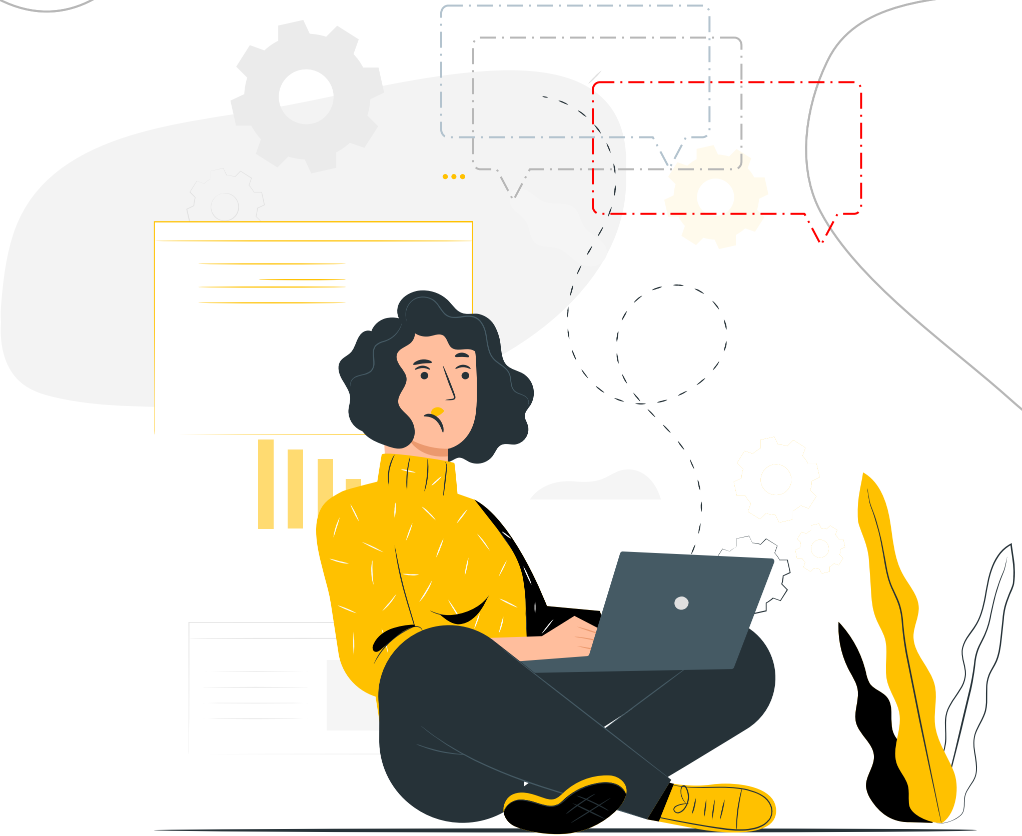


Discussion Detox

Multilingual Machine Learning algorithms to identify toxic comments on the internet

Author: Drenizë Rama

TABLE OF CONTENTS

Introduction 01

Data 02

Methods 03

Results 04

Recommendations 05

Future Work 06

> **“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."

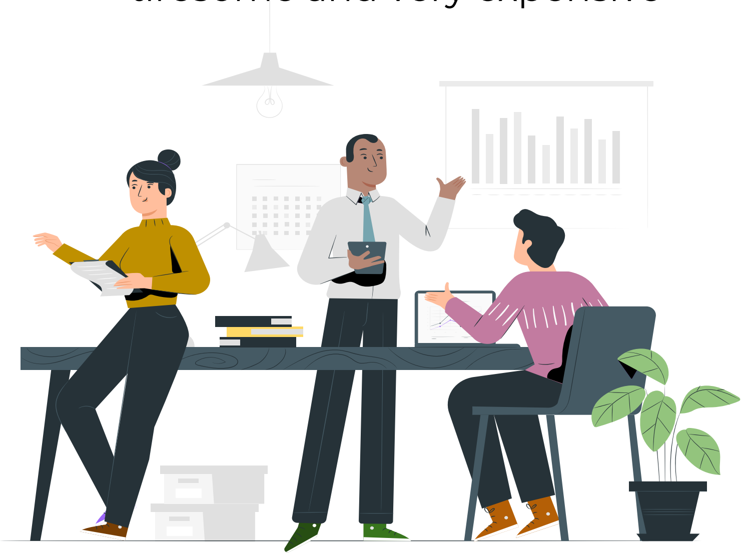
Example comment #3:

"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."

