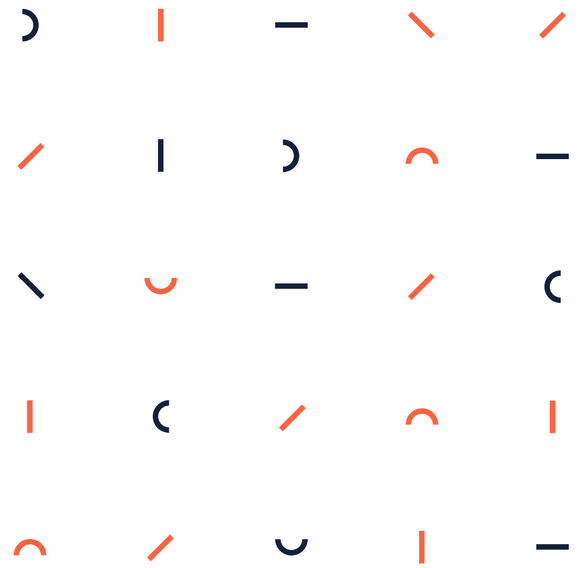
Hateful Meme Detection

A Text Classification Approach

Andrew Whitman
December 2021

Hateful Memes

- Multimodal
  - textual
  - visual
- Automatic detection



Business Problem

HATE SPEECH

- Violent or dehumanizing
- Inferiority
- Exclusion or segregation
- Slurs
- Mocking

