

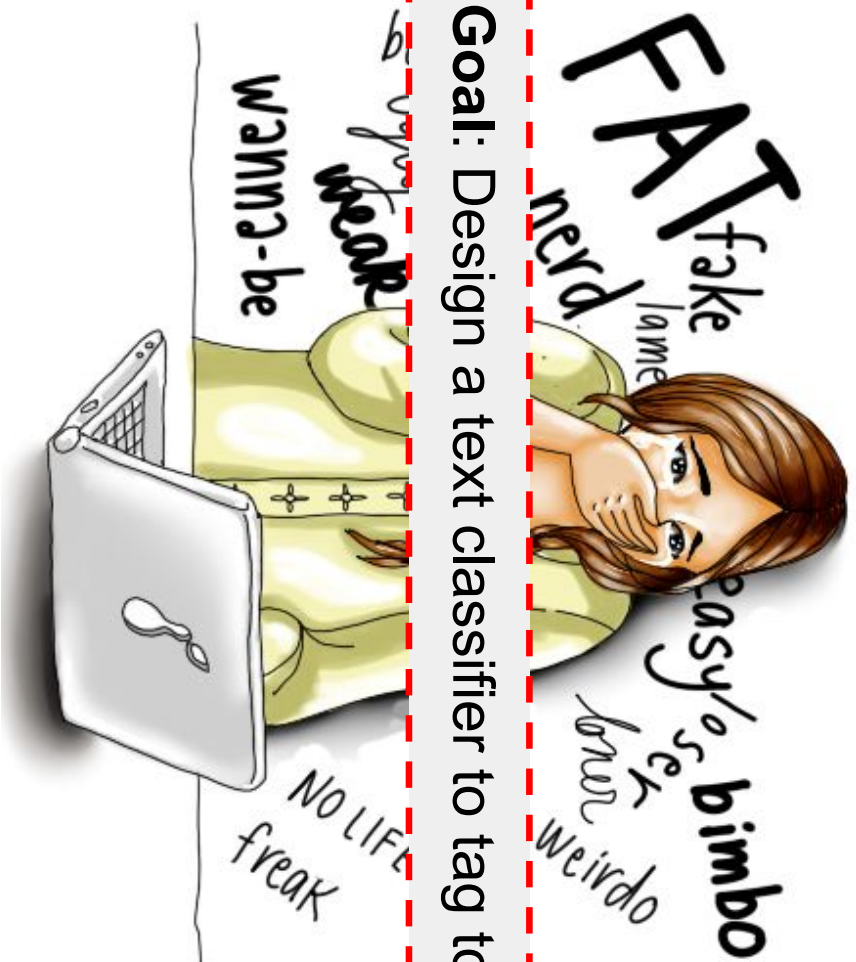
# Toxic Content Classification via Unsupervised Learning



Kundan Chaudhary

02-28-2020

How do we get rid of toxic contents on the internet?

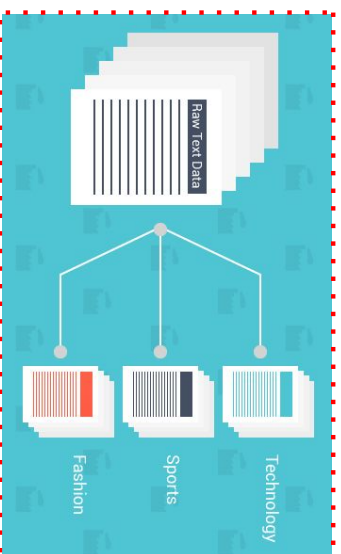


Goal: Design a text classifier to tag toxic contents on social media platform

50%

90%

# Getting from raw texts to coherent themes



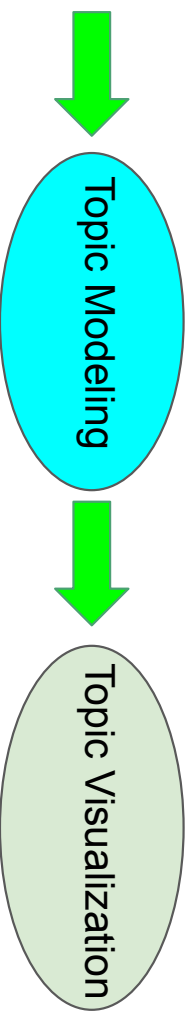
## Assumptions:

- Used nouns and adjectives only
- Ignored:
  - Abbreviations
  - Words w/ length less than 3
  - Misspelled words
  - n-grams

## Data Preprocessing

lower case → nouns/adjectives → lemmatization (stemming)

## Metrics: Coherence Score



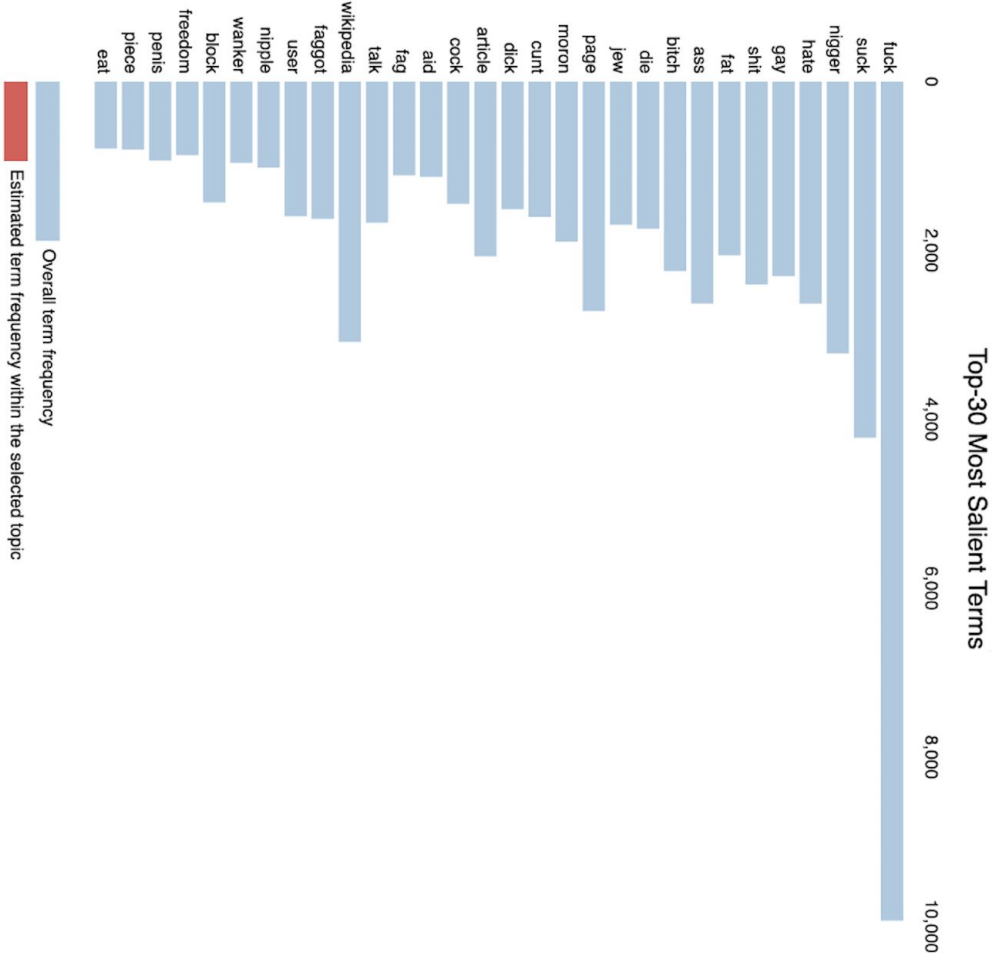
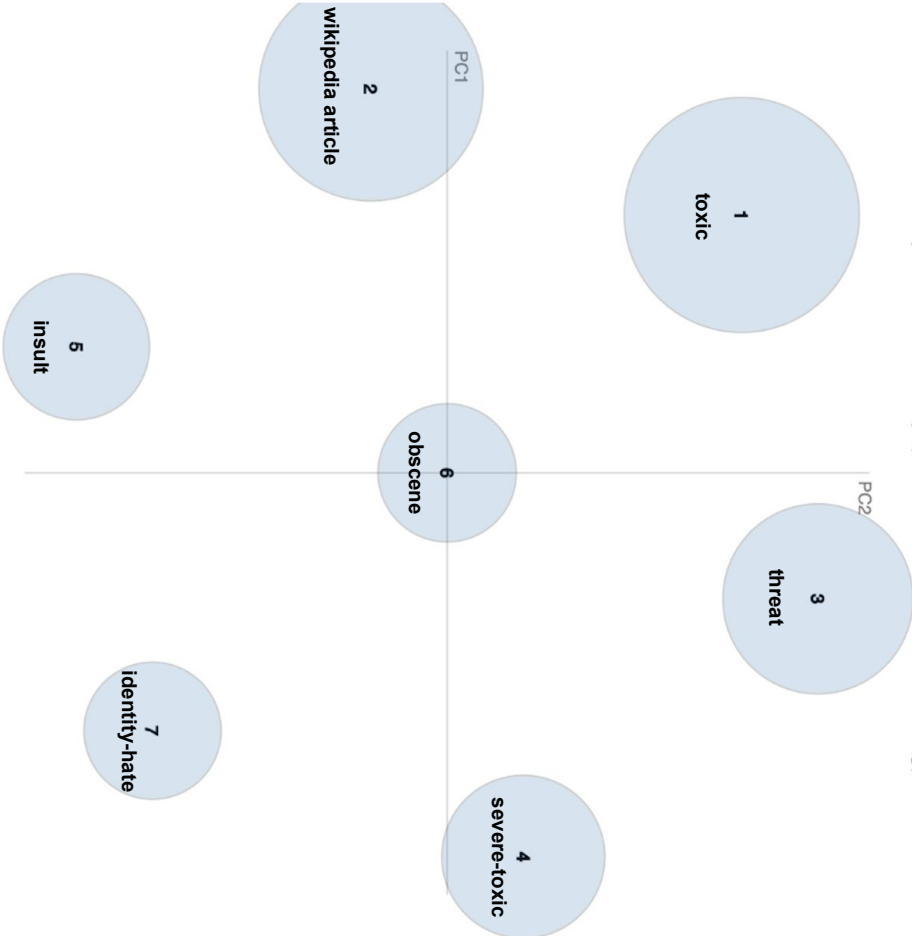
spacy

