The Internet Toxicology Test

By: Greg Eastman

The Internet and Toxicity

- Why does moderating matter?
  - Makes forums safer to talk on.
  - Encourages a specific culture.
  - Discourages gathering of hate groups.
- What is the trouble with moderating?
  - The number of forums are outgrowing the number of moderators.
  - Human moderators may have different opinions on what is acceptable discourse.
- The solution is to have a machine that is consistent evaluate these spaces.
- The Internet Toxicology Test is an AI made for that purpose.

Data

- This model was built on data of toxic comments from Wikipedia.
- There are over 30,000 observations.
- I prepared the data by
  - Balancing the outcome.
  - Tokenizing.
  - Casting to lowercase.
  - Removing stopwords.
  - Lemmatizing.

Model

- Neural Net
- Encoder-Decoder Architecture
- Uses Bidirectional Long Short-Term Memory.
- Final Layer is fully connected with a sigmoid activation function.
- It was built in Keras.

Results

Accuracy	Precision	Recall
.8511	.8114	.8811

In order of best statistic:

1. Recall is a measure of how well the model identifies a toxic comment when the comment is toxic.
2. Accuracy is total number of correct guesses over total guesses.
3. Precision is a measure of how often a comment is toxic when the model guesses it is toxic.

Demo Time!

Questions?

Thank You for Being a Great Audience!