Comment Moderation:

Identifying Toxic Comments

Kin-Yip

Comment Moderation

Problems:

- Scale: 11,000 comments are posted to the New York Times website each day.
- Consistency: Human moderators can be fickle.
- Different websites may want different standards of moderation with regards to swearing and tone.

Solutions:

- Computers are fast, consistent, and can handle different policies.
- 160,000 comments scored in minutes.

English Wikipedia Comments Corpus

160,000 comments labeled by up to 10 human annotators each through Crowdflower.

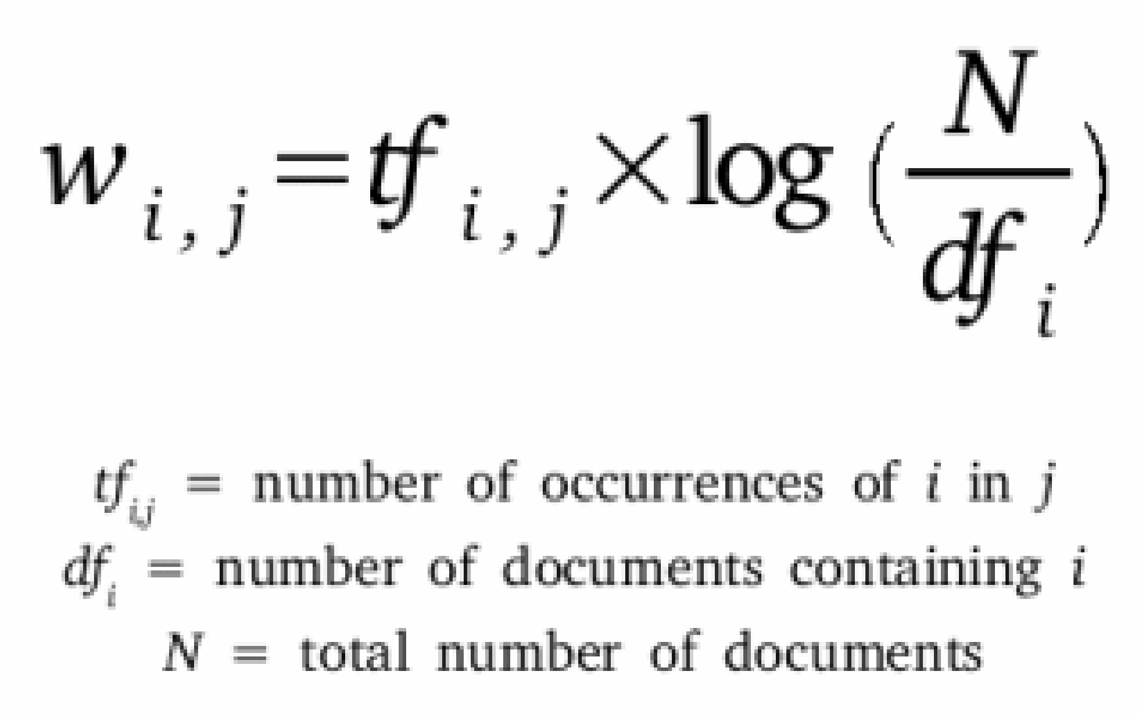
Multiple labels:

- Toxic (perceived as likely to make people want to leave the discussion)
- Subtypes: (Severe toxic, obscene, threat, insult, identity hate)

Goal:

- Build the best model in the world.

Term Frequency-Inverse Document Frequency



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j

df_i = number of documents containing i

N = total number of documents

Soft Voting Classifier

Ensemble of:

