CATCHING

CRUDE COMMENTS

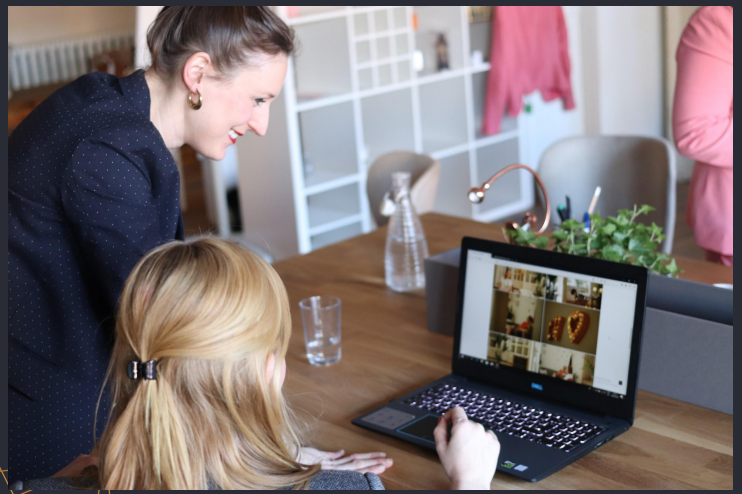
Identifying Toxic Internet Comments Using Natural Language Processing

Please help
All my improvements and corrections are being vandalized by some users in the Lion King pages. For example, I corrected that Kiara is Simab's heir and successor as stated in the movie. Then someone kept reverting it back so that Kiara is not heir. Can you help me?

I give up
Thanks for ruining the Lion King pages. Go ahead it fuck it up some more.

A TALE OF TWO COMMENTS

GOAL



Identifying toxic comments using machine learning is essential to creating safe internet forums at a large scale. Nuanced terminology and slang present a challenging NLP task.

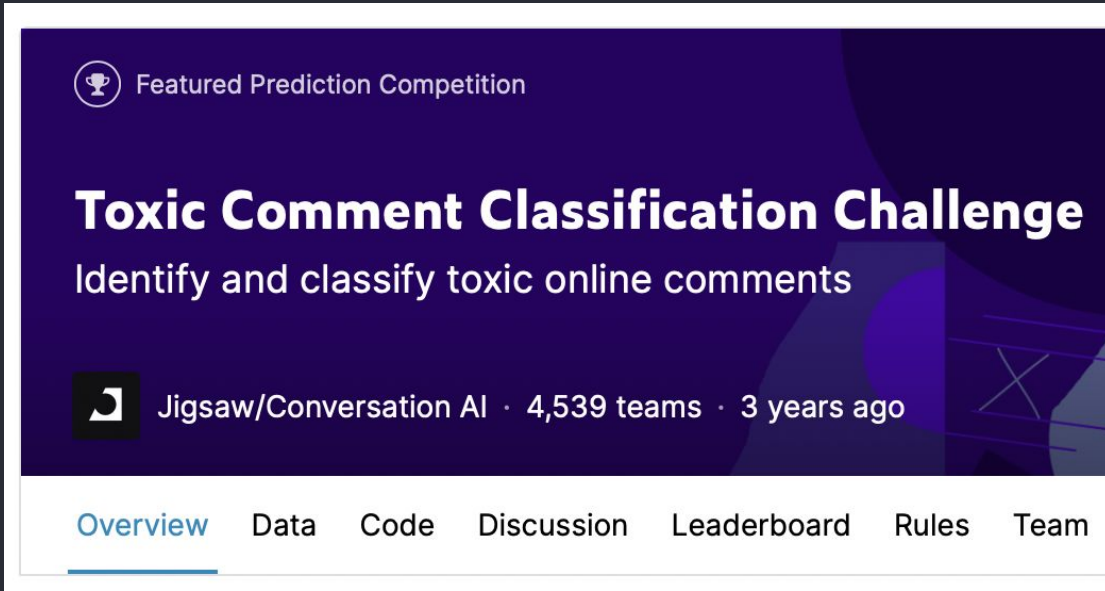
Goal: Correctly identify toxic comments using a combination of NLP, classification, and unsupervised learning.

DATA

Source
Jigsaw/Conversation AI
Kaggle competition

Size
150,000 Wikipedia
comments

Specific Features
Labeled for different toxicity
values (insult, threat, etc.)



