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

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.**"
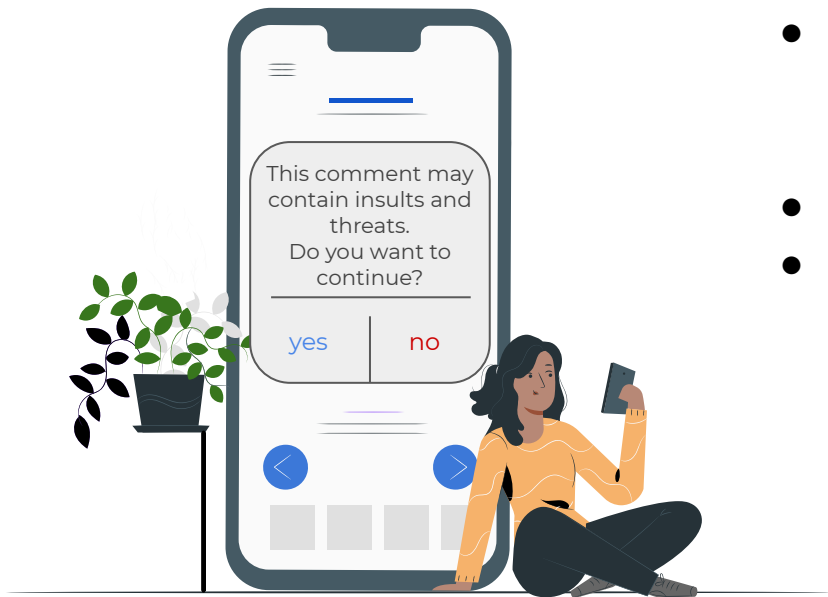
4

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
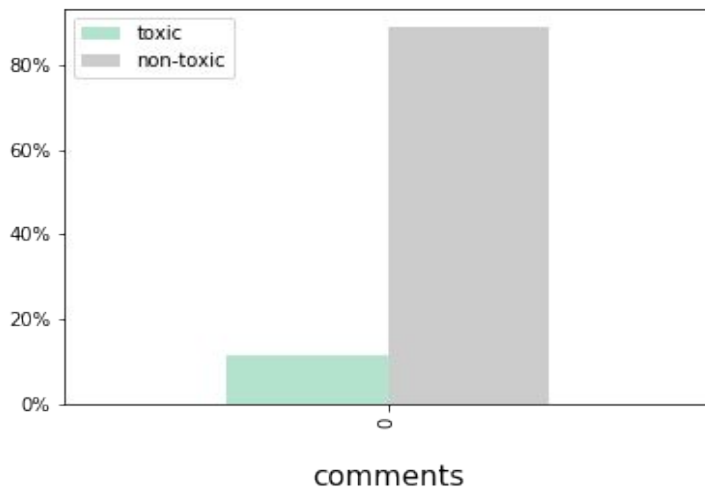
goal

