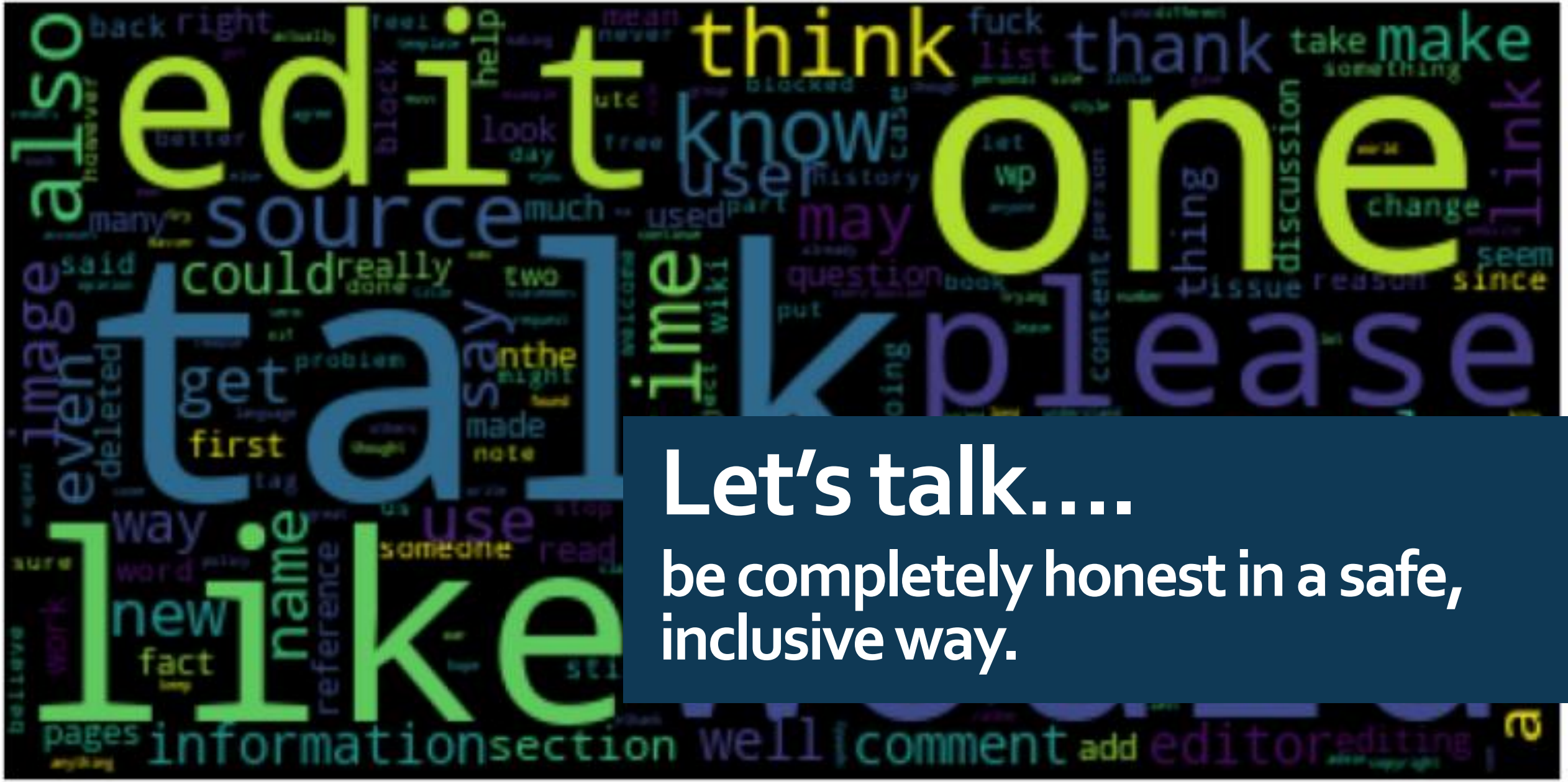
NoN-Toxic Communication

Using Natural Language Processing to Identify Toxic Language

By Andi Osika

WARNING: Offensive Language Identified



Let's talk….

be completely honest in a safe, inclusive way.