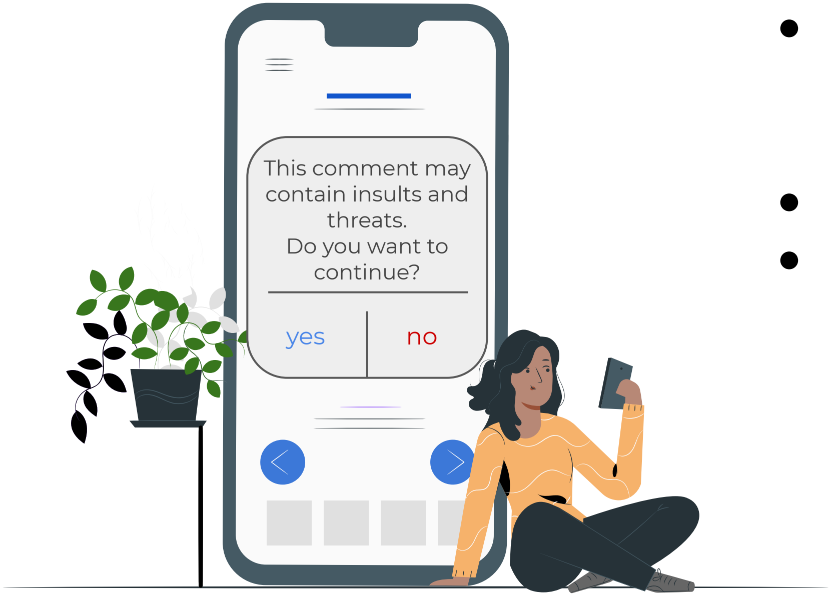
01 | introduction

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

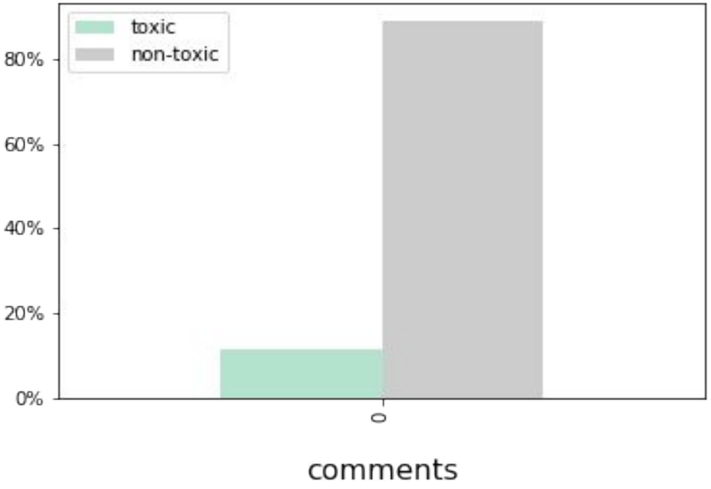

This comment may contain insults and threats.
Do you want to continue?

goal

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

02 | the data

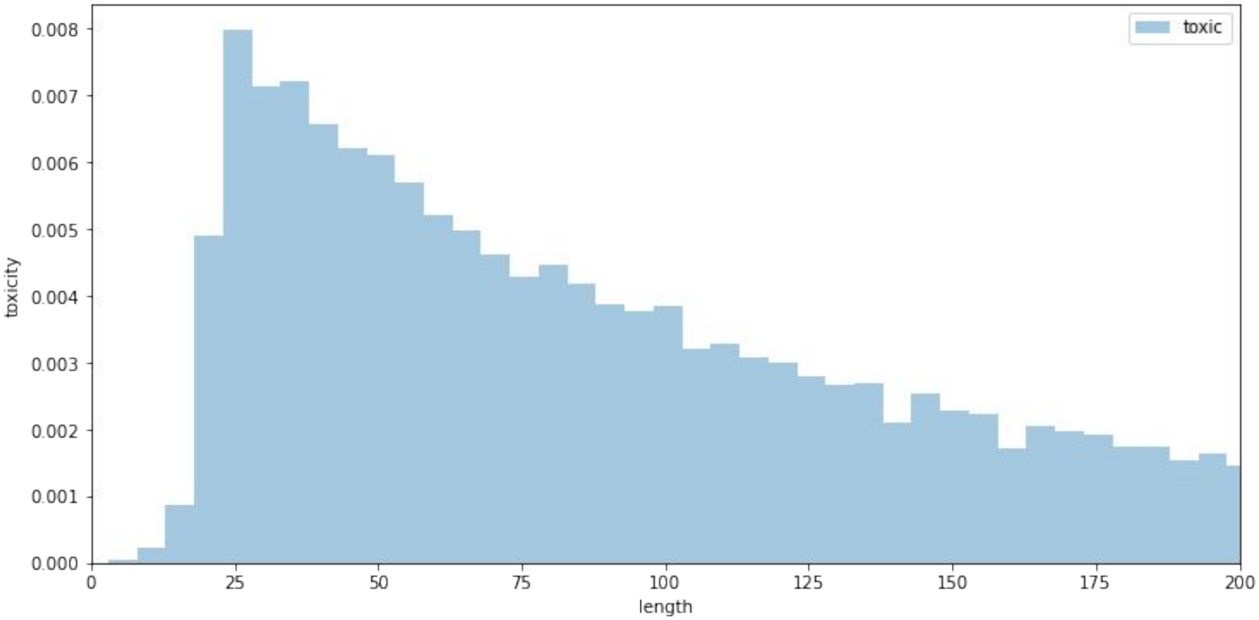
The overall amount of toxic comments is 11.4%



