

Toxic Comment Classification

The background of the slide features a dark blue grid. Overlaid on this grid is a decorative graphic consisting of a series of vertical bars of varying heights, creating a bar chart effect. A white line graph with circular markers is also overlaid, showing an upward trend with some fluctuations. The line starts at a low point on the left, rises to a peak, dips, and then continues to rise with several smaller peaks and valleys before ending at a high point on the right.

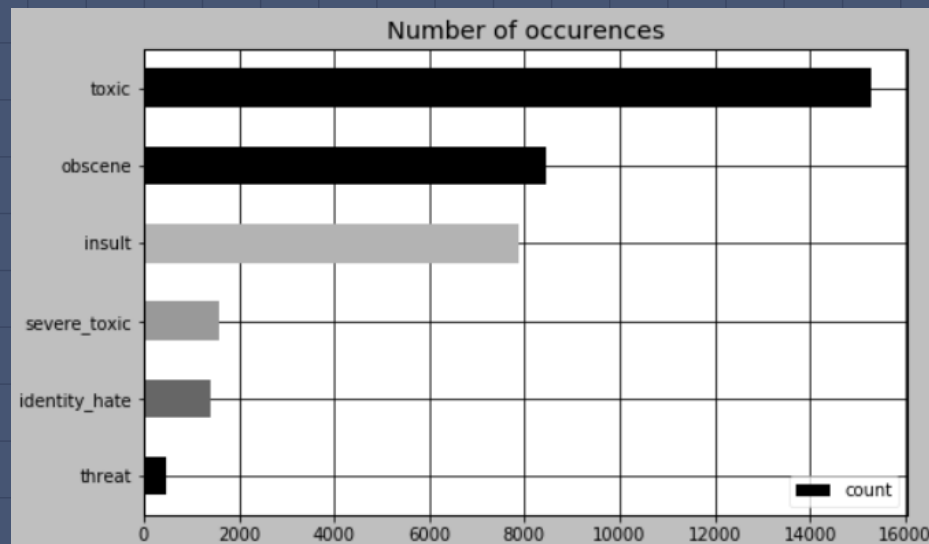
Niko Ganev

Can we improve
online
conversation
through machine
learning ?



The Data

- **Toxic Comment Classification dataset on Kaggle**
- **159,571 labeled comments from Wikipedia's talk page**
- **63,978 comments for model testing**
- **Six Labels :**
 - Toxic
 - Severely Toxic
 - Obscene
 - Identity Hate
 - Insult
 - Threat



Methodology: Deep Learning

- ▣ Keras Library
- ▣ Basic Neural Network
- ▣ Convolutional Neural Network
- ▣ Recurrent Neural Network



Test Data Performance

95.34% Accuracy - 11.55% Loss
Simple Neural Network

96.50% Accuracy - 8.78% Loss
Convolutional Neural Network

96.68% Accuracy - 8.23% Loss
Recurrent Neural Network

Recommendations:

Best Practice

- Recurrent Neural Networks are best suited for speech analysis

More Data

- NLP models require lots of data, increasing training data can have significant effect.

Regulate

- Ban utilization of words most often associated with toxic comments

Future Work:

Train Longer

- Training for more epochs will improve the performance but requires significant computational power

Try Using Pre-Trained models

- Pre trained models might be able to improve performance without requiring much computational power

Deploy on AWS

- This model can live on AWS in a very cost-efficient way and generate an alert everytime a certain type of comment (ex: threat) is detected

THANKS!

Any questions?

- ▣ [linkedin.com/in/niko-Ganev](https://www.linkedin.com/in/niko-Ganev)
- ▣ github.com/ganevniko
- ▣ ganevniko@gmail.com

