Erkennung von Hate Speech mit Twitter - Zwischenpräsentation

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

Artificial Intelligence Group,
University of Hagen, Germany

17. Dezember 2024

Overview

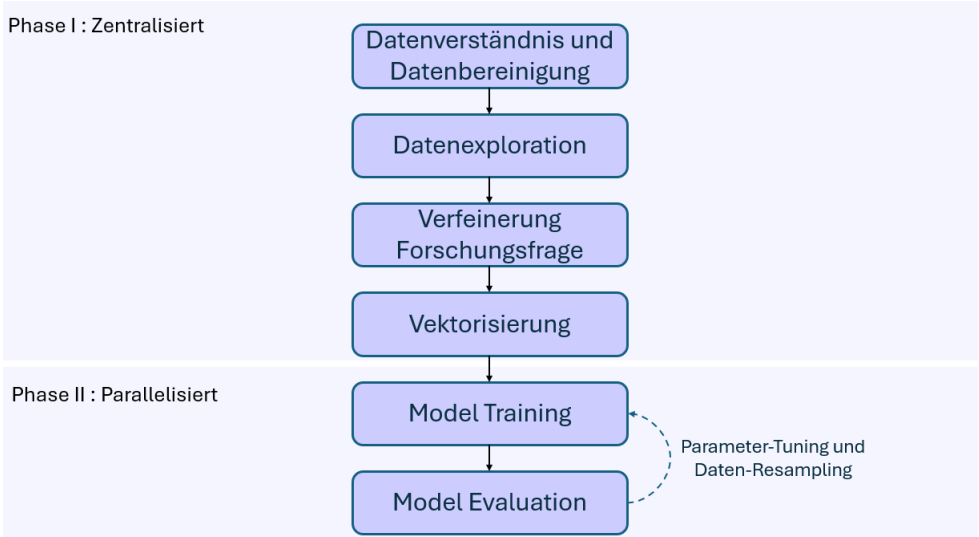
Fragestellung

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

Workflow


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

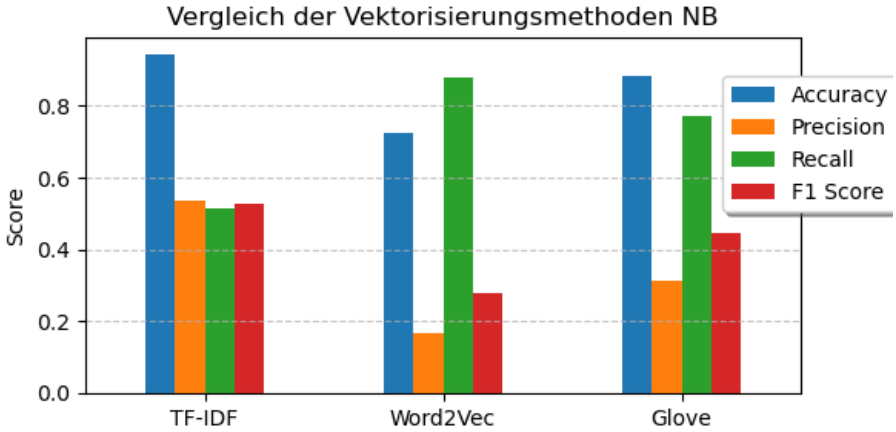
Close-up: Datenbereinigung

- Entfernung unnötiger Zeichen
- Entfernung von Erwähnungen
- Auftrennen zusammengeschriebener Wörter
- Entfernung von Stopwords und häufigsten Wörtern
- Konvertierung von Emojis zu Text
- Stemming und Lemmatisierung

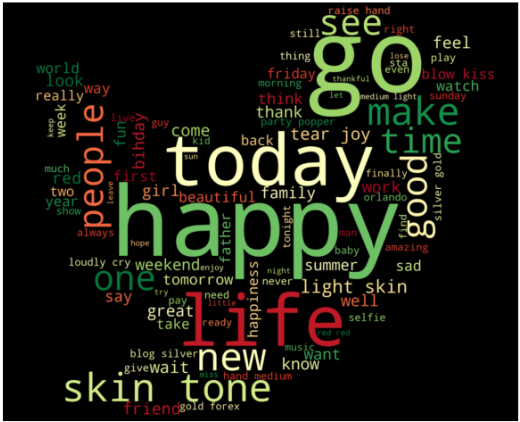
1. #nationalbestfriendsday #sisters #family #love thank you @user for being my best friend  
2. @user i will be there !!! hoping for a beautiful day ∼ clear skies with a breeze would be perfect !!! ;)

1. national good friend day sister family love thank friend sparkle heart grinning face
2. hope beautiful day ∼ clear sky breeze would perfect

Close-up: Vektorisierung



No Hate Speech Wörter bereinigte Daten



Hate Speech Wörter bereinigte Daten

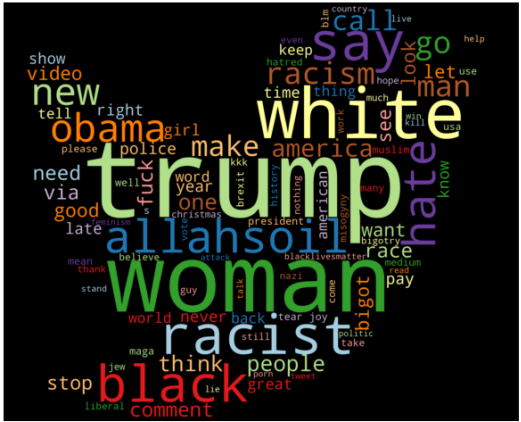


Abbildung: No Hate Speech und Hate Speech Tweets

Binäre Klassifizierung (nur Trainingsdaten)

Labels unterscheiden zwischen Hassrede und Nicht-Hassrede

→ Auswahl der Algorithmen und Metriken

Besonderheiten des Datensatzes

Binäre Klassifizierung (nur Trainingsdaten)

Labels unterscheiden zwischen Hassrede und Nicht-Hassrede

→ Auswahl der Algorithmen und Metriken

Unausgeglichener Datensatz

- 29720 Datensätze Nicht-Hassrede
- 2242 Datensätze Hassrede

→ Verzerrte Modelle und Metriken

Besonderheiten des Datensatzes

Binäre Klassifizierung (nur Trainingsdaten)

Labels unterscheiden zwischen Hassrede und Nicht-Hassrede

→ Auswahl der Algorithmen und Metriken

Unausgeglichener Datensatz

- 29720 Datensätze Nicht-Hassrede
- 2242 Datensätze Hassrede

→ Verzerrte Modelle und Metriken

Verwendete Sprache

- Rechtschreibfehler
- Abkürzungen
- Emoticons
- @-Erwähnungen
- Hashtags

→ Beeinflussung der Modellgenauigkeit
⇒ Vorverarbeitung

Verfeinerung Fragestellung

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?

Ausblick: Model Training

- Naïve Bayes
- Support-Vektor-Maschine
- Ensemble Learning
- RNN-LSTM
- RNN-GRU

Erkennung von Hate Speech mit Twitter - Zwischenpräsentation

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

Artificial Intelligence Group,
University of Hagen, Germany

17. Dezember 2024