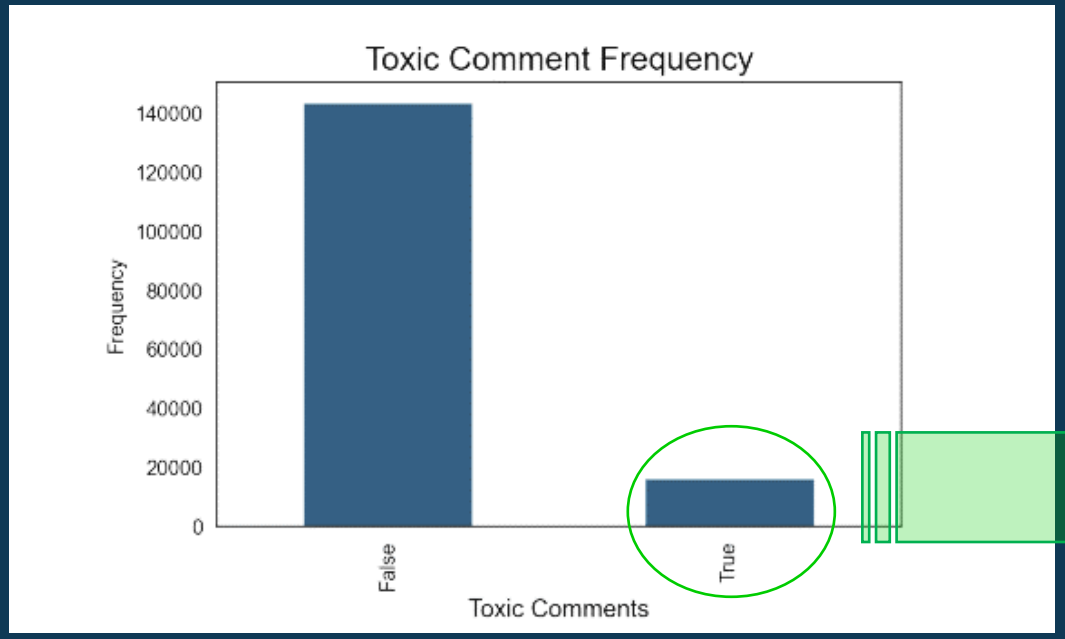
Problem:

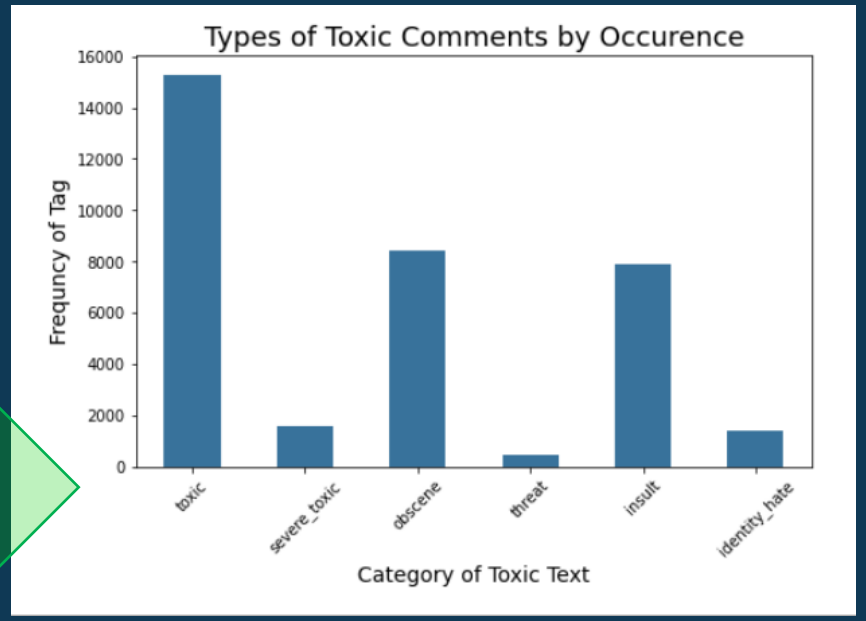
Freedom of speech is … sometimes toxic

The dataset was provided by Conversation AI in hopes to improve online discussion.

Wikipedia Talk Pages 150K + samples rated by humans for toxic effect varying in range from