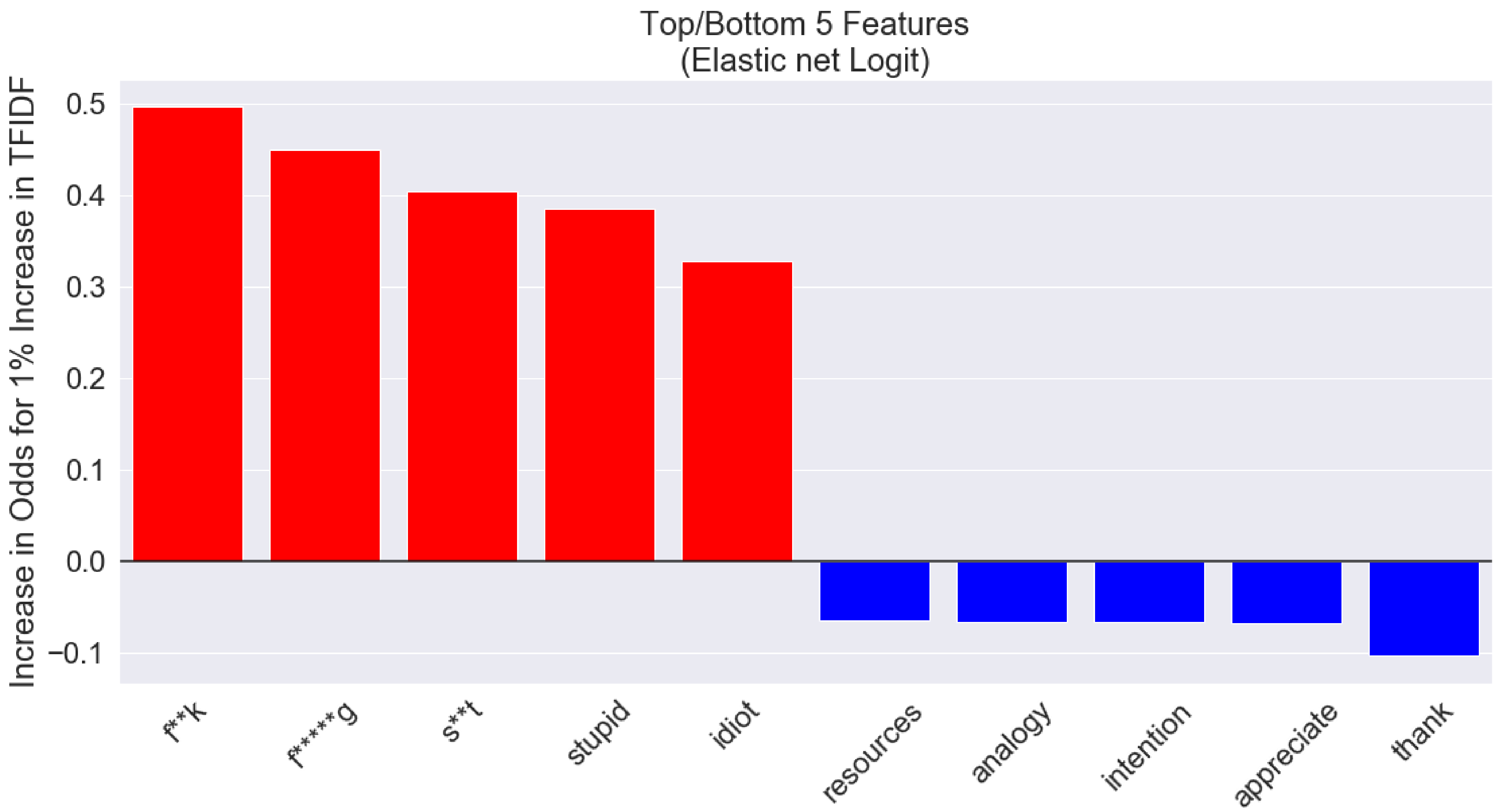
- Multinomial Naive Bayes
- Elastic net Logistic Regression
- Scikit-learn Random Forest
- LightGBM

