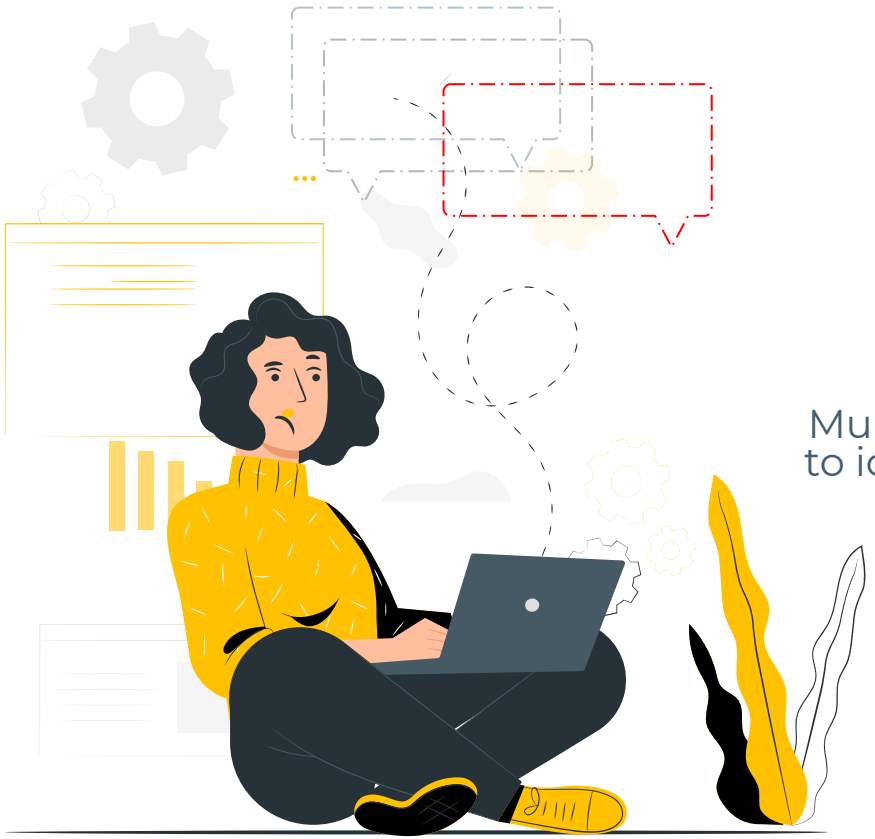


Discussion Detox

Multilingual Machine Learning algorithms to identify toxic comments on the internet

Author: Drenizë Rama

Hasskommentare
im Netz
identifizieren

Discussion Detox

App

by Drenizë Rama

TABLE OF CONTENTS

“INTERNET RULE #1:
Never read the comments.”

— WIRED

02 | the data

Example comment #1:

"What a motherfucking piece of crap those fuckheads for blocking us!"

Example comment #2:

"Hey, faggot.
You fucking retard. You better quit undoing my vandalism, bitchboy."

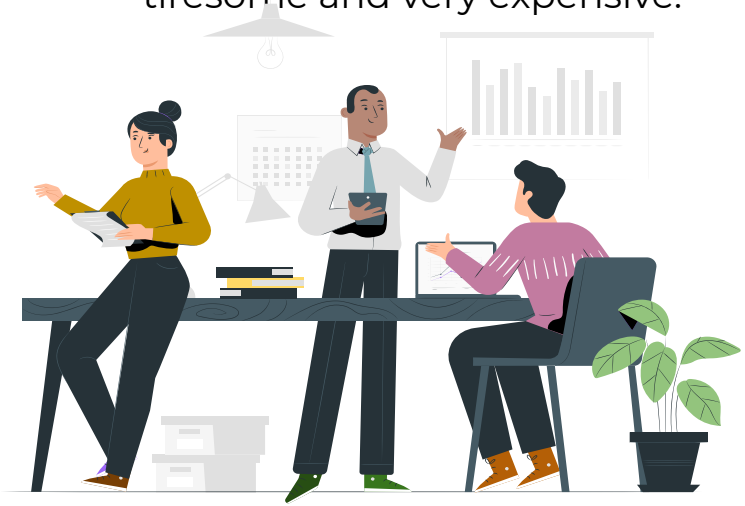
Example comment #3:

"but ew
He was a fαg which is against nature and is the most disgusting thing. Youre not a woman are you? Sexism is wrong. Being wrong is for women."

