

The background is a dark blue-grey color. It features several thin, gold-colored lines that form abstract, angular shapes. These lines radiate from the central text box, extending towards the corners of the frame. The lines vary in length and orientation, creating a dynamic, geometric pattern.

# CATCHING CRUDE COMMENTS

Identifying Toxic Internet Comments Using  
Natural Language Processing

Please help

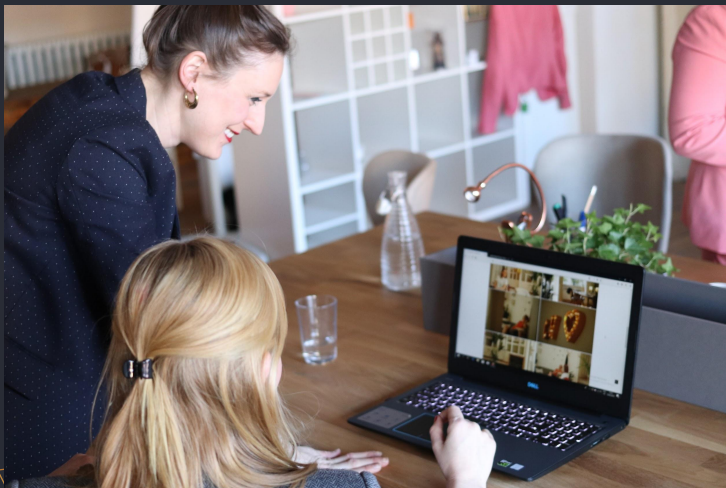
All my improvements and corrections are being vandalized by some users in the Lion King pages. For example, I corrected that Kiara is Simba's heir and successor as stated in the movie. Then someone kept reverting it back so that Kiara is not heir. Can you help me?

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

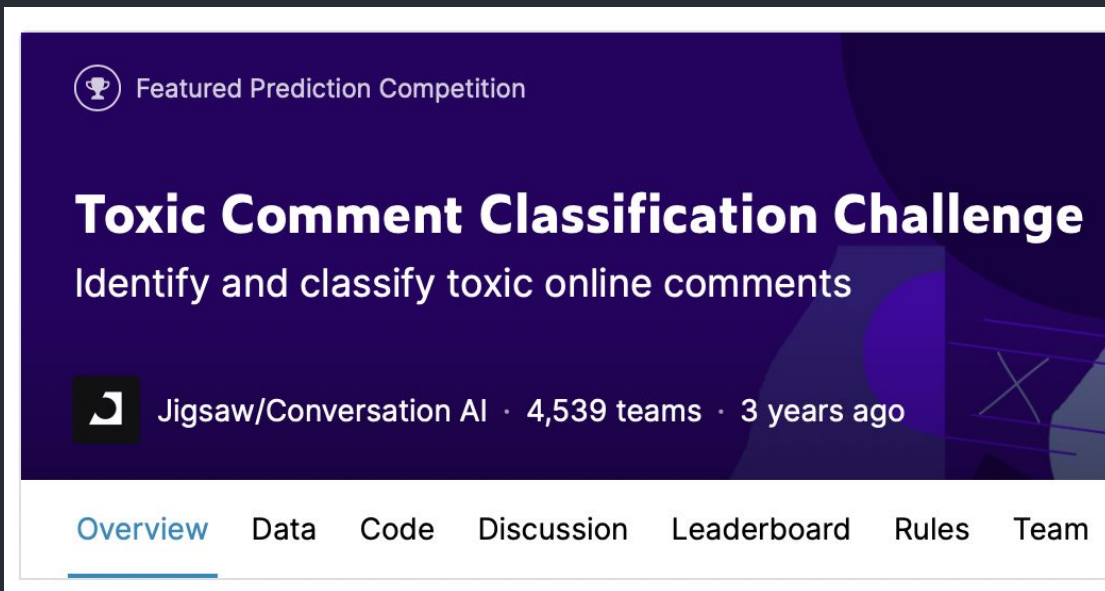
Jigsaw/Conversation AI  
Kaggle competition

