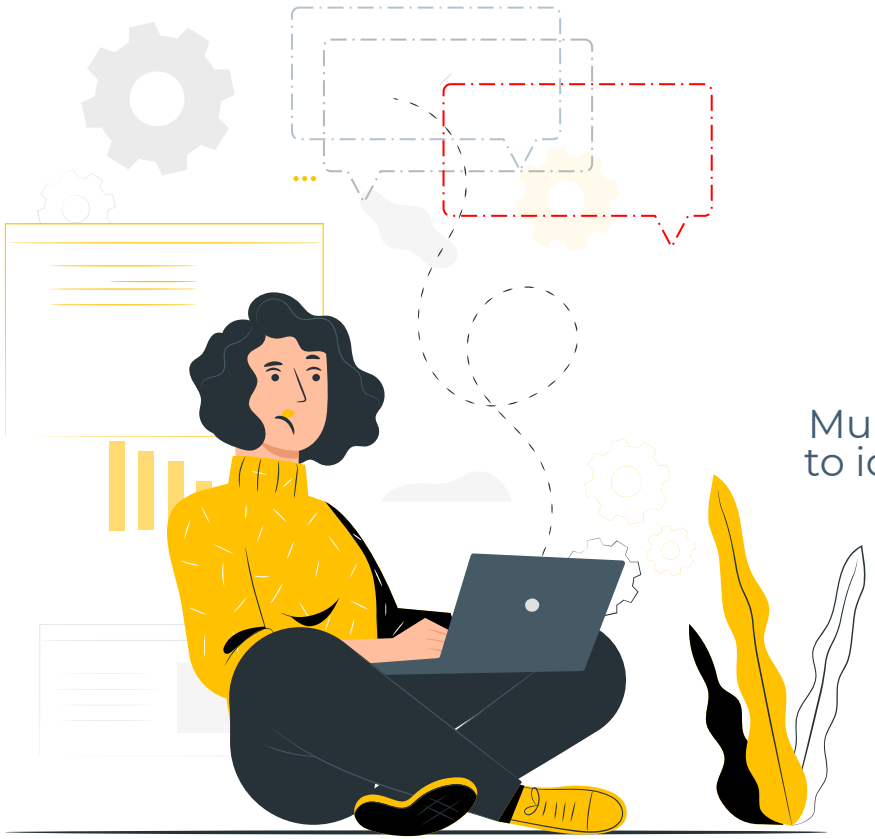


Discussion Detox

Multilingual Machine Learning algorithms to identify toxic comments on the internet

Author: Drenizë Rama

TABLE OF CONTENTS

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

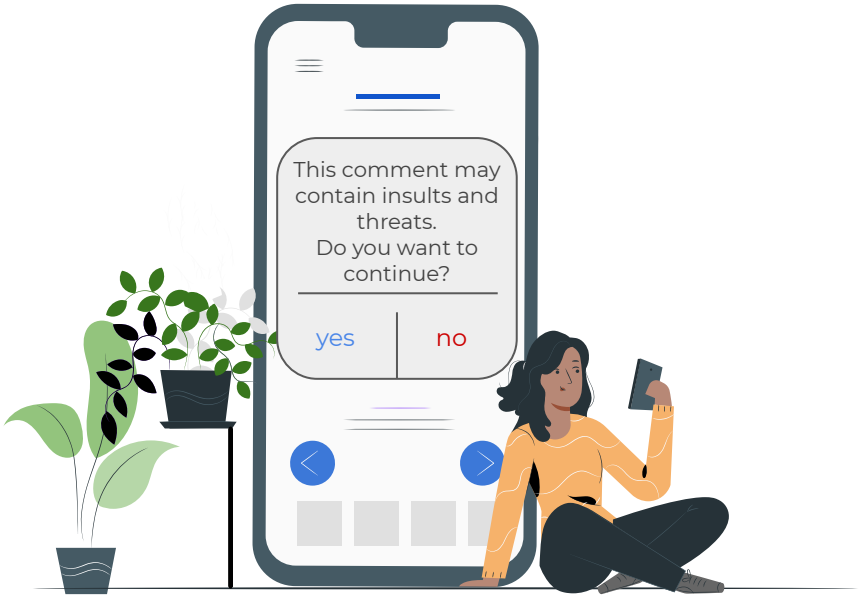
— WIRED

01 | introduction

An online newspaper wants to keep the discussions under each article clean and respectful.