Methodology

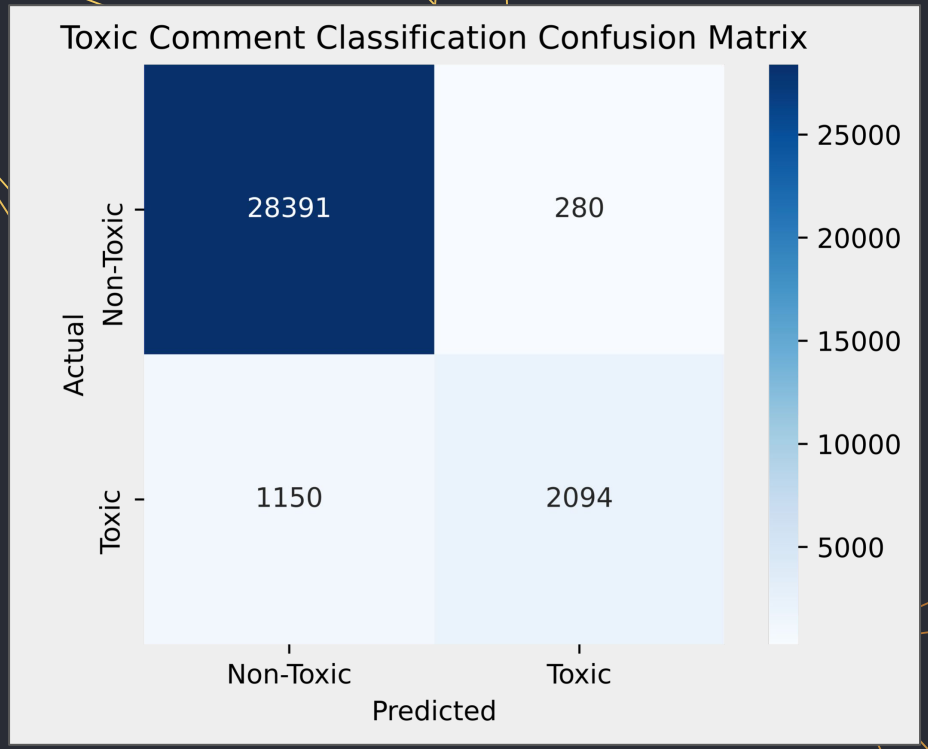
INITIAL EDA
Identify general patterns in dataset using pandas

PREPROCESSING
Clean, tokenize, and vectorize using spacy, sklearn, and more

DIMENSION REDUC.
Isolate key topics using SVD, NMF and LDA

CLUSTER/SUPERVISED
Classify comment toxicity using sklearn

TEXT GENERATOR
Comment generator using LSA and Markov Chains

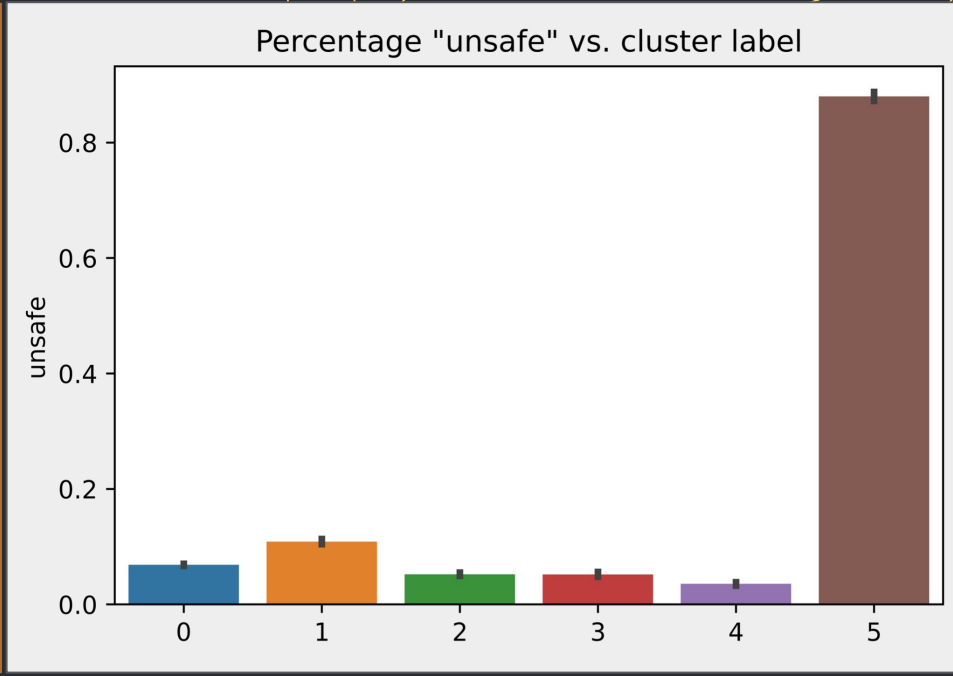


Classification

- Accuracy: .96
- Low recall
- Many missed comments:
  - Very short
  - Contain spelling errors
  - Have significant punctuation

Topic Modeling

- Cluster labels very successful at identifying toxic comments
- 7 out of 10 terms of topic 5 are curses or slurs
- Topic 1 includes “vandal”, “block” and “stop”



COMMENT CLASSIFICATION

Toxic Correct

Your comment on your edit proves you are ignorant of Islam.

False Negative

I don't care what you say here. I don't believe one sentence anymore.

False Positive

The Alpha version had multiplayer. The Beta version did not.

Non-Toxic Correct

Oh ya, I have one last simple request. Could you delete my account. I don't want to be in anyway associated with Wikipedia.

'Lambs wikipedia dickhead dickhead so hard ass so u again.'

COMMENT GENERATOR

'Fall into the last chance in this motherfucker shit fuck.'

FUCK YOU WIKIPEIA
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK
FUCK

Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady
Ummmmmm Sex Lady

Guess which is real!

'Supermarioman be back dirty wanker and no life lose interest.'

what a sad, lonely life you must have where you get a boner by playing Master of Wikipedia, deciding whos edits stay and whos get reverted.

RESULTS

- Nuanced, short text is difficult to catch
- Slurs and curse words make up majority of classification
- Spam can throw off filter

NEXT STEPS

PREPROCESSING

Capture ngram characters, punctuation, etc.

OUTSIDE DATA SETS

Better inform models using other hate speech/toxic comment data sets

MODELING

Additional model tuning and ensembling

THANKS!