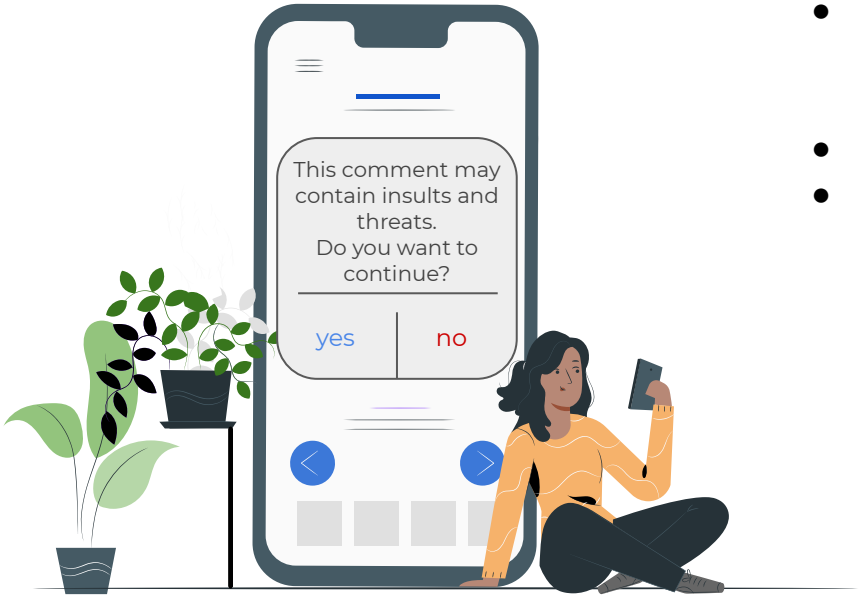
01 | introduction

An **online newspaper** or a **social media web host** wants to keep the discussions under each article clean and respectful.

However, going through every comment manually is tiresome and very expensive.



01 | introduction

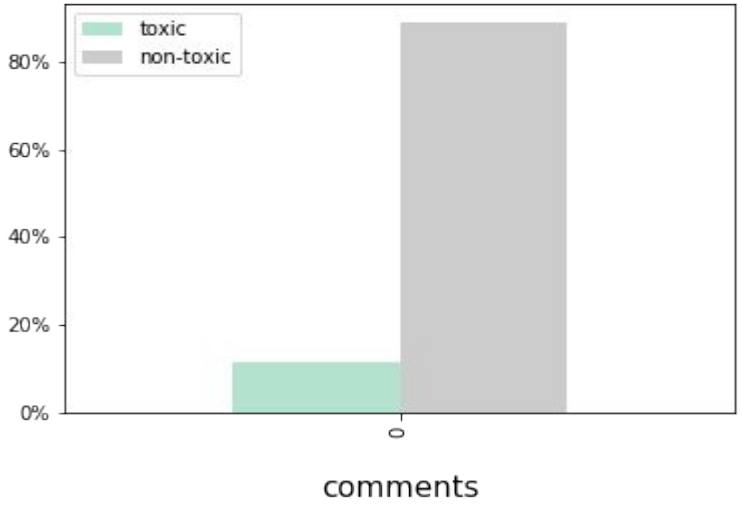


goal

- build a natural language algorithm that classifies social media comments into toxic and non-toxic categories
- at a low cost
- In different languages