gensim

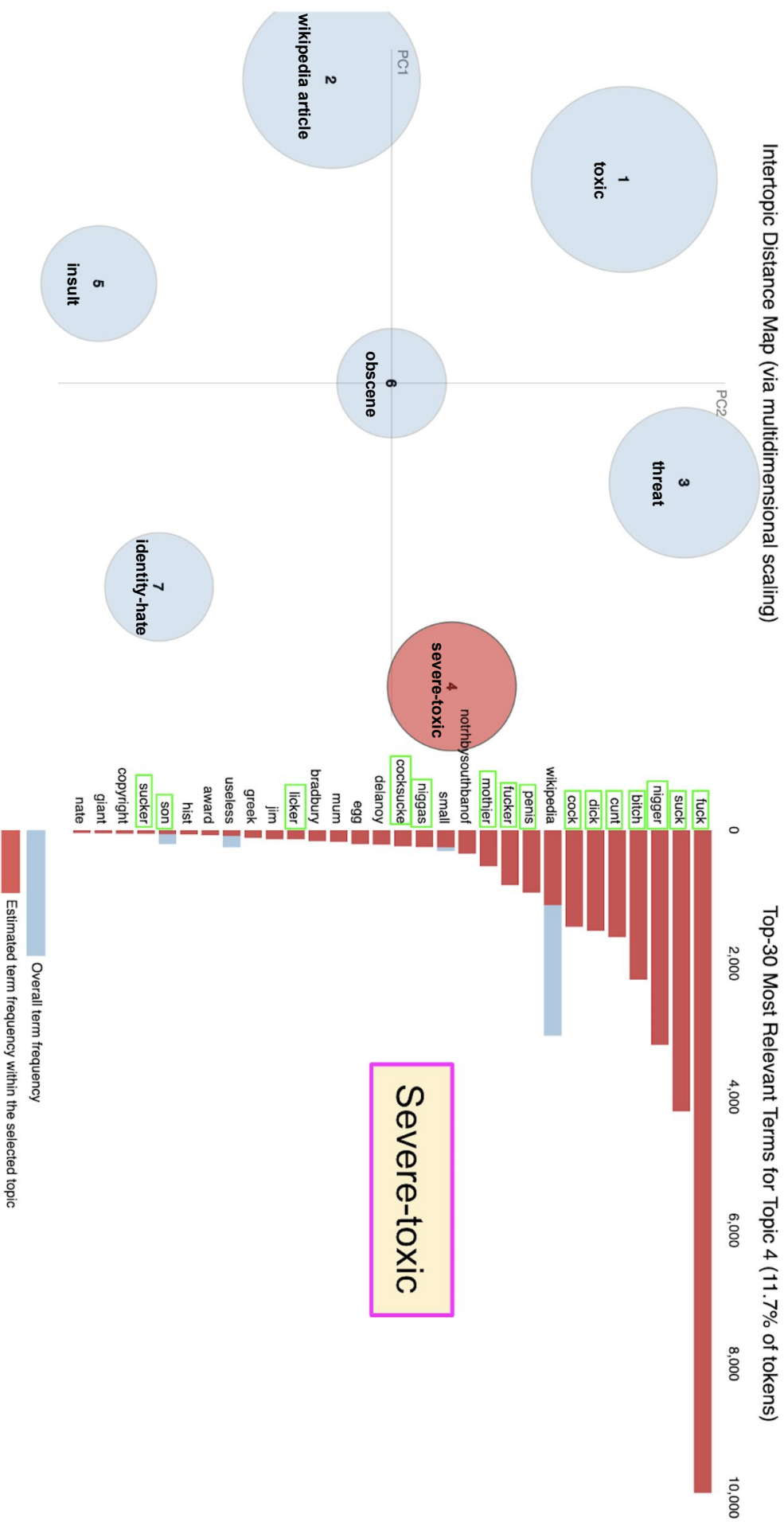
matplotlib

# Topics from LDA model

Intertopic Distance Map (via multidimensional scaling)

