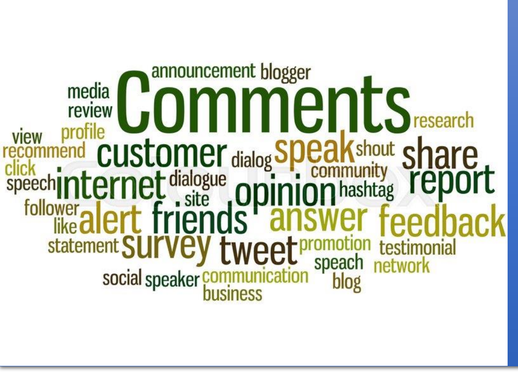


Toxic Comment Classifier

Mathews P. Mani (MSA)
Mayur Mahajan (MSA)
Nirmal Venkatachalam (MSA)
Sinchan Mukherjee (MSA)

Summary

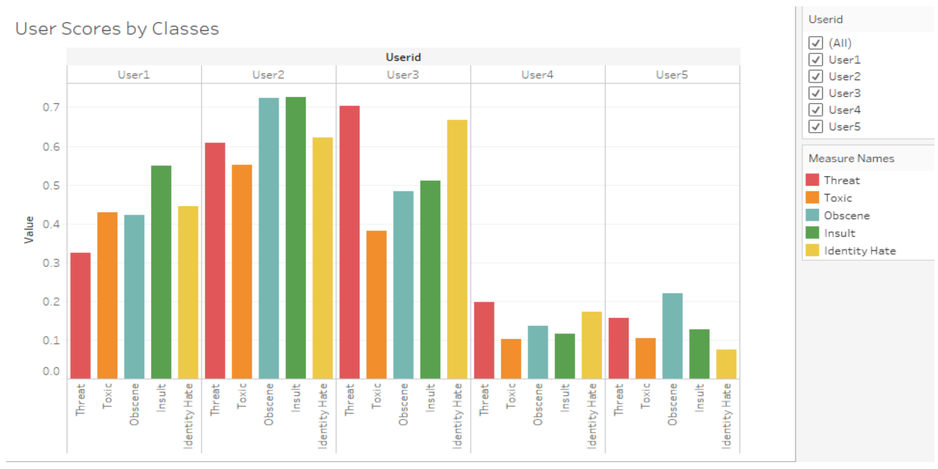
Effectively handling toxic comments is important to improve the quality of online discussions. We address this problem by classifying comments across various labels of toxicity. An ensemble model was created using three different Neural network based classifiers. AUC of 97% was achieved by using train data augmentation, pre-trained vectors and stacking multiple models. Model was subsequently used to create prototypes for scoring comments and visualizations for better moderation

What's new in our approach?

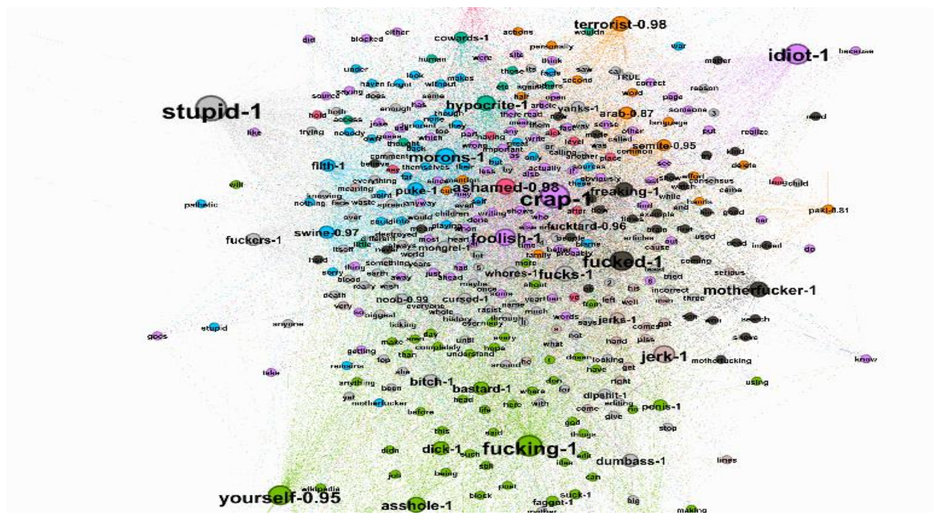
- Multi label toxic scoring (current approaches only do binary scoring - toxic vs non-toxic)
- Use of extended dataset (by translating to other languages - German, Spanish, French and then back to English) to make the model robust to perturbations
- Use of pre-trained word embeddings to make the model effective at classifying unseen words
- Stacking different neural networks based classifiers

Why is it important?

