Toxic Content Classification via Unsupervised Learning

Kundan Chaudhary

02-28-2020

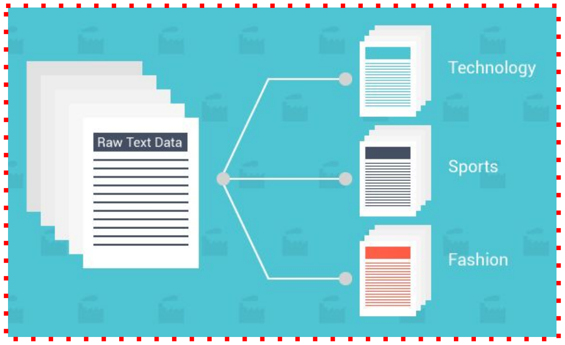
How do we get rid of toxic contents on the internet?

53 %

30 %

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

Getting from raw texts to coherent themes

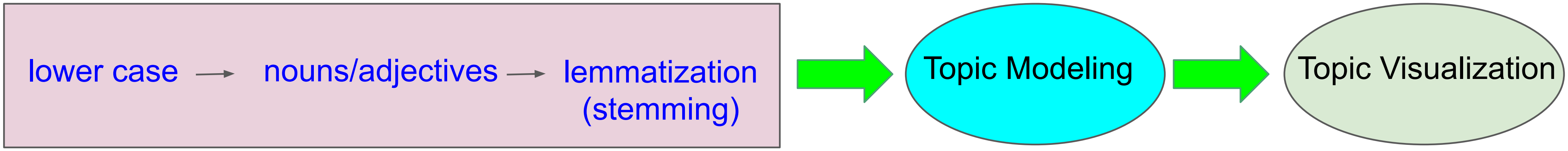


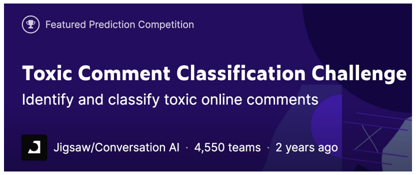
Assumptions:

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

Data Preprocessing

Metrics: Coherence Score




Featured Prediction Competition

Toxic Comment Classification Challenge

Identify and classify toxic online comments

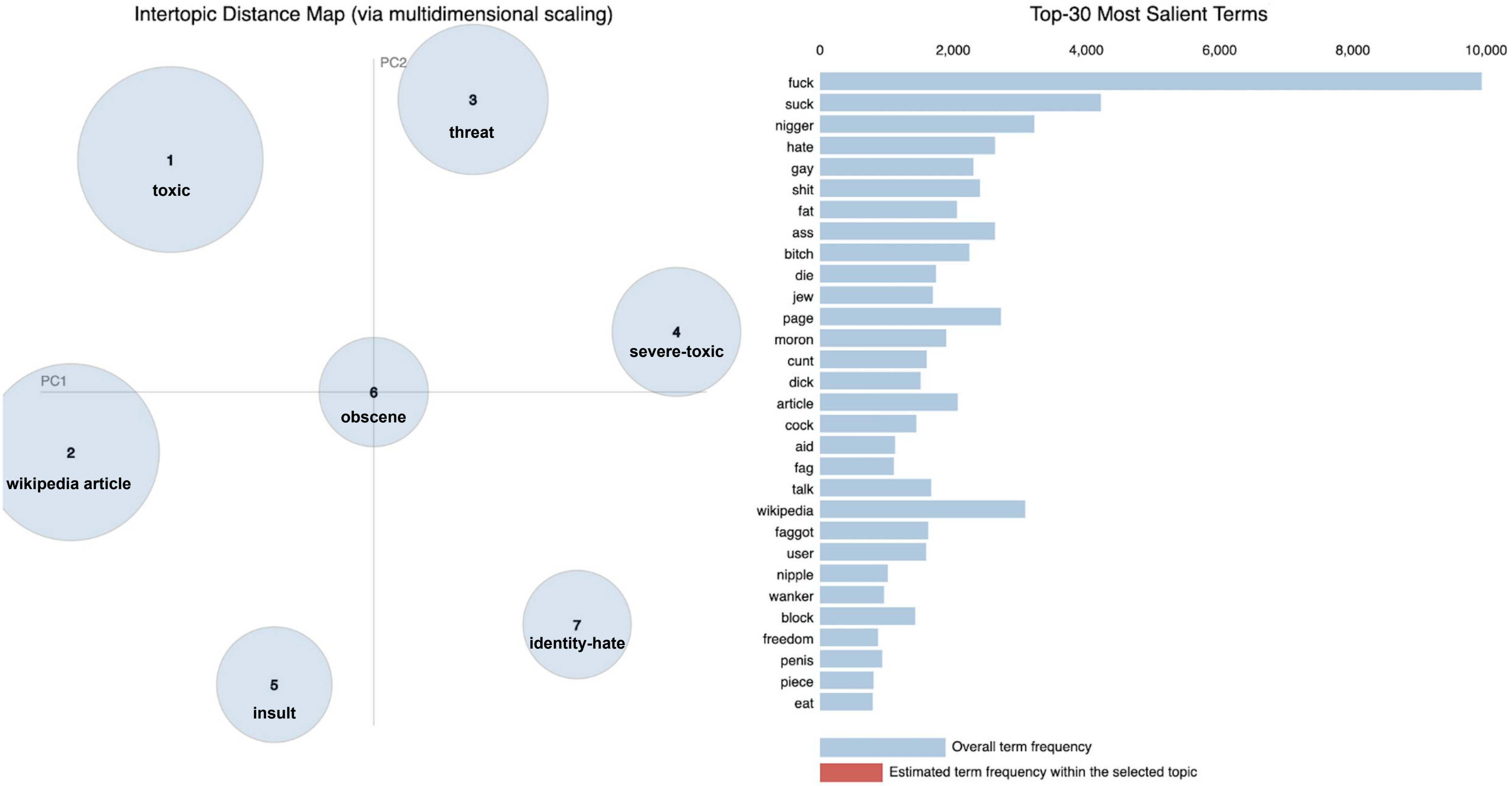
Jigsaw/Conversation AI · 4,550 teams · 2 years ago