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

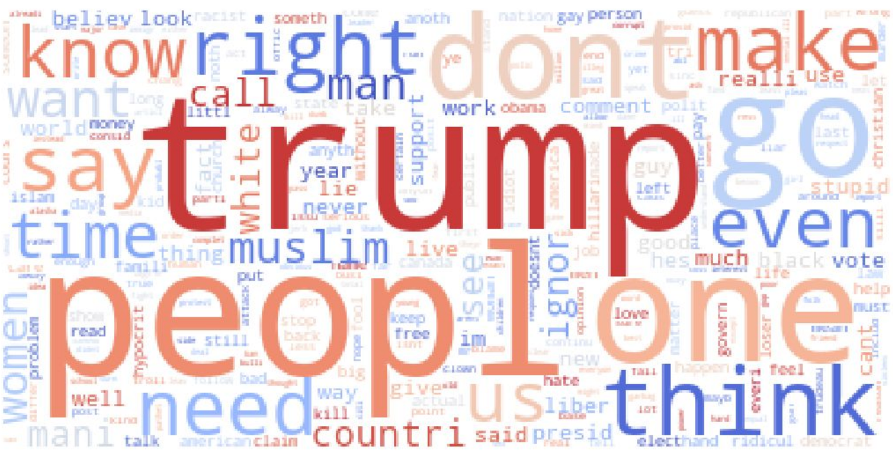
02 | the data

The more hate - the shorter the comments

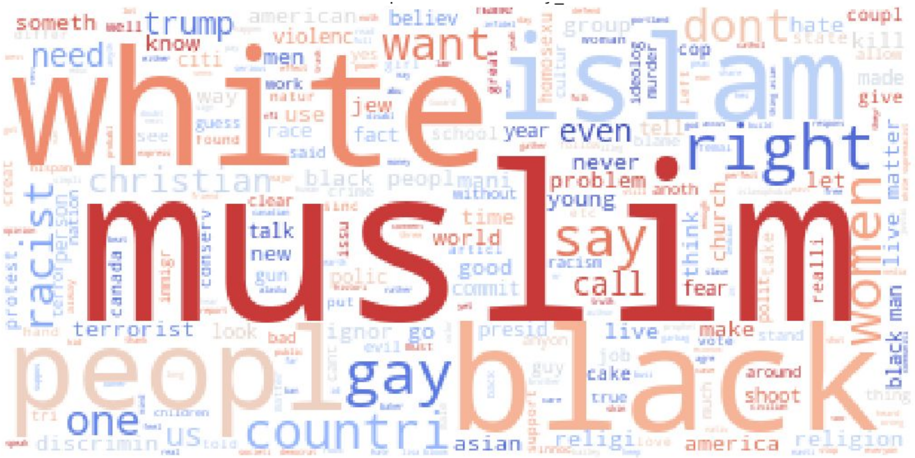


frequently used words

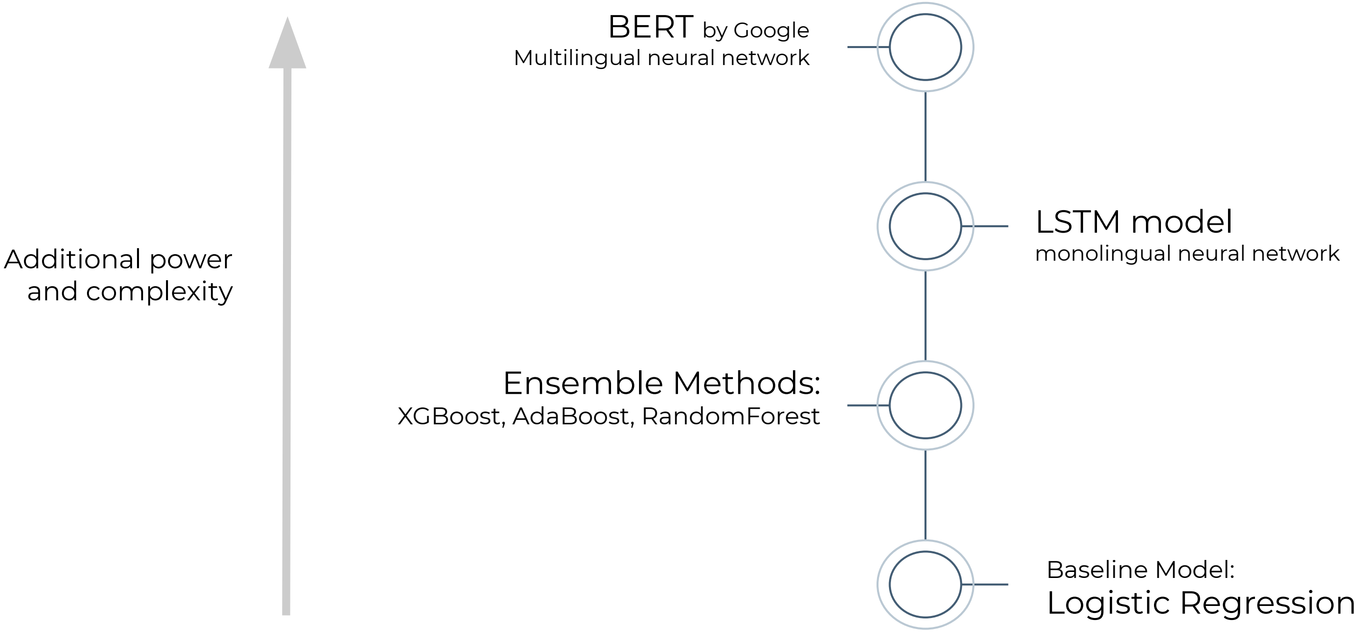
In toxic comments



In comments containing identity attack



03 | methods



BERT by Google
Multilingual neural network

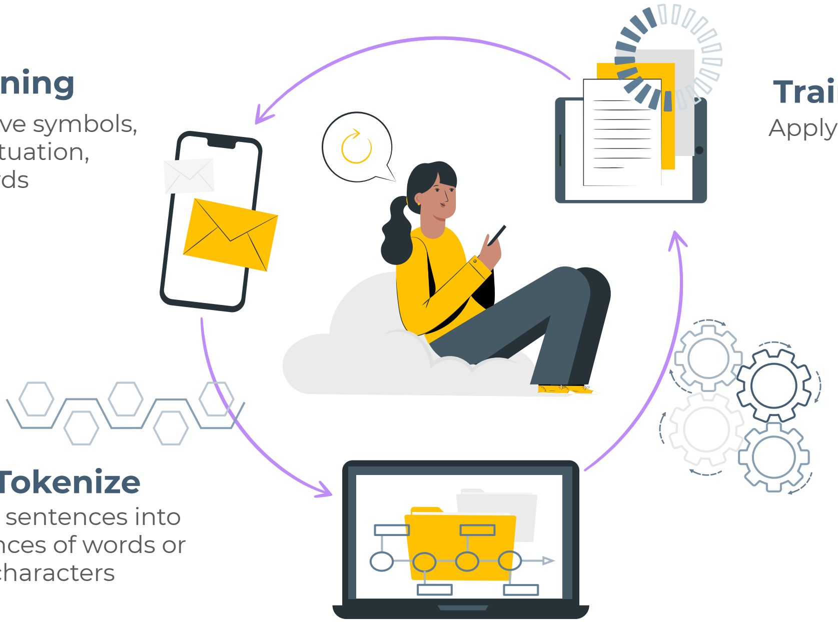
LSTM model
monolingual neural network

Additional power
and complexity

Ensemble Methods:
XGBoost, AdaBoost, RandomForest

Baseline Model:
Logistic Regression

How Natural Language Processing works



Text Cleaning

Stemming, remove symbols, emojis, punctuation, stopwords

Train Model

Apply the model in data

Tokenize

Divide sentences into sequences of words or characters

Build Model

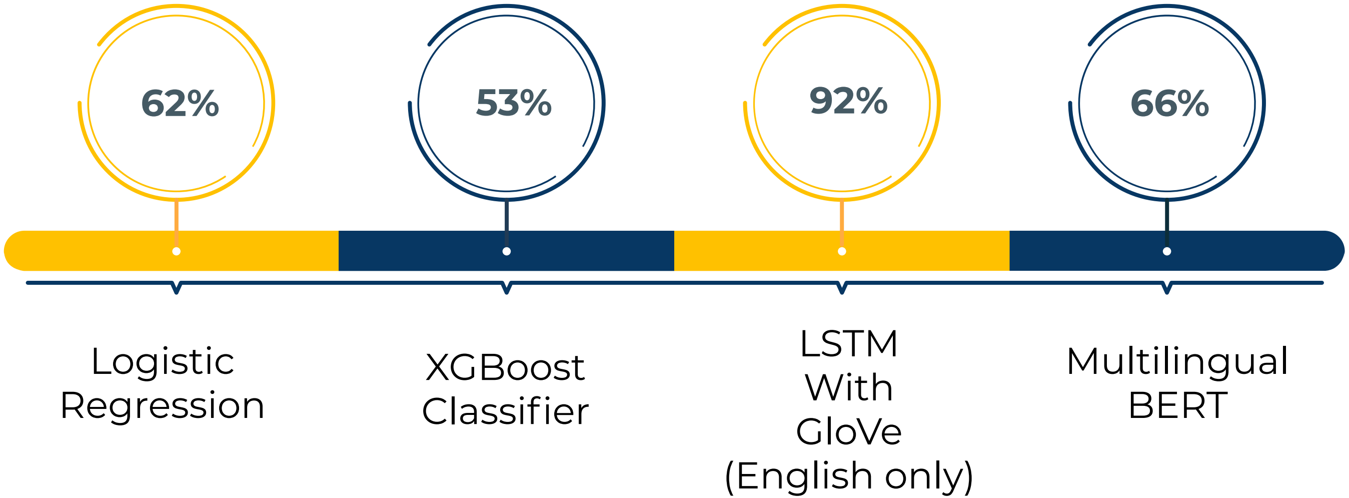
Adjust type, layers, epochs etc.

