LexiGuard

Capstone Project Presentation (December 19, 2023)

LexiGuard

Team

André Oliveira

Mechanical Engineer

Data Scientist

AI Enthusiast

Purvi Parmar

Software Developer

Data-Driven Insights, Problem Solving, Text Analysis

Michael Schickenberg

Linguist/translator with strong technical background going NLP

Eric Martinez

Industrial Engineer

Data, Strategy and AI Enthusiast

LexiGuard

Content Warning

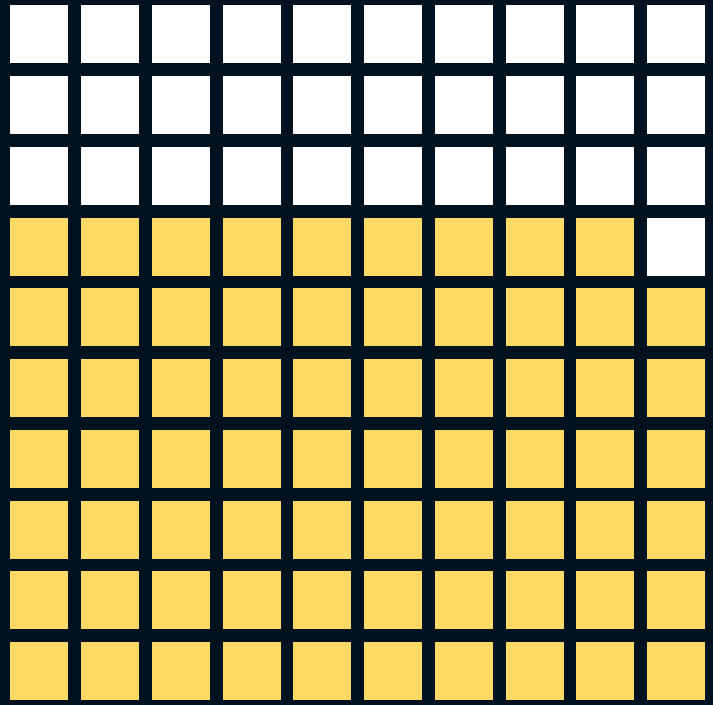
The following presentation is about detecting toxicity in user content published on the internet. It may therefore contain **offensive and hateful language** with regard to religion, gender, race and other kinds of identity and **may be disturbing**.

Agenda

- Why, what, how?
- Insights in our data
- Baseline model and performance metric
- Techniques and final NLP system
- Next steps and further improvements

Why is toxicity a problem?

69 % reported experiencing a risk last year (2022)

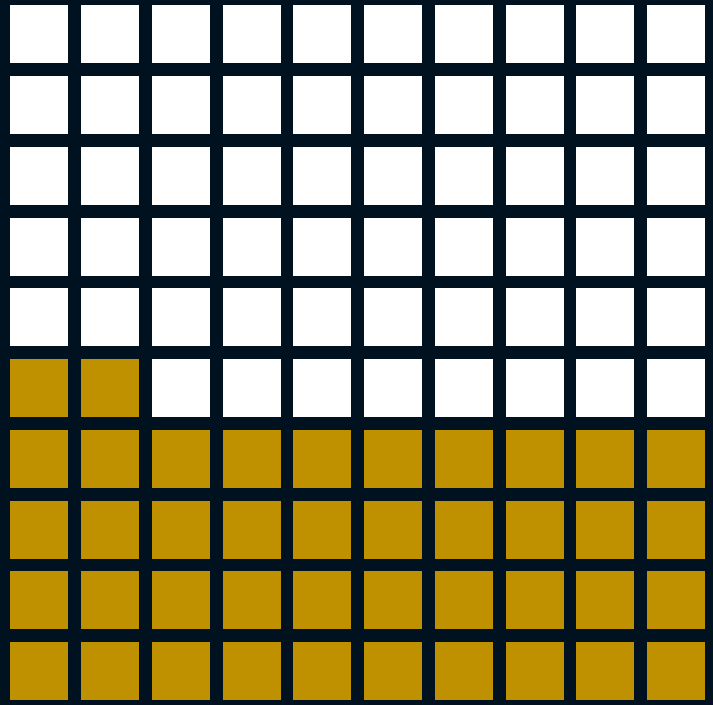


Microsoft Global Online Safety Survey (2023)

LexiGuard

Why is toxicity a problem?

42 % of people became less trusting



Microsoft Global Online Safety Survey (2023)

LexiGuard

LexiGuard

“Toxic online comments aren’t just hurtful for users, they’re also bad for business.”

Source: OpenWeb (2021)

LexiGuard

Our mission

Help to mitigate toxicity and ensure healthy dialogue online.