spaCy

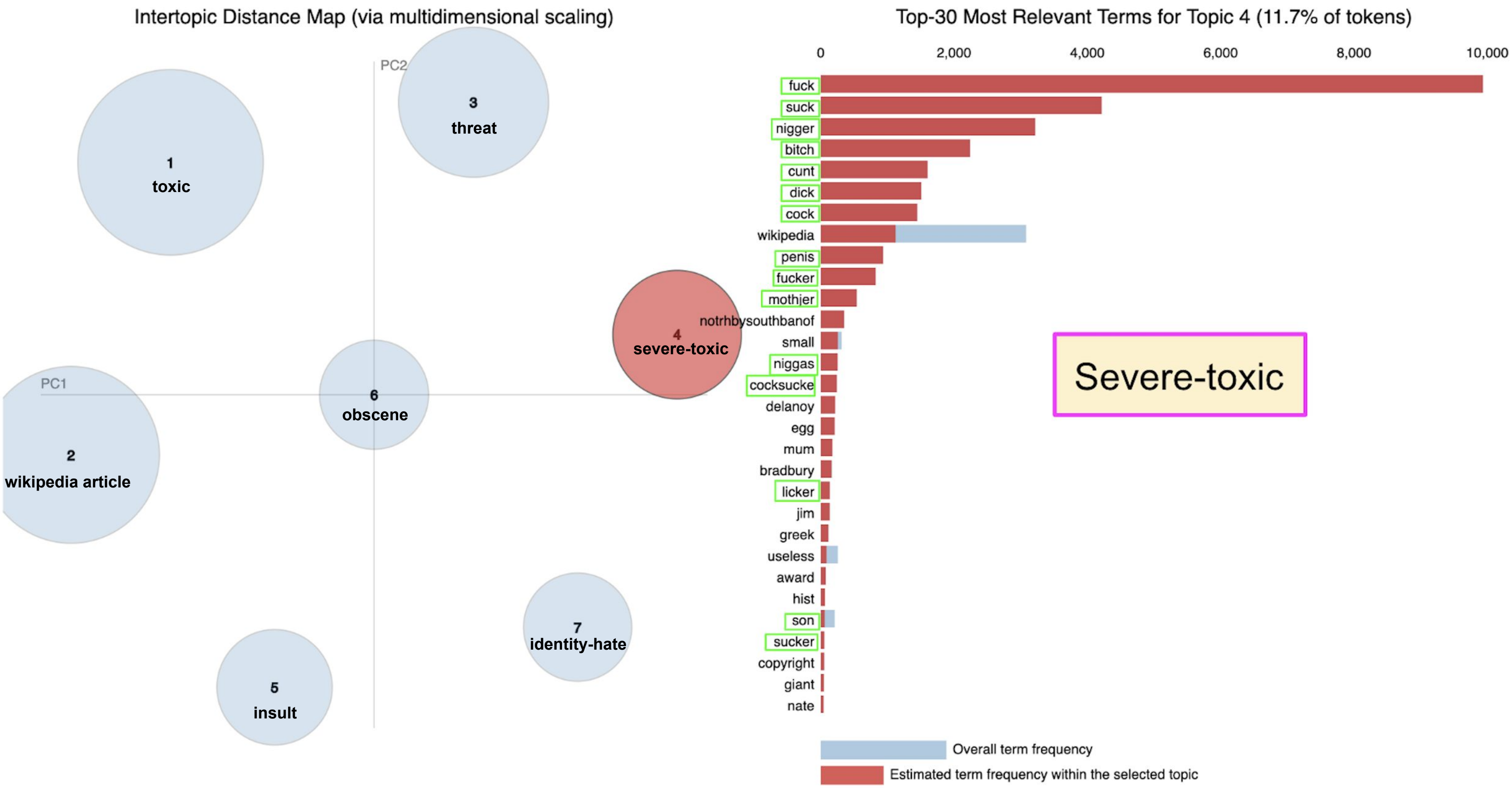
gensim

matplotlib

Topics from LDA model



Example Theme: 1 (Severe-toxic)

