



---

# Toxic-Free Environments

---

Luke Newman

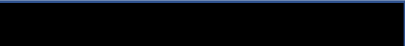
# Introduction



- Maintain civility in online conversations. For example,
  - Facebook




 Top Fan

 I've been waiting for this!  
Yay-Can't Wait!

Like · Reply · 1d · Edited



 He's trash and you're idiotic too

Like · Reply · 1d

# Methodology



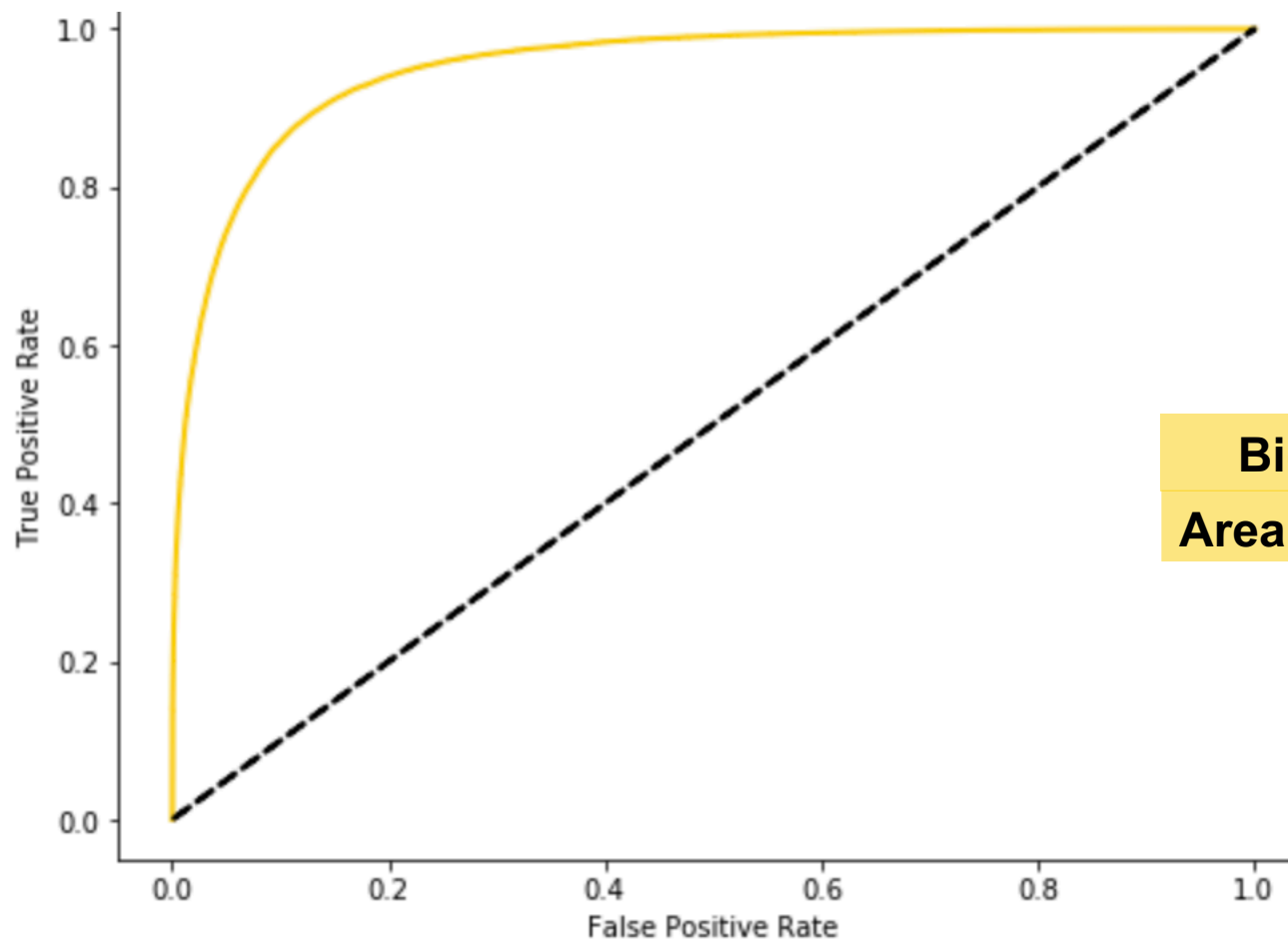
- 2 million online comments submitted to Civil Comments Platform
- Metric: AUC score
- Tools