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

01 | introduction

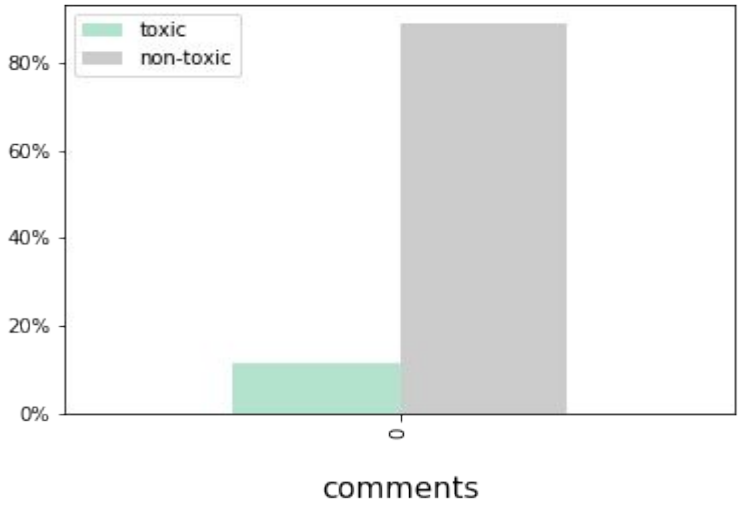


The **goal** of this project is to build a natural language algorithm that classifies text input (social media comments) into toxic and non-toxic categories in one or more languages.

In order to achieve this, I decided to go with pretrained RNN models like BERT or LASER, which include a variety of languages.