02 | the data

The overall amount of toxic comments is quite low

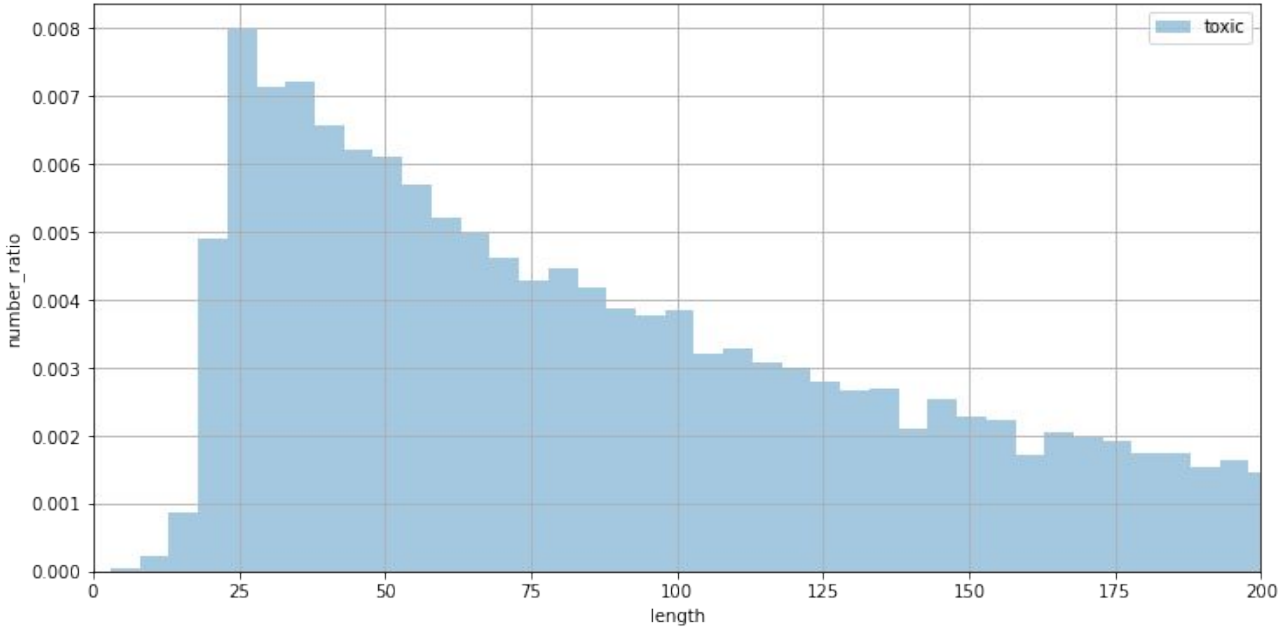


- Data was **provided by Google** and Jigsaw
- **Publication** dates of the comments range from **2015 to 2017**
- **223,549** comments in train set

Disclaimer: The dataset for this project contains text that may be considered profane, vulgar, or offensive.