- Toxic
- Severe Toxic
- Obscene
- Threat
- Insult
- Identity Hate

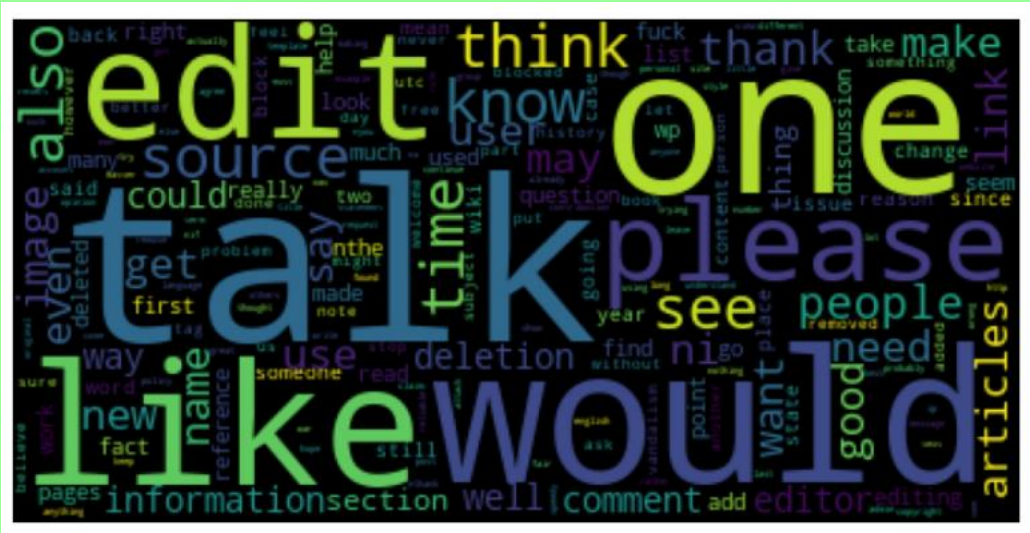




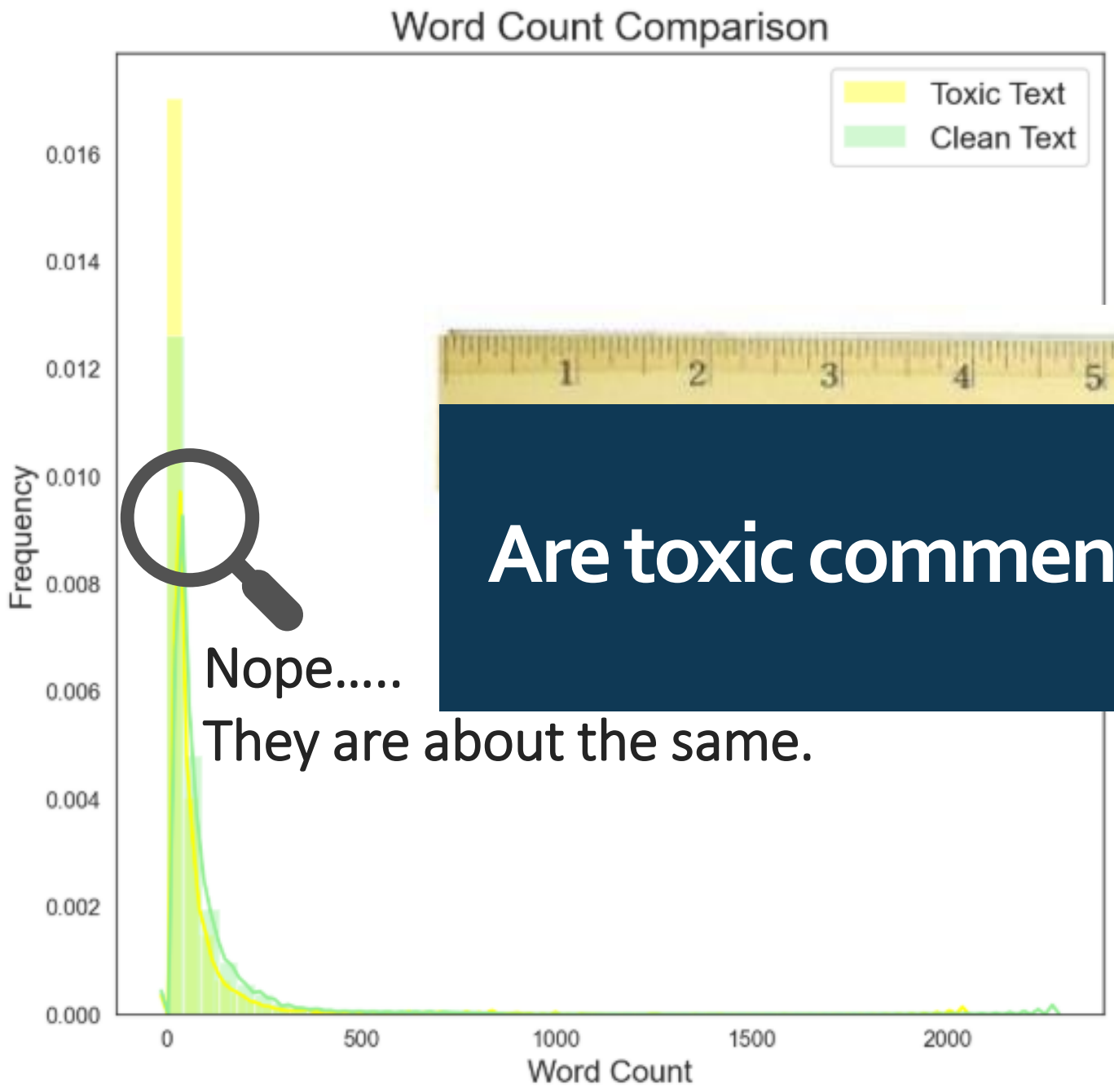
Examples:

Word Clouds display words in a collection. The larger the word, the more frequently used the word is.