02 | the data

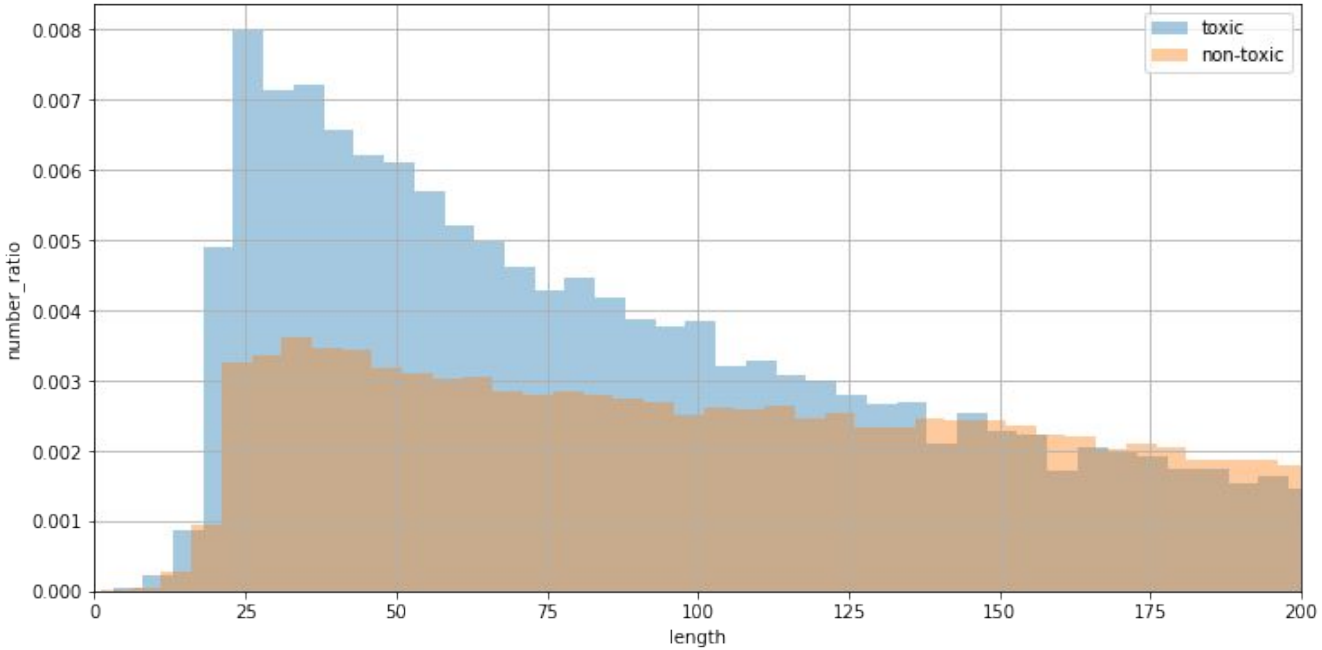
Distribution of Classes



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

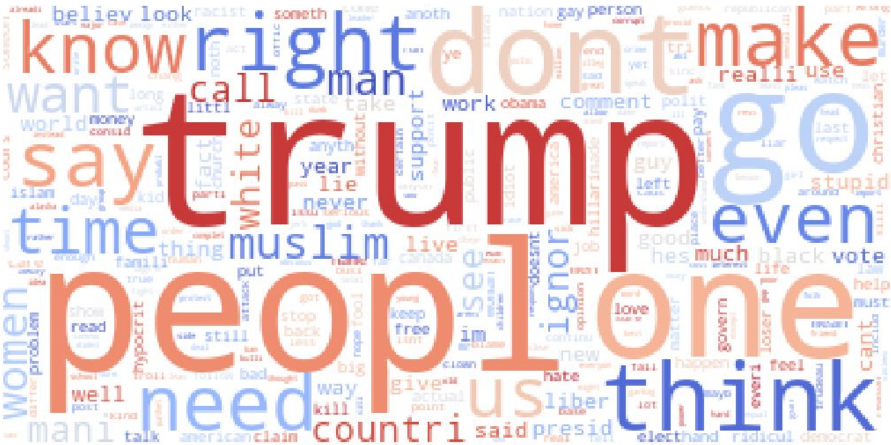
Disclaimer: The dataset for this project contains text that may be considered profane, vulgar, or offensive.

Length of comment per class



frequently used words

overall toxic comments



comments containing identity attack

02 | the data

Example comment #1:

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

Example comment #2:

"yeah so, whoever wrote that big shit about myles, im going to come round to your shitty little house and stuff a petrol bomb through you fucking shitty little cunt of a letterbox. and yes i do know who you and your inbred parent cunts are. i know where you live. i know your parents mobile phone numbers. i know where they work. be afraid, be very afraid.
and i will find a way to stick a fat off bottle of lit jack daniels thru your door.
so fuck off"

Example comment #3:

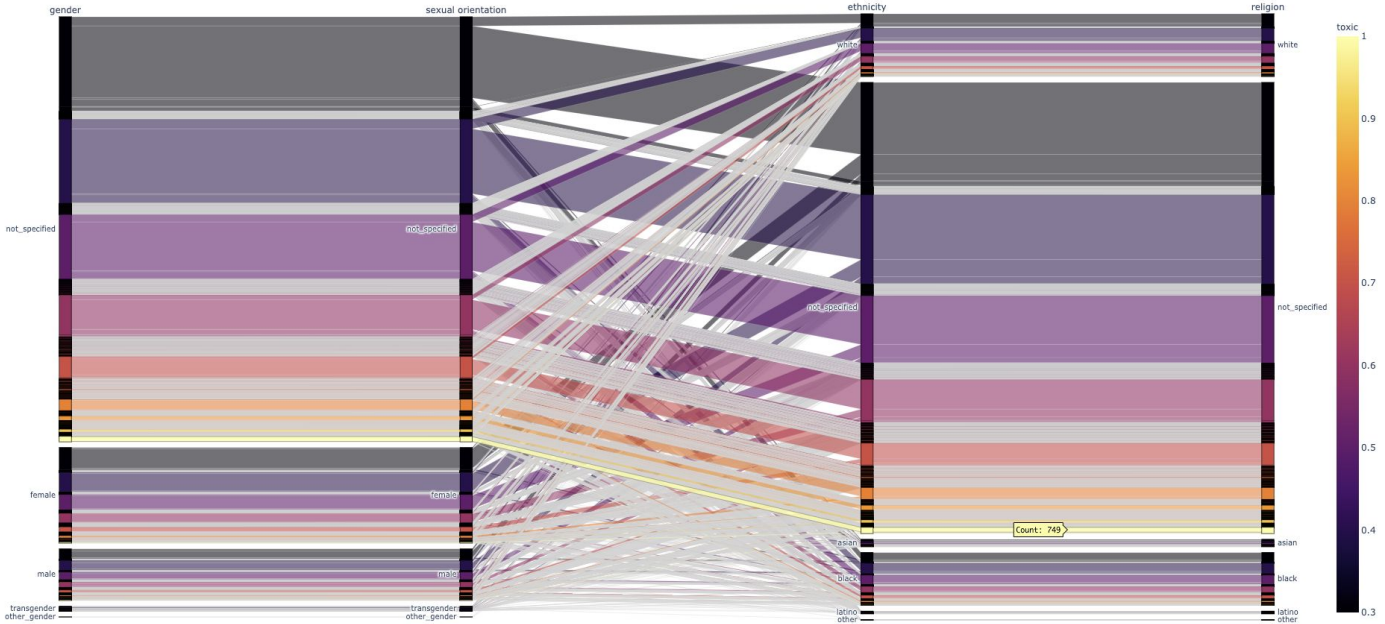
"Hey, faggot.
You fucking retard. You better quit undoing my vandalism, bitchboy."

Example comment #4:

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

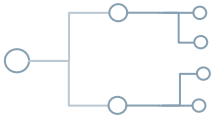
02 | the data

Parallel Categories by Toxicity of Comments

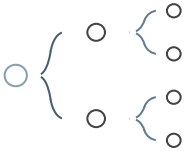


file:///Users/student/nf-may-2020/Capstone/Notebooks%20aus%20der%20Cloud/parallelcat3.html

03 | methods



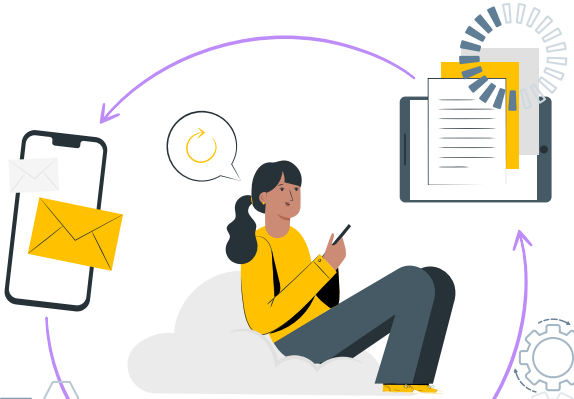
Long Short Term Memory neural network



Ensemble Methods

Recurrent Neural Network with BERT

Natural Language Processing



Text Cleaning

Stemming, remove symbols, emojis, punctuation, stopwords

Train Model

Build Model

Adjust type, layers, epochs etc.