Natural Language Analysis  
with Python NLTK



# ROC Curve



**Bi-Directional LSTM Transfer Learning**

**Area Under Curve: 0.951**

# Results



- Analyzing comments in real time

Probability of being toxic	Flagged
0.2%	No



# Results



- Analyzing comments in real time


Toxicity Percent	Flagged
0.2%	No
99.5%	Yes

 Top Fan

I've been waiting for this!  
Yay-Can't Wait!

 11

Like · Reply · 1d · Edited



He's trash and you're idiotic too

Like · Reply · 1d

# Conclusion

---




- How can we maintain civility in the digital world?
  - It depends on the online platform!
- Gaming



# Conclusion




- Facebook and Twitter
- Real-Time toxicity tracker



Toxicity: 0.00%

Hello, this is amazing|



Toxicity: 81.06%

You're too dumb to play in the NF|



# Further Work

---



- Build a model that is less biased towards marginalized groups
  - More data
- Toxicity tracker



---

# Questions?

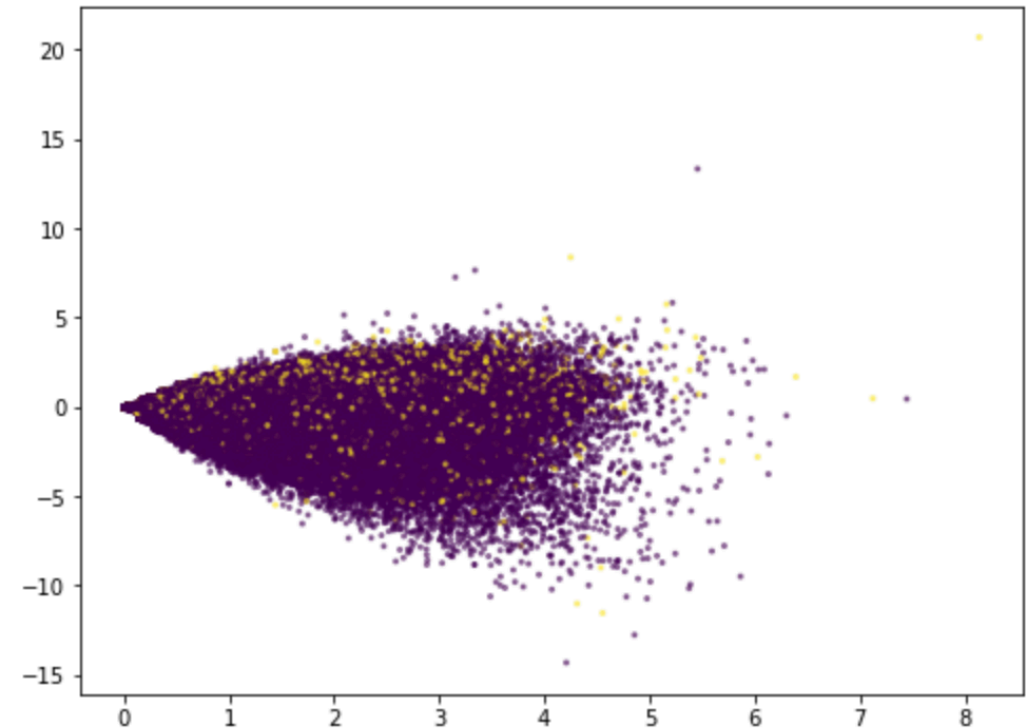
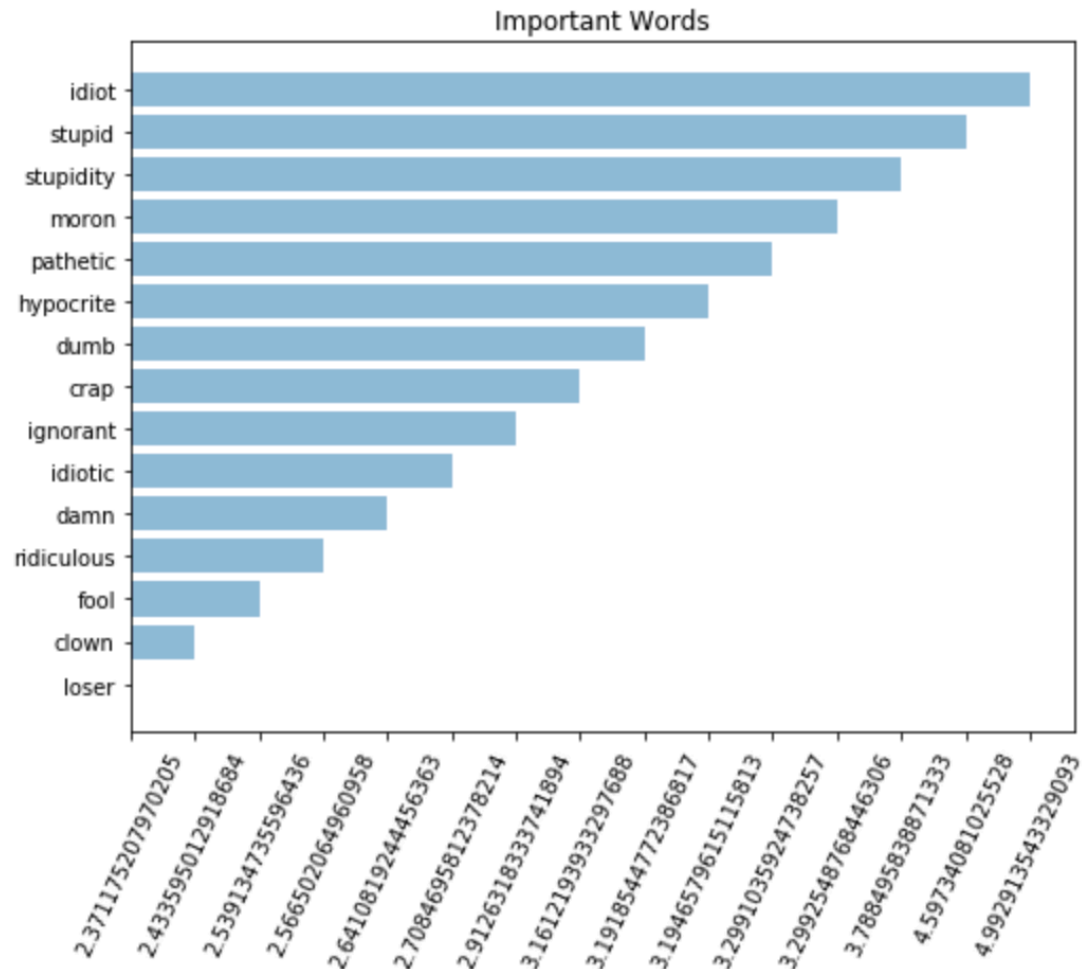
---

# Appendix



- Transfer Learning is a very powerful tool. 200 million words in my vocab I obtained pretrained embeddings for from google's Word2Vec. Leaving only 200,000 my model trained.
- Once you use Neural Networks models become harder to explain. For logistic regression and TF-IDF embeddings I was able to look at the words my model was using the most when predicting toxic comments

# Appendix: CountVectorizer



# Appendix: TF-IDF Vectorizer