Toxic

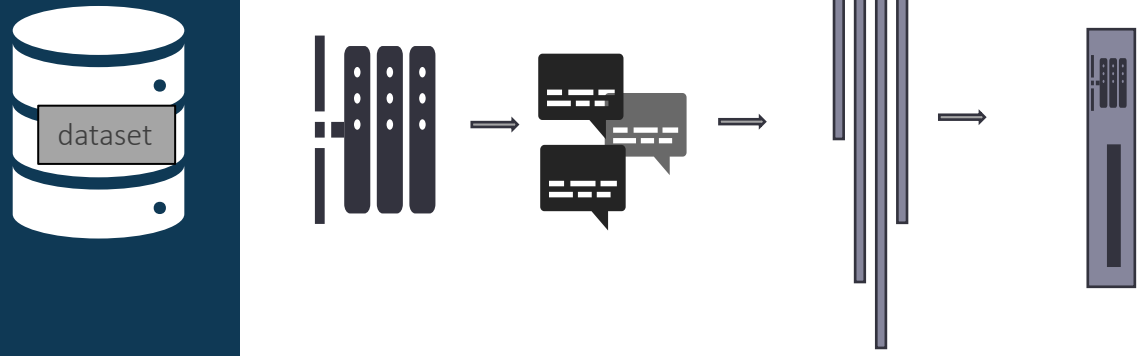


Clean



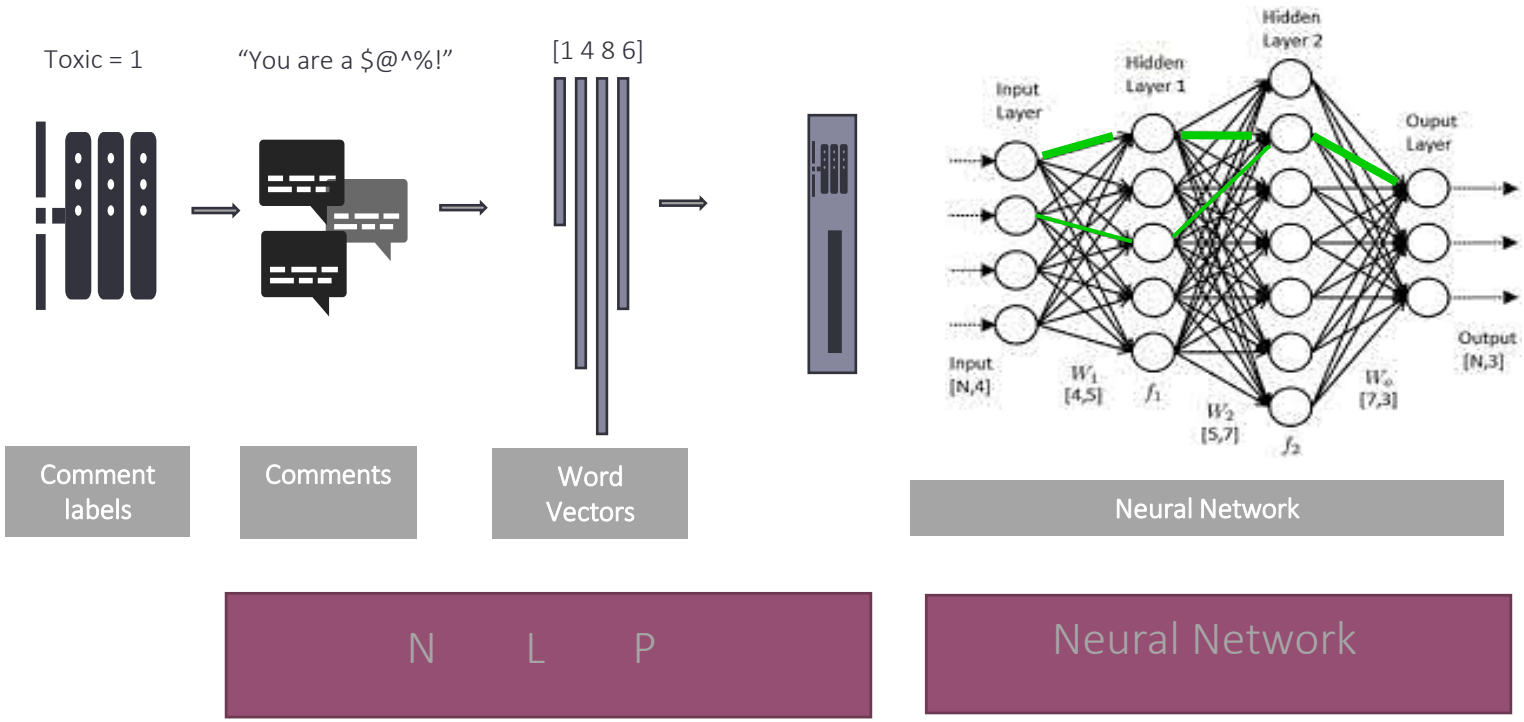
Are toxic comments longer than non-toxic?

