The Detoxifiers

Empower Fruitful Discussions

Ksenia Gerasimovich, Daniel Guhr,
Katharina Neumüller & Kristin Stöcker

The Detoxifiers

Dr. Daniel Guhr
Physicist
Data Scientist

My Background

PhD Physics, Univ. of Constance
System Engineer
Project manager

Looking for projects (freelancer) as

- **Data Scientist**
- **Data Analyst**

remote

The Detoxifiers

Ksenia Gerasimovich
Data Scientist
Consultant

My Background

Business Informatics B.Sc.
Finance MBA
Business analyst
Consultant: ERP, BI implementation

Looking for a job as

- **Data Scientist**
- **Data Analyst**

in Düsseldorf oder remote

The Detoxifiers

Katharina Neumüller
Data Scientist

My Background

Computational Linguistics B.Sc.
Web-Development
Project Management

Looking for a position as

- **Data Scientist**
- **Data Analyst**

in Cologne or remote

The Detoxifiers

Kristin Stöcker
Linguist
Data Scientist

My Background

M.A. Linguistics
Research Assistant at FU Berlin

Looking for a job as

- **Data Scientist**
- **Data Analyst**
- **Data Engineer**

} **in NLP**

in Berlin or remote

Attention

Trigger Warning

This presentation contains examples of profane, vulgar or offensive language that is likely to upset readers!

The Background

The perks of online communication

free speech/ anonymity

exchange of opinions

independent of time and location

The price of online communication

41% of US adults experienced online harassment*

30% of affected users stopped using an online service*

27% of witnesses refrained from posting online*

*Pew Research Center: The State of Harassment, 2021

Consequences for website operators

- content moderation presents enormous financial burden
- result: disabled comment sections
- New York Times: only 10% of articles allowed comments (before 2016)
- How can AI help to counter this problem?

JIGSAW

Conversation.ai

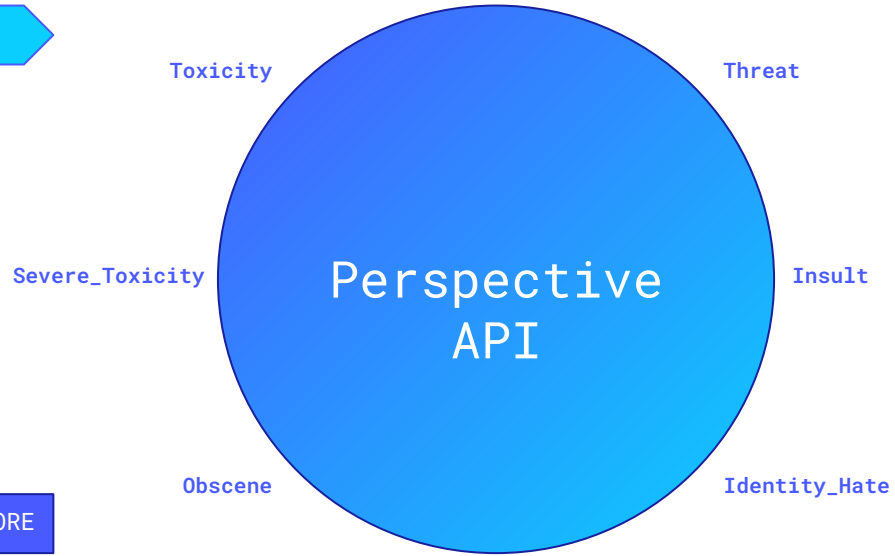
Application of Perspective API in online communication

"Shut up.
You're an idiot!"

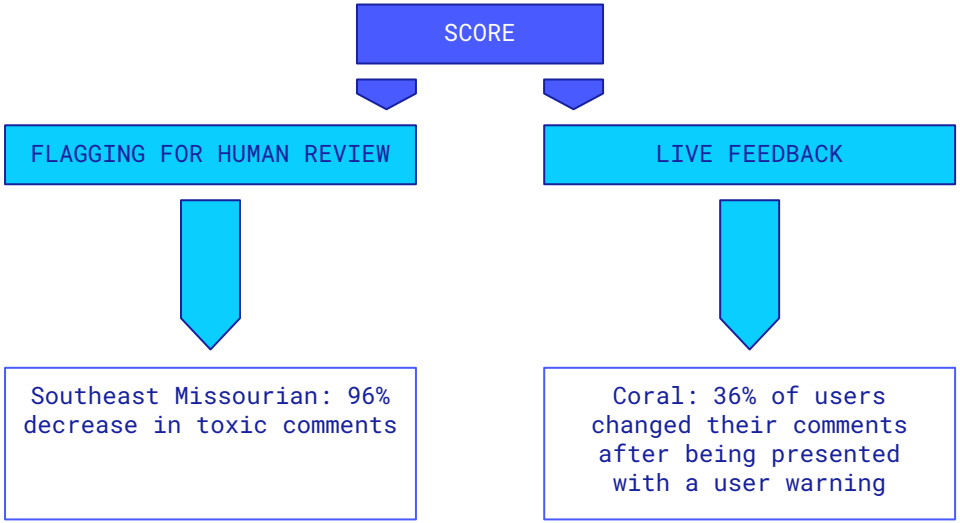
Toxicity	0.99
Severe_Toxicity	0.81
Obscene	0.20
Threat	0.09
Insult	0.97
Identity_Hate	0.02

