Erkennung von Hate Speech mit Twitter - Abschlusspräsentation

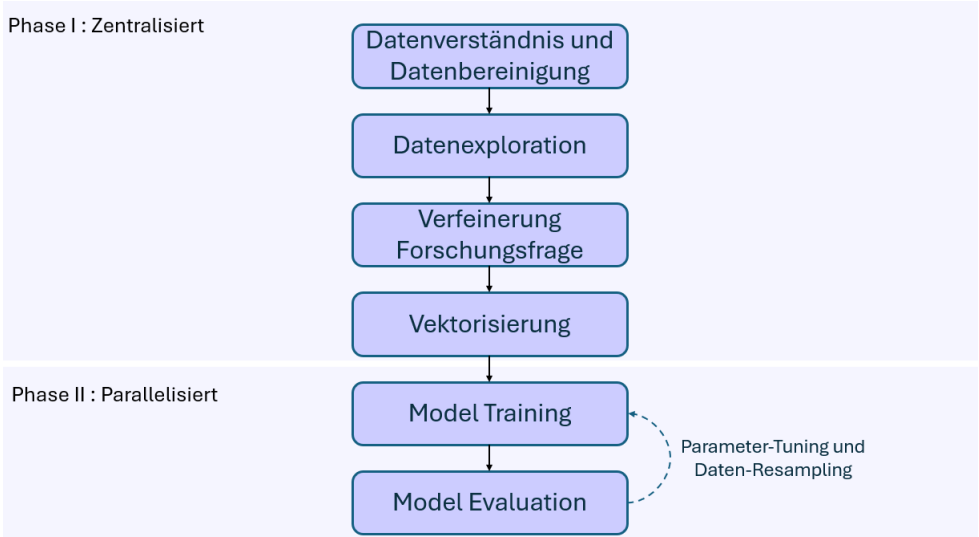
Elena Marion Friedrich, Nasiba Tuychieva, Sven Ole Wall, Imran Nteli Chalil, Christian Engels

Artificial Intelligence Group,
University of Hagen, Germany

18. März 2025

Overview

Workflow


Phase I : Zentralisiert
Datenverständnis und Datenbereinigung
Datenexploration
Verfeinerung Forschungsfrage
Vektorisierung
Phase II : Parallelisiert
Model Training
Model Evaluation
Parameter-Tuning und Daten-Resampling

Verfeinerung Fragestellung (Dez.)

Ausgangsfrage

Wie kann mithilfe von klassischen Verfahren des maschinellen Lernens sowie mit Deep-Learning-Verfahren Hate Speech auf sozialen Plattformen anhand des Beispiels Twitter erkannt werden?

- Klassifikation von Text / Sentiment Analysis

Methodische Detailfragen

- Welche Möglichkeiten zum Umgang mit Klassenungleichgewichten gibt es und wie ist deren Einfluss auf die Modellperformance?

Modellauswahl

Naive Bayes

- MultinomialNB & ComplementNB gut geeignet für unbalancierte Daten
- Erfolgreicher Einsatz in anderen Forschungs-arbeiten

SVM

- Gut geeignet für Klassifizierungs-probleme
- Robust gegenüber hochdimensio-naler Daten

Ensemble

- Breite Anwendbarkeit
- Guter Umgang mit unbalan-cierten Daten und Ausreißern

RNN (GRU & LSTM)

- LSTM: Speicherung globale Kontextinformationen über längere Se-quenzen hinweg mit 3 Gates
- GRU: Kombiniert Lang- und Kurzzeitspeicherung in einem einzigen Zustand mit 2 Gates
- Gut geeignet für Textklassifikation und Erkennung von sequentiellen Abhängigkeiten

Auswahl der Evaluationsmetriken

Warum Recall für Klasse 1 (Hate Speech)?

- **Hauptziel:** Kein Hate-Speech-Tweet soll übersehen werden.
- Definition: Anteil der korrekt erkannten Hate-Speech-Tweets an allen tatsächlichen Hate-Speech-Tweets.

Warum zusätzlich der F1-Score?