Encode

Tokenize

Divide sentences into sequences of words or characters

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

Ensemble methods

XGBClassifier:

Confusion Matrix :
[[150270 1382]
[8515 7494]]

Accuracy Score : 0.9409701719541217
Report :

	precision	recall	f1-score	support
0	0.95	0.99	0.97	151652
1	0.84	0.47	0.60	16009
accuracy			0.94	167661
macro avg	0.90	0.73	0.79	167661
weighted avg	0.94	0.94	0.93	167661

AdaBoostClassifier:

Confusion Matrix :
[[149526 2126]
[8388 7621]]

Accuracy Score : 0.9372901271017112
Report :

	precision	recall	f1-score	support
0	0.95	0.99	0.97	151652
1	0.78	0.48	0.59	16009
accuracy			0.94	167661
macro avg	0.86	0.73	0.78	167661
weighted avg	0.93	0.94	0.93	167661

RandomForestClassifier:

Confusion Matrix :
[[150664 988]
[9077 6932]]

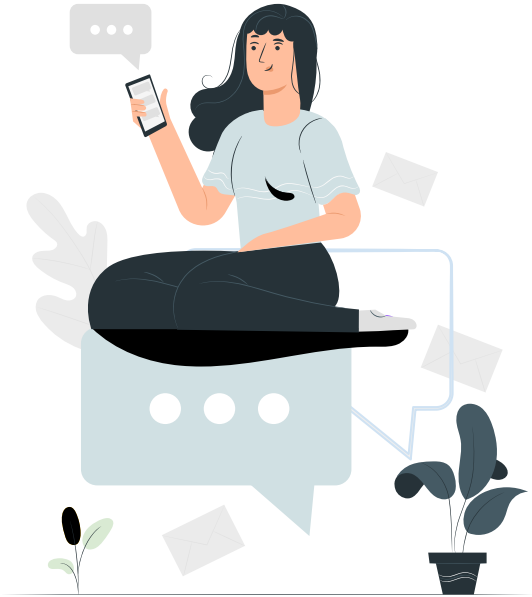
Accuracy Score : 0.9399681500169986
Report :

	precision	recall	f1-score	support
0	0.94	0.99	0.97	151652
1	0.88	0.43	0.58	16009
accuracy			0.94	167661
macro avg	0.91	0.71	0.77	167661
weighted avg	0.94	0.94	0.93	167661

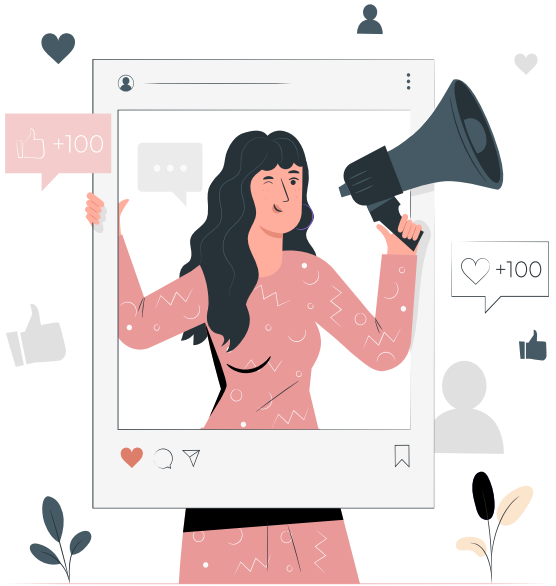
LSTM model

BERT model

05 | recommendations



06 | future work



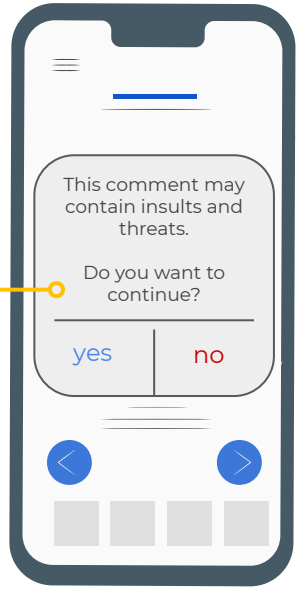
- time-series analysis of toxicity
- compare sentiments of trending topics among different languages
- Build an application that analyses comment threads
- Create web tool that recommends users to adjust their language before posting a comment
- Analyse psychological structures of language for different mental states