Solution

NLP system designed for detecting and flagging toxic comments by predicting a toxicity score.

LexiGuard

Raw Data

What's driving toxicity?

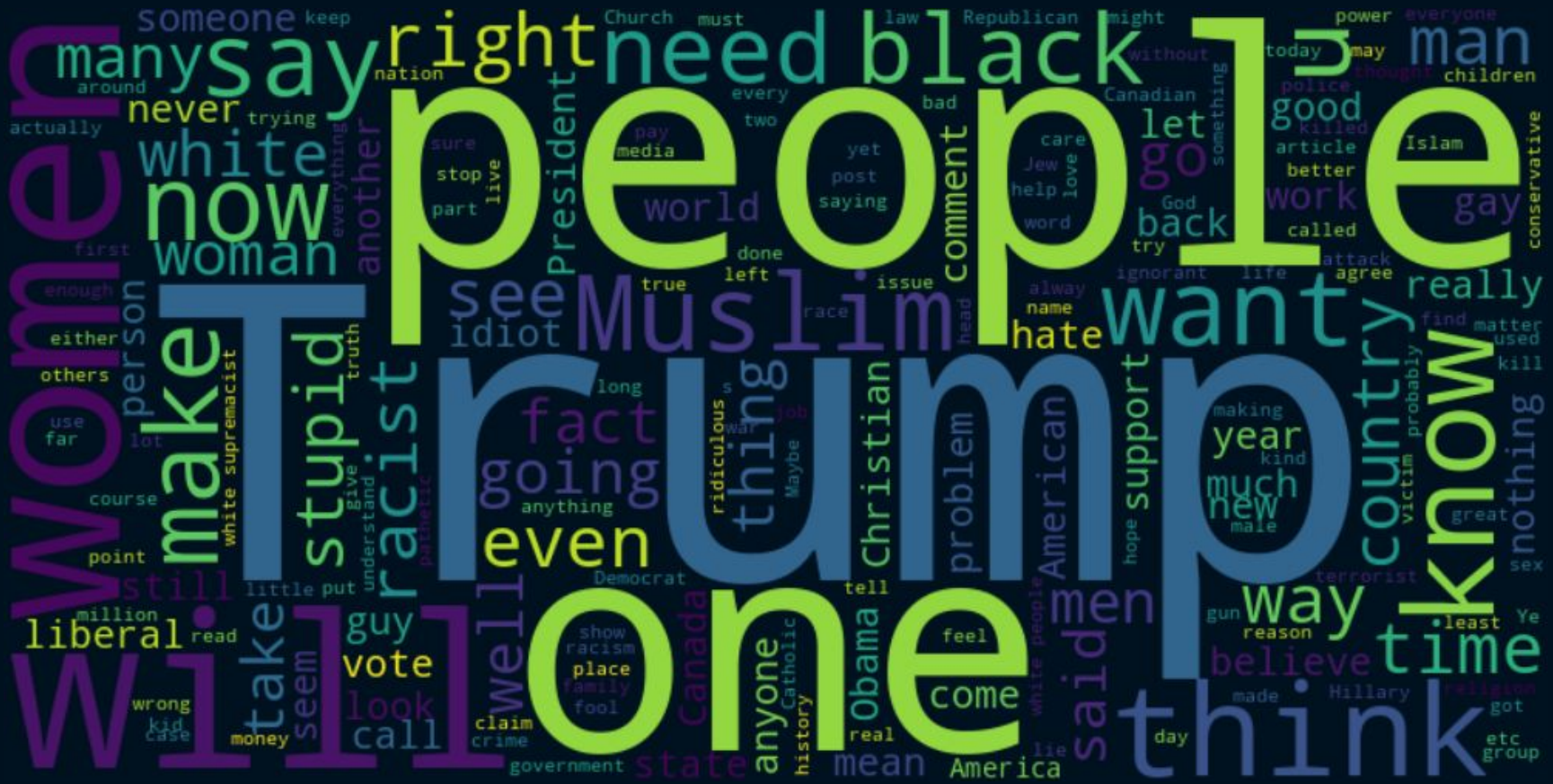
Black Lives Matter
2014

Charleston Church Shooting
2015

Trump elected US president
2016

Word Cloud

LexiGuard



Most Frequent Sequences in Toxic Comments

LexiGuard

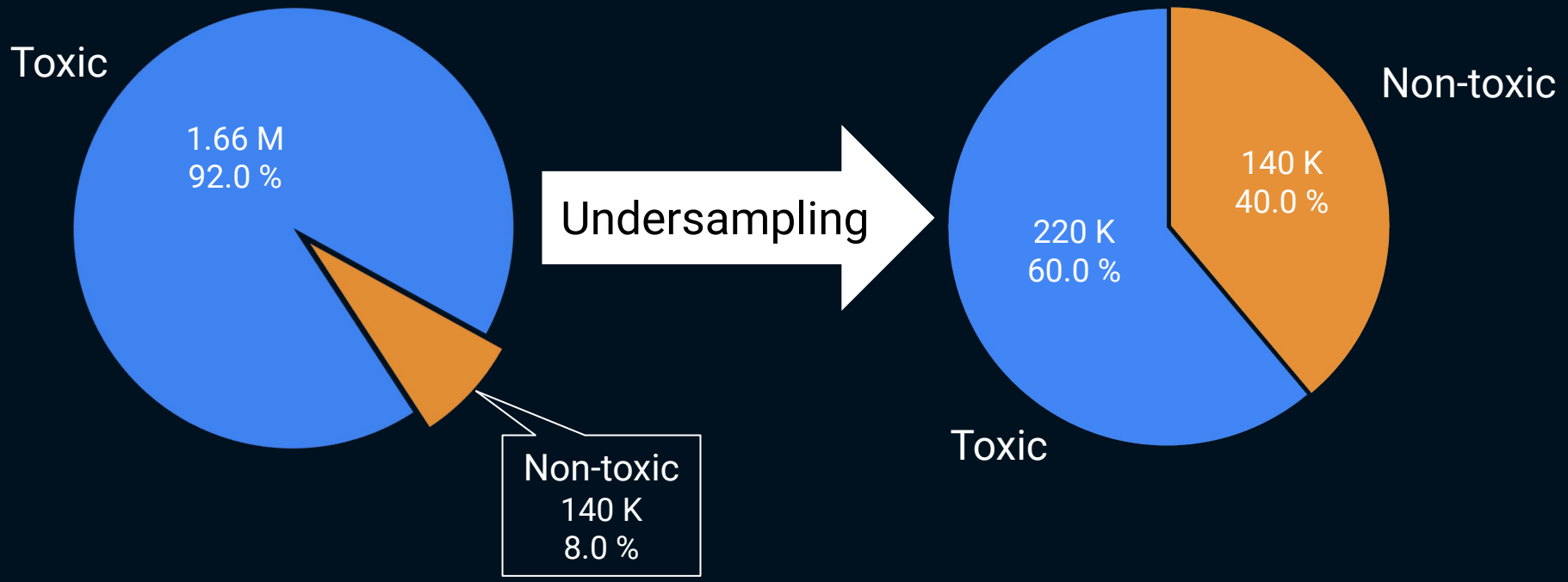
Word Pairs

white, supremacist
white, people
mental, ill
donald, trump
black, people
look, like
white, men
white, male
sexual, assault
white, house

Word Triplets

black, live, matter
sexual, assault, women
make, america, great
nazi, white, supremacist
president, unit, state
ha, ha, ha
racist, white, supremacist
could, care, less
lisa, bloom, also
liberal, mental, disorder

Toxic Comments vs Non-Toxic Comments



Baseline Model

Raw data

Text vectorization algorithm:
Bag of words

Classifier:
Logistic regression

Performance Metric

LexiGuard

Recall

(= how many of true toxic comments are detected in total)

Precision

(= how many of the comments predicted as toxic are really toxic)

F1 Score

(= harmonic mean of precision and recall)

