Detecting Cyberbullying with Natural Language Processing

Helping parents keep teens safe online

Greg Burgess
Data Scientist
Real Difference Data

Content Outline

1. Business Problem & Product Goals
2. Data & Methods
3. Results
4. Examples
5. What's Next?

Greg Burgess
Data Scientist
Real Difference Data

BUSINESS PROBLEM & PROJECT GOALS

Drugs/Alcohol

75.35% **of tweens and** 93.31% **of teens** engaged in conversations surrounding drugs/alcohol.

Self-Harm/Suicide

43.09% **of tweens and** 74.61% **of teens** were involved in a self-harm/suicidal situation.

Sexual Content

68.97% **of tweens and** 90.73% **of teens** encountered nudity or content of a sexual nature.

Violence

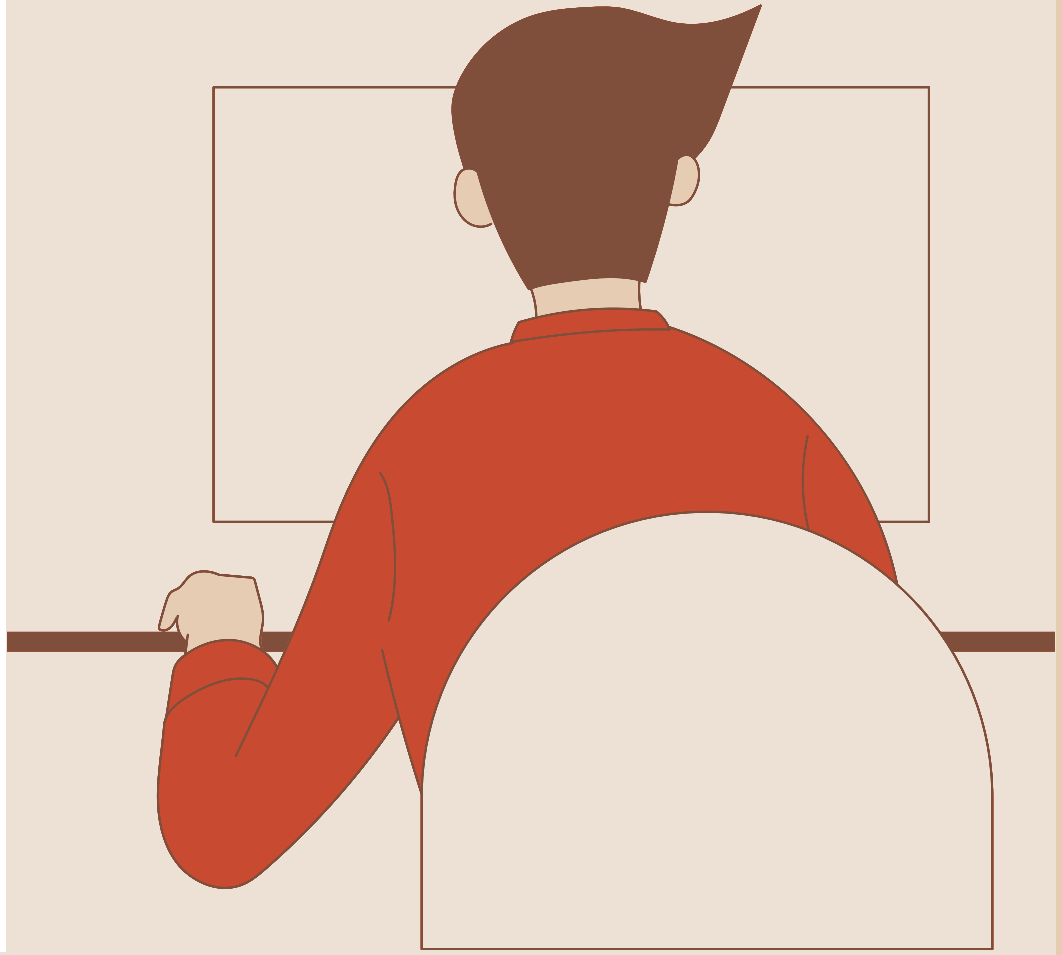
80.82% **of tweens and** 94.50% **of teens** expressed or experienced violent subject matter/thoughts.

Depression

32.11% **of tweens and** 56.40% **of teens** engaged in conversations about depression.

Bullying

72.09% **of tweens and** 85.00% **of teens** experienced bullying as a bully, victim, or witness.