Hate Speech on Facebook

CONTENT ACTIONED

79M pieces of content

- 1.2M restored

Hate Speech on Facebook

CONTENT ACTIONED

79M pieces of content

- 1.2M restored

PREVALENCE

0.03-0.06% of content views

Data

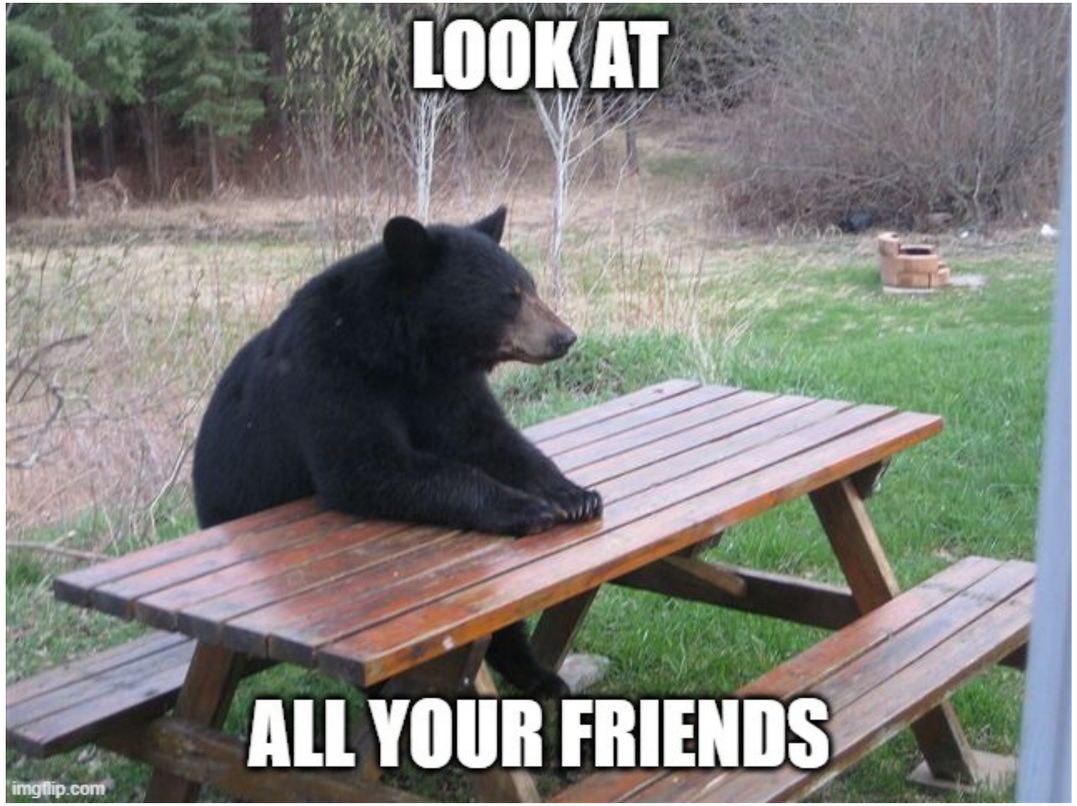
12,140 memes

Data

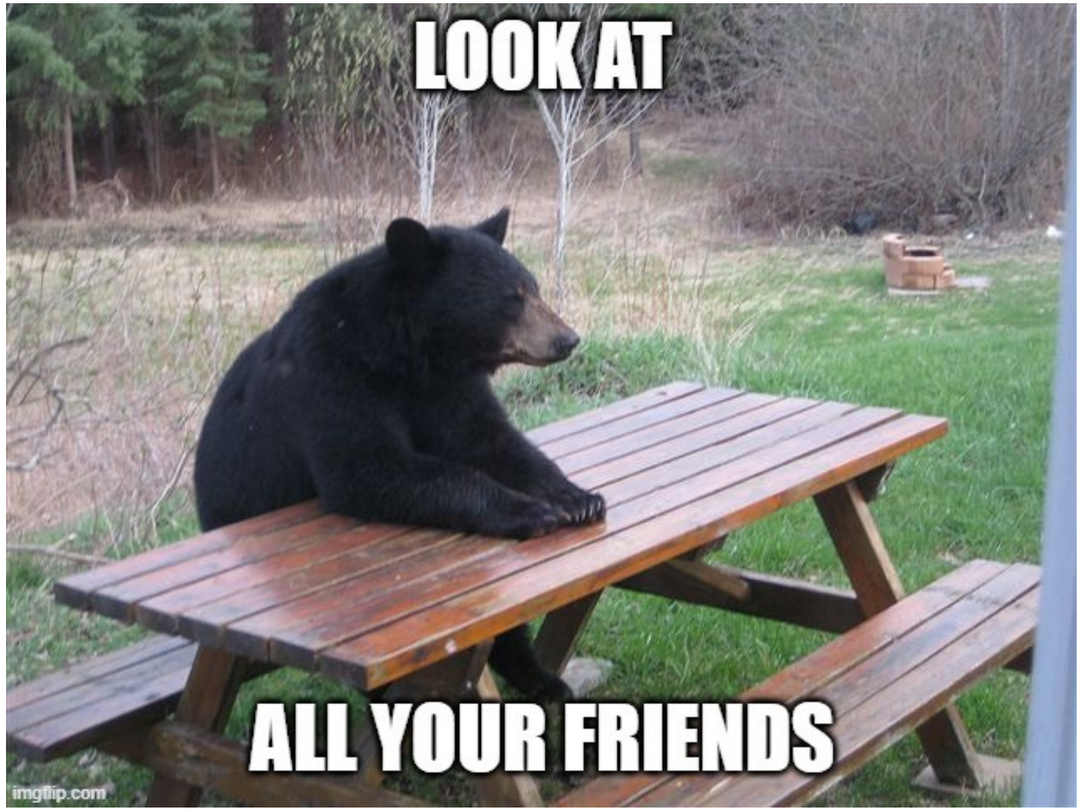
12,140 memes

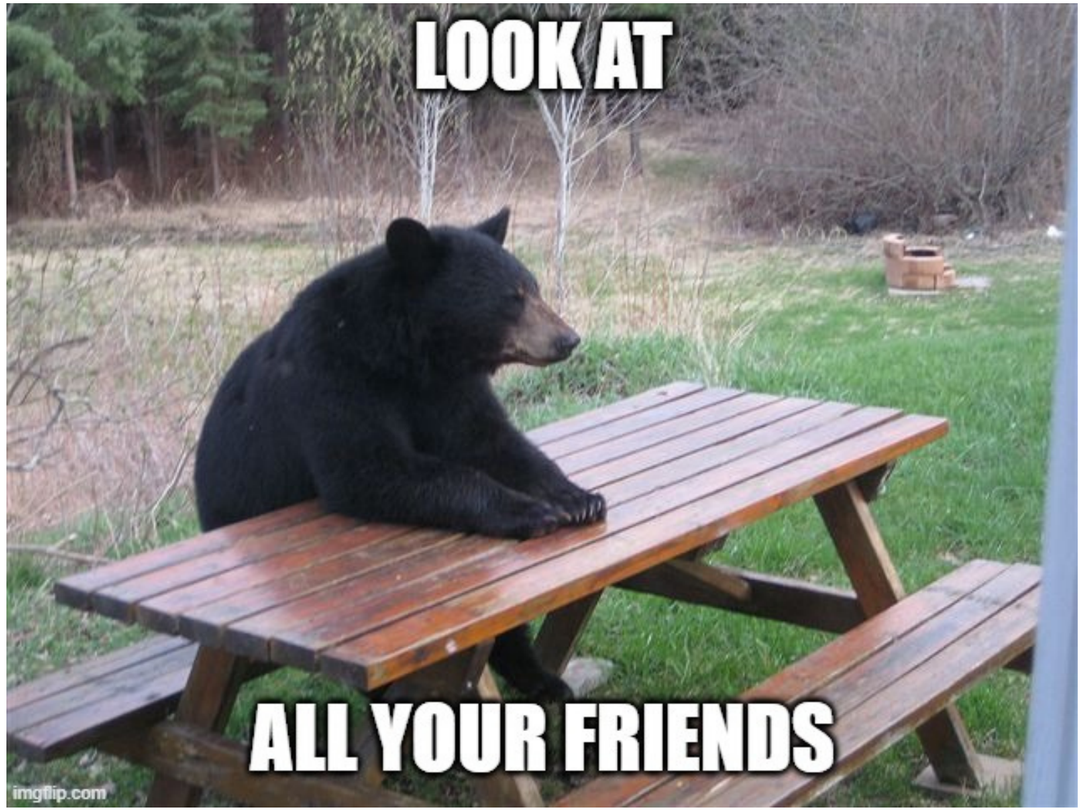
Meme Distribution

"Mean" Example

