Detecting Bannable Content on Reddit

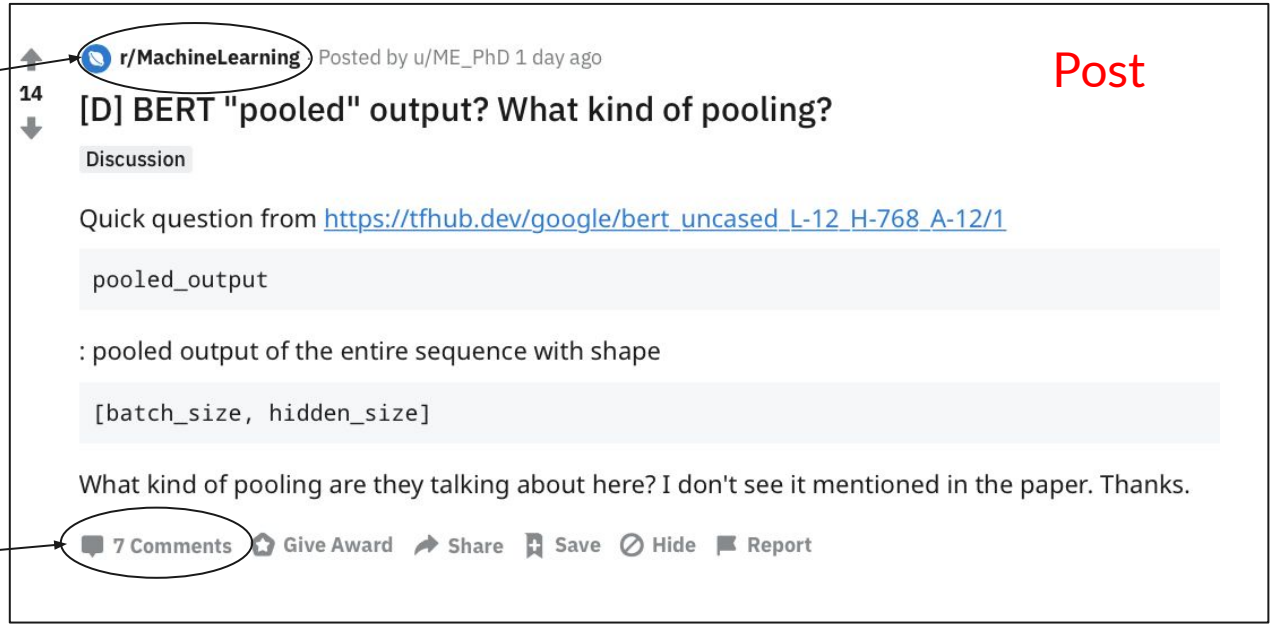
Kelly Gothard, David Matthews, Nick Hanoian, Colin Sandberg

Background

Subreddit

Post

Comments



Background: Examples

- **Banned:**

Hookers	"Services are free! F*** I'll fly there if the"
watchpeopledie	**"too bad he wasn't"**
fakeid	'420'
CringeAnarchy	'f*** is this g** s***'
fakeid	**"still haven't received my AK and U21 MS :("**

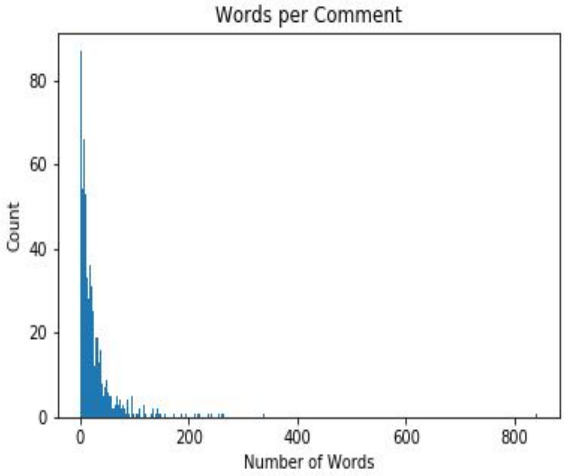
- **Not Banned:**

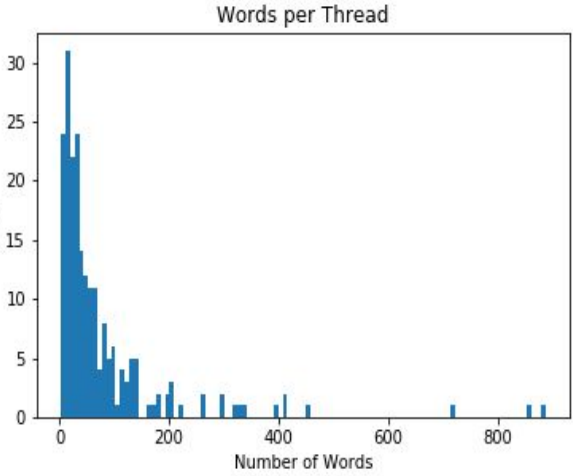
madisonwi	'When was the last time you were at Red Letter'
worldnews	'apparently PETA euthanize an enormous amount '
freemasonry	**"There's always been a demand for male-only sp"**