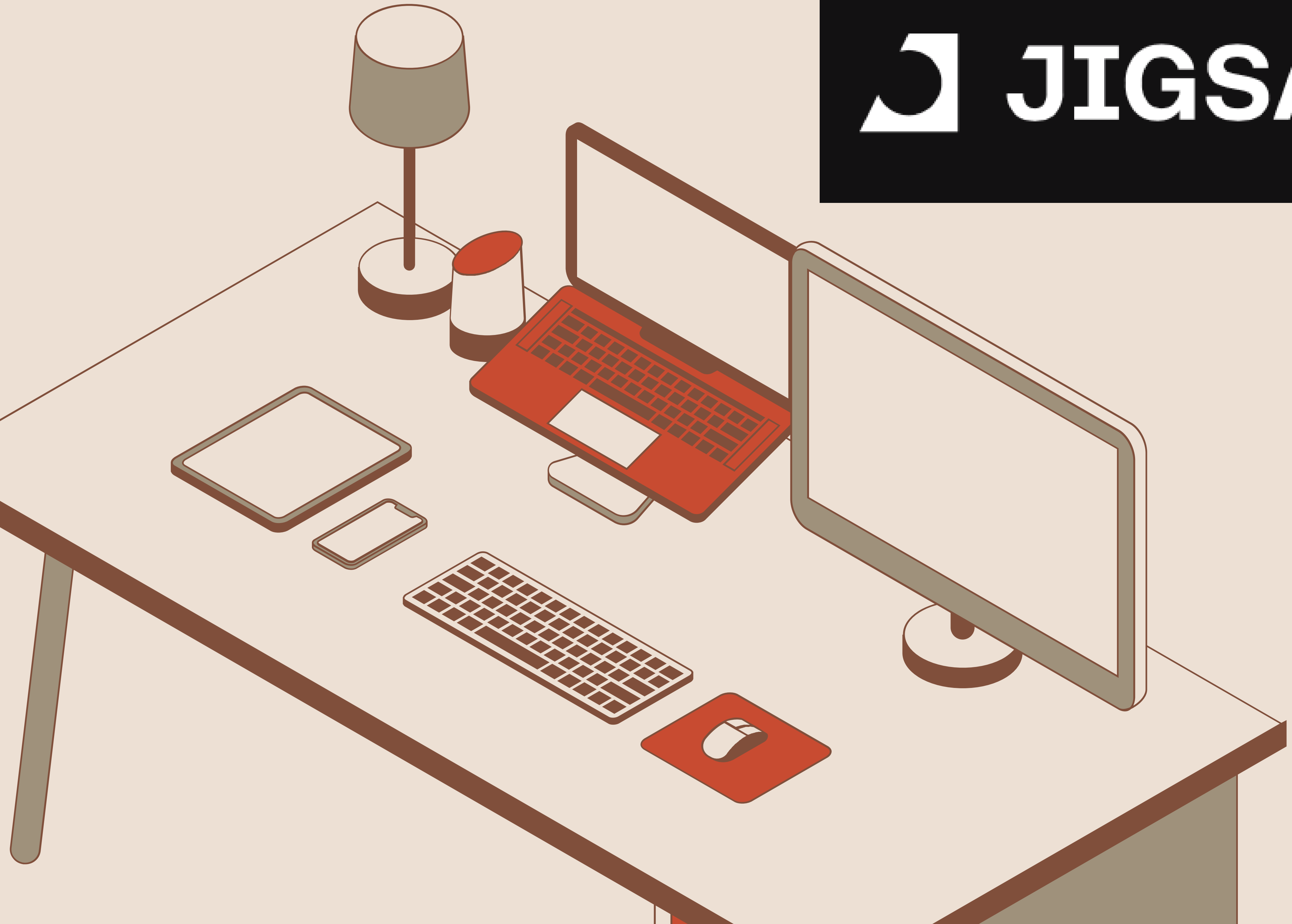


Source: Bark annual survey

Greg Burgess
Data Scientist
Real Difference Data

BUSINESS PROBLEM &
PROJECT GOALS

Products to help parents are limited



JIGSAW

Kindly

Kindly is the product of innovator Gitanjali Rao and UNICEF's collaboration

circle

Life360

bark

Greg Burgess
Data Scientist
Real Difference Data

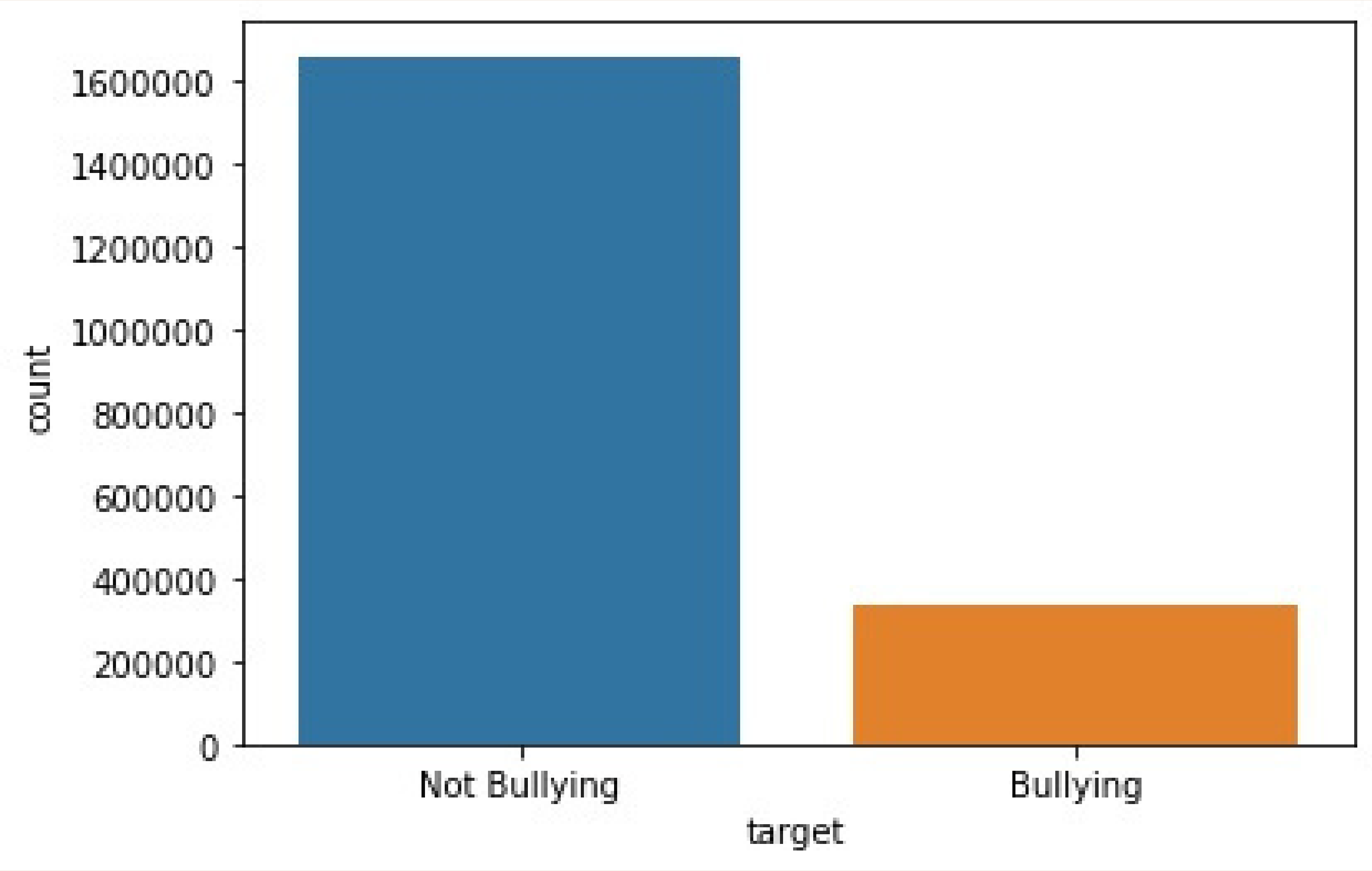
Data

- Two million online comments
    - Collected by Civil Comments
    - Curated by Jigsaw
    - Distributed on Kaggle
- Proportion of human raters endorsing:
    - toxicity
    - severe toxicity
    - obscene content
    - threats
    - insults
    - identity attack
    - sexually explicit content

Greg Burgess
Data Scientist
Real Difference Data

Methods

- Combined toxic subtypes into single target
- Standard "Bag of Words" preprocessing
- Undersampled "non-toxic" comments
- Naive Bayes and Logistic Regression Classifiers