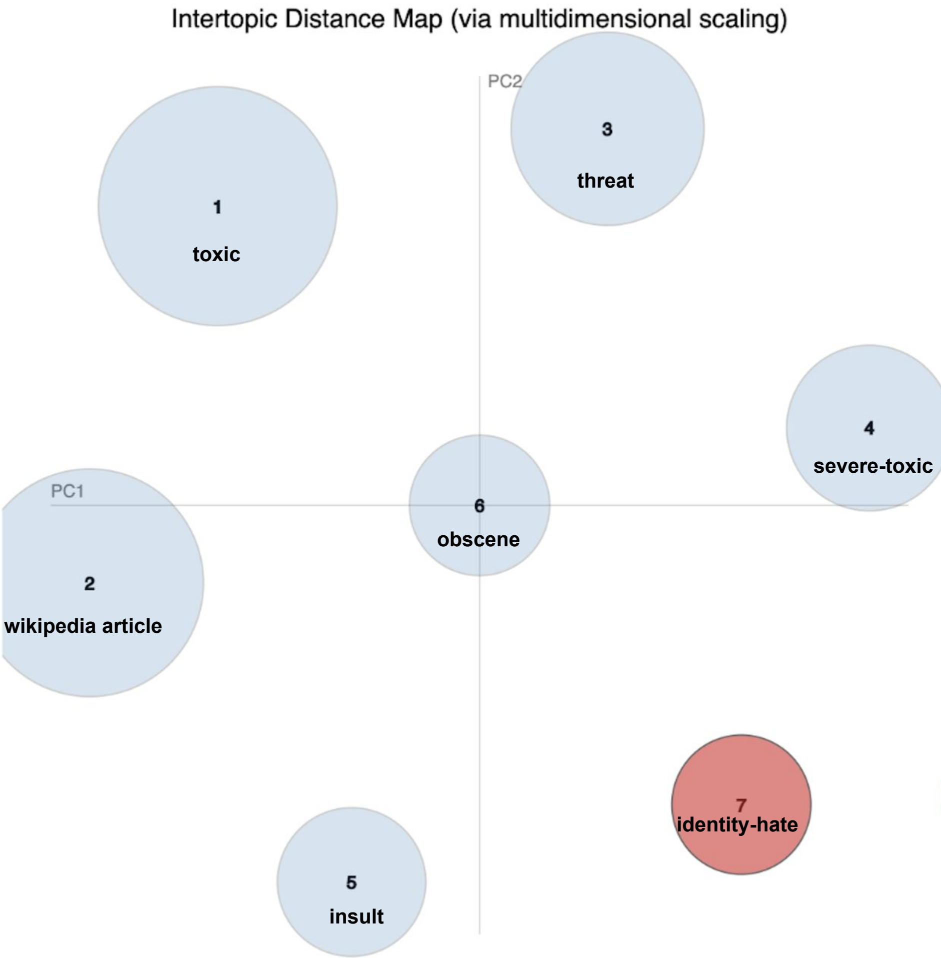

Intertopic Distance Map (via multidimensional scaling)

Top-30 Most Relevant Terms for Topic 4 (11.7% of tokens)

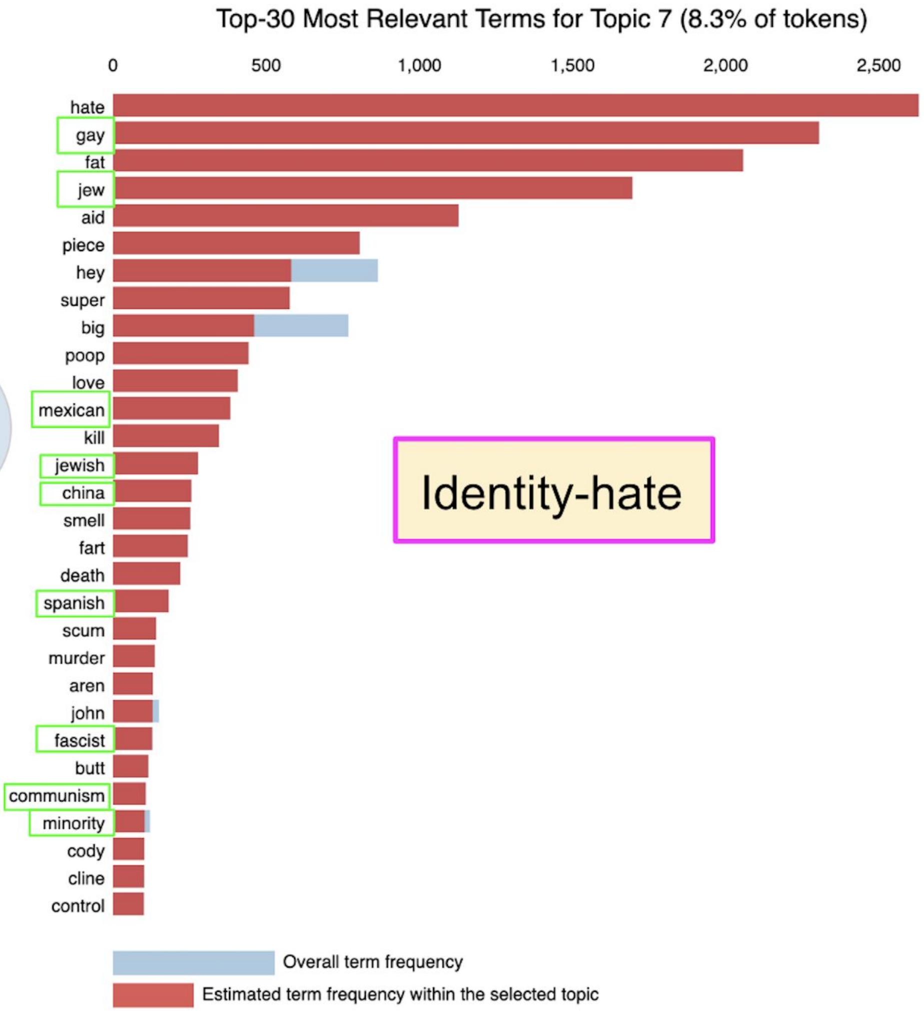
Severe-toxic

Example Theme: 2 (Identity-hate)

