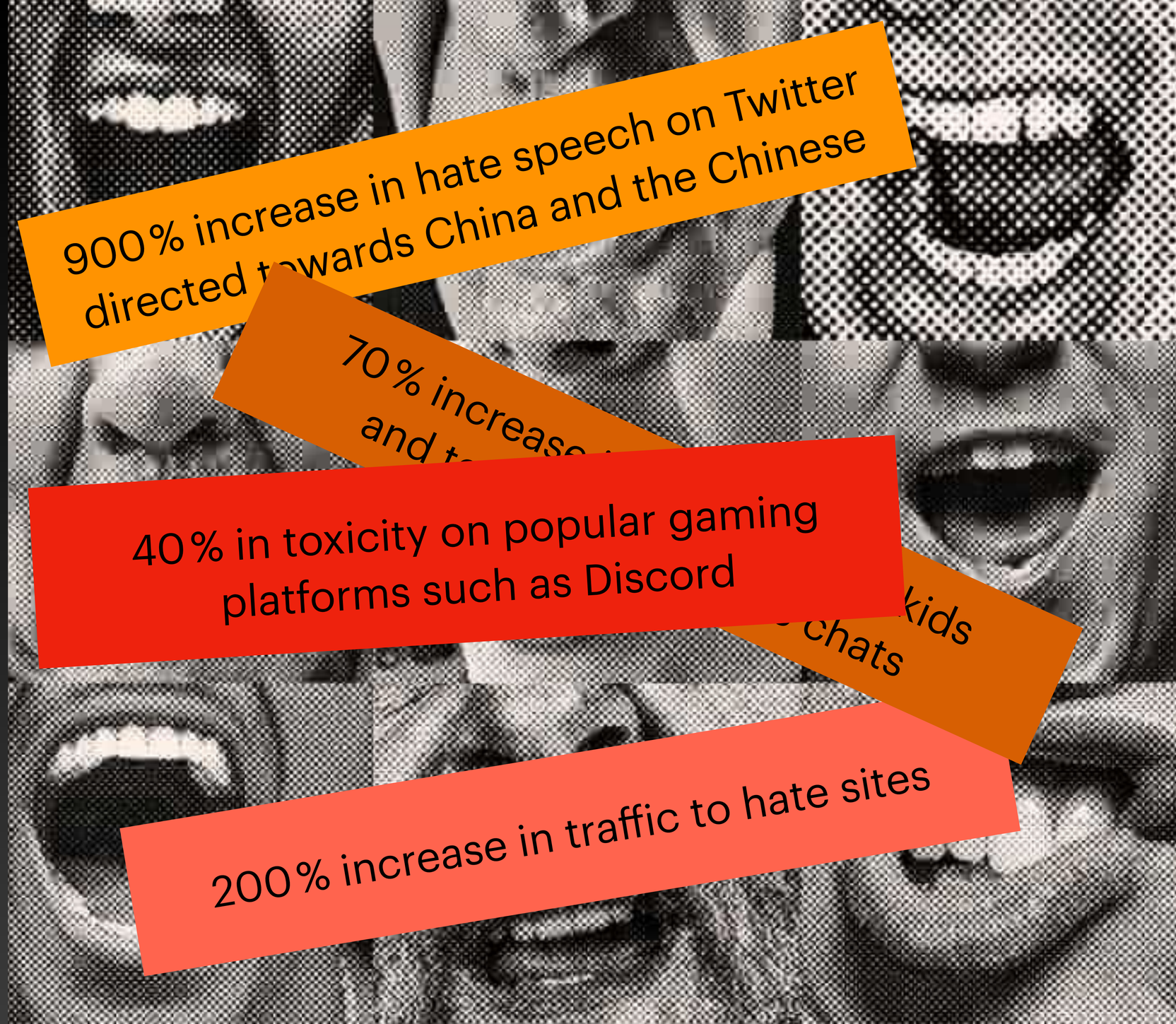
Toxic Rank

Ranking Toxicity in Wikipedia Talk Page Comments

Internet Toxicity

is on the rise

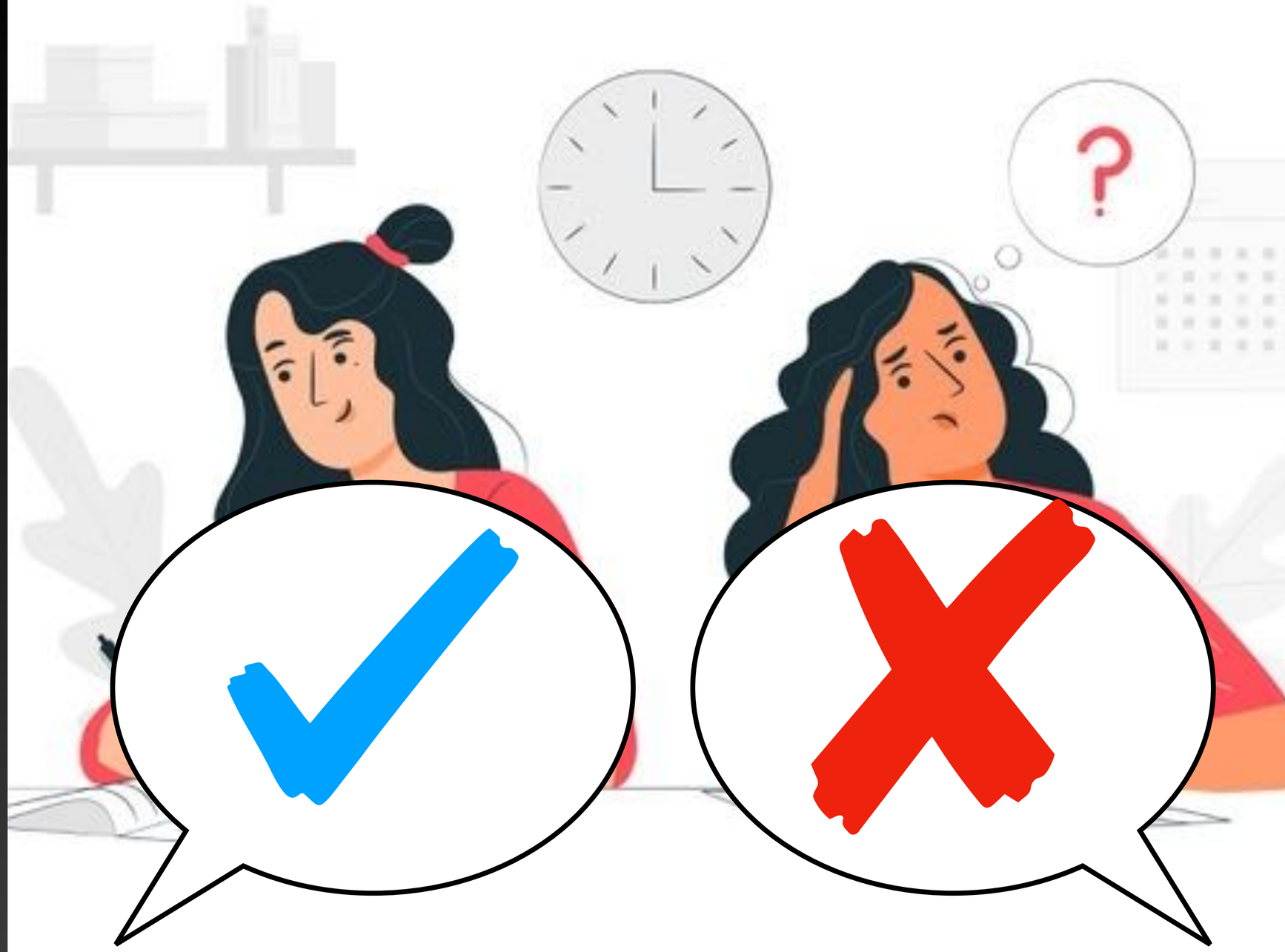


https://l1ght.com/Toxicity_during_coronavirus_Report-L1ght.pdf