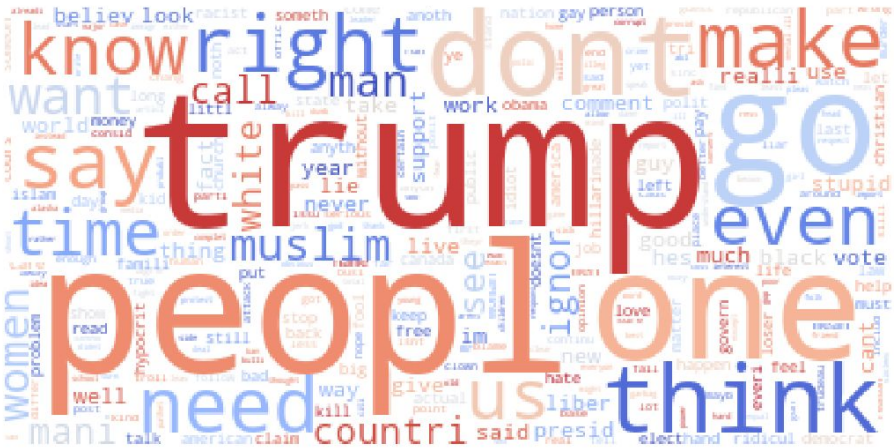
02 | the data

The more hate - the shorter the comments



frequently used words

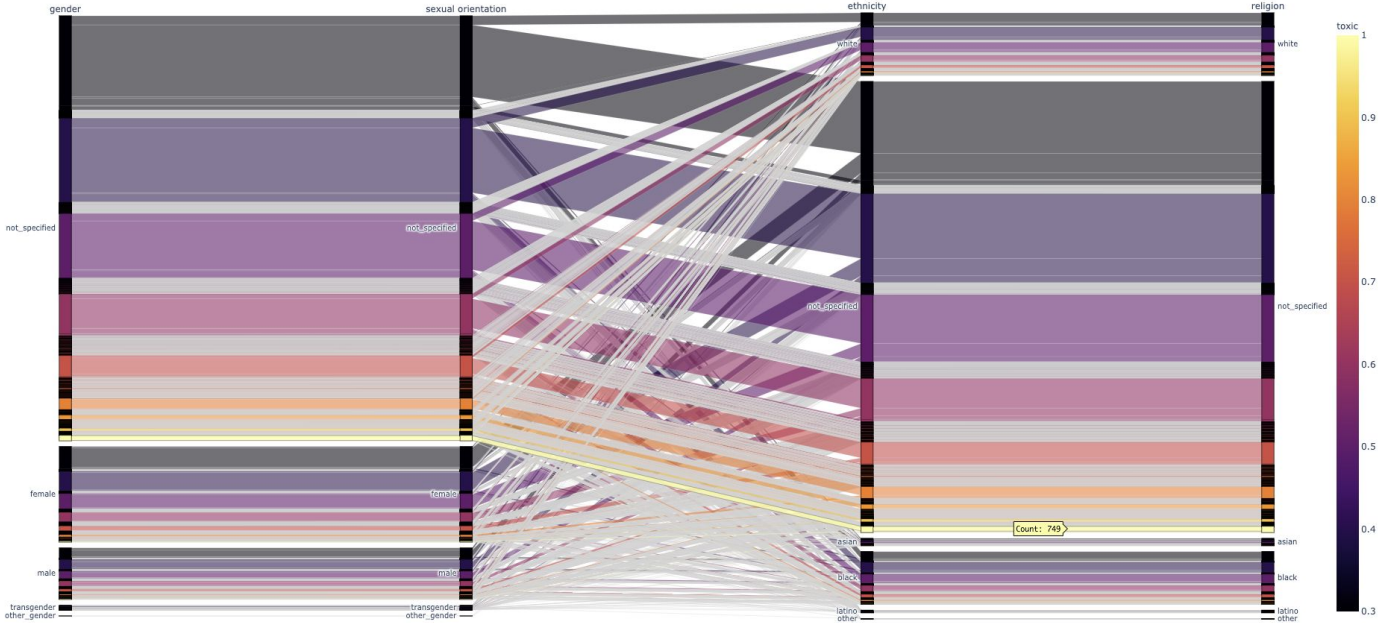
In toxic comments



In comments containing identity attack

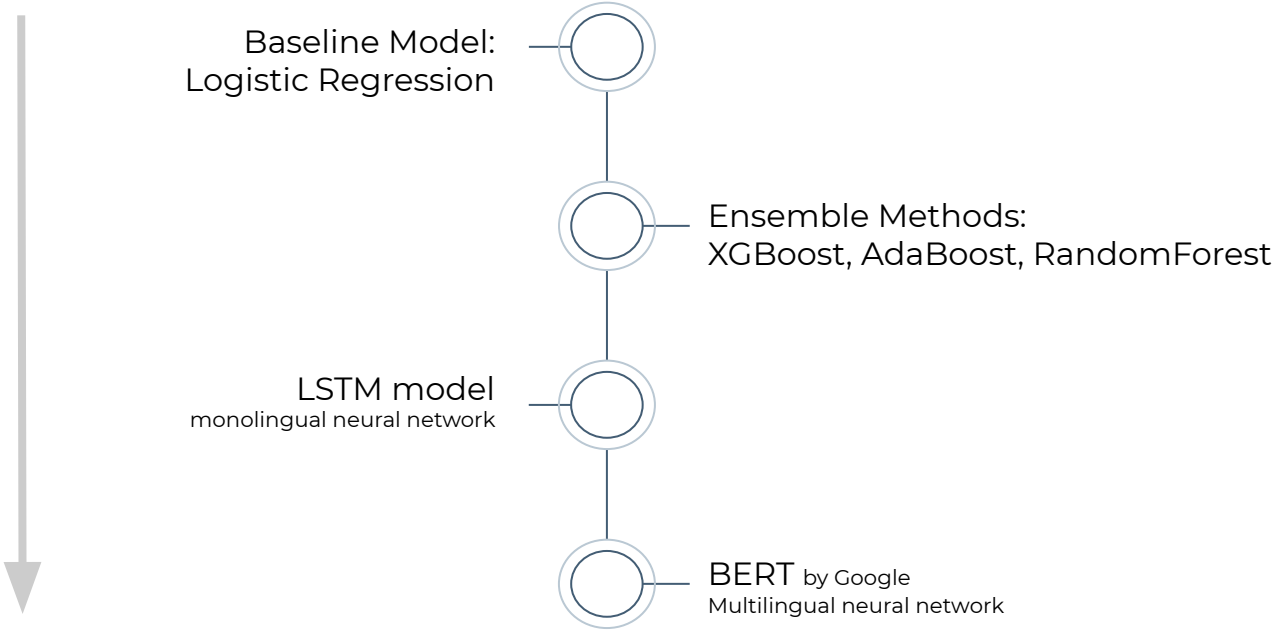
02 | the data

Who’s getting the most hate?