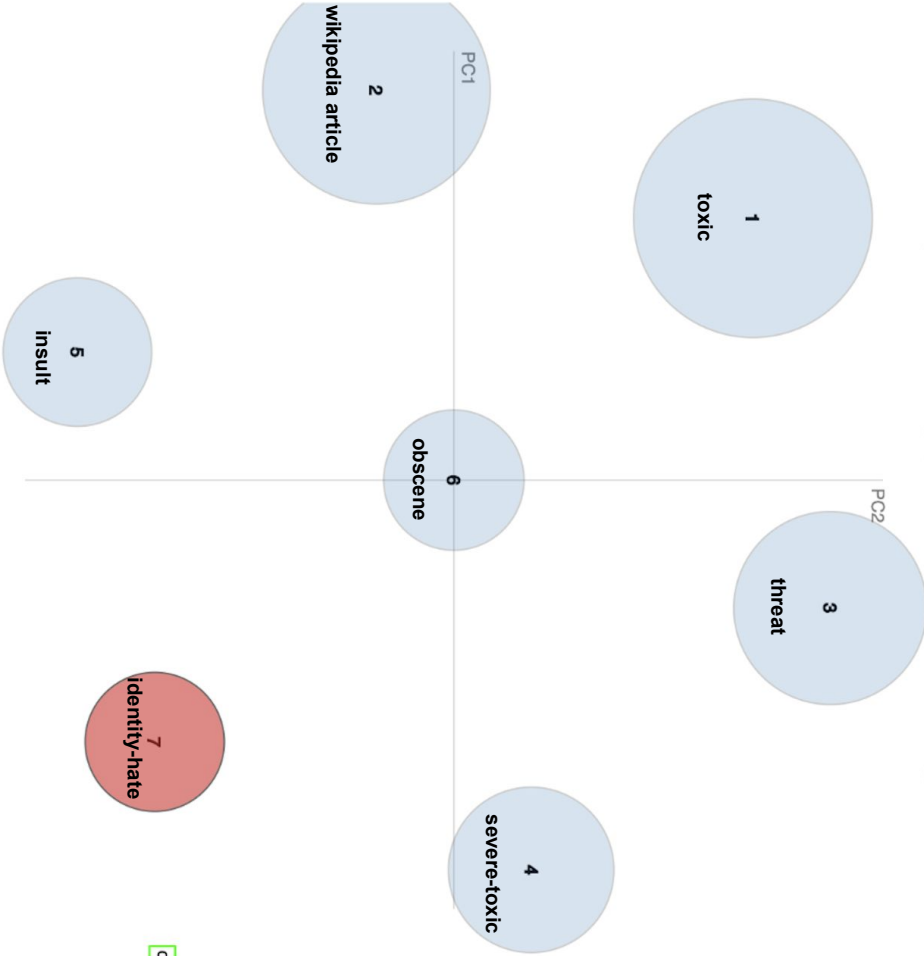


# Example Theme: 1 (Severe-toxic)

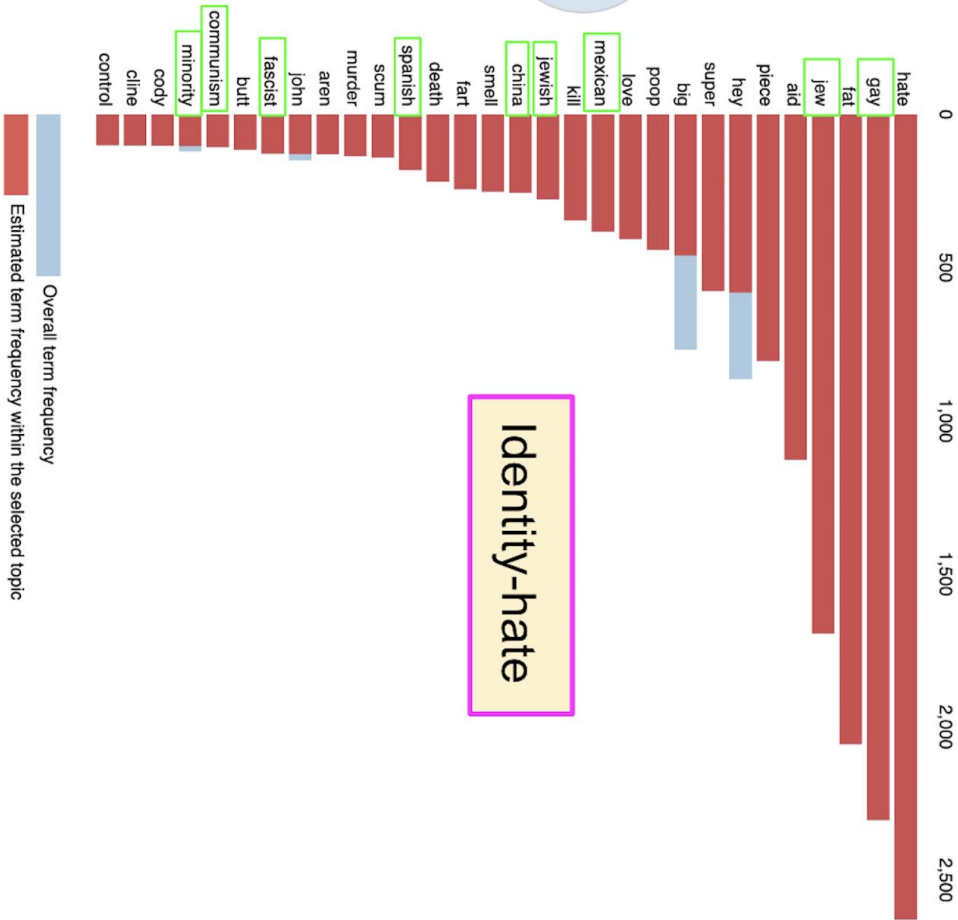


# Example Theme: 2 (Identity-hate)

Intertopic Distance Map (via multidimensional scaling)



Top-30 Most Relevant Terms for Topic 7 (8.3% of tokens)



Identity-hate

Overall term frequency

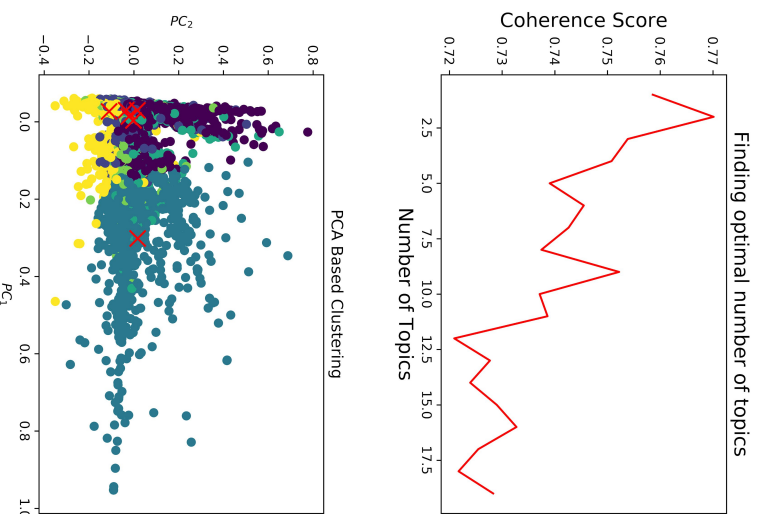