## Size

150,000 Wikipedia  
comments

## Specific Features

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




The screenshot shows the Kaggle competition page for the 'Toxic Comment Classification Challenge'. At the top, it is marked as a 'Featured Prediction Competition' with a trophy icon. The title 'Toxic Comment Classification Challenge' is prominently displayed in white on a dark purple background, followed by the subtitle 'Identify and classify toxic online comments'. Below this, the Jigsaw logo is shown next to the text 'Jigsaw/Conversation AI · 4,539 teams · 3 years ago'. At the bottom, a navigation bar includes links for 'Overview', 'Data', 'Code', 'Discussion', 'Leaderboard', 'Rules', and 'Team', with 'Overview' being the active link.

Featured Prediction Competition

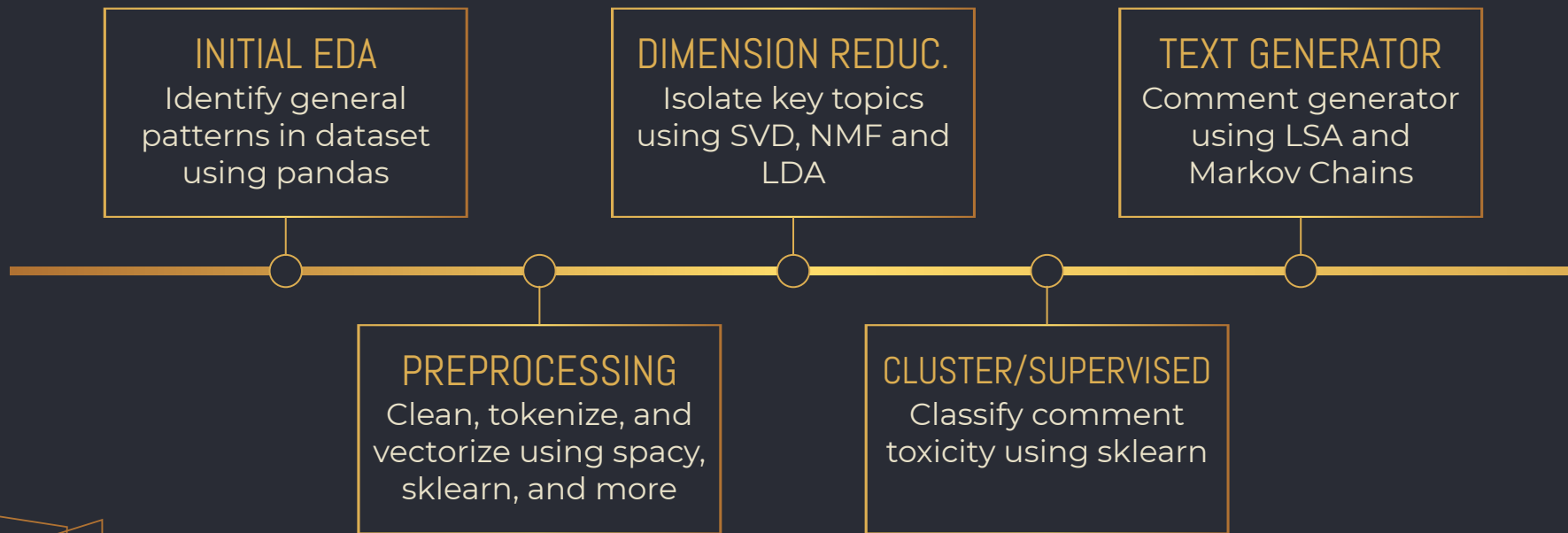
## Toxic Comment Classification Challenge