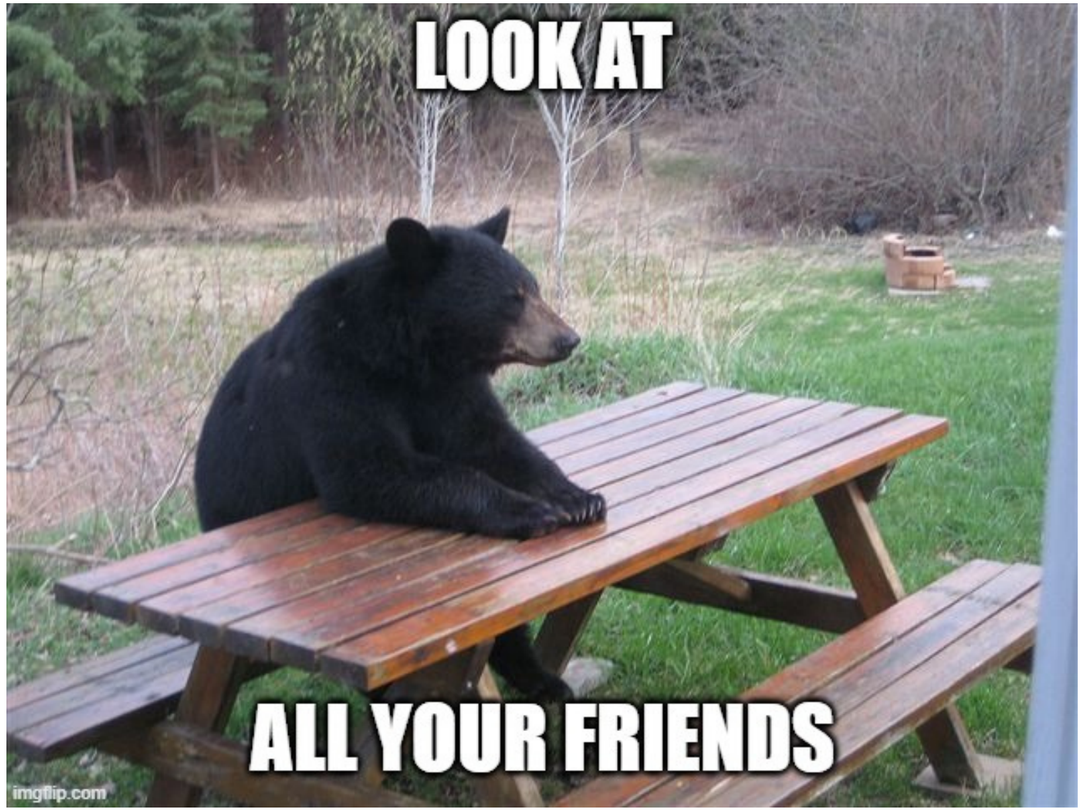


Benign Confounders





Benign Confounders



Benign Confounders

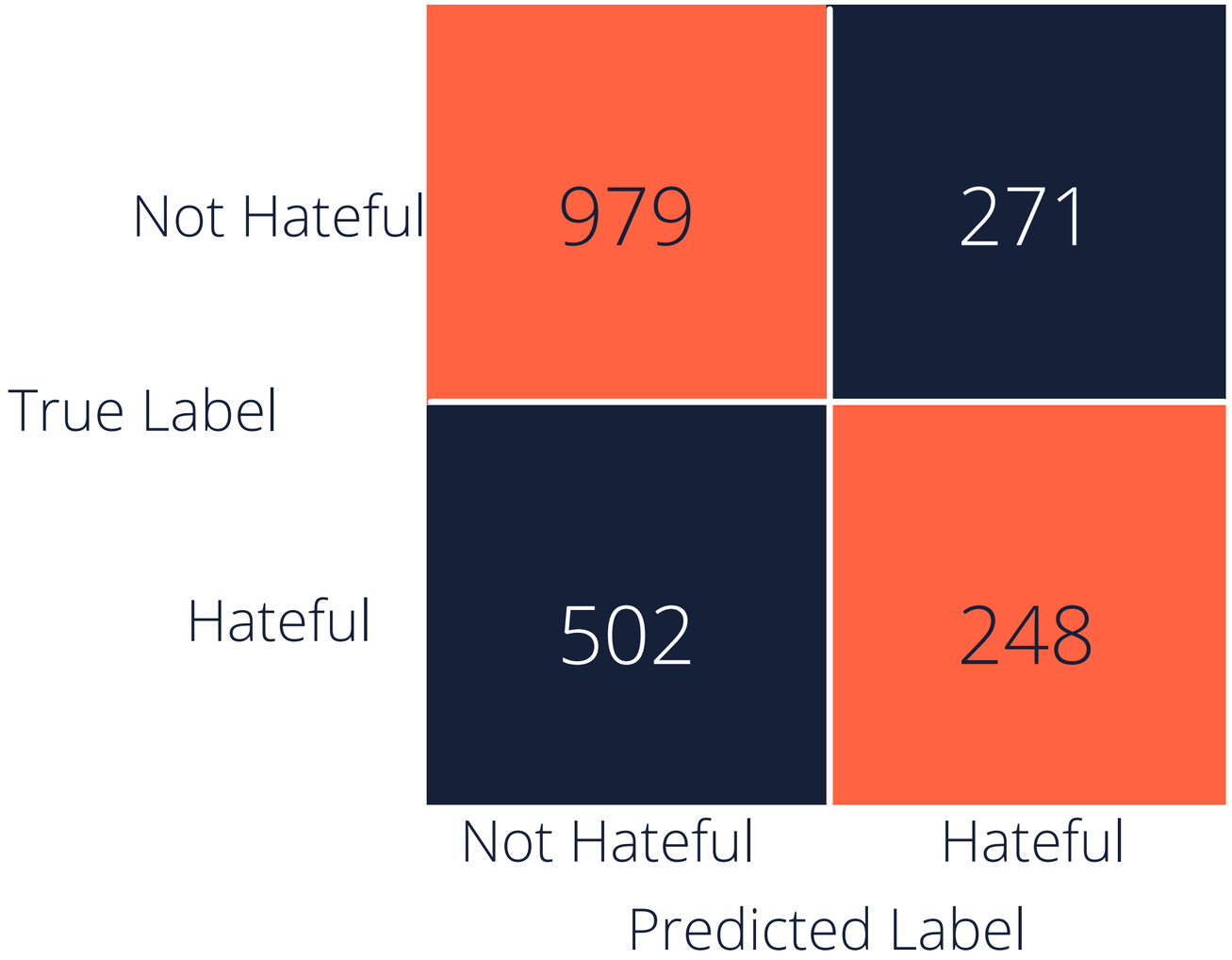
METHODS & RESULTS

Bag-of-Words Model (BoW)