- build a natural language algorithm that classifies social media comments into toxic and non-toxic categories
- at a low cost
- across different languages
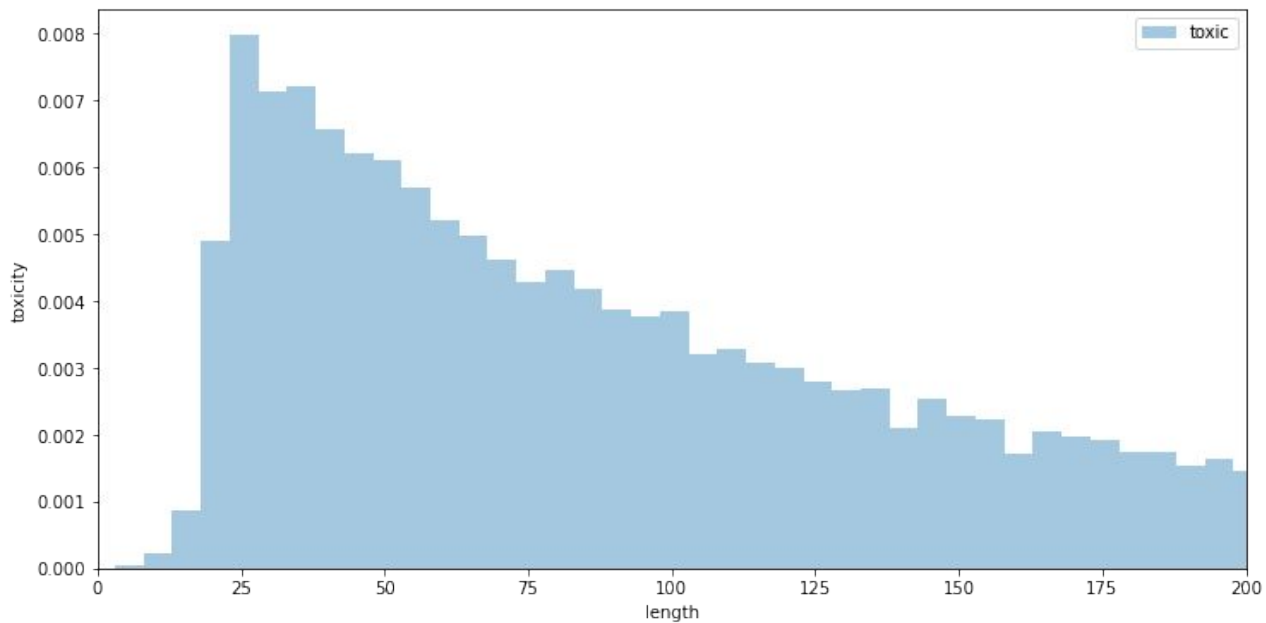
The overall amount of toxic comments is 11.4%

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

Disclaimer: The dataset for this project contains text that may be considered profane, vulgar, or offensive.
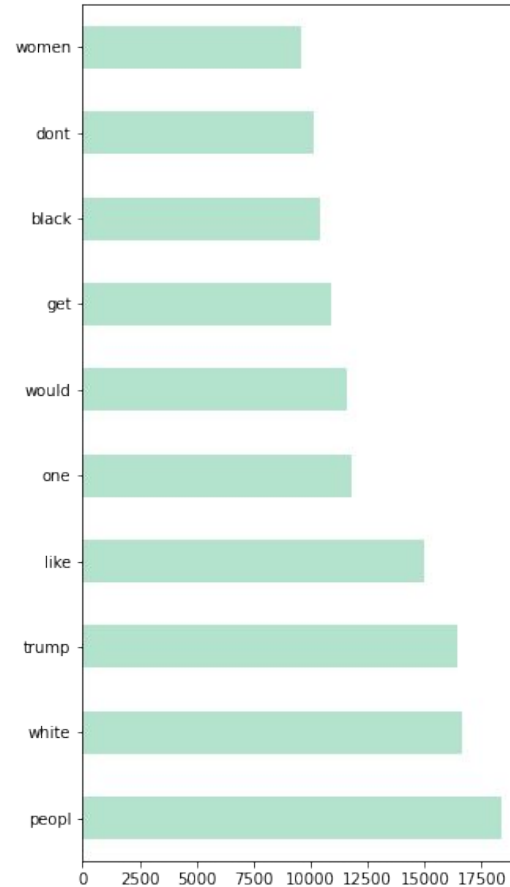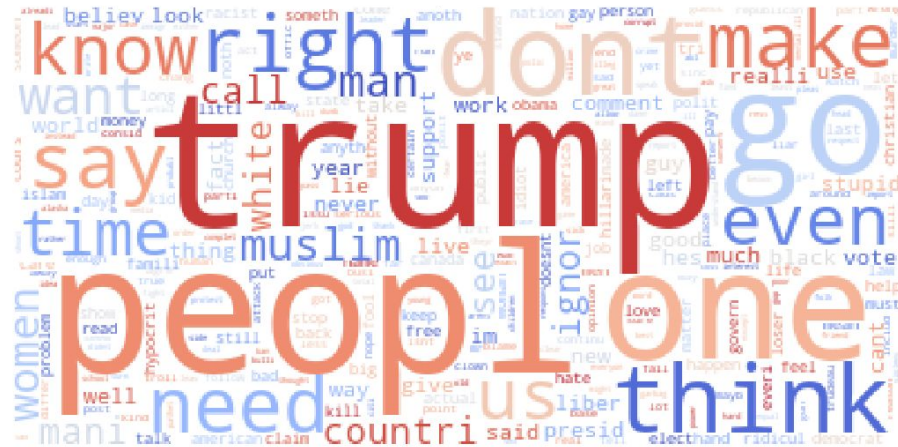
The more hate - the shorter the comments

