



Sinchan Mukherjee

(MSA)

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

Project Overview

-
- ```
graph TD; subgraph Training; DP[Data Preprocessing] --> FE[Feature Engineering]; FE --> T1[Train the Level 1 classifier (Cross Validation)]; T1 --> T2[Use Output values from Level 1 Classifier as input feature in Level 2 Classifier]; end; DP -.-> BGRU[Bidirectional GRU (fastText)]; FE -.-> NN[Neural Network (Word2Vec + GloVe)]; T1 -.-> LSTM[LSTM]; BGRU --> S[Stacking]; NN --> S; LSTM --> S; S --> L2C[Level 2 Classifier (Neural Net)]; L2C --> V[Visualization]; L2C --> P[Prototype (Toxicity Scorer & subreddit extension)];
```
- The flowchart illustrates the proposed toxicity detection system architecture. It is divided into two main sections: Training and Inference/Deployment.
- Training Phase (enclosed in a dashed box):**
- Data Preprocessing** leads to **Feature Engineering**.
  - Feature Engineering** leads to **Train the Level 1 classifier (Cross Validation)**.
  - Train the Level 1 classifier (Cross Validation)** leads to **Use Output values from Level 1 Classifier as input feature in Level 2 Classifier**.
  - Test its performance on test dataset** is a parallel step to the training process.
- Inference/Deployment Phase:**
- The **Use Output values from Level 1 Classifier as input feature in Level 2 Classifier** step feeds into the **Stacking** process.
  - The **Stacking** process receives input from three models: **Bidirectional GRU (fastText)**, **Neural Network (Word2Vec + GloVe)**, and **LSTM**.
  - The output of the **Stacking** process is fed into the **Level 2 Classifier (Neural Net)**.
  - The **Level 2 Classifier (Neural Net)** produces two outputs: **Visualization** and **Prototype (Toxicity Scorer & subreddit extension)**.



Reddit: Subreddit Toxicity

Subreddit Name

Check

Atleast 21.0% of comments in r/The\_Donald are toxic

Severly Toxic Comments: 0.0%

Obscene Comments: 13.0%

Insult Comments: 0.0%

Identity-Hate Comments: 1.0%

**Examples of Toxic comments:**

- \* F\*ck Mueller
- \* That's so f\*lling weak. Good f\*lling God.
- \* For real GM's lineup for sick ass!!.
- \* Black people are stupid! Act stupid so they will like you Tell them they carry hot sauce
- \* "Black gay?": Can you explain to me again how that works? "White liberal": We were kingz n s\*llt I swear to god man I'm f\*lling family member from each iniquity I was an ancestor egypt man you telling me that their embems over there are black' look at they skin man they black. "Black gay?": ???
- \* I don't f\*lling hilarious, these n\*llz dont theme themselves down, because apparently minorities can't understand big words.
- \* Ahaahahah! I used to have to do that. Then I got divorced. Now I smoke weed and please and drink milk right out of the goddamn carton.
- \* You can Wikipedia the term Nazis' yeah. THAT's not going to have false information lol.
- \* Really? because usually they are f\*lling delusional with hope.
- \* What a f\*lling rude mat

## Prototype – Website for scoring comments & subreddits

- 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