- Precision bleibt relevant: Unnötige False Positives sollen vermieden werden.
- F1-Score als harmonisches Mittel: Stellt sicher, dass nicht nur eine der beiden Metriken optimiert wird.

Warum nicht Accuracy als Hauptmetrik?

- Klassenungleichgewicht: Hate Speech ist viel seltener als Nicht-Hate-Speech (7% der Datensätze).

Modell	F1	Recall
BalancedRandomForest TFIDF	0.555	0.439
SVM GloVe rbf	0.572	0.738
SVM TFIDF rbf	0.633	0.610
NB TFIDF	0.325	0.323
NB GloVe	0.337	0.417
LSTM Fasttext	0.234	0.936
LSTM GloVe	0.581	0.638
GRU GloVe	0.651	0.603

Tabelle: Konfiguration und Ergebnisse für die am besten performenden Modelle

ID	Tweet	Label Original	Label Team
805	keep up the opposition to @user #endthenation	1	0
14306	@user i'm blessedt iconic #lovebeingalegend	1	0
18	retweet if you agree!	1	0
2994	aloha and peace symbol on earth symbol we oppose fascism and all warsforoil	1	0

Tabelle: Beispiele für unzureichende Qualität der originalen Labels

↓

Entscheidung gegen Resampling und stattdessen für Datenerweiterung

Verfeinerung Fragestellung (Jan.)

Ausgangsfrage

Wie kann mithilfe von klassischen Verfahren des maschinellen Lernens sowie mit Deep-Learning-Verfahren Hate Speech auf sozialen Plattformen anhand des Beispiels Twitter erkannt werden?

- Klassifikation von Text / Sentiment Analysis

Methodische Detailfragen

- Gibt es Modelle mit besserer Performance z.B. BERT/RoBERTa?

Modell	F1	Recall
BalancedRF TFIDF	0.564	0.702
NB TFIDF	0.513	0.754
NB W2V	0.469	0.712
SVM TFIDF Linear	0.539	0.779
SVM GloVe rbf	0.506	0.823
LSTM GloVe	0.524	0.803
GRU GloVe	0.563	0.602

Tabelle: Konfiguration und Ergebnisse für die am besten performenden Modelle

Training Transformer-Modelle

BERT/RoBERTa

- Erfassung lokaler als auch globaler Kontexte mittels Self-Attention → präzisere Sprachrepräsentation
- Bi-direktionales Training → verbessertes Verständnis von Satzkontexten
- Auf großem Bestand von Textdaten vortrainiert → verbesserte Generalisierungsfähigkeit

Modell	F1-Score	Recall
BERT	0.59	0.678
RoBERTa	0.593	0.619

Tabelle: Ergebnisse der besten Modelle

Modell	F1	Recall
BalancedRF	0.842	0.774
NB	0.669	0.989
SVM	0.753	0.785
LSTM	0.739	0.907
GRU	0.745	0.751
BERT	0.781	0.797
RoBERTa	0.847	0.814

Tabelle: Konfiguration und Ergebnisse für die am besten performenden Modelle mit manuell gelabelten Datensätzen

Live Demo

Fazit und Ausblick

Herausforderungen

- Datensätze mit Klassenungleichgewicht.
- Fehlklassifizierte Labels beeinflussten die Ergebnisse.
- Hoher Anteil an Rechtschreibfehlern in Social-Media-Texten erschwert die Klassifikation.

Erkenntnisse

- Vortrainierte Vektorisierungsmethoden performen meistens besser (GloVe vs. Word2Vec).
- Getestete Modelle haben unterschiedliche Stärken.
- Manuelles Labeling führte zu erheblicher Performance-Verbesserung.

Ausblick

- Zukünftige Arbeiten mit hochwertiger gelabelten Datensätze.
- Einbeziehung zusätzlicher Merkmale (z. B. Hashtags, Emojis).

Erkennung von Hate Speech mit Twitter - Abschlusspräsentation

Elena Marion Friedrich, Nasiba Tuychieva, Sven Ole Wall, Imran Nteli Chalil, Christian Engels

Artificial Intelligence Group,
University of Hagen, Germany

18. März 2025