frequently used words

frequently used words

In toxic comments

In comments containing identity attack

03 | methods

BERT by Google
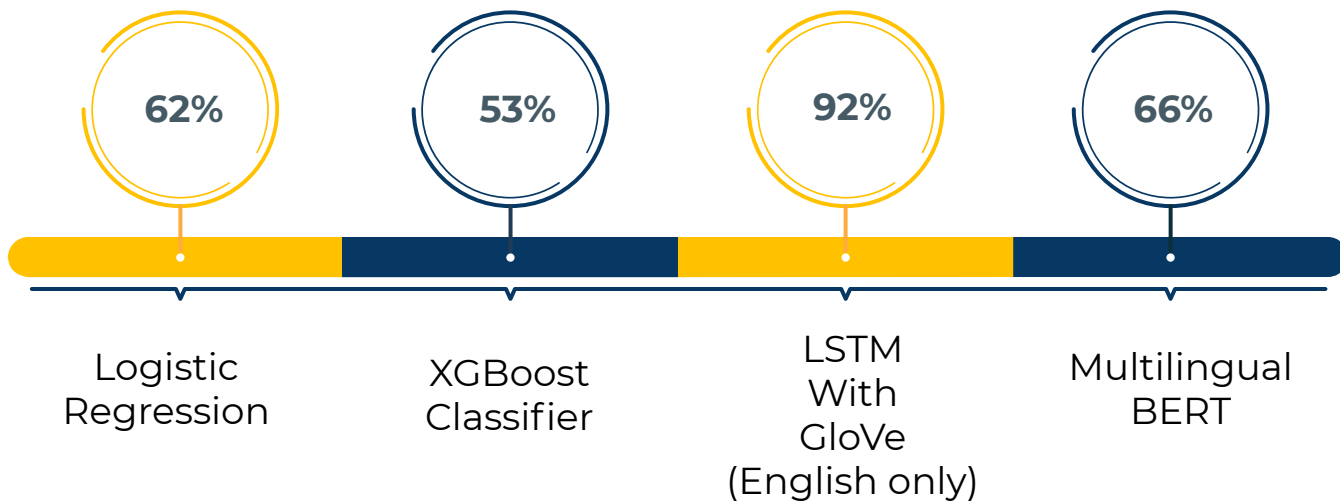Multilingual neural network

LSTM model
monolingual neural network

Ensemble Methods:
XGBoost, AdaBoost, RandomForest

Baseline Model:
Logistic Regression

Additional power
and complexity
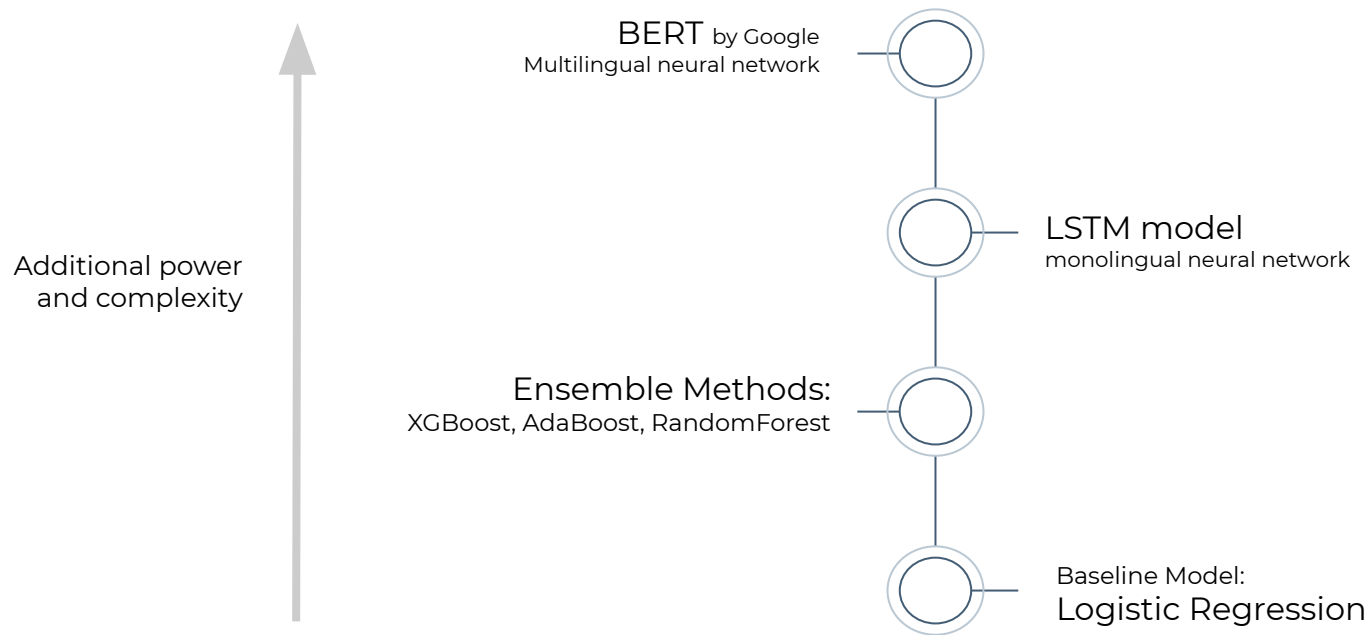
How Natural Language Processing works