Estimated term frequency within the selected topic

# Summary & Future Outlook

Summary      unsupervised → supervised learning

- Online chat platforms are highly toxic

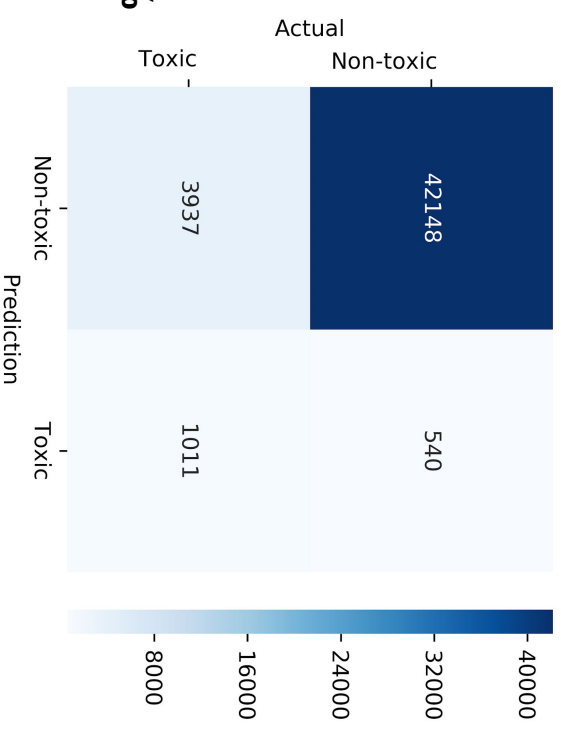
Future Outlook



Text Features

Logistic Regression

Accuracy: 91%  
Precision: 65%



- Incorporate abbreviations and misspelled words
- Explore further stop-words to be added
- Use n-grams