Identify and classify toxic online comments

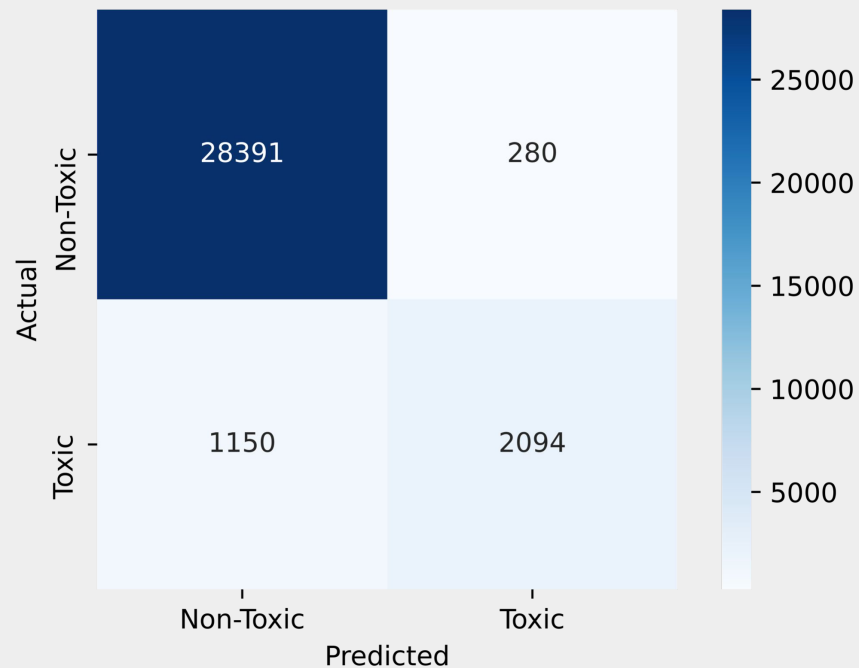
 Jigsaw/Conversation AI · 4,539 teams · 3 years ago

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#)

# Methodology



Toxic Comment Classification Confusion Matrix



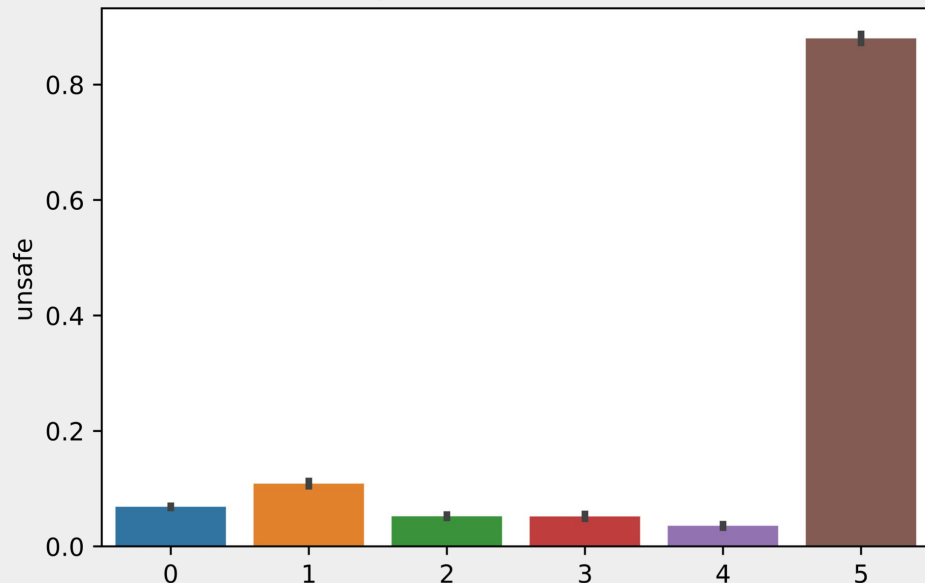
## Classification

- 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”

Percentage "unsafe" vs. cluster label





# 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.'

[illegible]

# Guess which is real!

[illegible]

'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

The background is a dark navy blue. It features several thin, gold-colored lines that form abstract geometric shapes, including triangles and polygons, scattered across the frame. A prominent gold-colored rectangular border is positioned around the central text. The word "THANKS!" is written in a large, bold, gold-colored, sans-serif typeface, centered within the rectangular border.

THANKS!