- ✅ Word Frequency
- ❌ Word Order

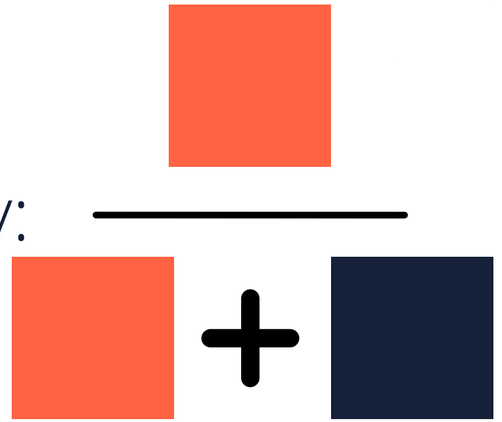
METHODS & RESULTS

METHODS & RESULTS

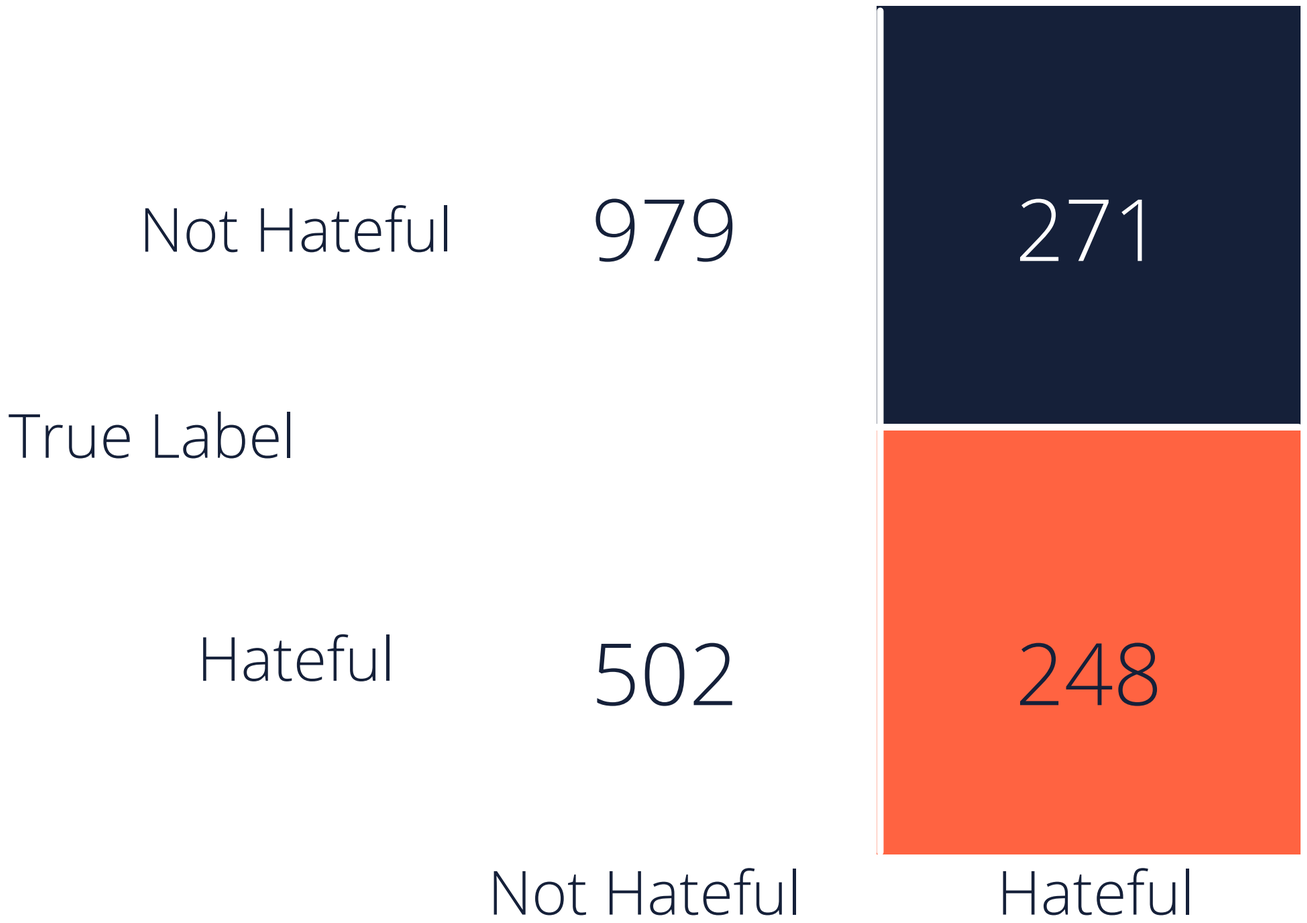


61%

Accuracy:



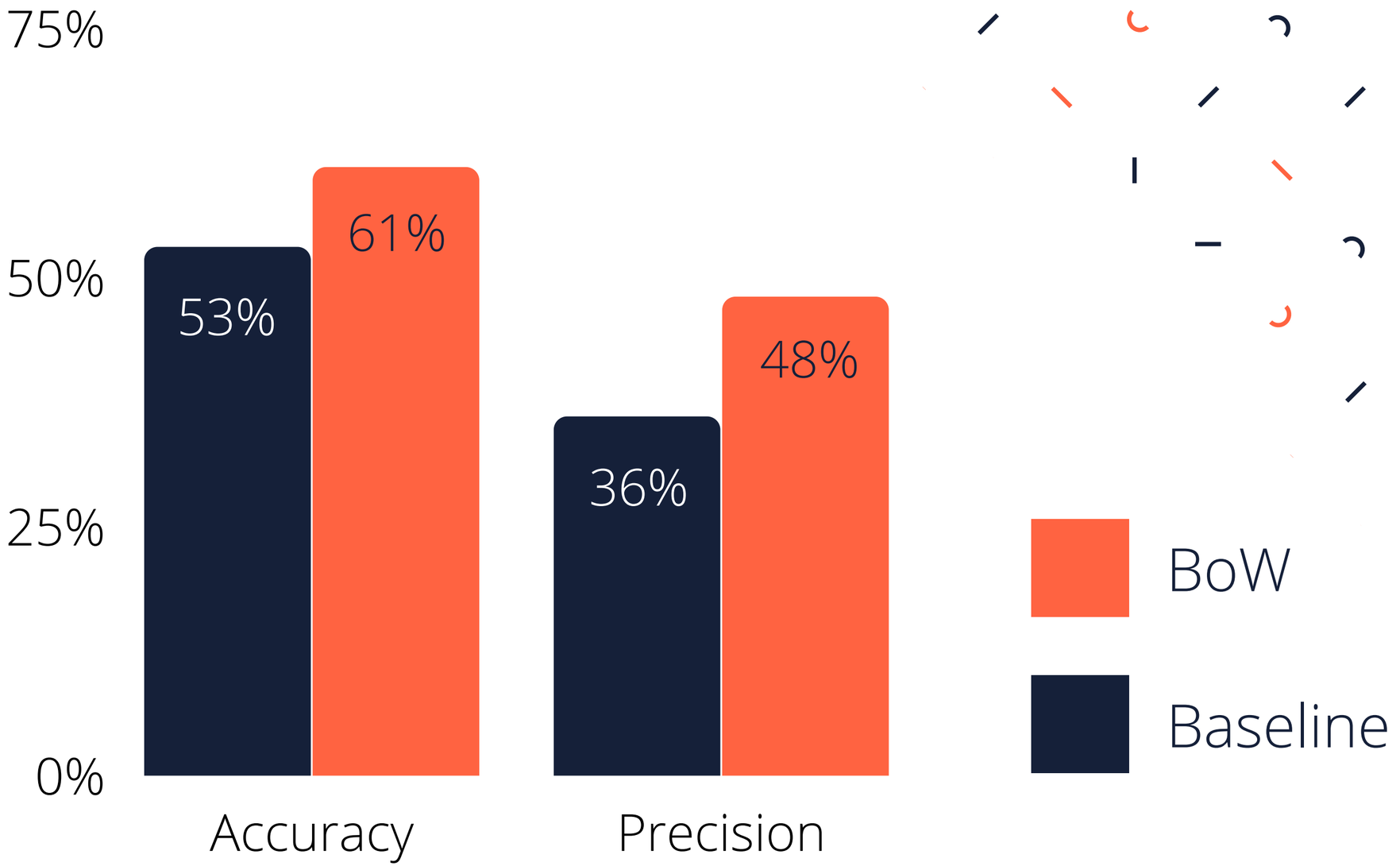
METHODS & RESULTS



48%

Precision:

METHODS & RESULTS



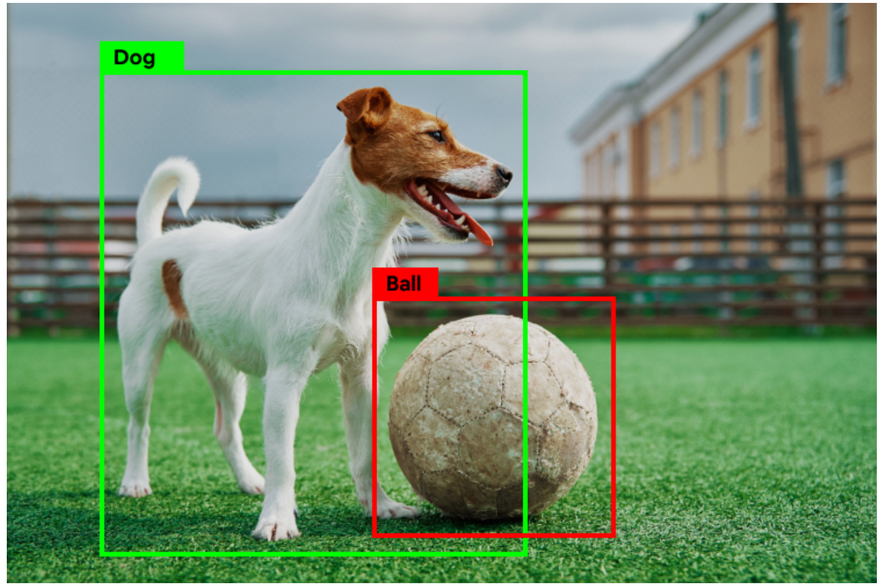
Recommendations

TAG PREDICTED HATEFUL MEMES

REVIEW PREDICTED HATEFUL MEMES

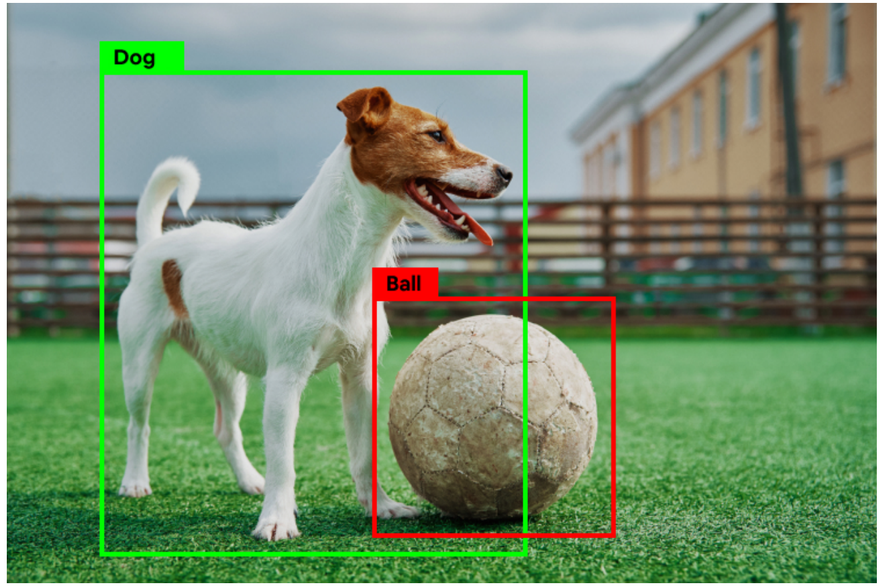
TUNE DECISION THRESHOLD

Next Steps



Object Detection

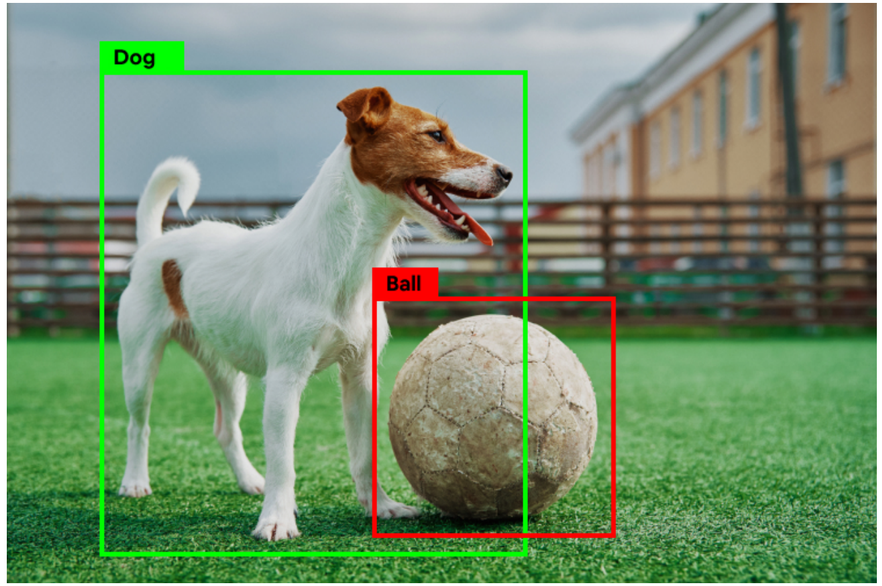
Next Steps



Object Detection

Optical Character Recognition

Next Steps

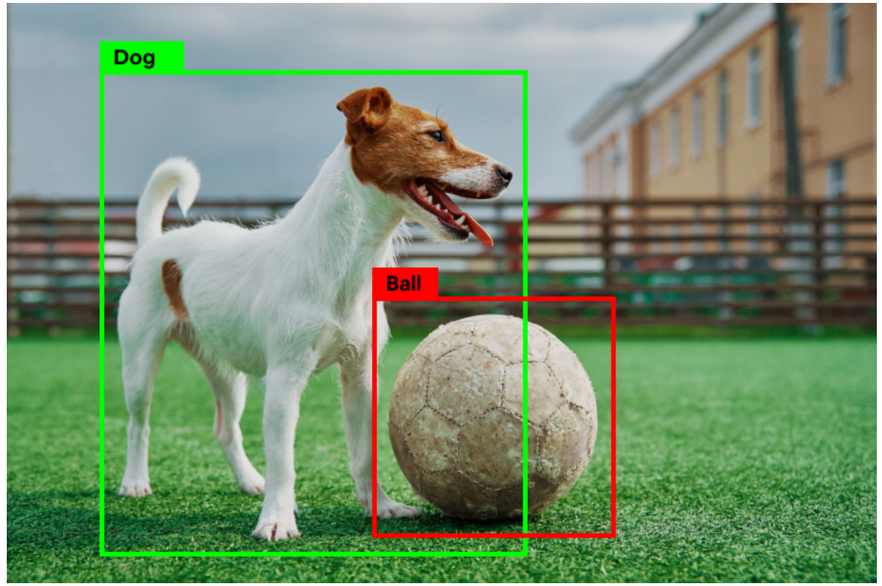


Object Detection

Optical Character Recognition

Advanced Language Models

Next Steps



Object Detection

Optical Character Recognition

Advanced Language Models

Language Diversity

Andrew Whitman

andrewwhitman

andrew-whitman

Appendix: Attribution

SLIDE 3

Photo by Jon Tyson on Unsplash

SLIDES 4-5

Photo by Barefoot Communications on Unsplash

Statistics from https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/

SLIDE 17

Photo by Dan Edge on Unsplash

SLIDES 18-21

https://deeplobe.ai/wp-content/uploads/2021/06/object-detection-2.jpg