identify Toxicity using Algorithms

Rate
severity of
Toxic Comments

Rate
severity of
Toxic Comments



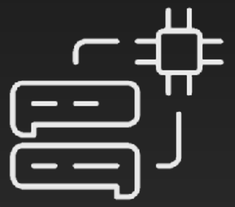
Methodology

Kaggle Competition

- 14,000 +
- unique
- comments
- Collected from Wikipedia Talk Pages

Collect Data

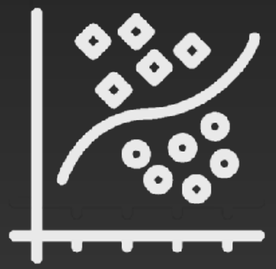
Prepare Coment Data



- Text Cleaning
- NLTK Tokenizer
- Snowball Stemmer
- Truncated SVD

Preprocess

Evaluate Toxicity

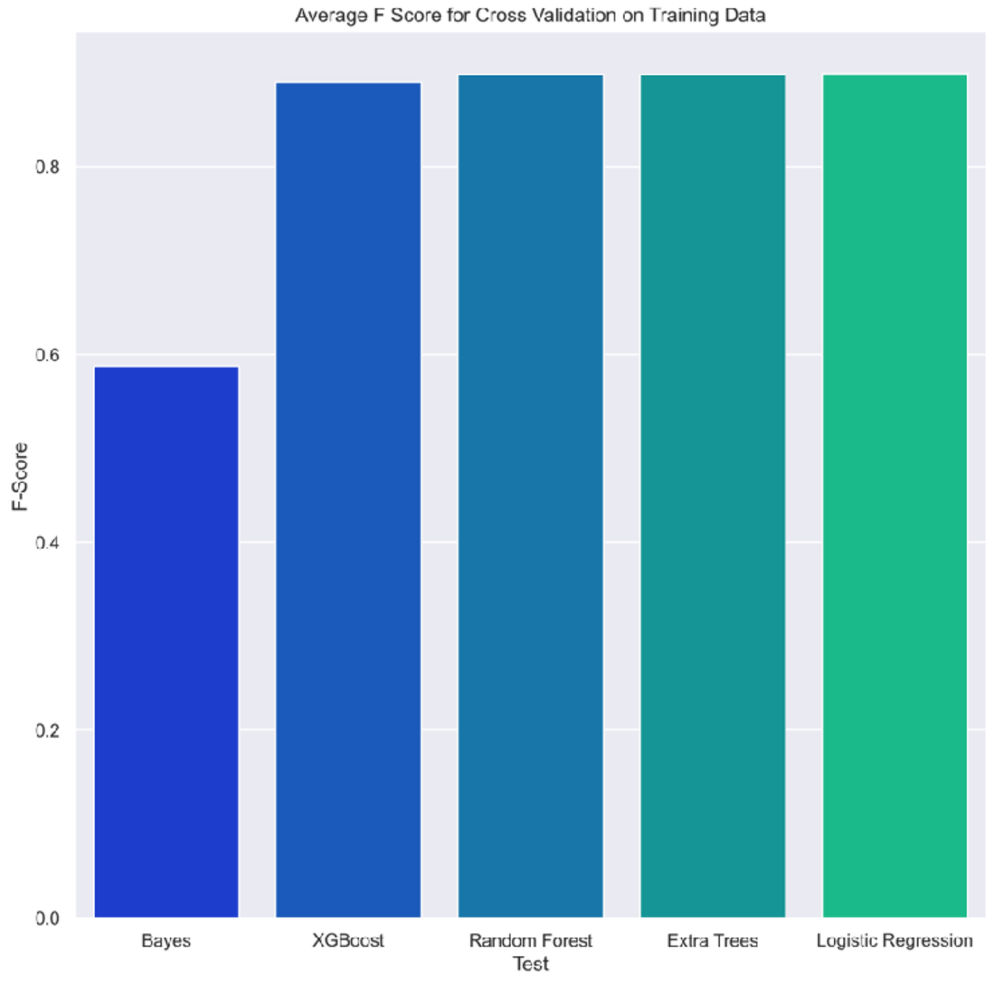


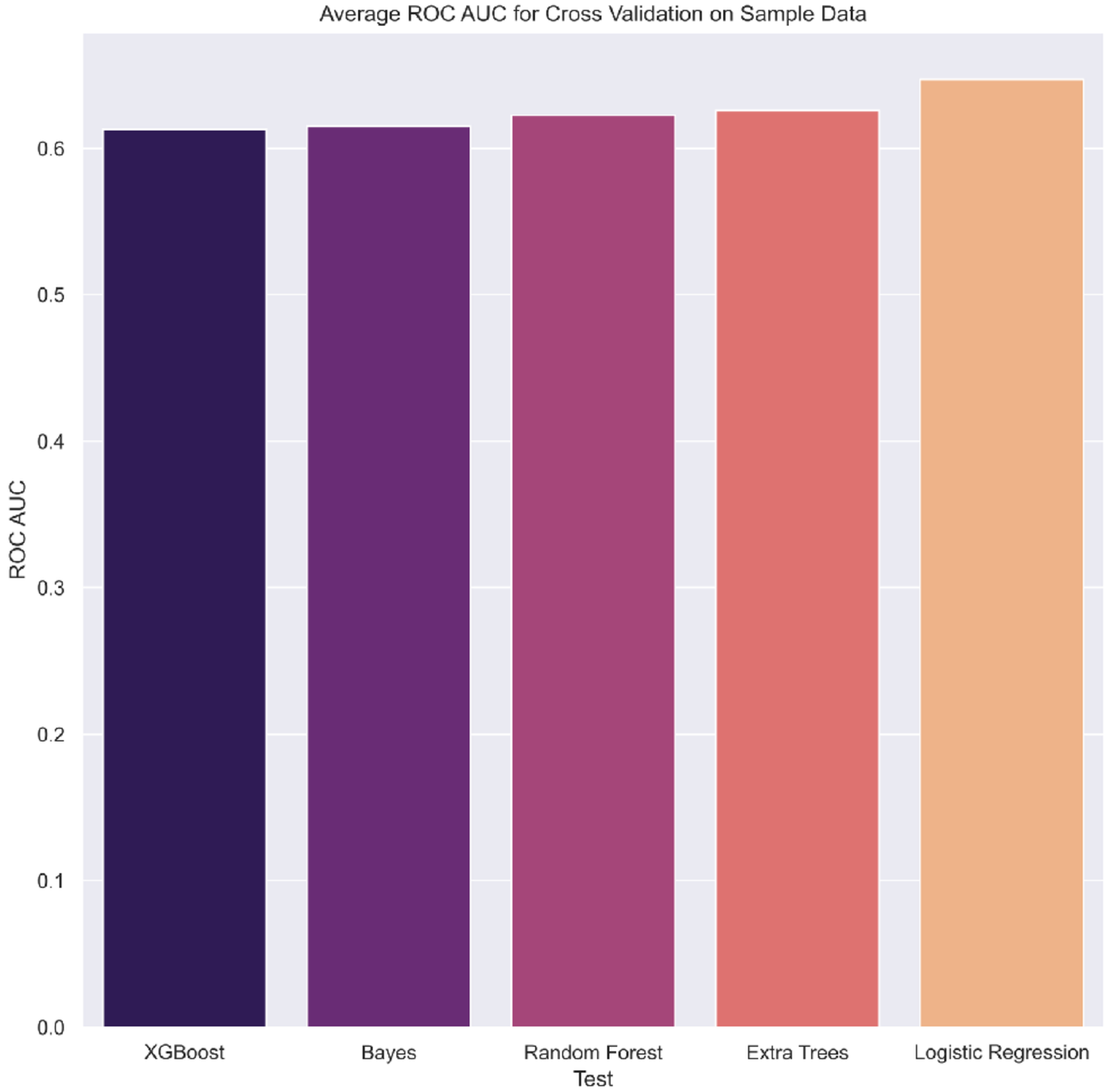
- Assign Toxicity based on Class Probability

Model

Logsitic Regression

Assign Toxic Probability Score





Less Toxic

More Toxic

Less Toxic

“Please stop your disruptive editing. If your vandalism continues, you will be blocked from editing Wikipedia.”