Intertopic Distance Map (via multidimensional scaling)



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



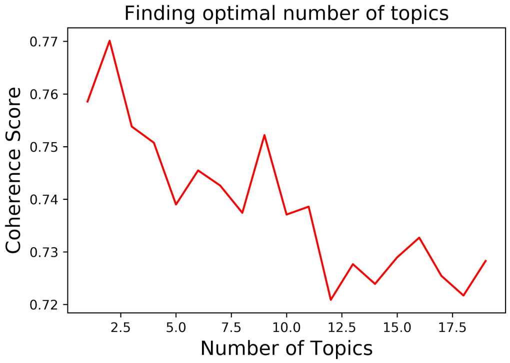
Identity-hate

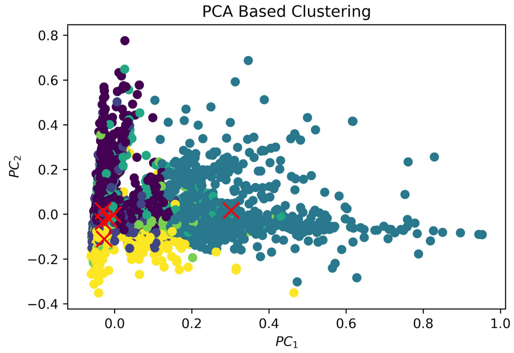
Summary & Future Outlook

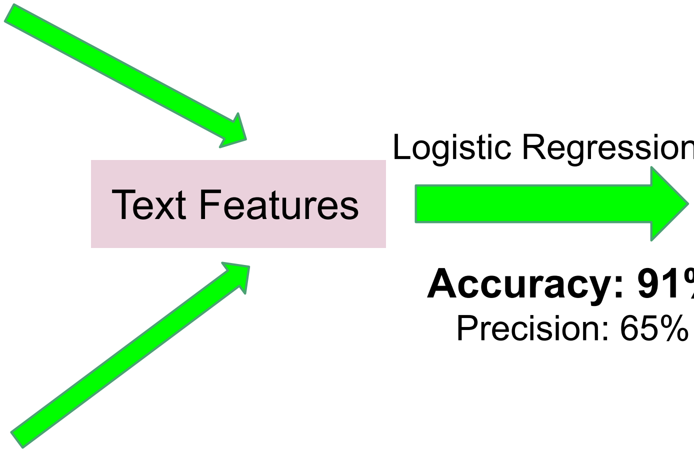
unsupervised ➤ supervised learning

Summary

- Online chat platforms are highly toxic







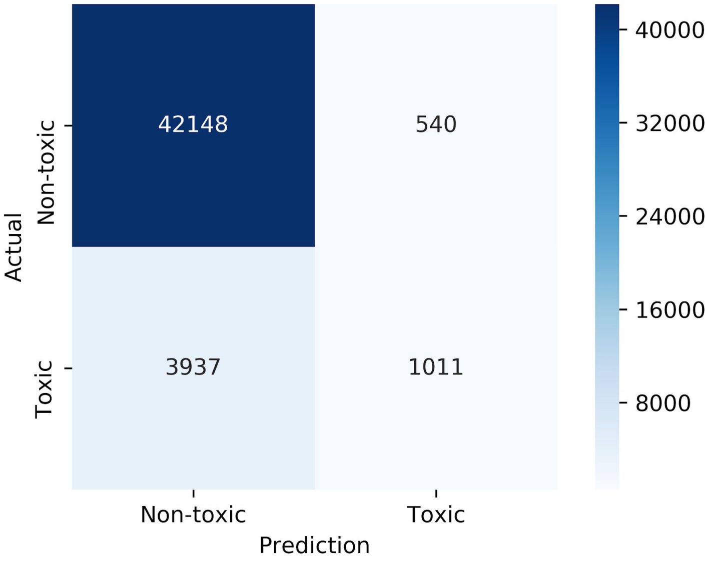
Text Features

Logistic Regression

Accuracy: 91%

Precision: 65%

Future Outlook



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