Encode

Create vectors for word representation

04 | results

At what rate do the models predict toxic comments correctly? (recall)



62%

Logistic Regression

53%

XGBoost Classifier

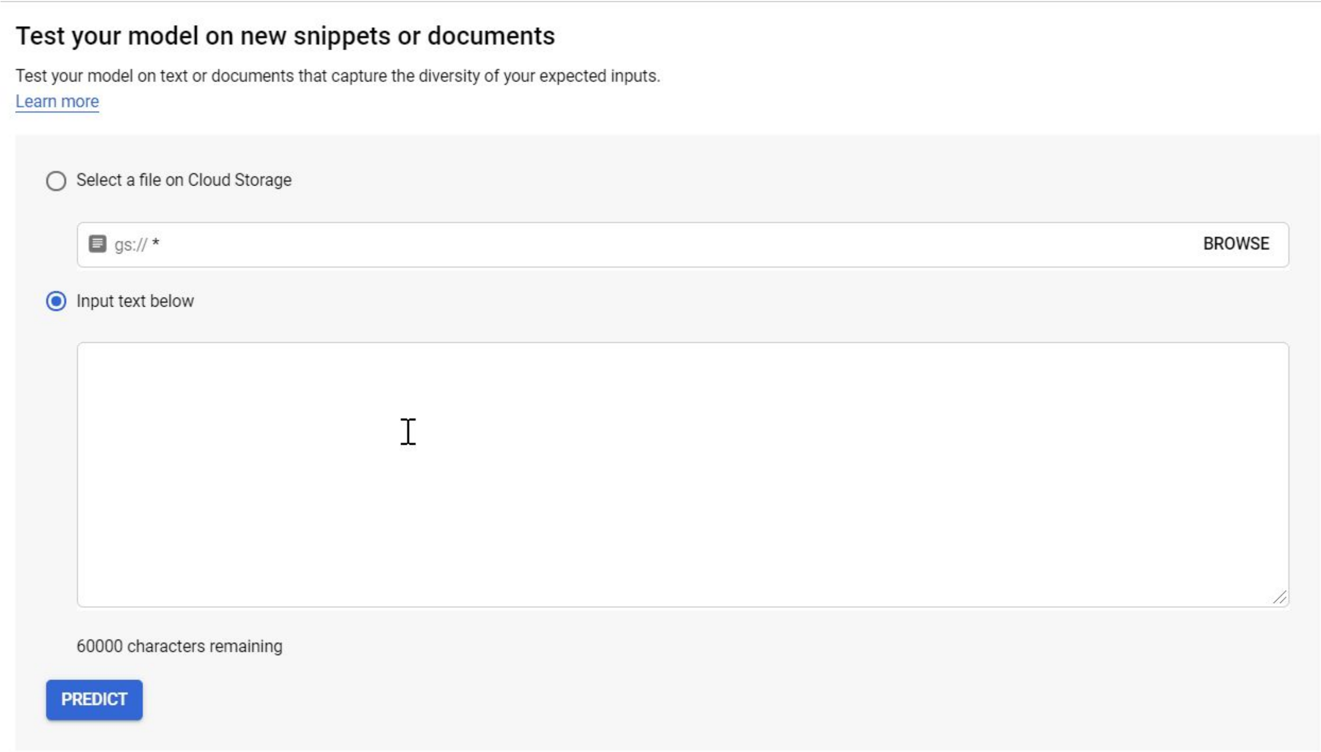
92%

LSTM With GloVe (English only)

66%

Multilingual BERT

Multilingual Toxic Comments Classifier on Google Cloud Platform



Test your model on new snippets or documents

Test your model on text or documents that capture the diversity of your expected inputs.
Learn more