Text Cleaning

Stemming, remove symbols, emojis, punctuation, stopwords

Train Model

Apply the model in data

Build Model

Adjust type, layers, epochs etc.

Tokenize

Divide sentences into sequences of words or characters

Encode

Create vectors for word representation

04 | results

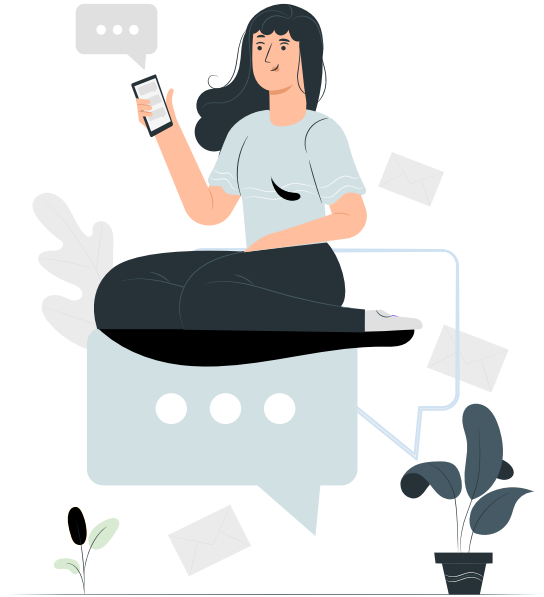| 62%	53%	92%	66%
Logistic Regression	XGBoost Classifier	LSTM With GloVe (English only)	Multilingual BERT

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
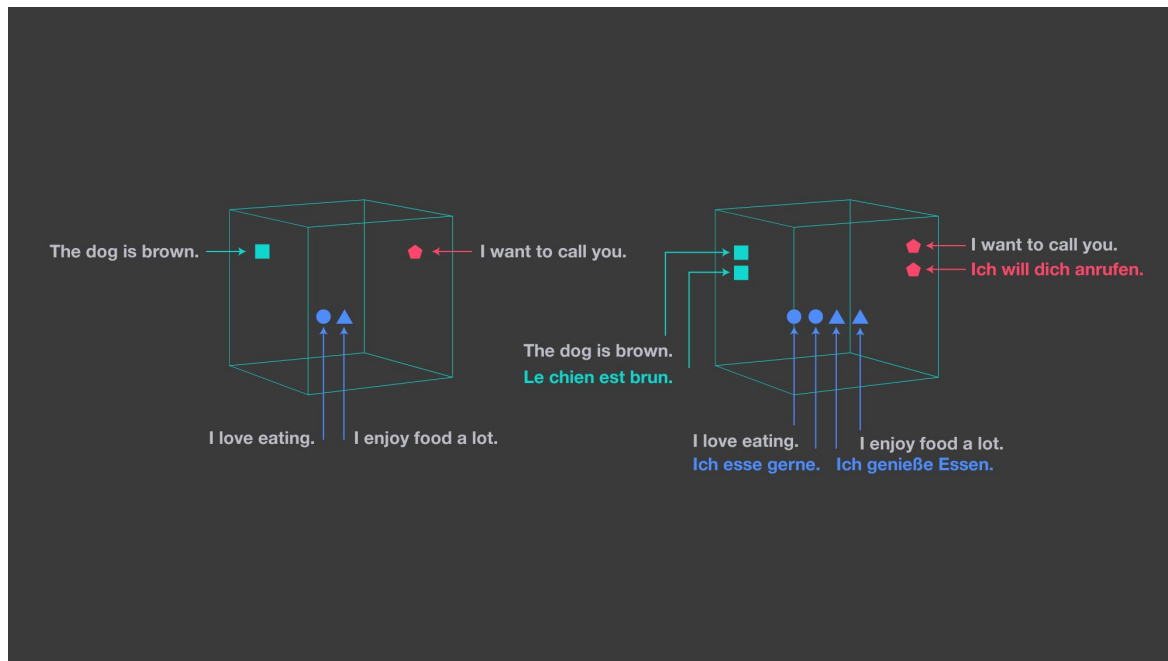
- Optimize model to reduce bias and recognize ironie and implicit aggression