Nope.....
They are about the same.



Methodology:

Deep Learning using Natural Language Processing and Neural Networks



Natural Language Processing (NLP):

- "Tokenizes" text
- Uses resulting tokens to create **vectors** as data

+

Neural Networks:

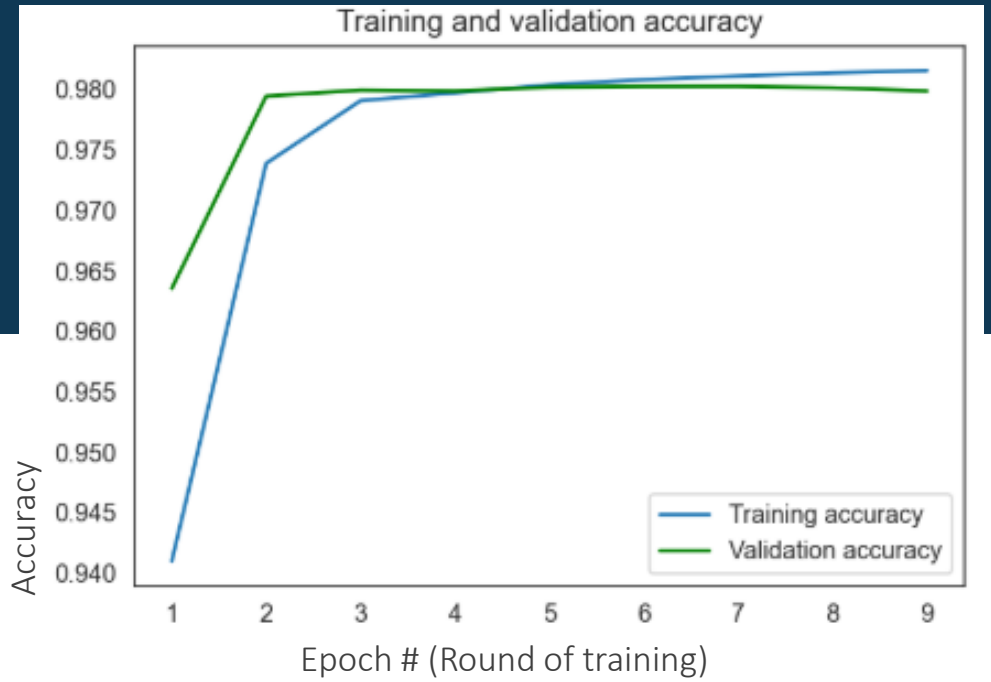
- multilayer perceptron
  - Neurons
  - Synapses
  - Input and output
- Artificial Neural Networks
  - Updating '**weights**' as it learns patterns

Results:

98% Accuracy Rate

99- 100% of the time identified clean text

Better at predicting classes where there is more data… a lot better.



Category	Rate of Accurately Predicting Cases or 'Recall"
Toxic	70
Severe toxic	32
Obscene	75
Threat	not enough samples
Insult	60
Identity Hate	not enough samples

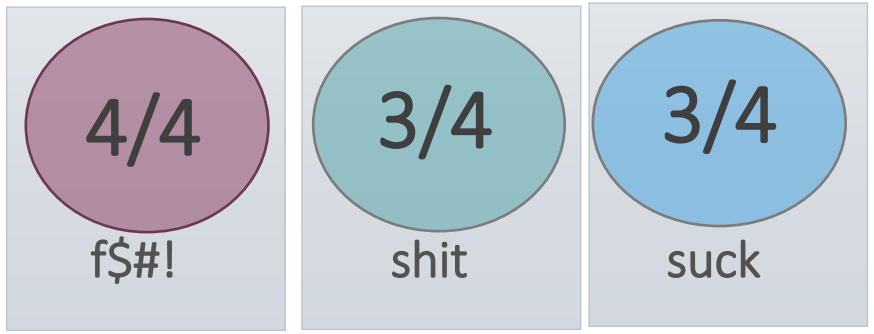
Findings:

Patterns Emerged

Top 5 Words Per Category:

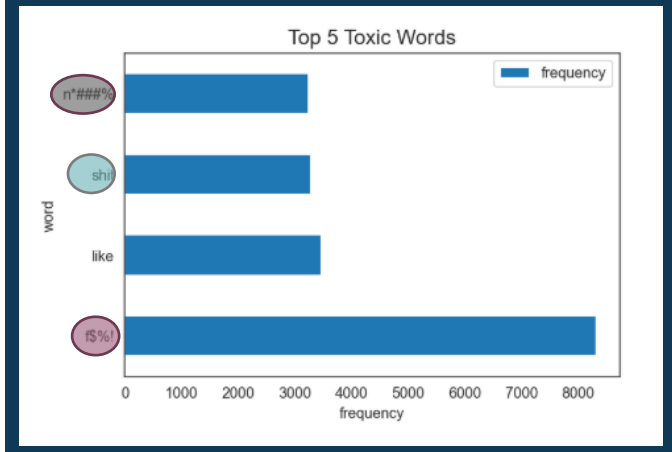
The same words appeared in 4 categories:

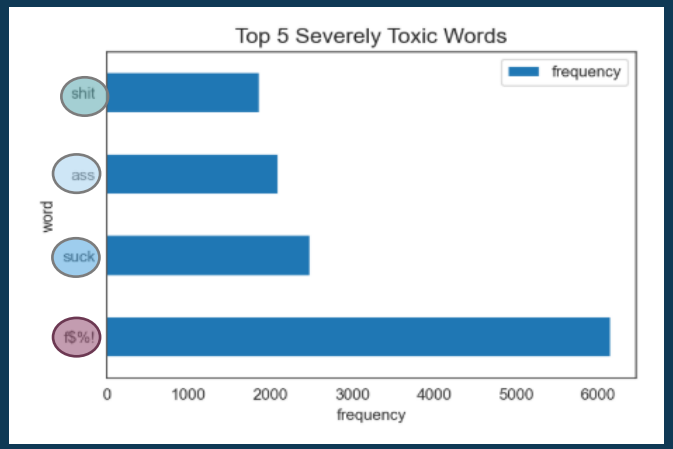
Toxic, Severely Toxic, Insulting & Obscene

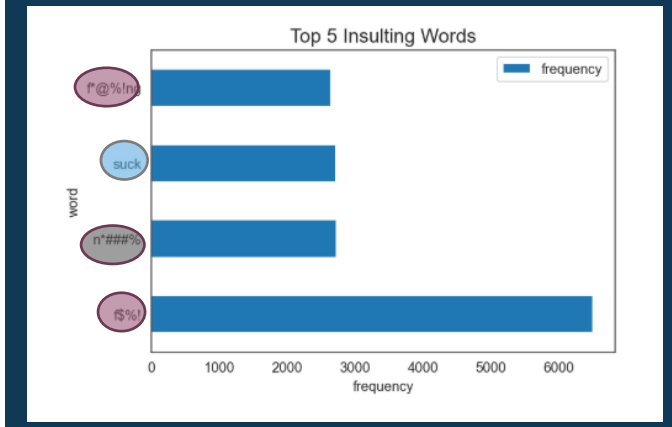


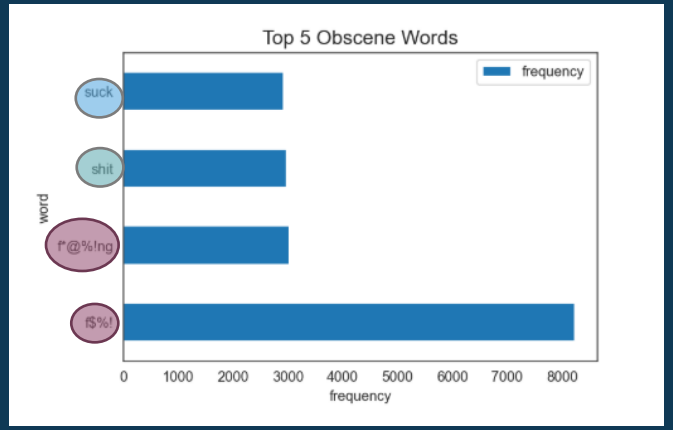
Two categories had more distinct top 5 words:

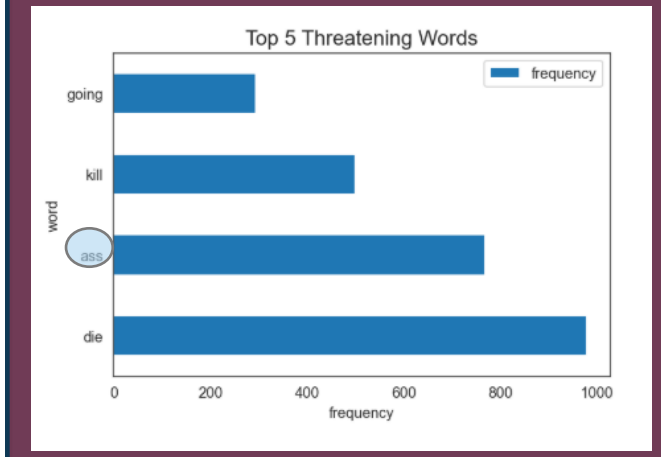
Threatening & Identity-Based Hate

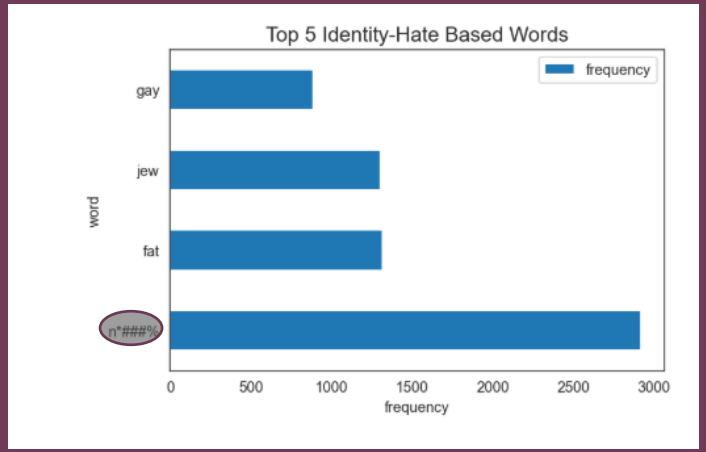




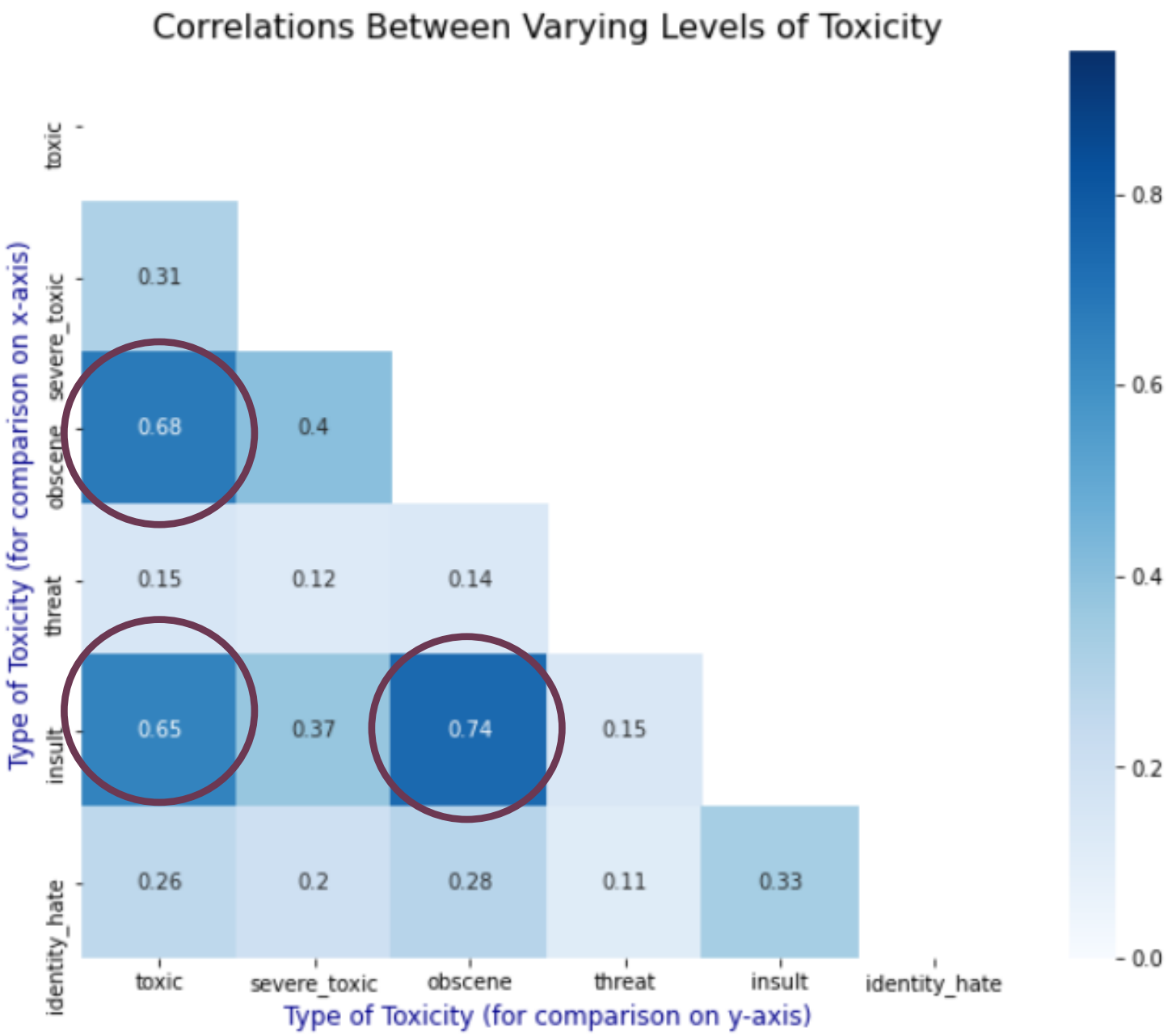








Correlations:



Recommendations

So now what?:

Develop metrics and actionable plans for varying levels of toxicity

Use model to identify varying levels of toxicity to **promote brand loyalty** by aligning with ideals of free speech AND creating a safe culture where true threats and hate aren't tolerated

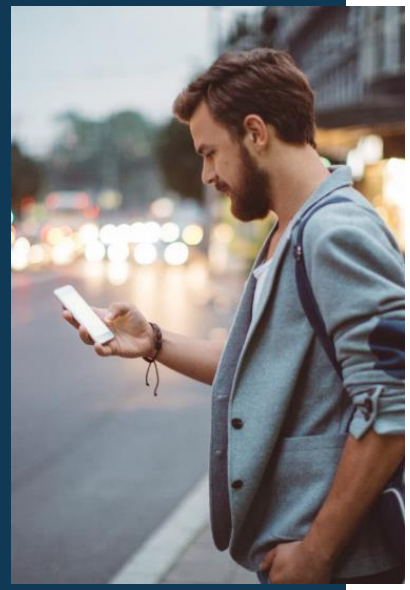
Research **and implement best practices** so that everyone feels comfortable sharing thoughts.

