


Don't Hate, Appreciate

Loren Hinkson
Andrea Koch
Natasha Mathur

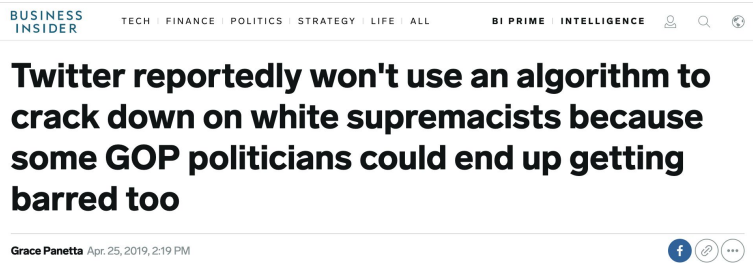




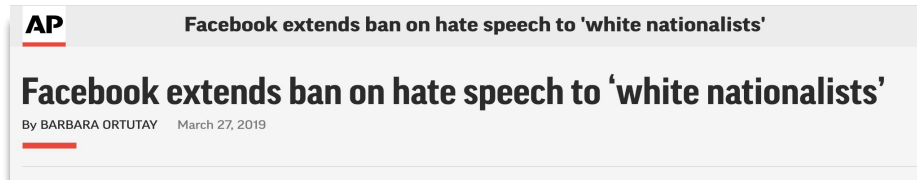
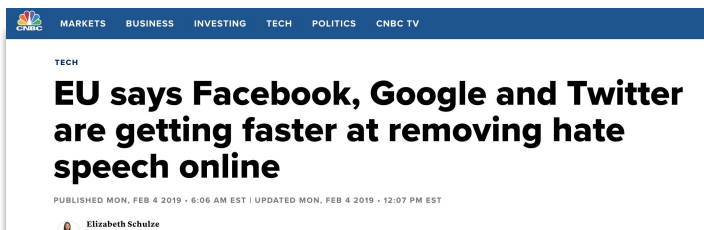
Could you identify
hate speech?

The Data & Policy Problem

The Data & Policy Problem



Facebook restores censored nude 'napalm girl' photo due to "historical importance"



Published on February 7, 2019

Facebook Has a Right to Block 'Hate Speech'—But Here's Why It Shouldn't

written by Brian Amerige



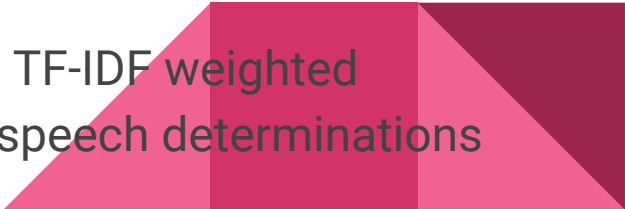
What Others Have Done

What others have done

Twitter (Aditya Gaydhani et al, 2018; Punyajoy Saha et al, 2018): Researchers used n-grams, BOW, and TF-IDF to classify tweets into three classes: hateful, offensive and clean focusing on racism and sexism; and to identify misogynistic language or misogynistic intent targeting a group or person, respectively

Reddit (Eshwar Chandrasekharan, 2015): Researchers evaluated the variation in hate speech on Reddit after the ban of two problematic subreddits, including platform churn and hate speech in other subreddits using Sparse Additive Generative Model (SAGE)

YouTube (Karthik Dinakar et al, 2011): Researchers used TF-IDF weighted bi-grams/ unigrams to assign binary and multiclass hate speech determinations



The Data

Who labeled the data? How were the labels verified?

- Each comment was shown to up to 10 human raters/annotators
 - They provided the toxicity scores as well as identity labels
 - Capture range of opinions re: toxicity and identity
- Label quality/verification:
 - 10% of comments show to each rater were labeled with the correct label
 - Raters who missed too many of these verification labels were DQ'd



Jigsaw Unintended Bias in Toxicity Classification

- Civil Comments platform:
 - Comments from 50 English-language news sites around the world, 2015-2017
- Data sets:
 - Training: 1.8M
 - Testing: 97K
- Each comment is rated for how toxic it is
 - Further labeled with toxicity subtypes and “identity mention” attributes



Unbalanced Classes & Resolutions

Unbalanced classes, mitigation, & considerations

- Optimizing the models performance for neutral comments about targeted identities may decrease its performance on hate speech.
 - **Option 1:** ID the neutral comments and duplicate them in the training set
 - **Option 2:** Bootstrapping: Train preliminary models to identify comments that reference targeted identities, and then evaluate their confidence levels on predictions for those comments to inform hyperparameter tuning
 - **Option 3:** Train models on labeled data from other social media platforms
- Additional hate speech indicator to weight 'target' score (toxicity rating)





Our Approach & Considerations

Our approach

Baseline:

- Run baseline models to see what sorts of examples we're getting wrong

Experimenting with the train set:

- Increase concentration of target comments (ex: non-hateful comments about a protected class)
 - Bringing in additional, labeled sets of comments vs. duplicating neutral messages in train set

Evaluation:

- Precision, recall, false positive rate, false discovery rate performance
 - Priority: models with better performance in false positive rate and false discovery rate, as we do not want to incorrectly flag topical comments that are not hateful

Techniques:

- RNN (Recurrent Neural Network), Logistic Regression, SGD
- 



Encouragements and Suggestions