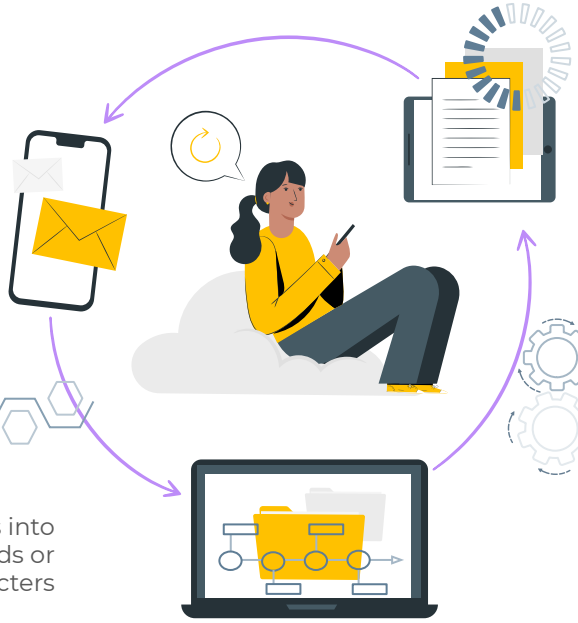


03 | methods

Additional power and complexity



How Natural Language Processing works



Text Cleaning

Stemming, remove symbols, emojis, punctuation, stopwords

Train Model

Build Model

Adjust type, layers, epochs etc.

Tokenize

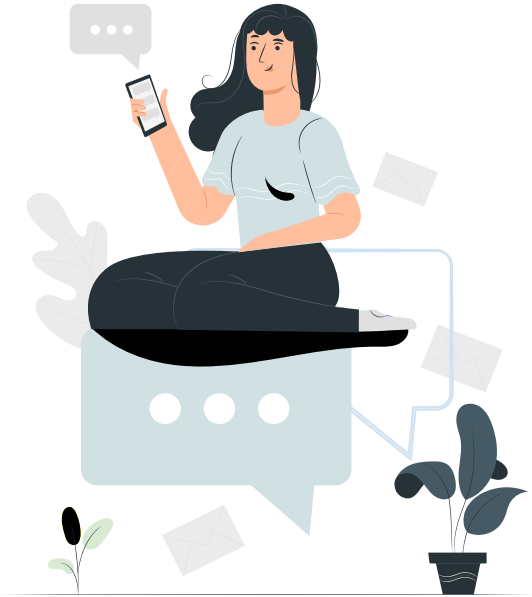
Divide sentences into sequences of words or characters

Encode

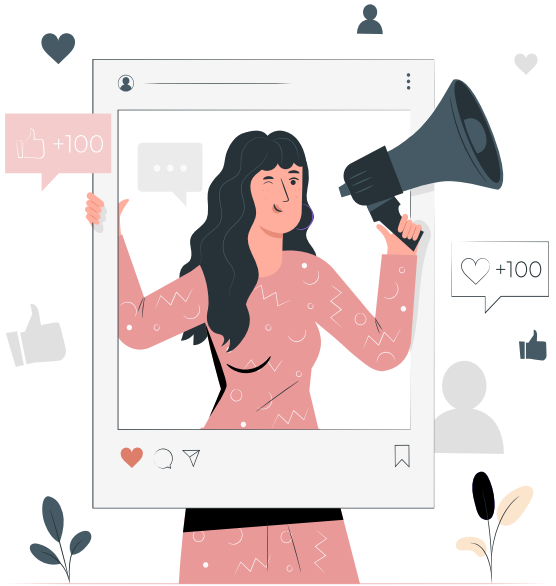
04 | results

Which one is the most efficient model?

