

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



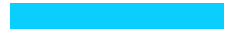
# Trigger Warning

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\*

# 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?



# Application of Perspective API in online communication

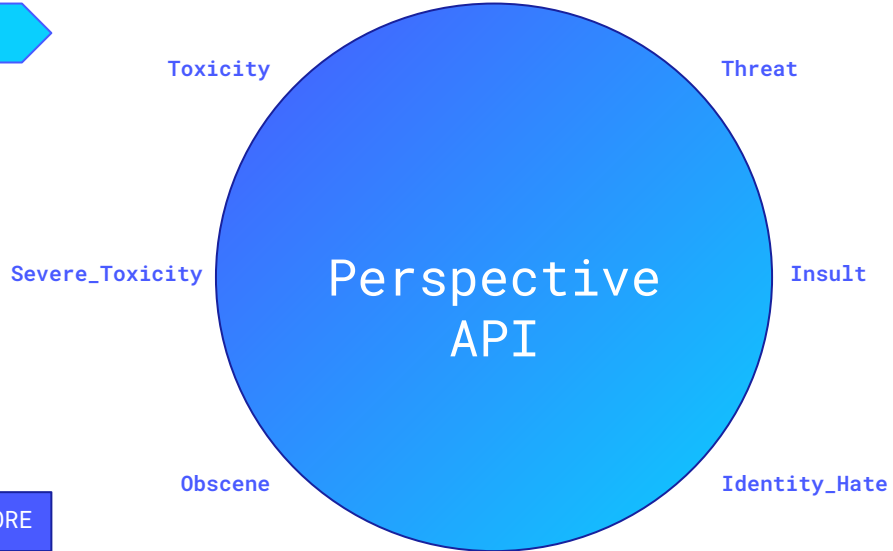


"Shut up.  
You're an idiot!"

INPUT: TEXT

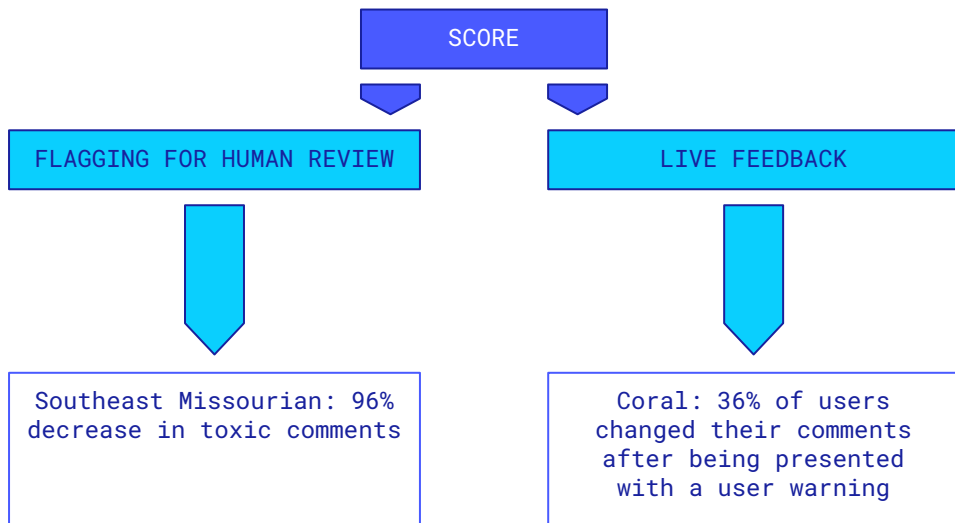
Toxicity	0.99
Severe_Toxicity	0.81
Obscene	0.20
Threat	0.09
Insult	0.97
Identity_Hate	0.02

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

A promotional banner for the Toxic Comment Classification Challenge. The top half has a dark purple background with white text. It features a trophy icon, the text 'Featured Prediction Competition', '\$35,000 Prize Money', the challenge title 'Toxic Comment Classification Challenge', and the subtitle 'Identify and classify toxic online comments'. The bottom half is white with the JIGSAW and Conversation.ai logos.