- 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
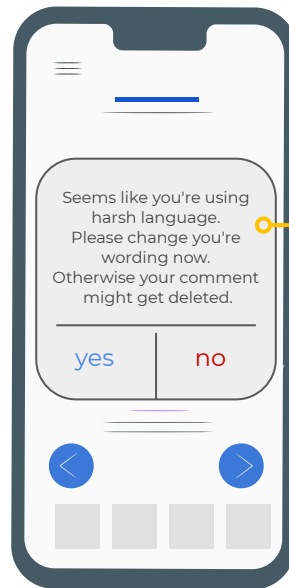
NLP throughout different languages

06 | future work

Approach 1

Approach 2

protect users from toxic content

This comment may contain insults and threats.

Do you want to continue?

yes | no

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

yes | no

mirror toxic behavior in advance

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:// *	BROWSE

● Input text below

60000 characters remaining

PREDICT

Thank you

Drenizë Rama

Data Scientist

https://drenize.github.io/
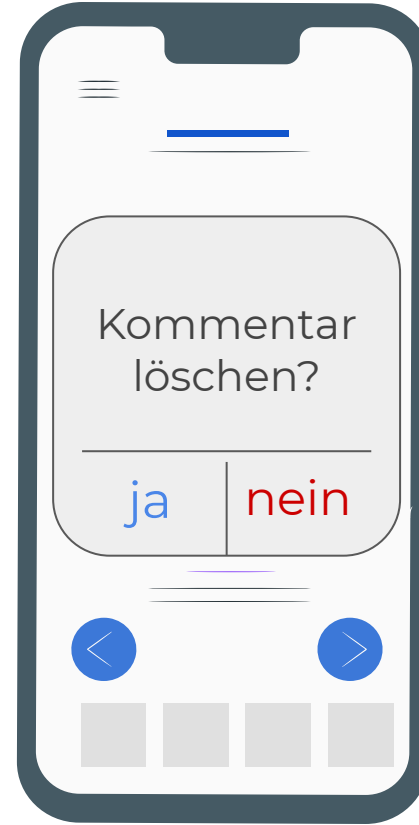
https://www.linkedin.com/in/d reniz%C3%AB-rama-6121a4157/

Hasskommentare
im Netz
identifizieren

Discussion Detox
App

by Drenizë Rama

Kommentar
löschen?