05 | recommendations



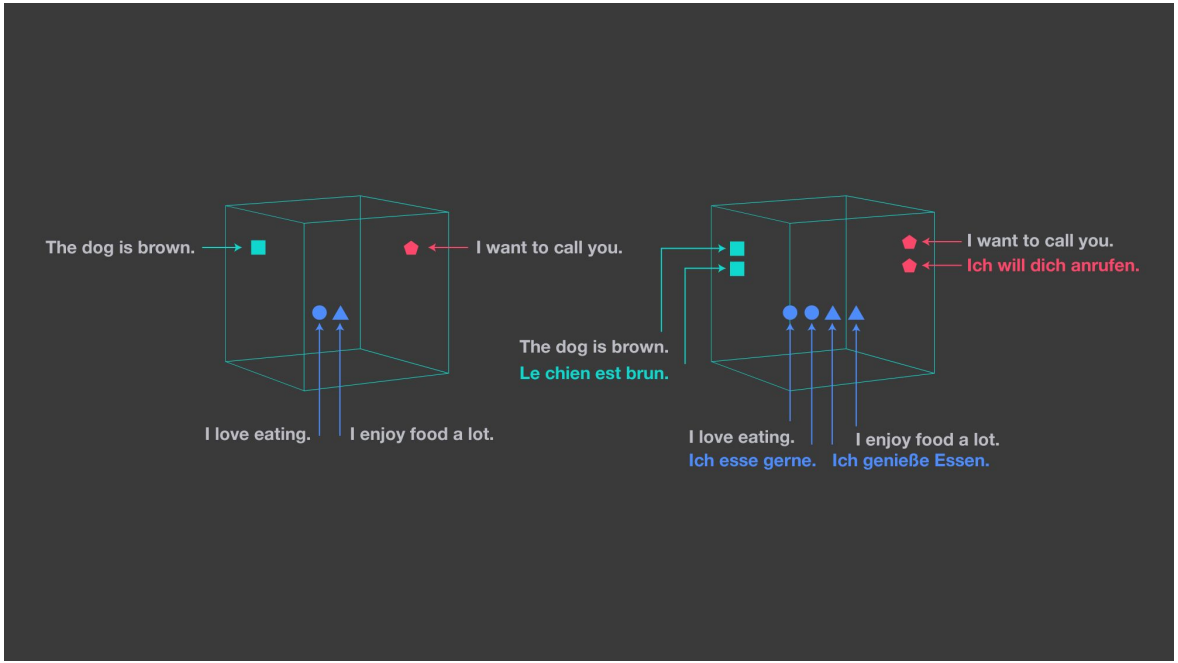
06 | future work



- time-series analysis of toxicity
- compare sentiments of trending topics among different languages
- Create web tool that recommends users to adjust their language before posting a comment

Future work

text vectorization throughout different languages

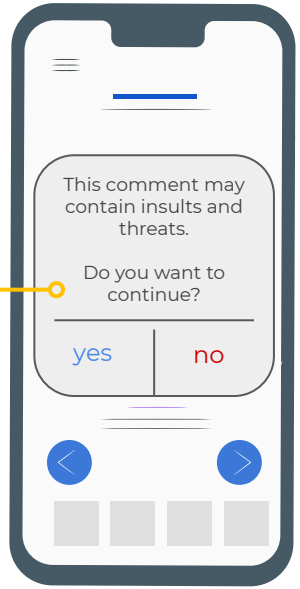


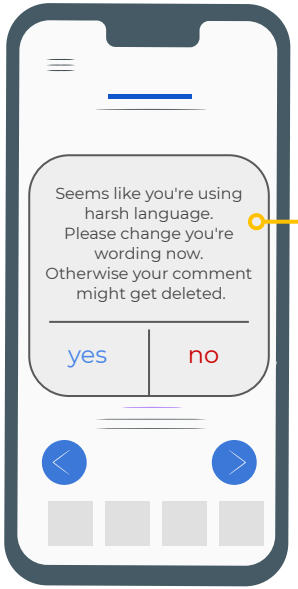
06 | future work

Approach 1

Approach 2

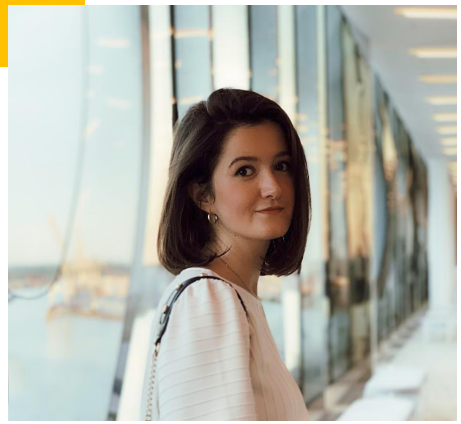
protect users from toxic content





mirror toxic behavior in advance

Thank you



Drenizë Rama

Data Scientist

https://drenize.github.io/

https://www.linkedin.com/in/dreniz%C3%AB-rama-6121a4157/

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, infographics & images by Freepik and illustrations by Stories

Appendix

Wie funktioniert

- NLP grafiken
- BERT
- LASER
- zwischenergebnisse

Baseline Model: Logistic Regression

```
[[49808   853]
 [ 1991  3236]]
```

	precision	recall	f1-score	support
0	0.96	0.98	0.97	50661
1	0.79	0.62	0.69	5227
accuracy			0.95	55888
macro avg	0.88	0.80	0.83	55888
weighted avg	0.95	0.95	0.95	55888