Featured Prediction Competition

\$35,000 Prize Money

**Toxic Comment Classification Challenge**

Identify and classify toxic online comments

 **JIGSAW**  **Conversation.ai**

# The Data



## Toxic

give better dick nothing think time talk  
another well comes still leave really trying edit fuck  
piece need piece blocked please idiot  
good hate tell shit block doe delete  
every gonta cunt much year something cock stop  
faggot damn page take little know said  
killen edits life going stupid  
right thing anything even editing user look message read fuckin  
hell wikipedia bitch  
never dont asshole love back someone article

## Severe Toxic

fucker time asshole tell shut  
cock said right wrong keep retarded idiot  
dumb know hate want edit page face stop  
piece little life head want think where  
made hell back block dont well pussy ugly take  
mother please faggot care kill article stupid  
never editing make loser prick mean talk  
delete thing hope give lick people  
leave blocked wikipedia going  
dick guess dirty gonna find nothing penis stick mother fucker  
cunt still fucked fuckin bastard dare deleting mother fucking

## Obscene

guy read user wikipedia better  
said thing pussy look kill something fucker  
blocked really doe hate dont please  
last really tell cock made fuckin right  
anything nigger stop piece block talk little dick  
back never stop piece still delete idiot mean leave life  
faggot love fucked want give well stupid  
nothing take faggot mother going need dumb someone piss  
comment penis nothing edit personal much keep trying article  
another bullshit

## Threat

even piece give look blood keep hunt mother dick  
asshole destroy find back pathetic know hope fuckin  
need stop murder last beat life hate cancer  
shoot will fuckin  
make edit little punch faggot wikipedia death article  
take right user shut dont face message please wish house suck  
want right good shit delete better every never  
attack come thing shit delete better every never  
bitch stupid dead family block think really fuck time  
page cunt suggest edits  
fire head live hell gonna watch raped burn must  
child next nazi hello deleting said

## Insult

fucker shut wikipedia people  
message dumb kill read gonim another editing  
blocked edit want even mother back asshole faggot  
cock bitch block year good look  
tell going dont stupid will give things  
time user head delete little piece life page nothing  
edits delete pathetic hate still hell said name ugly please  
idiot know shit stop  
made much fuckin wiki article make dick

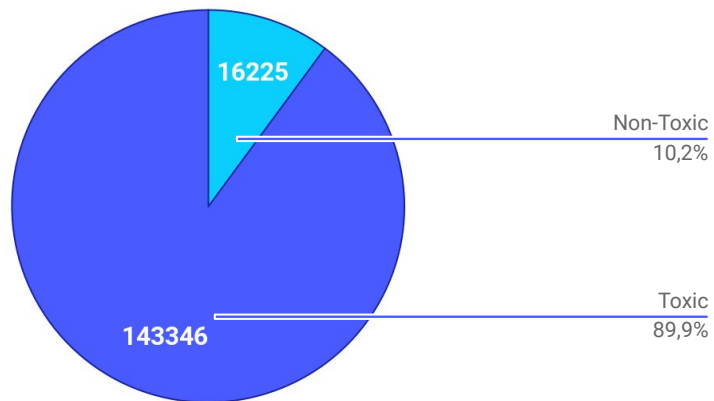
## Identity Hate

stupid hell white think cunt hate  
page said blocked comment asshole  
will muslim idiot delete cock really take  
even nigger delete american arab nazi doe  
homosexual fuckin racist mother  
rape keep good talk always nothing time real huge  
jew well user right pussy homo kill wikipedia  
please people bitch give know right dick shit  
edit fucked going retarded dumb shut make dirty probably still bastard  
never penis hope going shit indian thing must everyone  
stop need dont cant faggot come  
piece need suck back love want

