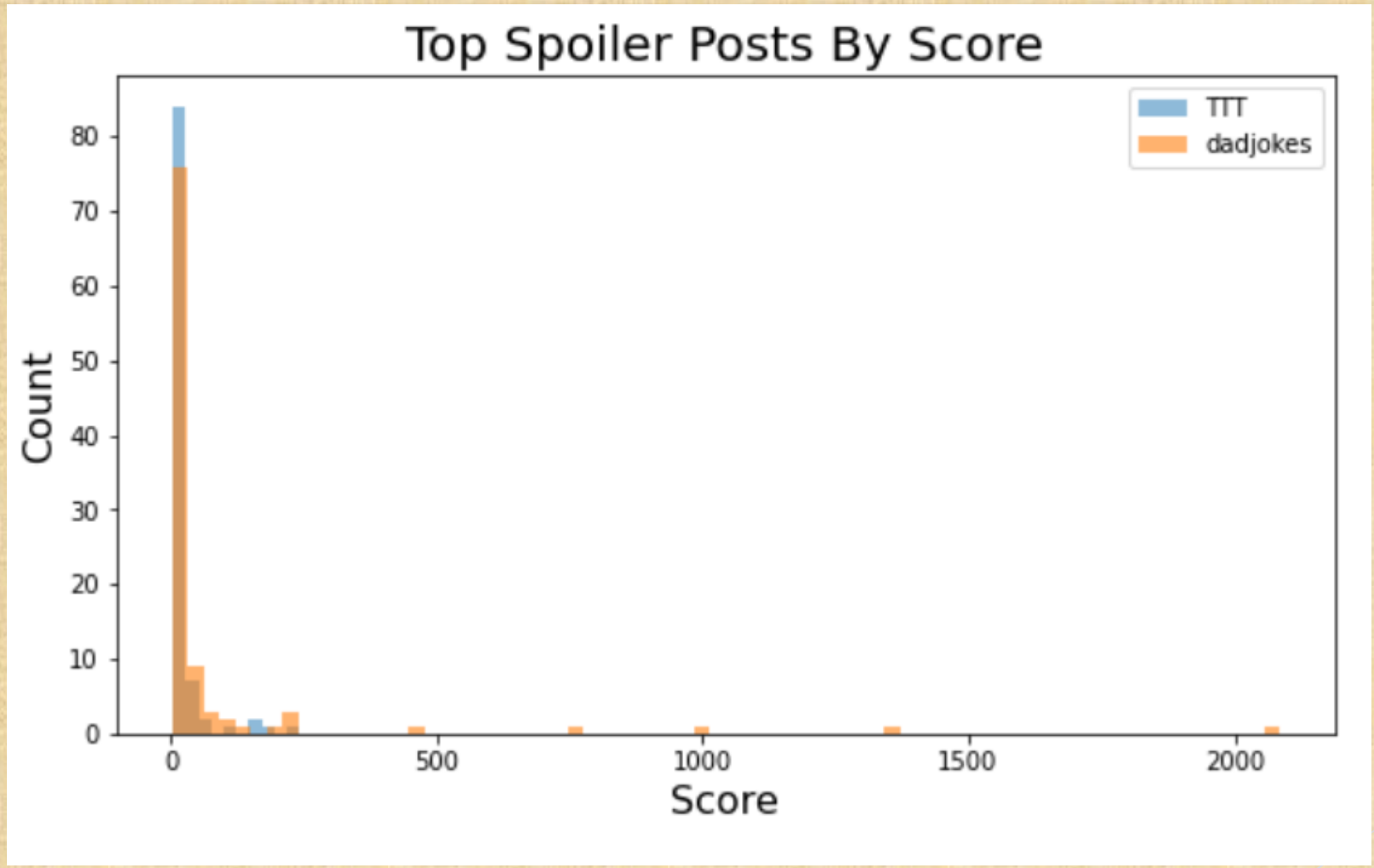

Top 100 NSFW Posts By Score
TTT
dadjokes
Count
Score
0
20
40
60
80
100
0
5000
10000
15000
20000
25000
30000
35000
40000

Spoiler Outlier

15.1k Don't open, it's a spoiler.

spoiler

Spoiler


Top Spoiler Posts By Score
TTT
dadjokes
Count
Score
0
10
20
30
40
50
60
70
80
0
500
1000
1500
2000