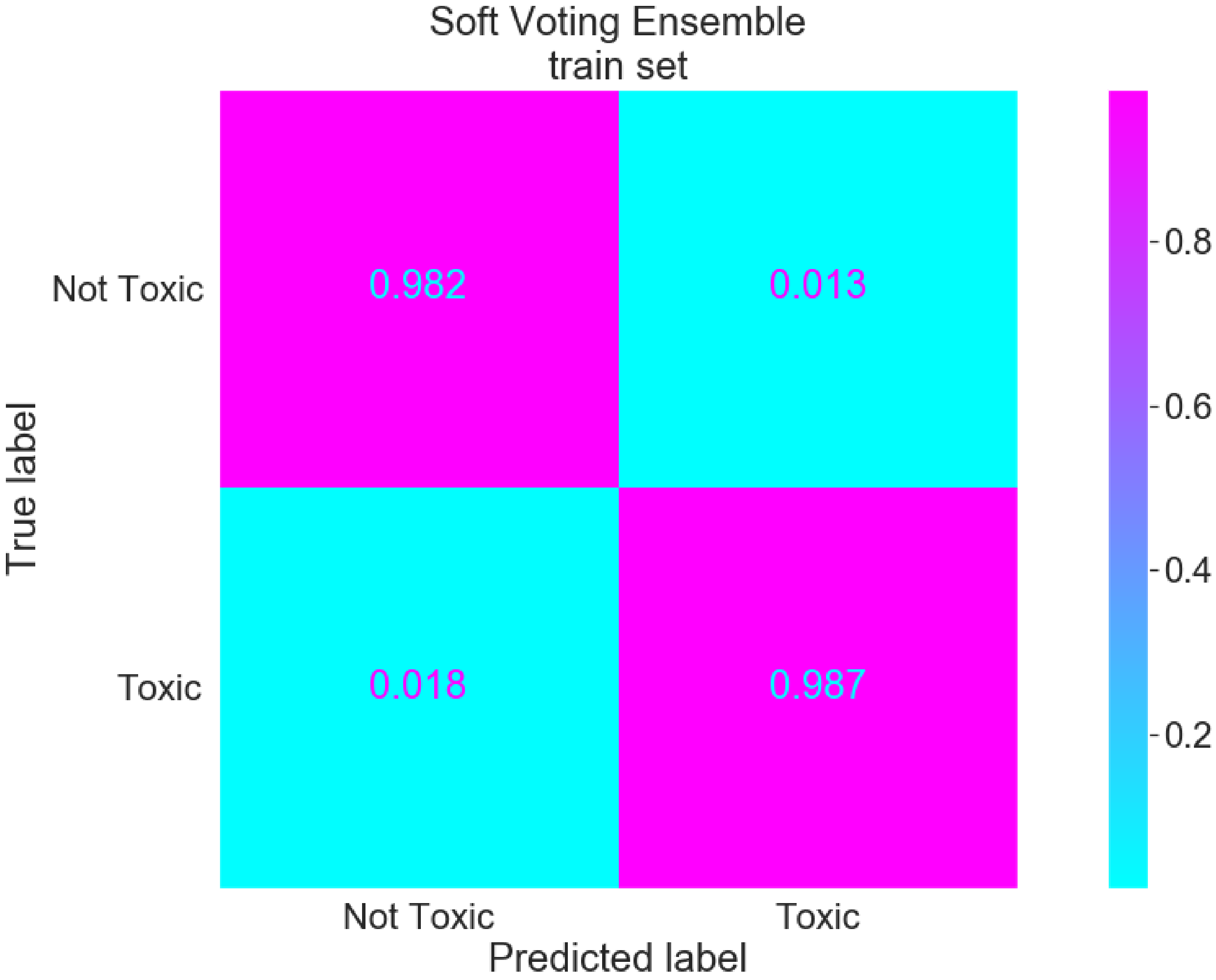
Averages predicted probabilities from each base learner.

- Theoretically, an ensemble formed from uncorrelated models will reduce bias and variance.