ja | nein

Appearance of languages in test data
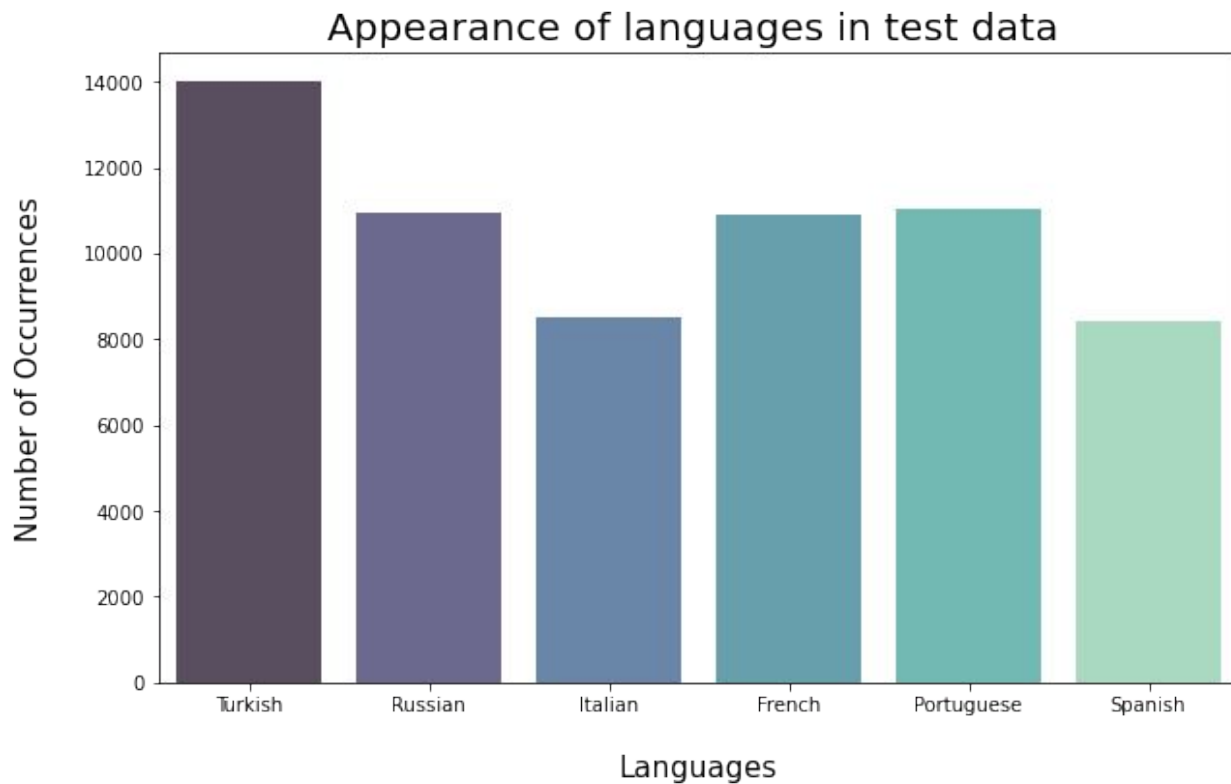
Baseline Model: Logistic Regression

```
[[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
```
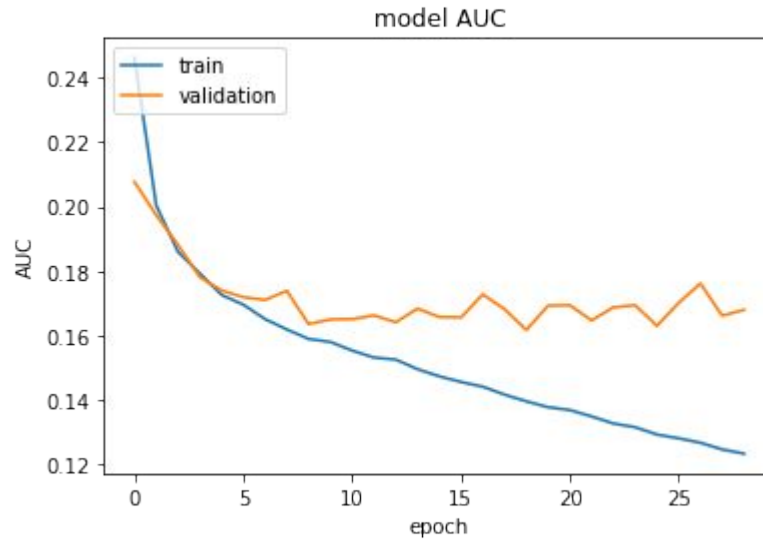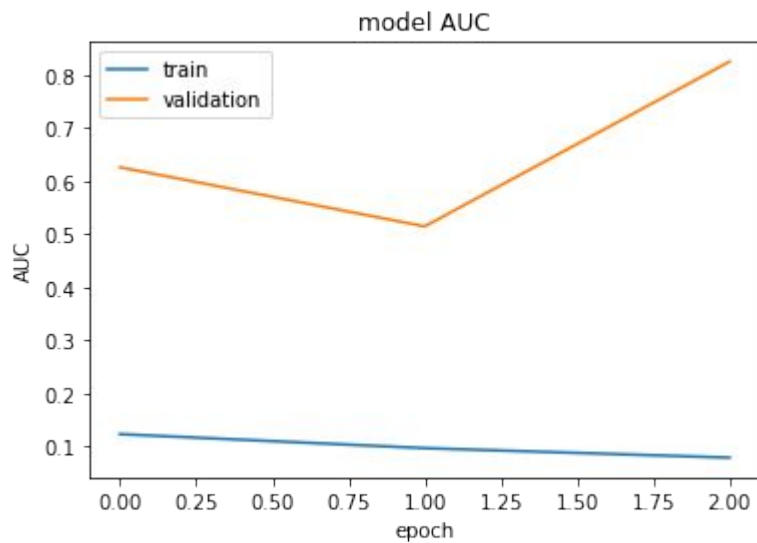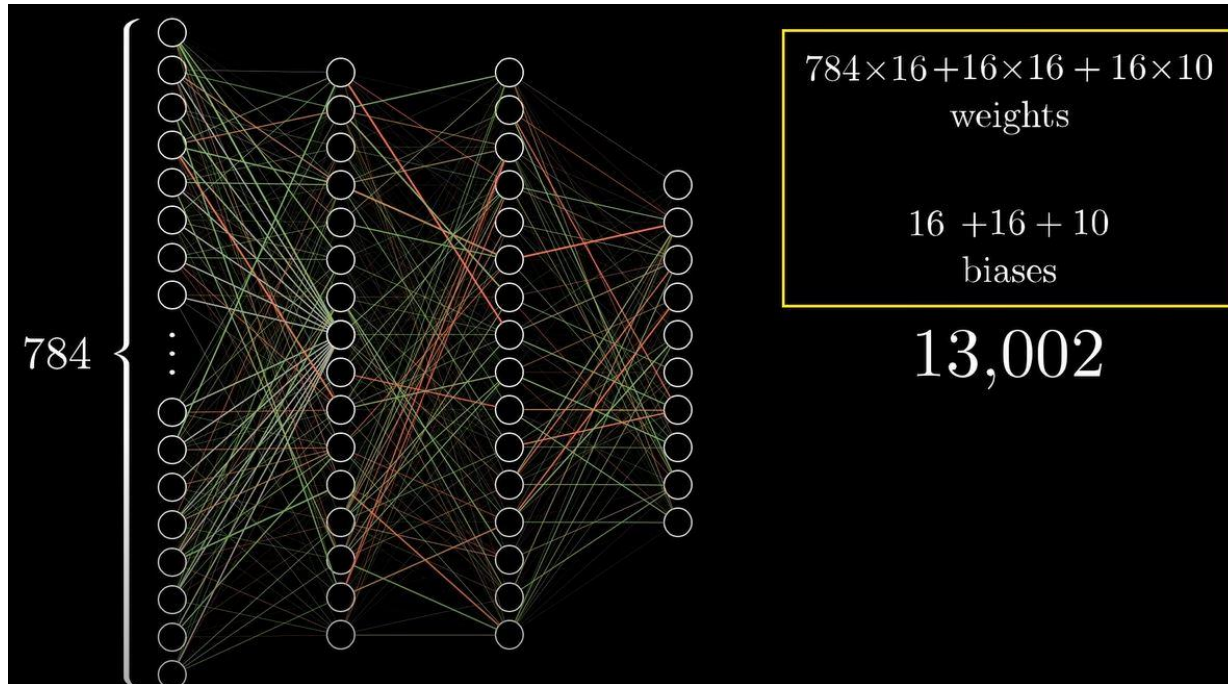
23

LSTM model with GloVe

BERT multilingual model

How does a Neural Network work?

$$784 \times 16 + 16 \times 16 + 16 \times 10$$
weights

$$16 + 16 + 10$$
biases

$$13,002$$

784

GloVe: Global Verctors for Word Representation

man - woman | company - ceo | city - zip code | comparative - superlative