Baseline Model Performance

Recall
0.441

Precision
0.681

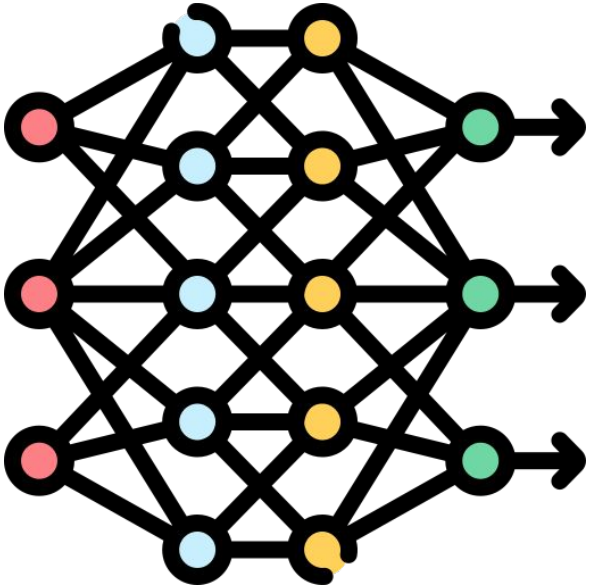
F1
0.536

Data Cleaning and Pre-Processing

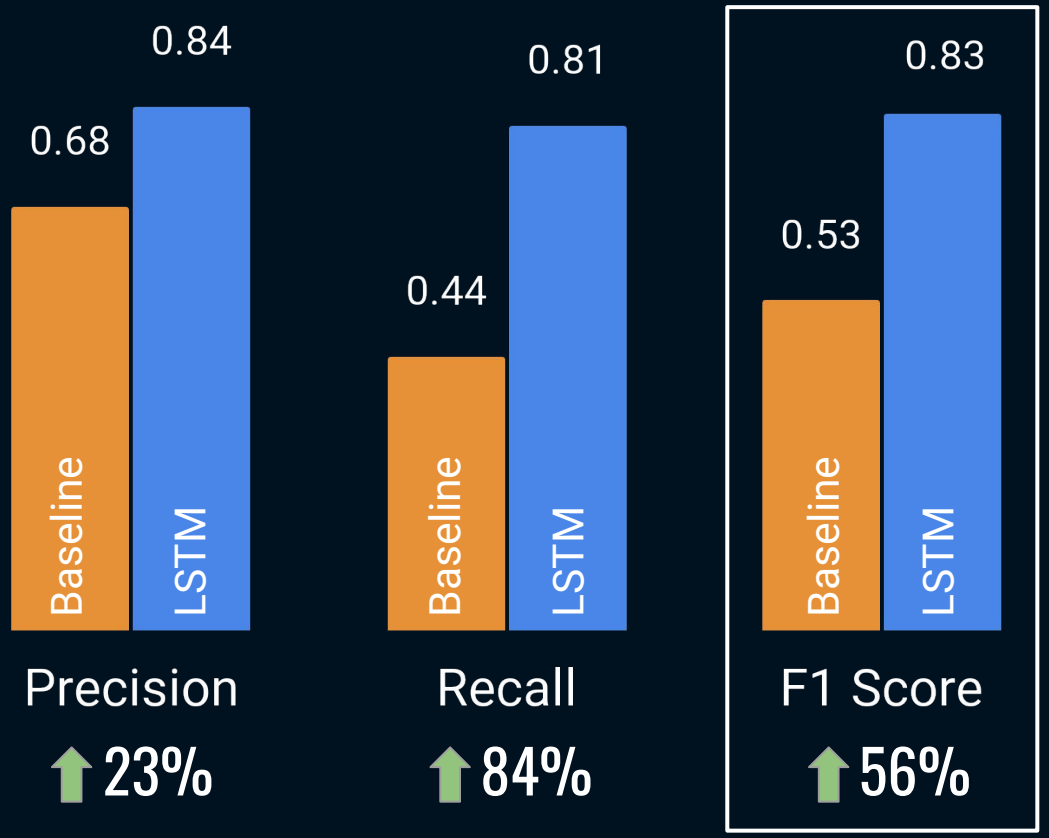
- Standard procedure: remove HTML tags, URLs, punctuation, special characters etc.
- “Normalize” creative spellings:
  yuuuge → huge
  f*ck → fuck
- Lemmatize words

LSTM
Long Short-Term Memory

Captures complex sequential patterns in data, while logistic regression handles simpler linear relationships between features.



LSTM Performance



LSTM outperforms logistic regression.

F1 score is boosted by 56 %.

LexiGuard

Next Steps and Further Improvements

LexiGuard

1. Integrate trained model into target applications.

2. Address potential biases and unintended consequences using **more complex models** which might capture **textual nuances and bias** more **effectively**.

3. **Combine predictions** from multiple models to **improve prediction performance and reliability**.

Dashboard Prototype (Streamlit)

LexiGuard

Toxicity Prediction!

LexiGuard

Real corruption exists - but it's not at these fundraisers. Private meetings take place all the time, with NO money changing hands. Lobbyist meet privately as a matter of practice (it's their job to get access).

Predict

Toxicity prediction: 13.22 %

There are bunch of losers!

Predict

Toxicity prediction: 49.65 %

I'm all for Sharia law. Finally women will have to shut the hell up. I want to have multiple wives by which all I need to get one is trade a donkey. Trudeau bringing in all this scum from the middle east will be his downfall I say bring it. It's only going to cause violence and then the whole world will see how weak and stupid Canadians really are. Enjoy being the next Sweden. I'm moving to USA.

Predict

Toxicity prediction: 93.6 %

LexiGuard

Thank you!

Obrigado!

André Oliveira

धन्यवाद!

Purvi Parmar

Danke!

Michael Schickenberg

¡Gracias!

Eric Martinez