Greg Burgess
Data Scientist
Real Difference Data

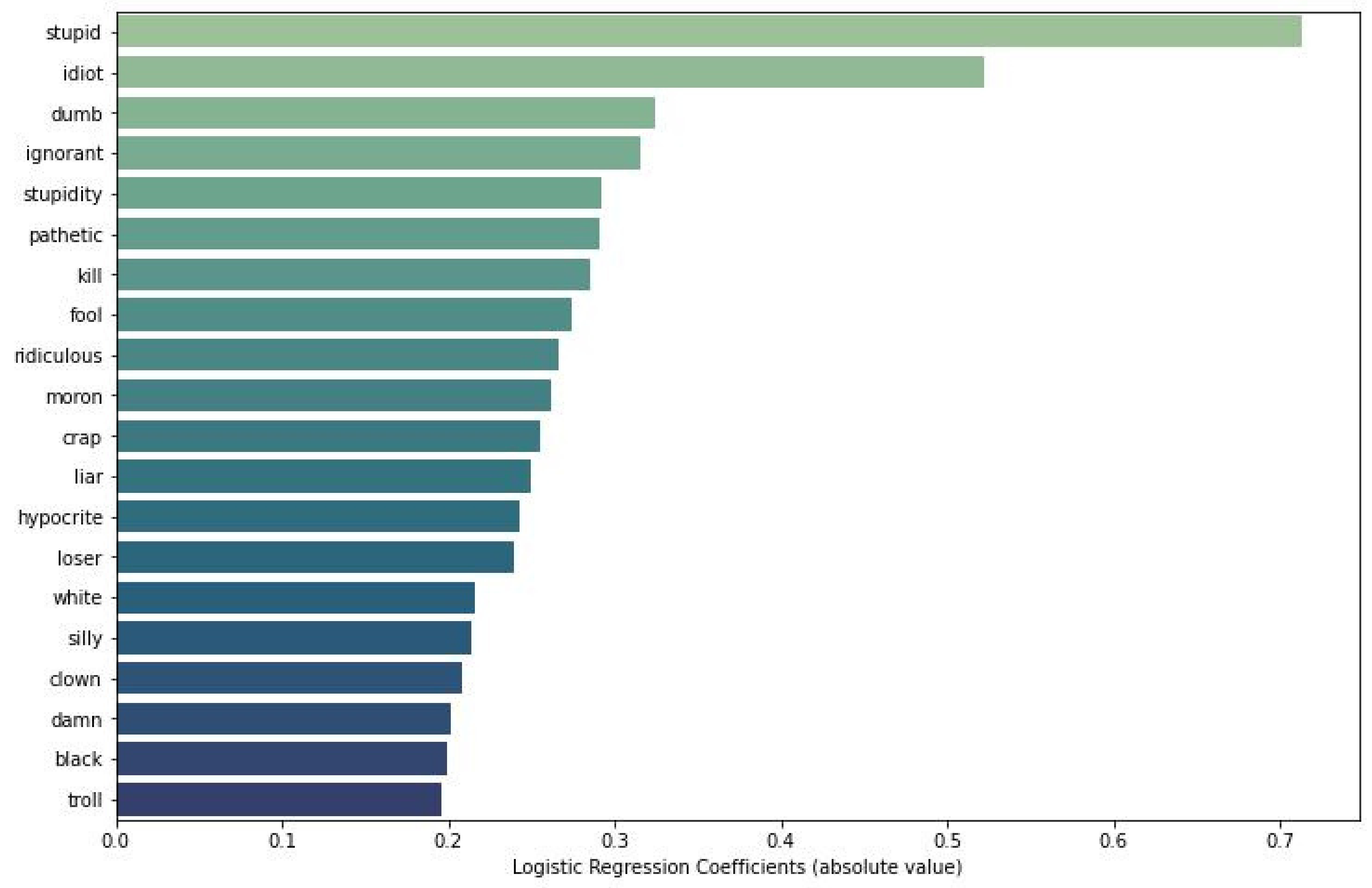
RESULTS

Toxic comments

Non-toxic comments

Greg Burgess
Data Scientist
Real Difference Data

Strongest Predictors of Toxic Comments



Greg Burgess
Data Scientist
Real Difference Data

EXAMPLES

A comment that should be classified as 'toxic' by using some of the terms with the strongest coefficients in the logistic regression.

```
detect("You're a stupid idiot")
```

```
toxic comment
```

Another comment that, on its face, should be classified as 'non-toxic'

```
detect('You are awesome and I love you')
```

```
not a toxic comment
```

Lastly, a comment that was created to be intentionally ambiguous.

```
detect('Damn, I love you, silly')
```

```
toxic comment
```

Greg Burgess
Data Scientist
Real Difference Data

WHAT'S NEXT

Phase 1: Develop an API

- Connect to users' accounts
- Highlight potential toxic comments
- Provide parental alerts

Phase 2: Acquire more diverse data

- Different groups of users
- Content by app

Phase 3: Improve models

- Additional feature engineering
- Advanced models (BERT and GPT)

Greg Burgess
Data Scientist
Real Difference Data

Thank you for listening!

Please reach out with questions!

https://RealDifferenceData.com

Greg Burgess
Data Scientist
Real Difference Data