Is_self

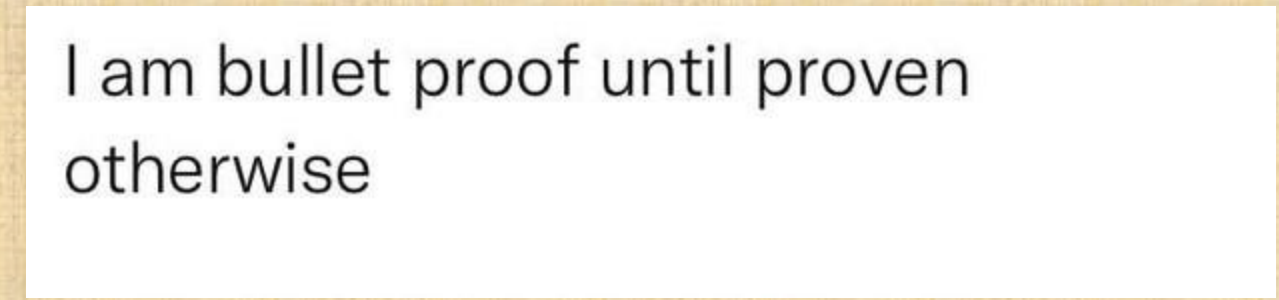
This feature alone could be used to build a very accurate classifier.

Dataset	True Is_self
TTT top 100	0%
TTT NSFW	6%
TTT Spoilers	5%
Dadjoke top 100	97%
Dadjoke NSFW	100%
Dadjoke Spoilers	99%

But, that’s not in the spirit of the problem

Data Cleaning/Processing

- For posts with images of text, we used pytesseract to convert the images to strings. E.g.



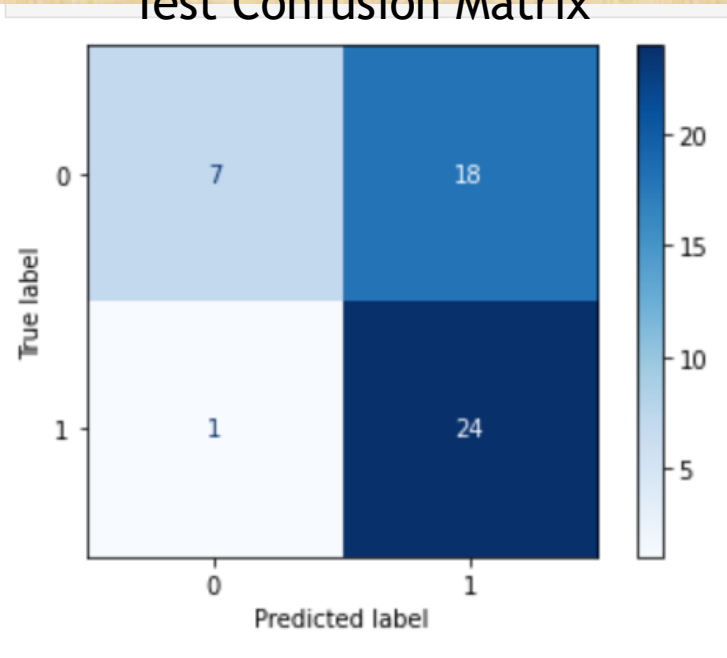
Becomes “| am bullet proof until proven \n otherwise”

- CountVectorize text for the resulting posts.
- Train/Test Split

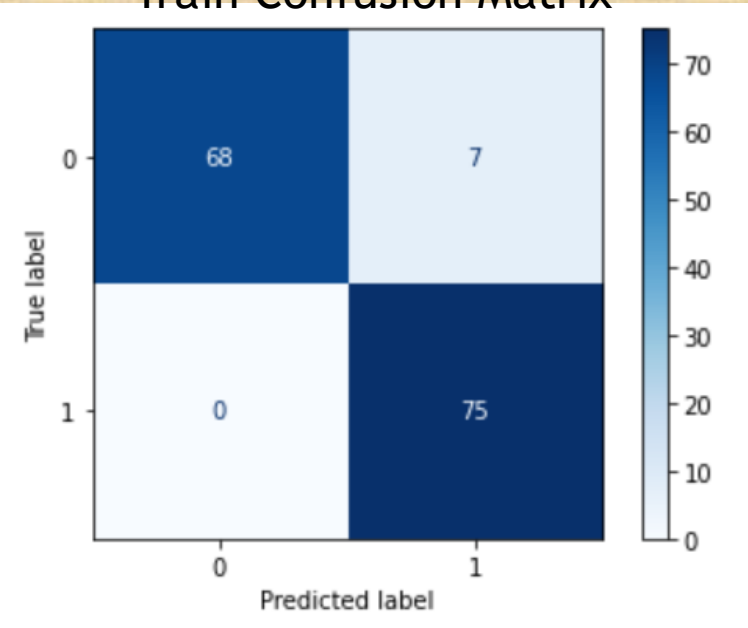
Modeling

- Binary Classification (is_dadjoke)
  - Logistic Regression
  - 62% accurate on test data But, 95% accurate on the training data
- Baseline model: 50% accuracy score

Test Confusion Matrix



Train Confusion Matrix



Conclusion

- Posts on r/dadjokes and r/technicallythetruth are formatted differently
- The language used on these subreddits is different
  - Though similar

- Future research could
  - Use different data (e.g. the average, not just the hits)
  - Further process the data.
  - Use a combination of other models.

Thank you

- Any Questions or Comments?