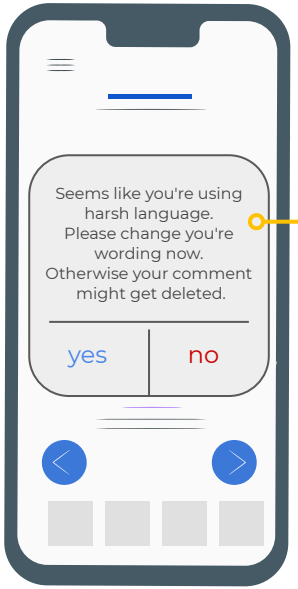
06 | future work

Approach 1

Approach 2

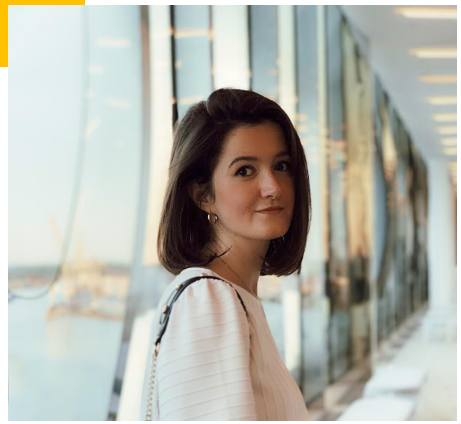
protect users from toxic content





mirror toxic behavior in advance

Thank you



Drenizë Rama

Data Scientist

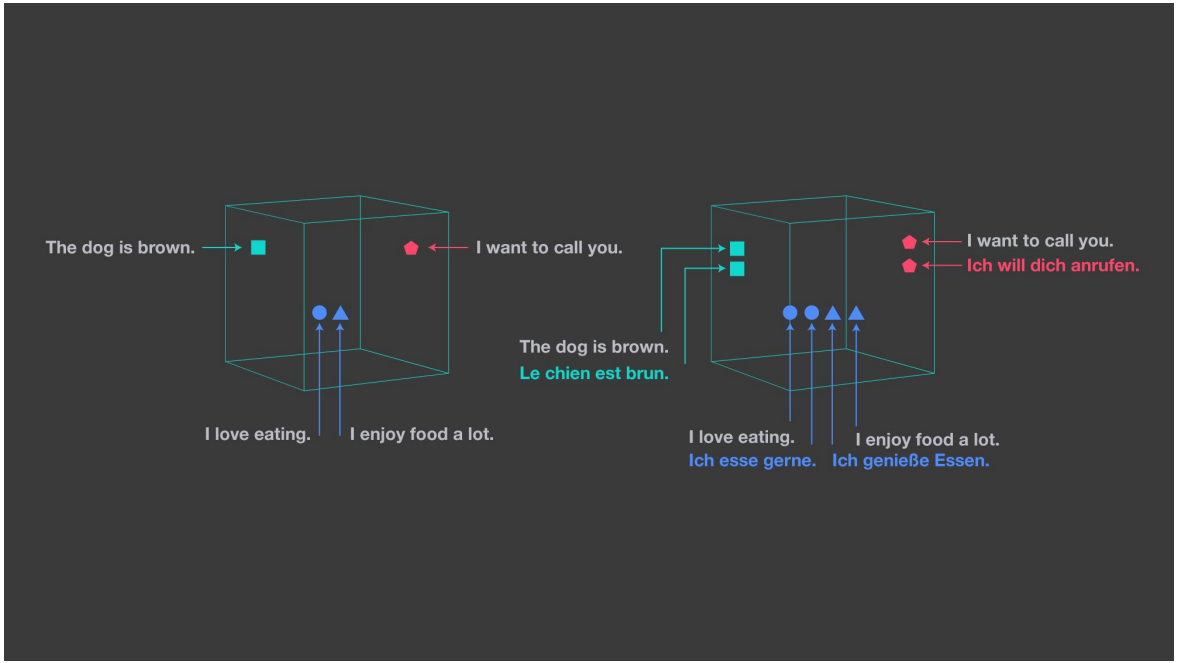
https://drenize.github.io/

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

text vectorization throughout different languages



Appendix

Wie funktioniert

- NLP grafiken
- BERT
- LASER
- zwischenergebnisse