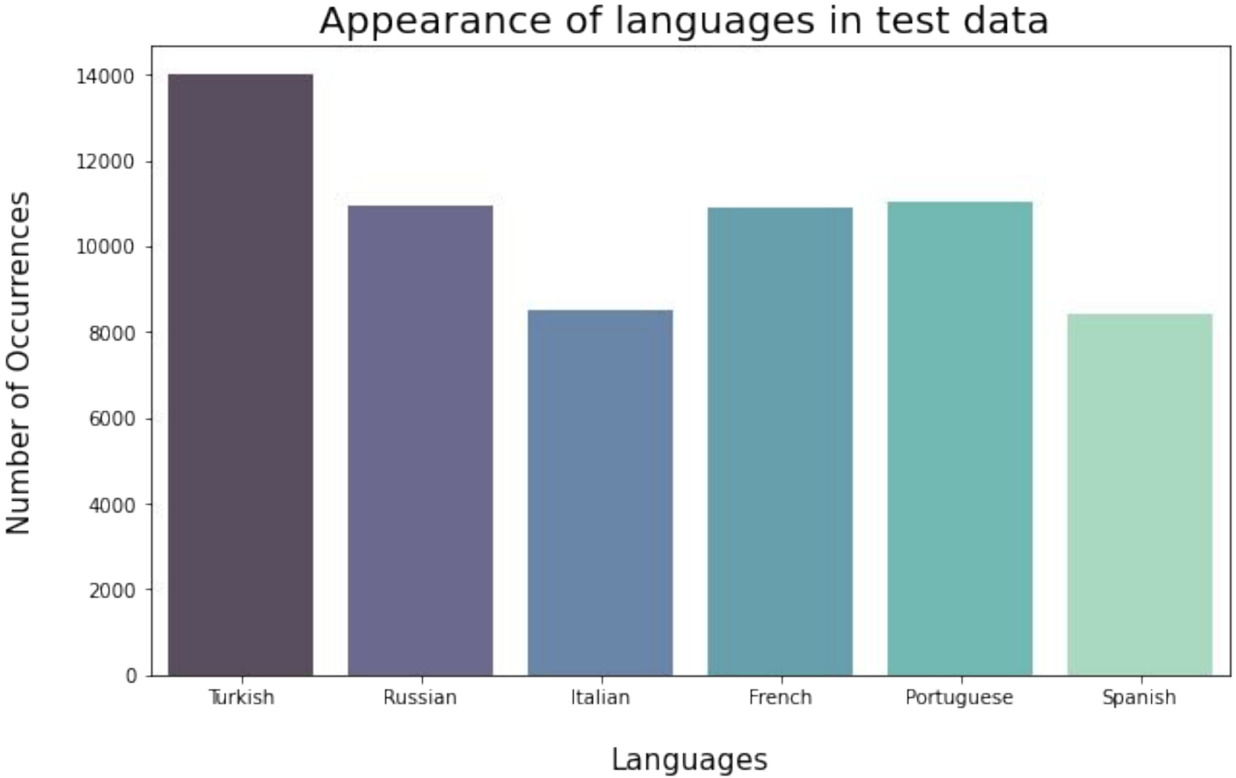
○ Select a file on Cloud Storage

gs:// *

● Input text below

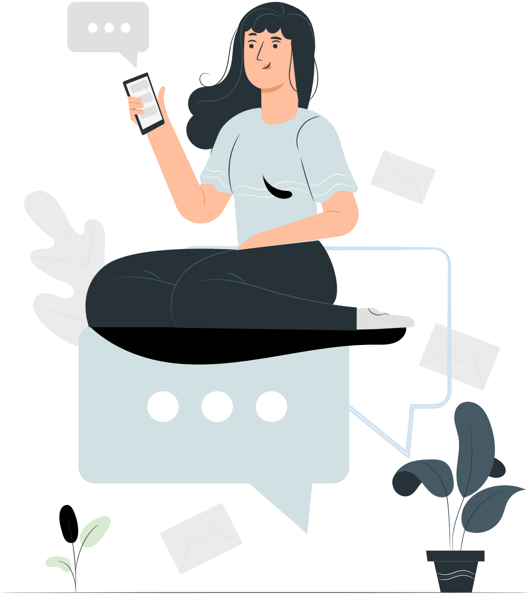
60000 characters remaining

Appendix


Appearance of languages in test data
Number of Occurrences
14000
12000
10000
8000
6000
4000
2000
0
Turkish
Russian
Italian
French
Portuguese
Spanish
Languages

05 | recommendations

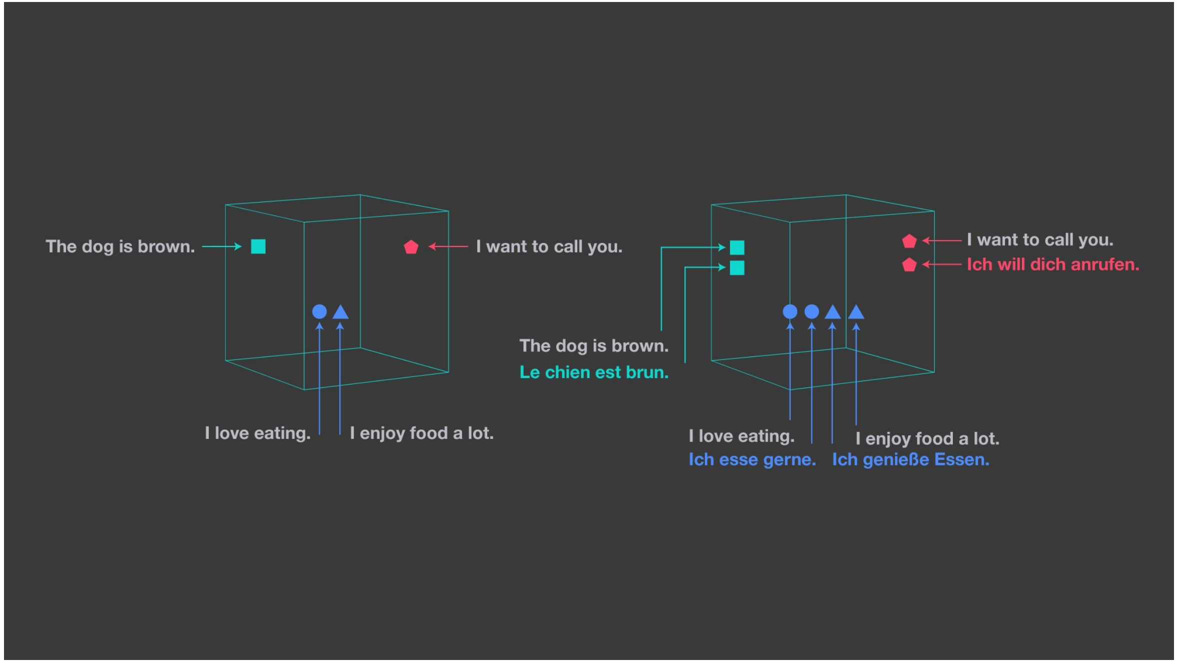
- reduce costs of identifying toxic comments with simple models
- Use pre-trained word embeddings to improve model performance
- Invest in training multilingual models to secure competitiveness in the future



06 | future work

- Optimize model to reduce bias and recognize irony and implicit aggression
- Create a web tool that recommends users to adjust their language before posting a comment
- time-series analysis of toxicity online
- Work with new tools like LASER by Facebook

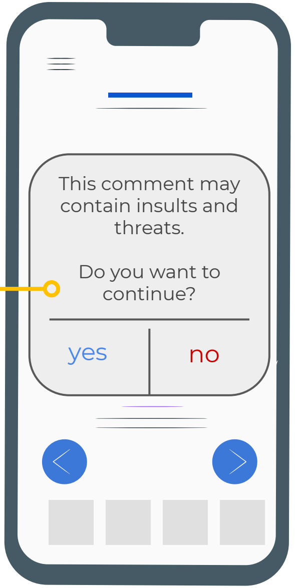
NLP throughout different languages



06 | future work

protect users from toxic content

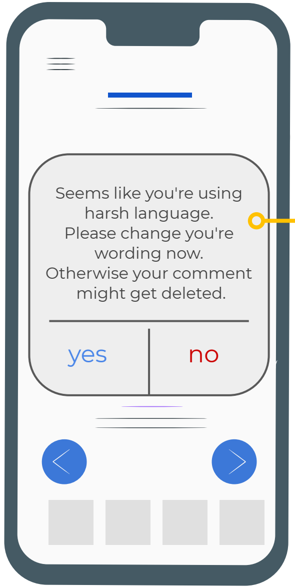
Approach 1



This comment may contain insults and threats.