Toxic Probability: 0.27

More Toxic

“b**** \nyou are a f****** h***. you s*** d*** you big a** h****. you are g** you f**** a** b***. you can go to h*** you b**** a** m**********, suck a big d*** a*****…”

Toxic Probability: 0.99

Text

"You just make me laugh \n\nHAHAHAHAHAHA so you are some aussie b**** hiding behind a computer and changing what i write, im half abo half l*** so i am the truth u cannot face but your girlfriend faces every night ;)"

Label: Less Toxic

Toxic Probability: 0.82

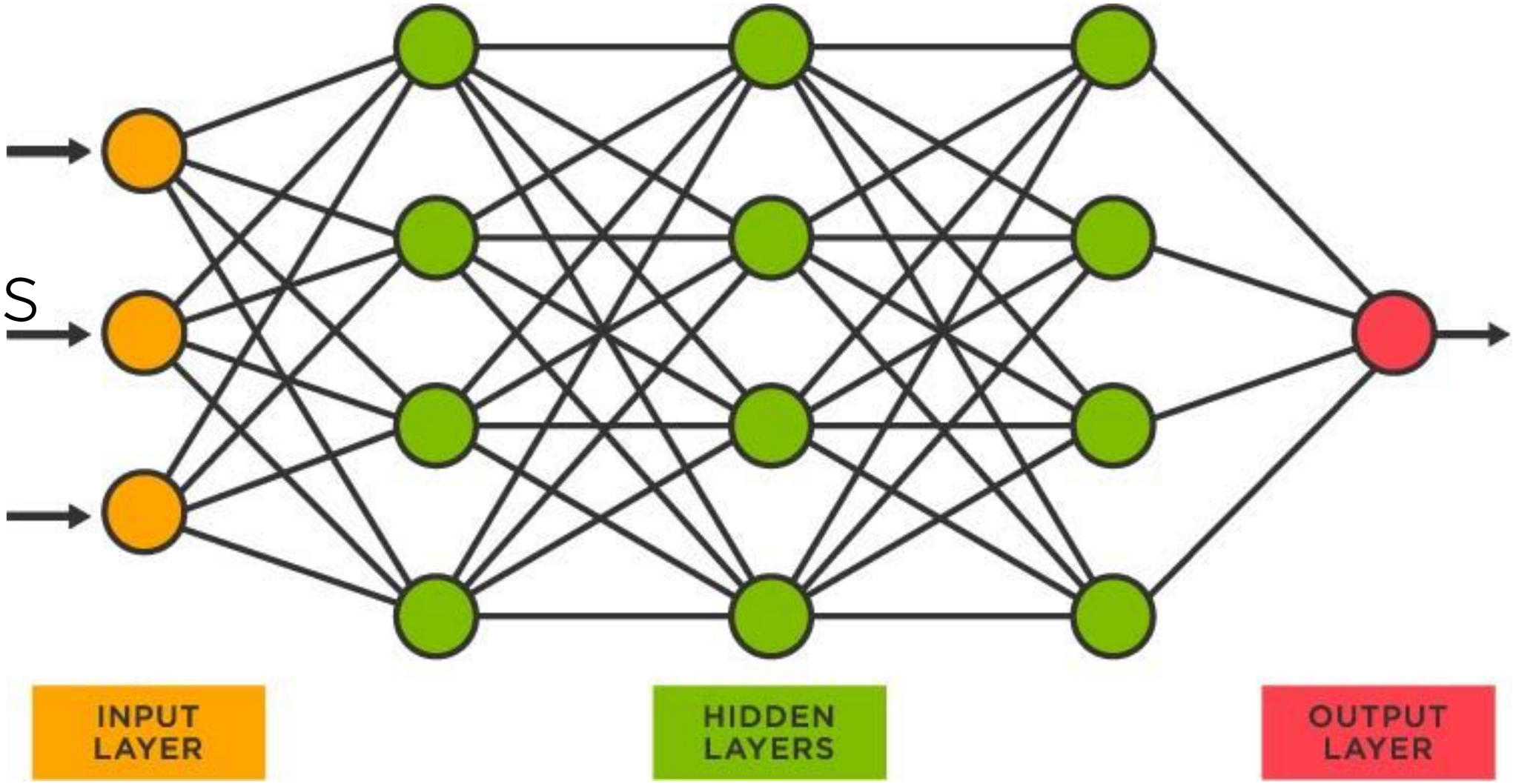
Final Score

Logistic Regression Probability

- ROC : .64
- F1: .89
- Competition Score: .724

Next Steps

- Neural Nets
- Transformers



Toxic Rank

Questions?