Model Selection

model	**fit_time**	**fit_std**	**score_time**	**score_std**	**AUC**	**AUC_std**
Multinomial NB	2.521	0.172	0.509	0.090	0.951	0.002
Elastic Net Logit	2.842	0.286	0.475	0.062	0.968	0.002
sklearn RF	112.597	0.401	1.271	0.038	0.929	0.003
LightGBM Classifier	129.962	0.806	0.894	0.141	0.967	0.002
Soft Voting Ensemble	242.908	7.552	9.956	5.467	0.972	0.001


Top/Bottom 5 Features
(Elastic net Logit)
Increase in Odds for 1% Increase in TFIDF
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
f**k
f*****g
s**t
stupid
idiot
resources
analogy
intention
appreciate
thank



Soft Voting Ensemble
train set
True label
Not Toxic
Toxic
0.982
0.013
0.018
0.987
Not Toxic
Toxic
Predicted label
0.8
0.6
0.4
0.2

Toxic Comments Classified as Not Toxic

Models cannot understand tone:

“... In my humble opinion this draft is a pathetically watered-down rendition of what's been discussed and usually agreed upon by everyone except you-know-who. Denazification reduced to half a sentence?...”

Comments referencing other non-toxic comments:

I have to put this as sittign here cackling my head off. Are you peopel on leave from a psyche hospital? I just found the following post from prissy Dame Ewart on some other user's talk page…

“Ms Gundagai, normally I would be thrilled to be the centre of so much attention (so far all your edits, except those to your userpage, have been devoted to me)...”

Not Toxic Comments Classified as Toxic

Models only check for offensive words and cannot understand context:

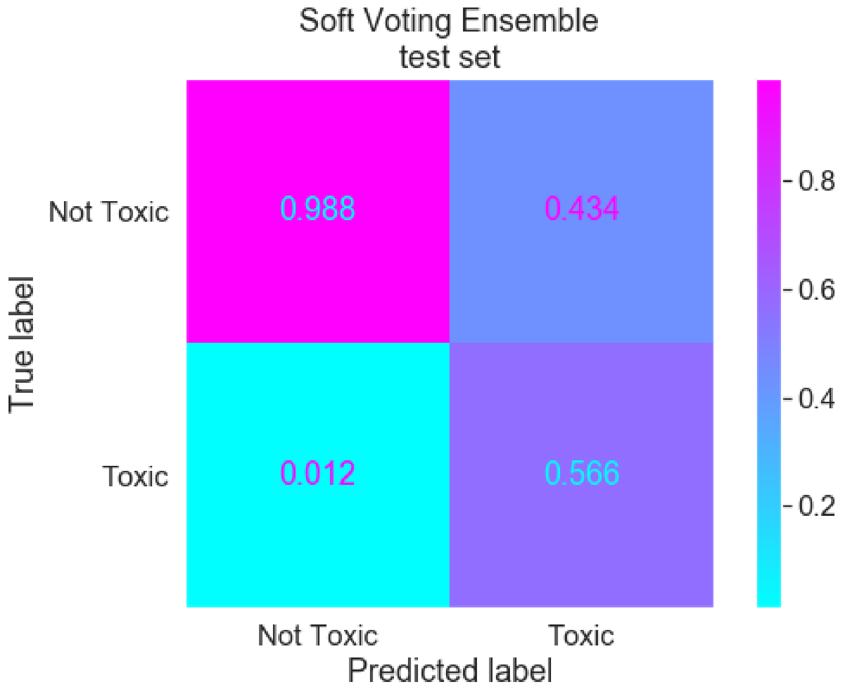
“But keep in mind I am a gay Jew.”

Human raters are flawed:

“hey dude i d like to know who is in charge of all of this deleting that is going on, i did a thing on the harly drags, and some punk ass f****t deleted it and i really dont appreciate that very much at all, and im kinda very extremely pissed and i think that since it wasnt good enough for yall nerds then yall should write abput it , but yall dont have any clue what it is cuz you aint been you no p***y gettin guy.... feel free to write me back dude”

Soft Voting Ensemble

AUC: 0.972



Future Directions

Extract more features: parts-of-speech

Hyperparameter tuning:

- other methods to handle imbalanced classes, character n-grams, word n-grams, model hyperparameters, prediction thresholding, stacking ensembles, etc.

Predict subclasses

Address unintended bias

QUESTIONS?