Do you want to continue?

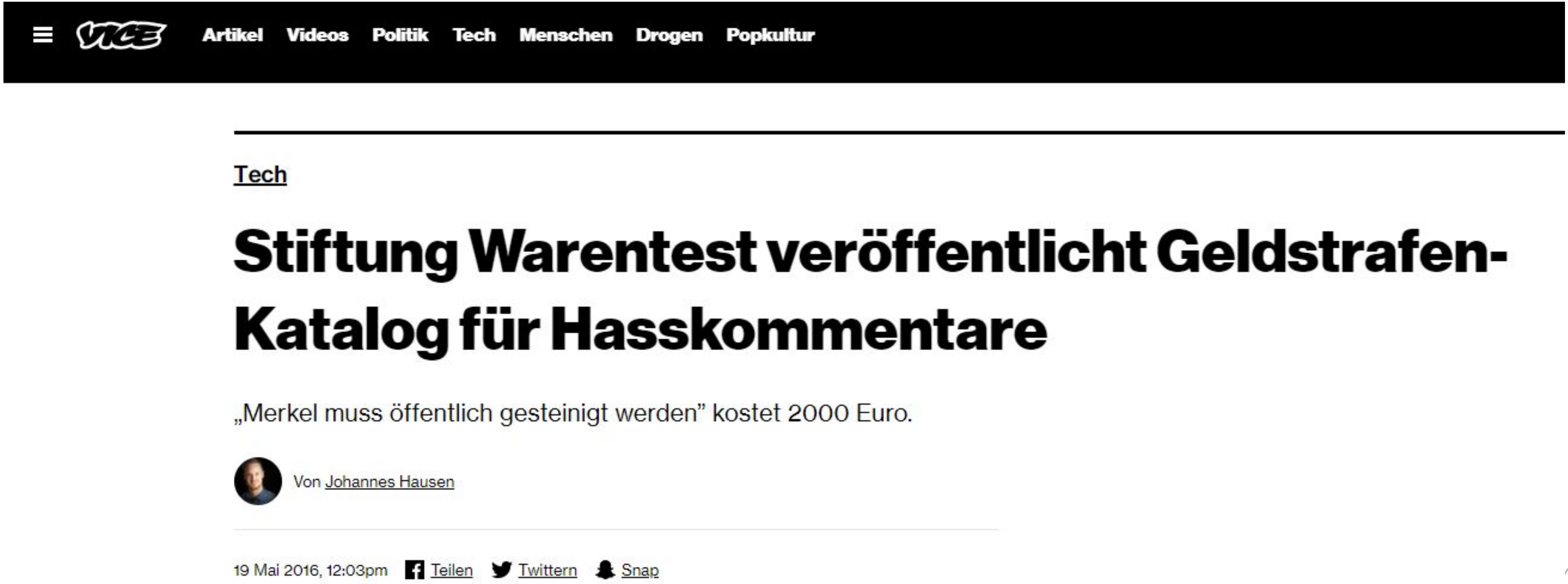
Approach 2



Seems like you're using harsh language. Please change you're wording now. Otherwise your comment might get deleted.

mirror toxic behavior in advance

Ja, und?



Tech

Stiftung Warentest veröffentlicht Geldstrafen-Katalog für Hasskommentare

„Merkel muss öffentlich gesteinigt werden" kostet 2000 Euro.