# Data

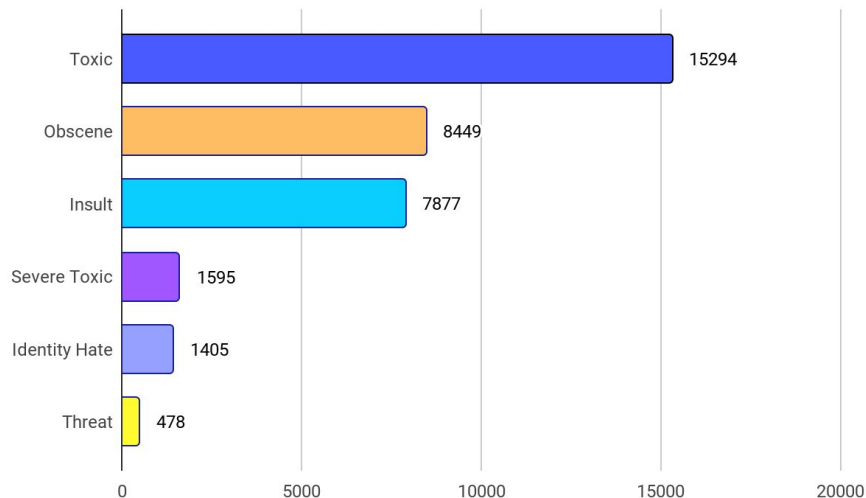


Toxicity Distribution



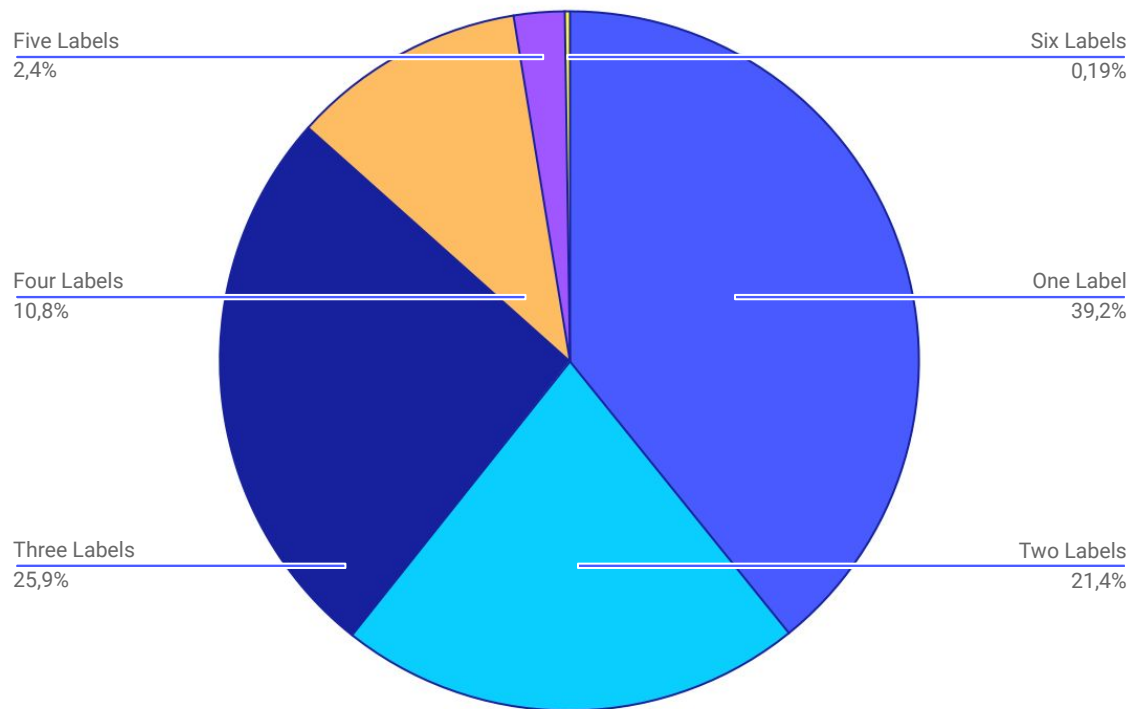
Dataset of 159571 Samples

Category Counts





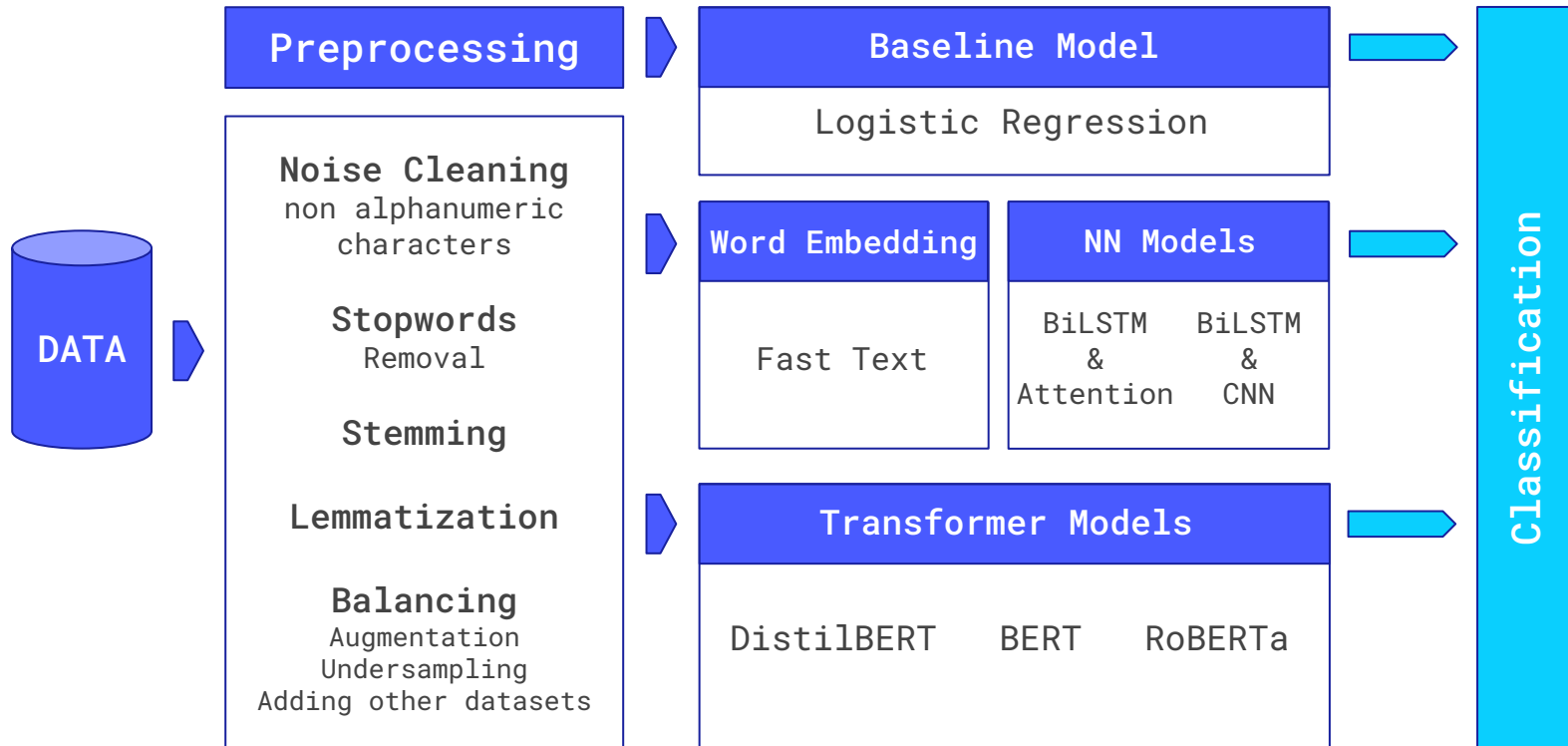