- 66% (110 million) of adult internet users have seen someone being harassed & 40% have personally experienced online harassment
  With our model, we will be able to
- Improve overall experience of all online users who face toxic comments on social media platforms
- Save time for moderators in filtering out comments
- Increase user engagement on online platforms

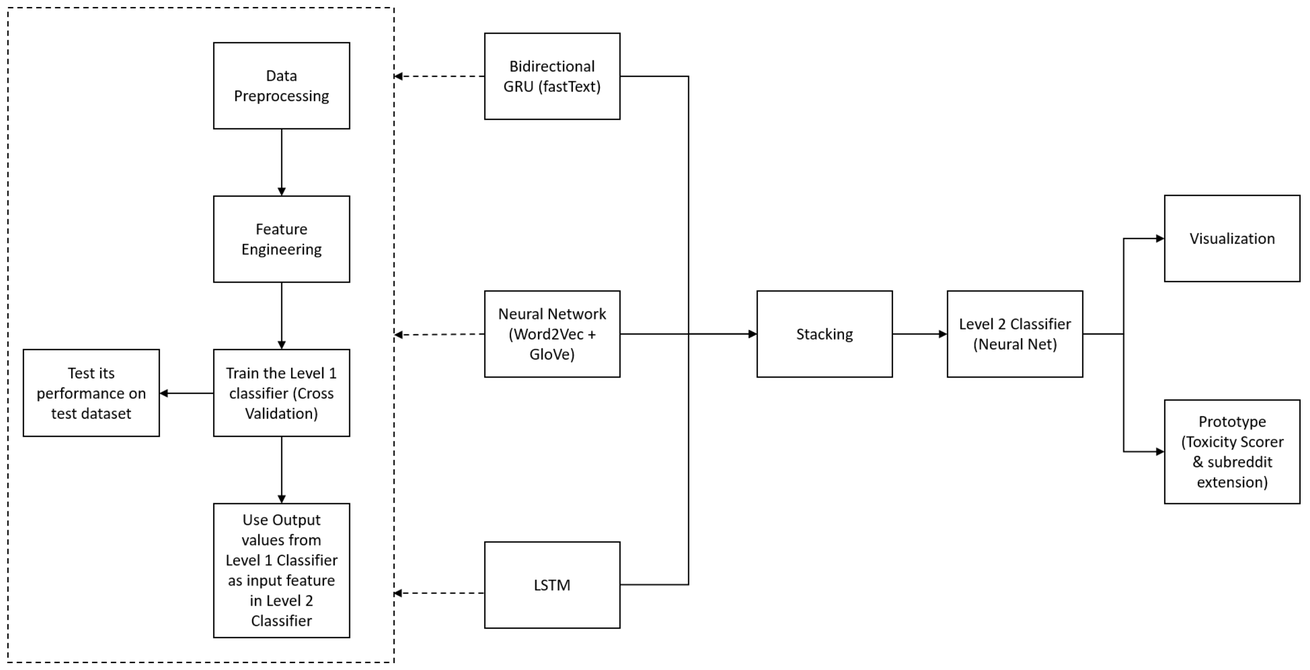


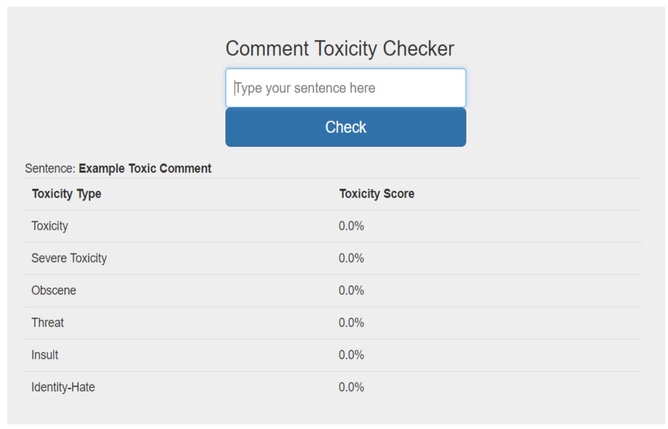
Interactive moderator dashboard

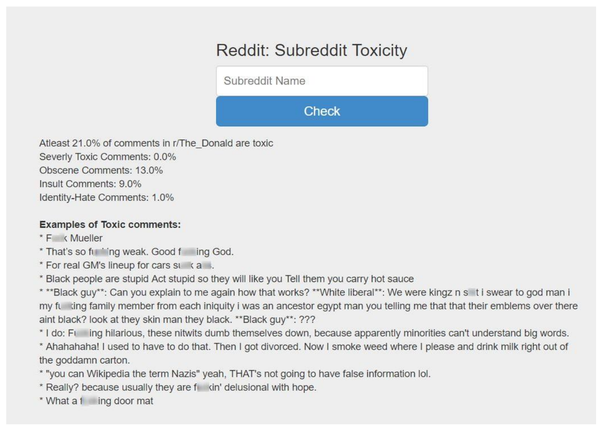


Force directed graph of scored similar toxic words

Project Overview







Prototype – Website for scoring comments & subreddits

Evaluation

- High AUC-ROC (97%) on train & test datasets
- High Accuracy (97%) on train & test datasets
- 0.5-1% performance improvement due to stacking
- Low response time for prediction in the final model

Current Approaches

- Basic blacklists, bag of words approach augmented with TF-IDF frequencies, SVMs, CNNs, and RNNs
- Use of word embeddings
- Perspective API (by Google & Jigsaw)
- Drawbacks of current approaches: computationally expensive; don't incorporate relationship between words & context; not robust to small perturbations