Von Johannes Hausen

19 Mai 2016, 12:03pm

Thank you

Drenizë Rama

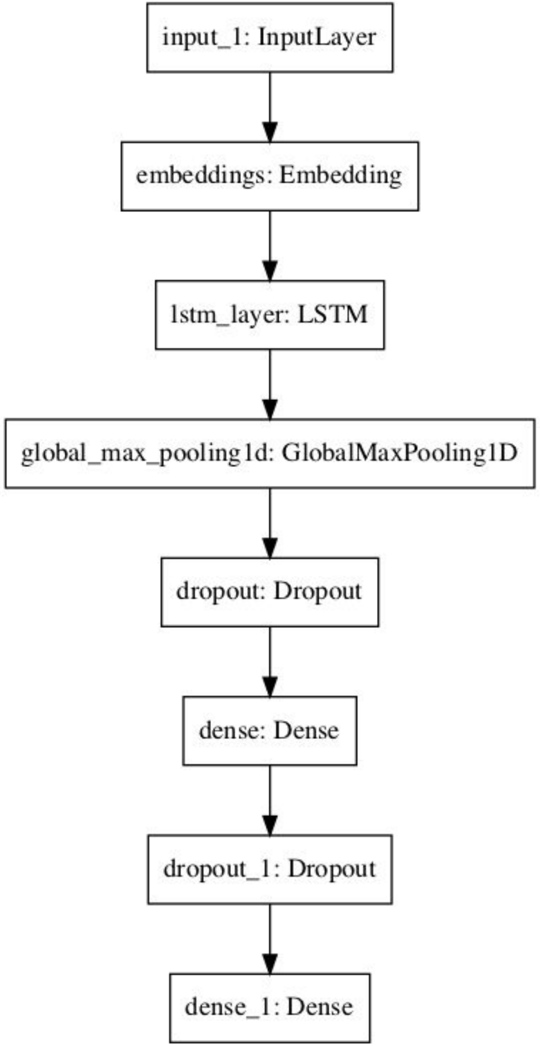
Data Scientist

https://drenize.github.io/

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

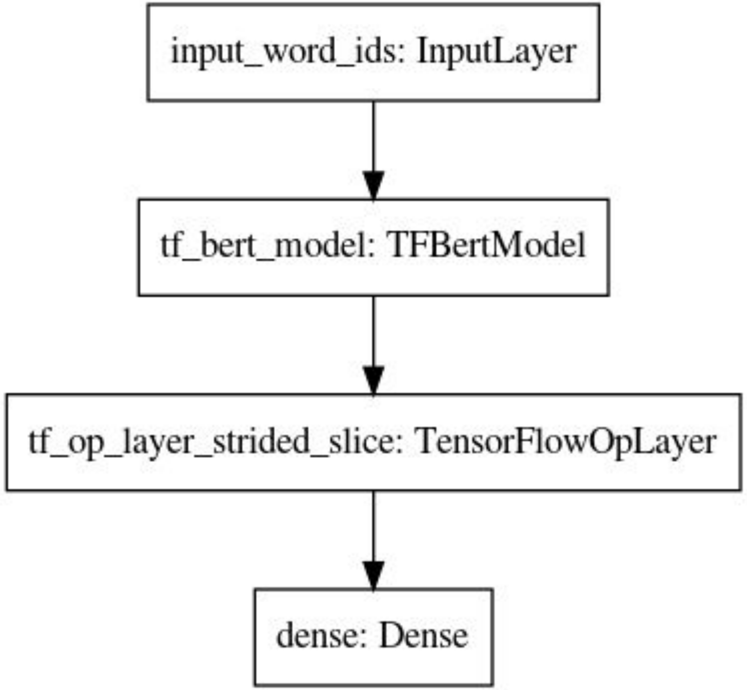
Appendix

LSTM architecture



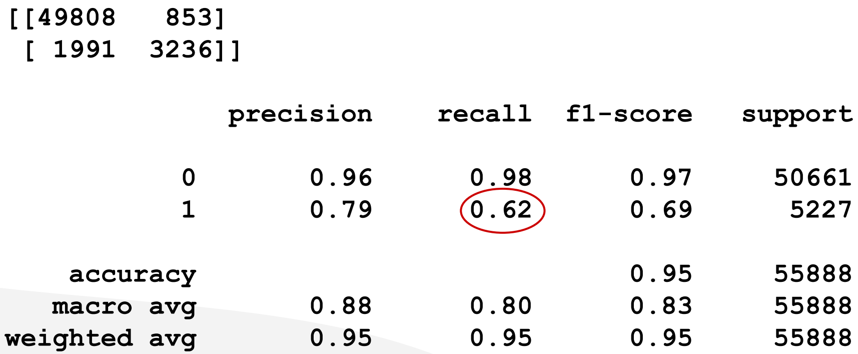
Appendix

BERT architecture


input_word_ids: InputLayer
tf_bert_model: TFBertModel
tf_op_layer_strided_slice: TensorFlowOpLayer
dense: Dense

Appendix

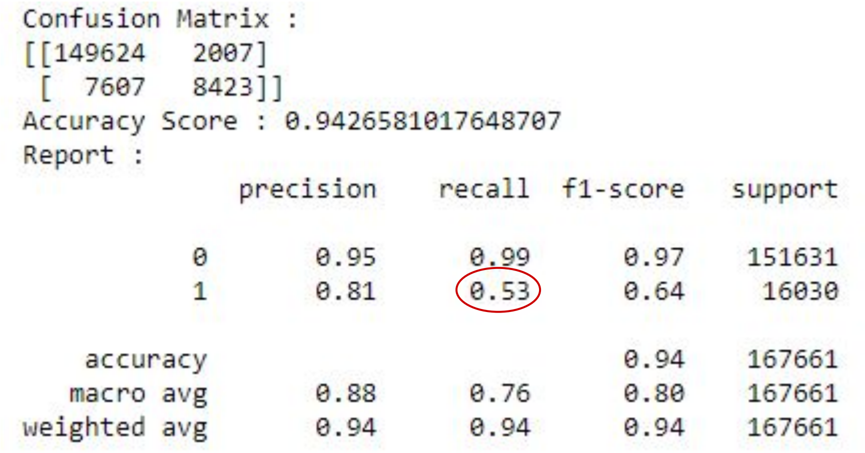
Baseline Model: Logistic Regression



```
[[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
```

Appendix

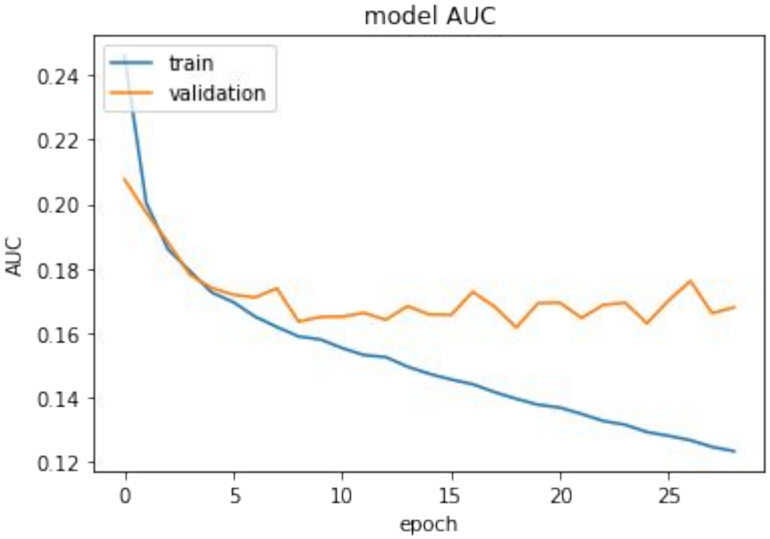
XGBoost Classifier



```
Confusion Matrix :
[[149624   2007]
 [  7607   8423]]
Accuracy Score : 0.9426581017648707
Report :
              precision    recall  f1-score   support

           0       0.95      0.99      0.97    151631
           1       0.81      0.53      0.64     16030

    accuracy                           0.94    167661
   macro avg       0.88      0.76      0.80    167661
weighted avg       0.94      0.94      0.94    167661
```