# Comments with multiple labels



# Approach & Results



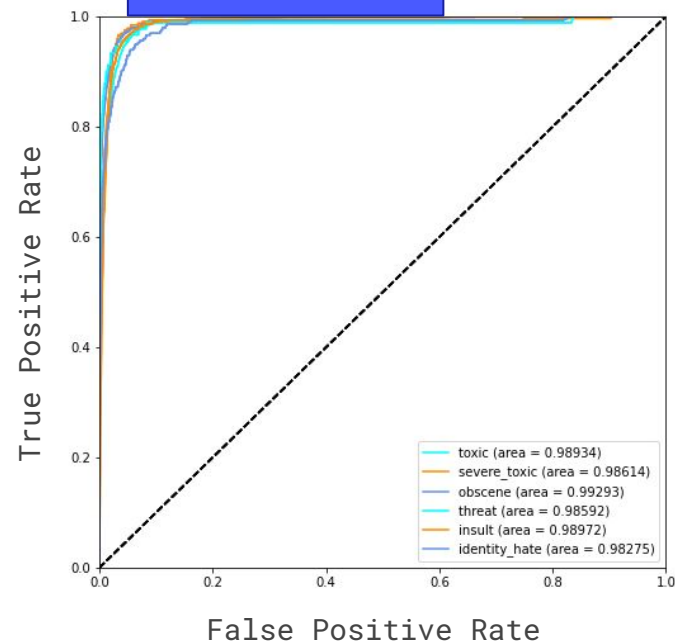
# A Modular Approach



## RoBERTa Classification Results

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