INPUT: TEXT

OUTPUT: SCORE



One API,
many applications



Toxic Comment Classification Challenge

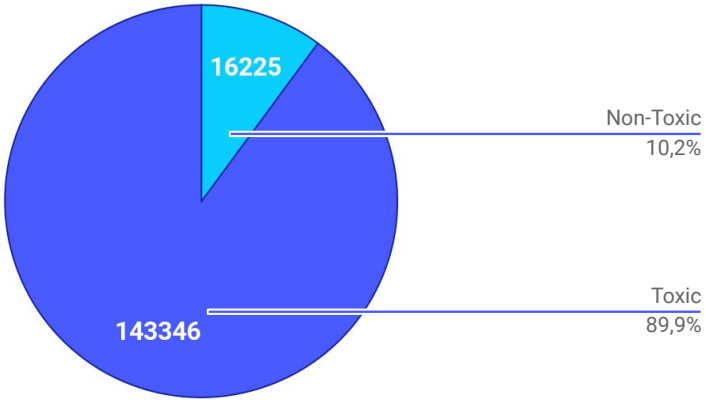
- goal: multi-headed model capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate
- dataset of 160.000 comments from Wikipedia's talk page edits
- train data labelled by humans

The Data

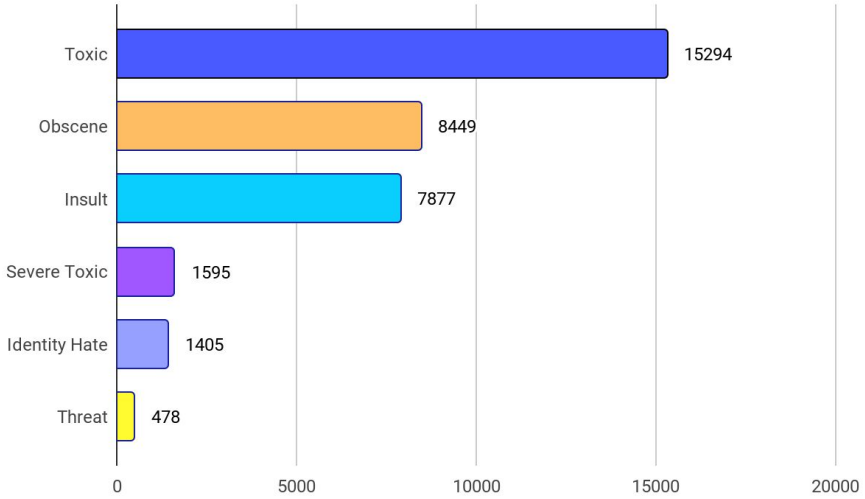

Toxic
fuck
fucking
suck
will
stop
know
wikipedia
Severe Toxic
fuck
fucking
bitch
shit
suck
will
asshole
dick
Obscene
fuck
fucking
shit
suck
bitch
will
stupid
Threat
will
kill
fucking
fuck
going
die
Insult
fuck
fucking
suck
bitch
stupid
will
shit
Identity Hate
fuck
nigger
fucking
faggot
bitch
shit
suck