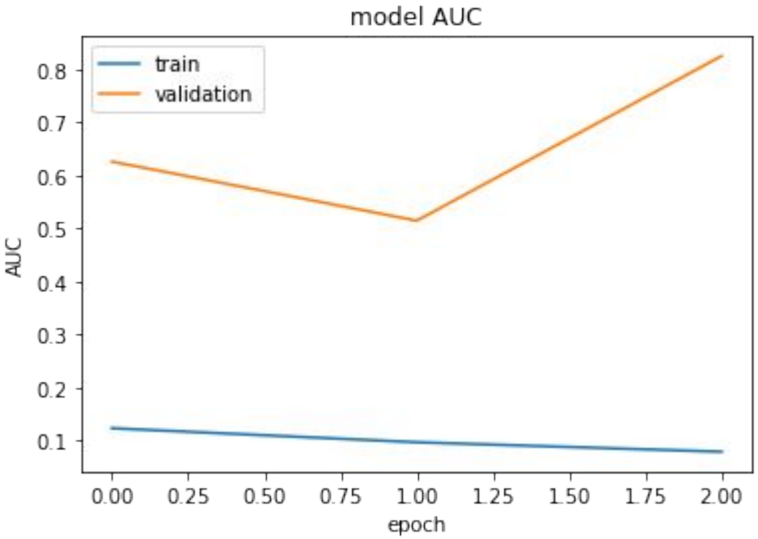
Appendix

LSTM model with GloVe



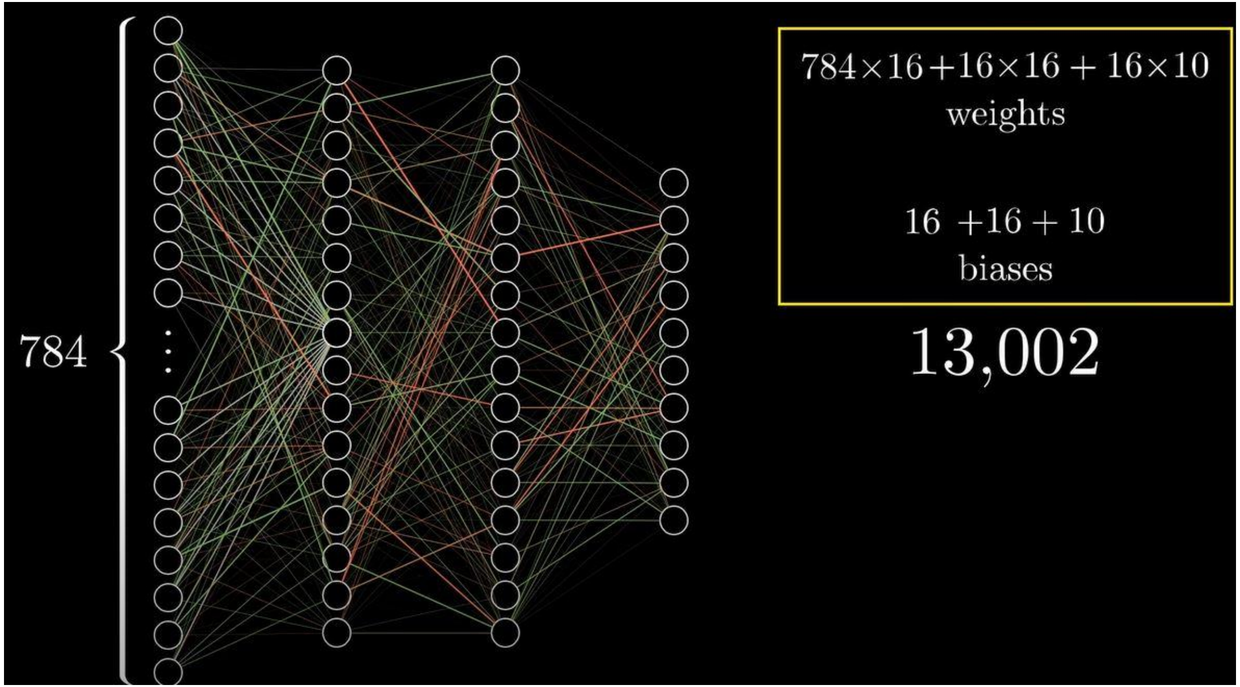
Appendix

BERT multilingual model

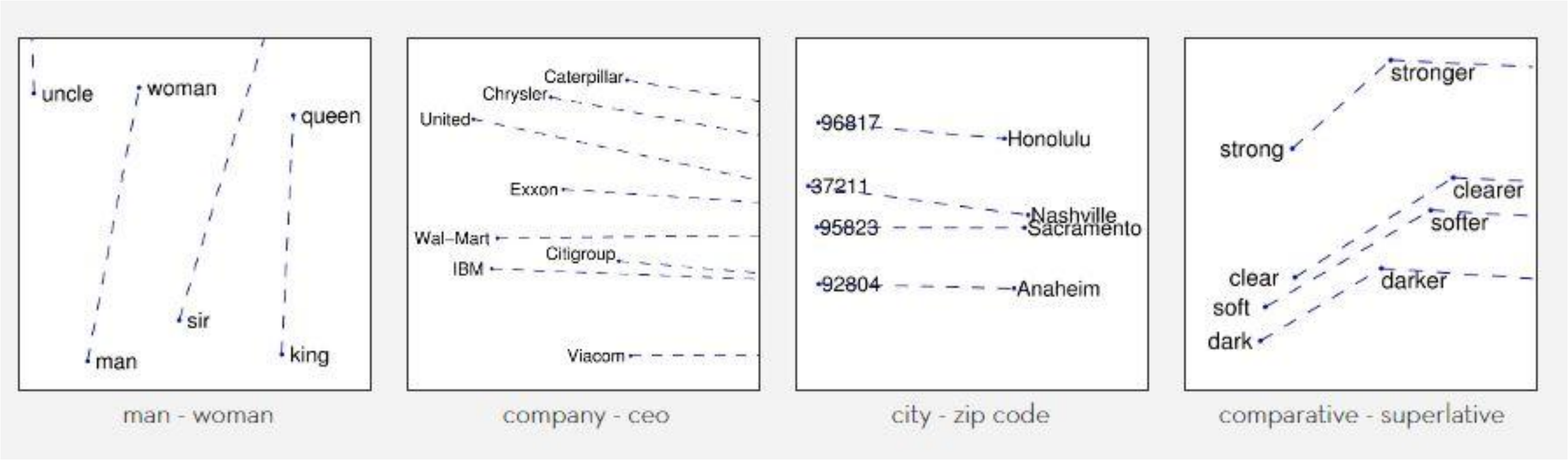


Appendix

How does a Neural Network work?



GloVe: Global Verctors for Word Representation



Hasskommentare
im Netz
identifizieren

Discussion Detox

App

by Drenizë Rama