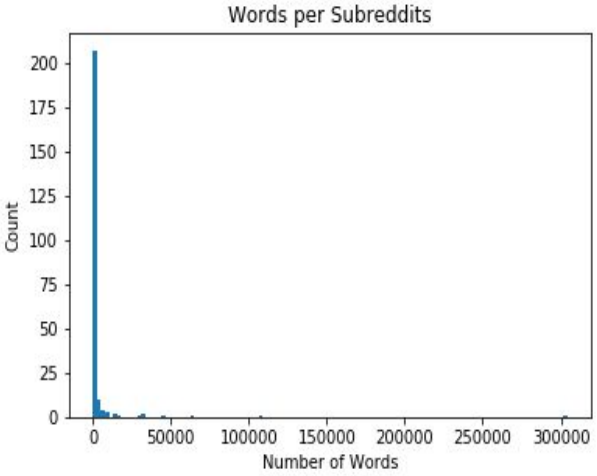
- **Total Data Not Banned / Banned Ratio : 183:1**

Data Types

- Comments, Threads, 200 Words, Subreddits
- 200 Word distribution is uniform

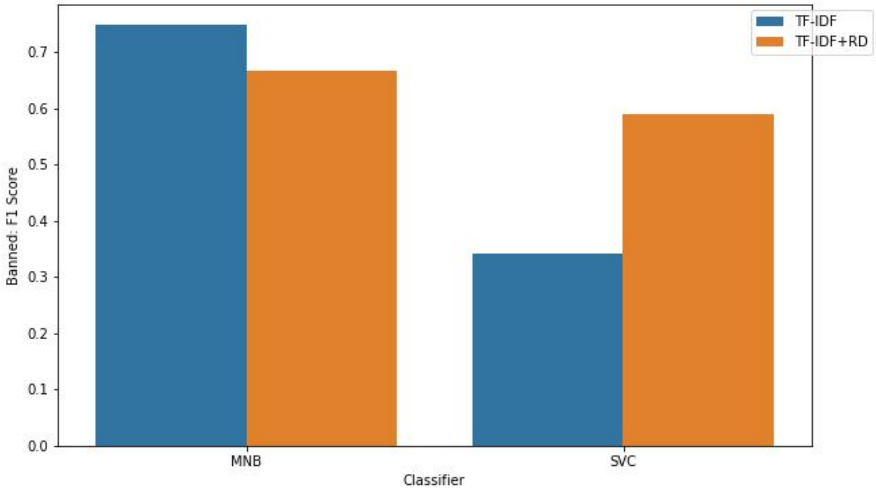






Support Vector Machine and Naive Bayes

200 Words corpus, n = 2000, balanced



Confusion Matrix:

	Not Banned	Banned
Not Banned	896	104
Banned	341	659

We can do better!

Rank Divergence

r/Disney Corpus

Word	Count	Rank
the	10000	1
fun	2000	130

r/MachineLearning Corpus

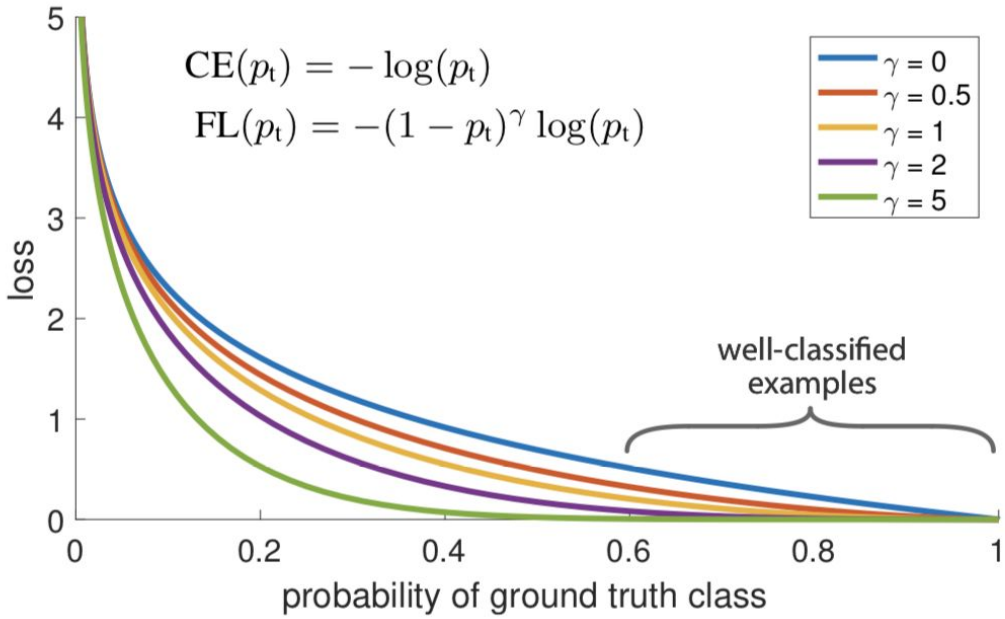
Word	Count	Rank
the	10000	2
fun	200	2345