Invest in future work to further develop existing models to specifically identify severe forms, **targeting threats and identity-based hate specifically**

Future Work:

THREATS &

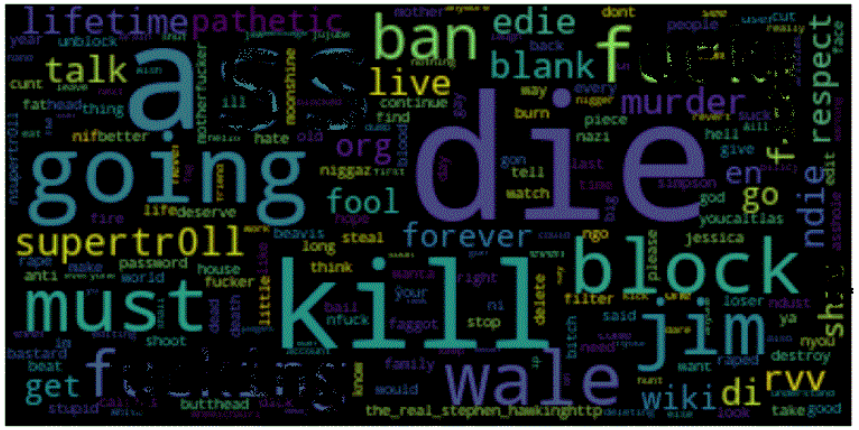
SENTIMENT ANALYSIS



Thankfully the most severe types of toxic comments are less frequent. In some situations, speech can constitute a crime, such as in the case of criminal threats.

Hopefully, collective work can help everyone express themselves in more meaningful ways.

- Collect more data around these more severe types of toxic comments to improve recognition.
- Make threatening content main target



- Comparative analysis on sentence sentiment rather than word.

Thank you

Appendix:

LSTM