Data

Toxicity Distribution

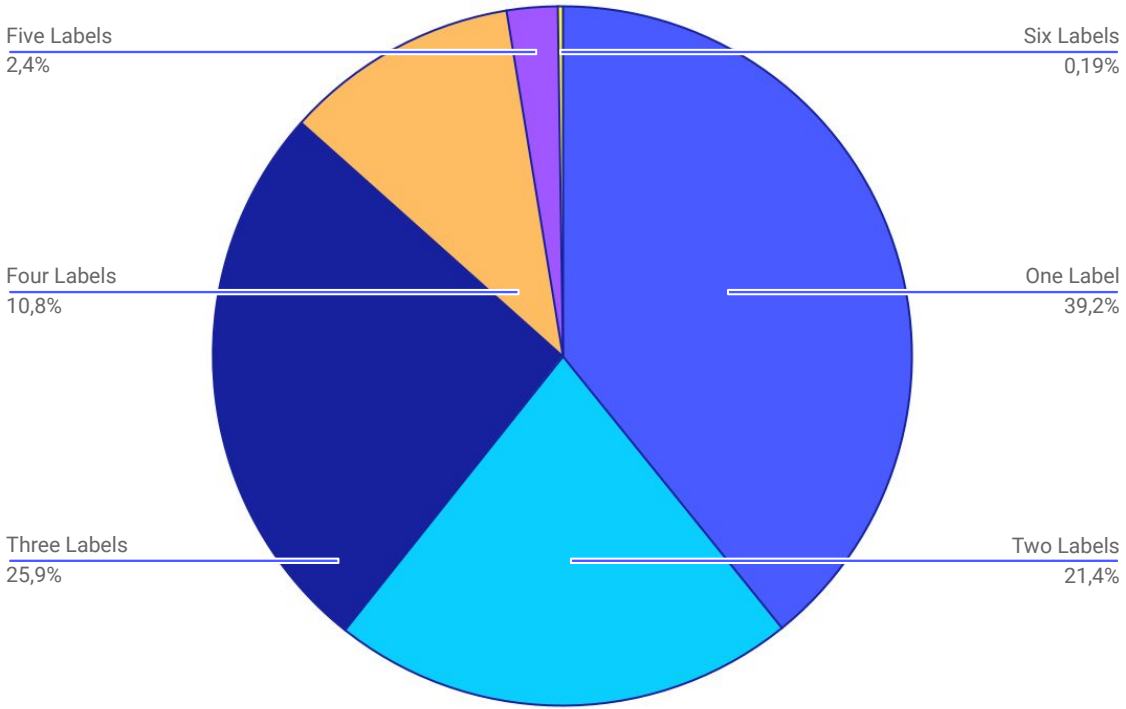


Category Counts



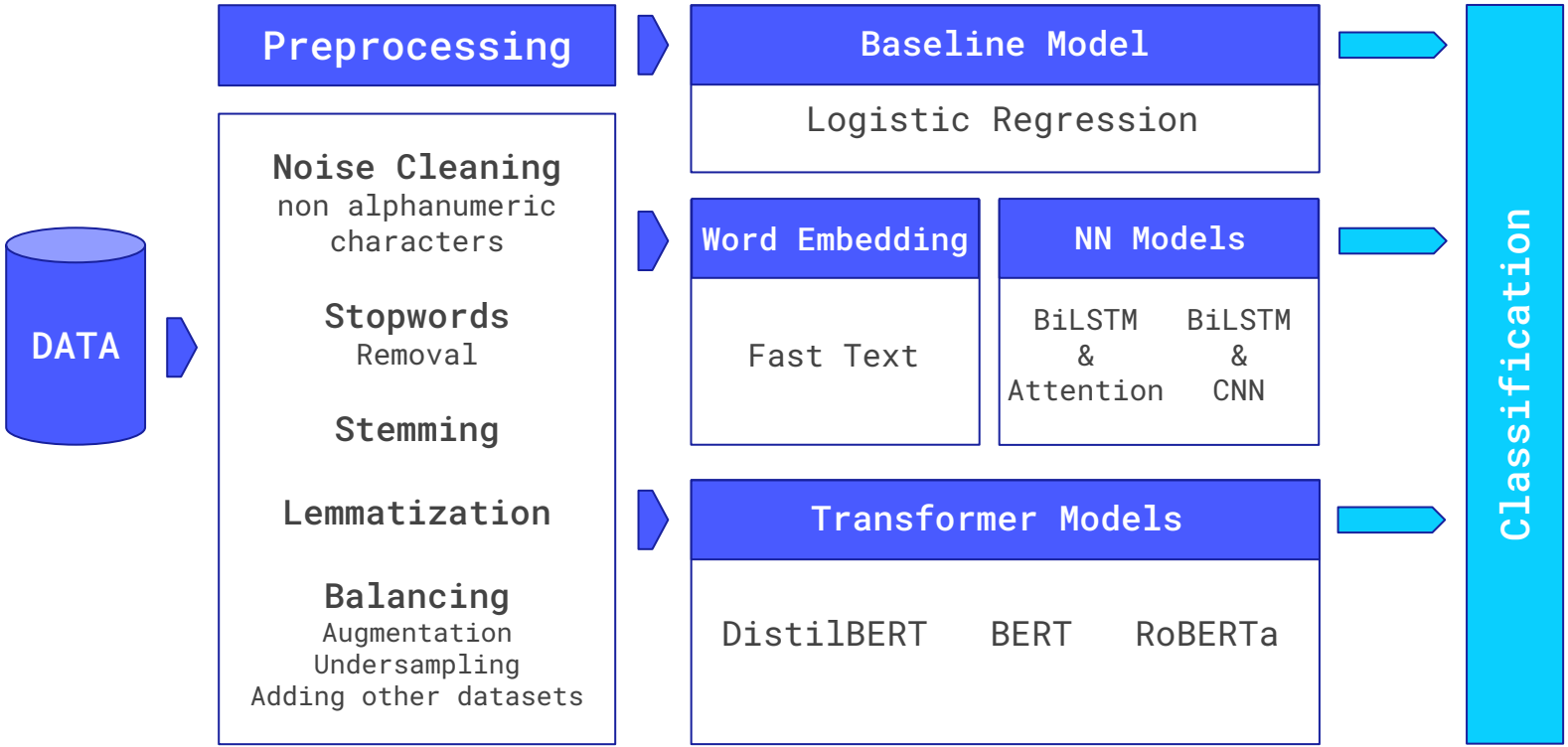
Dataset of 159571 Samples

Comments with multiple labels



Approach & Results

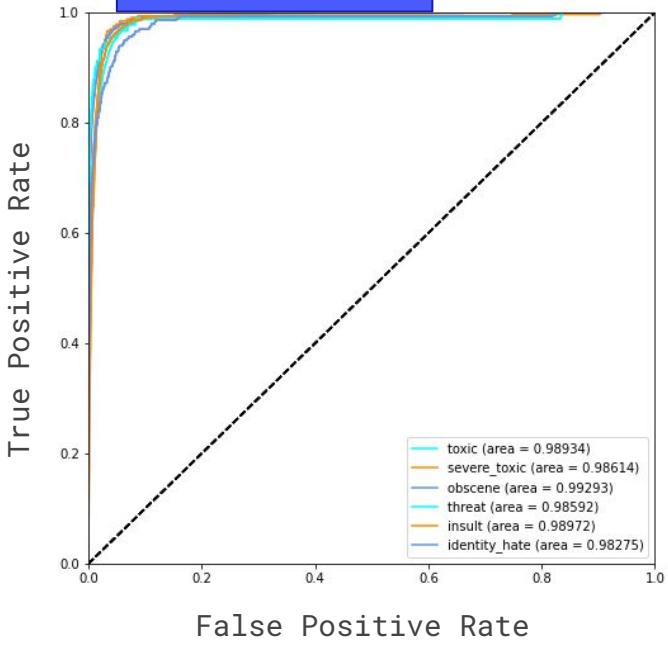
A Modular Approach



RoBERTa Classification Results

labels	precision	recall	f1-score	support
toxic	0.94	0.90	0.92	3102
severe_toxic	0.60	0.45	**0.51**	**345**
obscene	0.86	0.91	0.88	1772
threat	0.64	0.69	**0.66**	**102**
insult	0.78	0.88	0.83	1613
identity_hate	0.65	0.72	**0.69**	**277**
micro avg	0.85	0.87	0.86	7211
macro avg	0.75	0.76	**0.75**	7211

ROC AUC



mean column-wise ROC AUC		
	train	0.98780
	test	**0.98482**

Error Analysis

Contextual Issues

“I’m glad you have gone do not come back”

Category	True	Predicted
toxic	1	0
severe_toxic	0	0
obscene	0	0
threat	0	0
insult	0	0
identity_hate	0	0

PROBLEM Lack of context in training samples

SOLUTION Additional labeling of comments with frequent word repetitions and retraining

Context Awareness

Subjectivity in labelling

“And I know I am a dickhead”

Category	True	Predicted
toxic	1	1
severe_toxic	0	0
obscene	1	1
threat	0	0
insult	1	0
identity_hate	0	0

PROBLEM Correct model predictions

SOLUTION Pseudolabelling

Label Assignment Error

Lack of profanity and context to train the model

“islams you motha fers”

Category	True	Predicted
toxic	1	1
severe_toxic	1	0
obscene	1	0
threat	0	0
insult	1	1
identity_hate	0	1

PROBLEM Context awareness exists, but in insufficient volume

SOLUTION More samples for training

Context Awareness: Insults on political and religious grounds

Outlook

- Try out other Transformers - XLNet
- More training data (from other sources like Twitter)
- Further fine-tuning
- Ensemble models
- Out-Of-Vocabulary Words
  - Training own embeddings

Thank you

Questions, comments and discussions are welcome.