- RankDiv(the) = 1 - 2 = -1
- RankDiv(fun) = 134 - 79 = 50

Data Blocking

- Comments, threads, and subreddits were not performing well
- New dataset
  - 1st 10 M Reddit Comments from all of October 2016
  - Split into banned and not banned by 200 word segments
  - Test set differs from training set

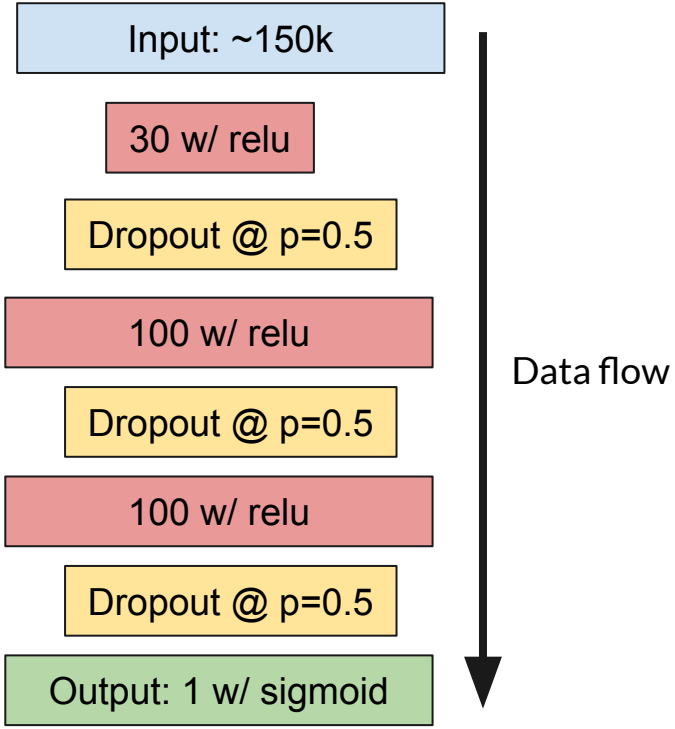
Focal Loss



Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

Neural Network

- 200-word samples -> bag of words -> TF-IDF -> MLP neural network
- Choice of loss function
  - Binary Cross-entropy
    - Training / validation on balanced data -> Good
    - Validation after training on unbalanced -> Bad
  - Focal Loss
    - Training / validation on balanced data -> Still Good
    - Validation after training on unbalanced -> Better
- To predict a subreddit:
  - Given set of all 200-word samples from that subreddit
  - Calculate average output of MLP for all samples
  - Predict banned/not banned based on a threshold



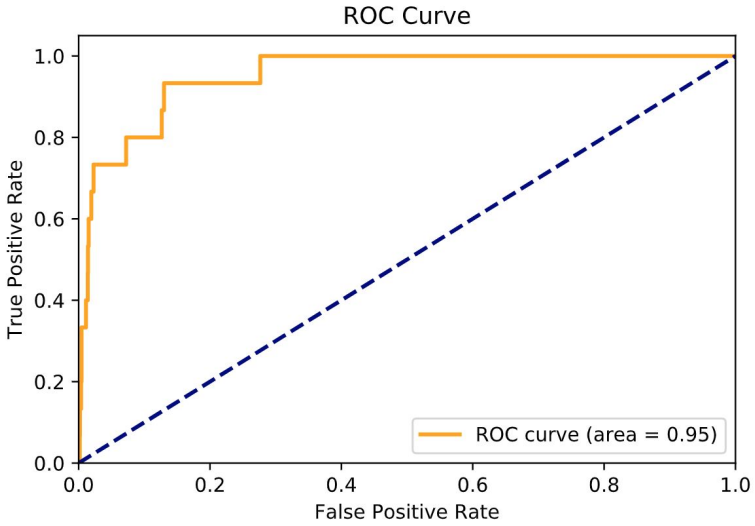
MLP Architecture

Neural Network Results

Network training results

	Class Balance	F1 score
Training	20:1	0.936
Validation	1:1	0.524

Picking a threshold



Neural Network Results (cont.)

End Result

Validation after picking threshold

```
              precision    recall  f1-score   support

       False       0.91      1.00      0.95      2692
        True       0.87      0.05      0.09       269

   micro avg       0.91      0.91      0.91      2961
   macro avg       0.89      0.52      0.52      2961
weighted avg       0.91      0.91      0.88      2961

[[2690    2]
 [ 256   13]]
```

Final Testing

```
              precision    recall  f1-score   support

       False       0.92      1.00      0.96      2700
        True       0.80      0.05      0.09       262

   micro avg       0.91      0.91      0.91      2962
   macro avg       0.86      0.52      0.52      2962
weighted avg       0.90      0.91      0.88      2962

[[2697    3]
 [ 250   12]]
```

Conclusion and Future work

- We are able to recall bannable subreddits (true positives) over 90% of the time
- At-risk subreddits would be forwarded to a content moderation team for review

- Future work
  - Improve neural network architecture so that we can train on the true class balance (~183:1)
  - Possible models to investigate
    - 1D convolutional models
    - WaveNet
    - ResNet1D