Wenn selbst Neuronale Netzwerke beleidigt sind

Johannes Schäfer

Betreuer: Professor Dr. Ulrich Heid

Universität Hildesheim
Institut für Informationswissenschaft
und Sprachtechnologie

23. November 2018

Erkennung von beleidigender Sprache (bS)

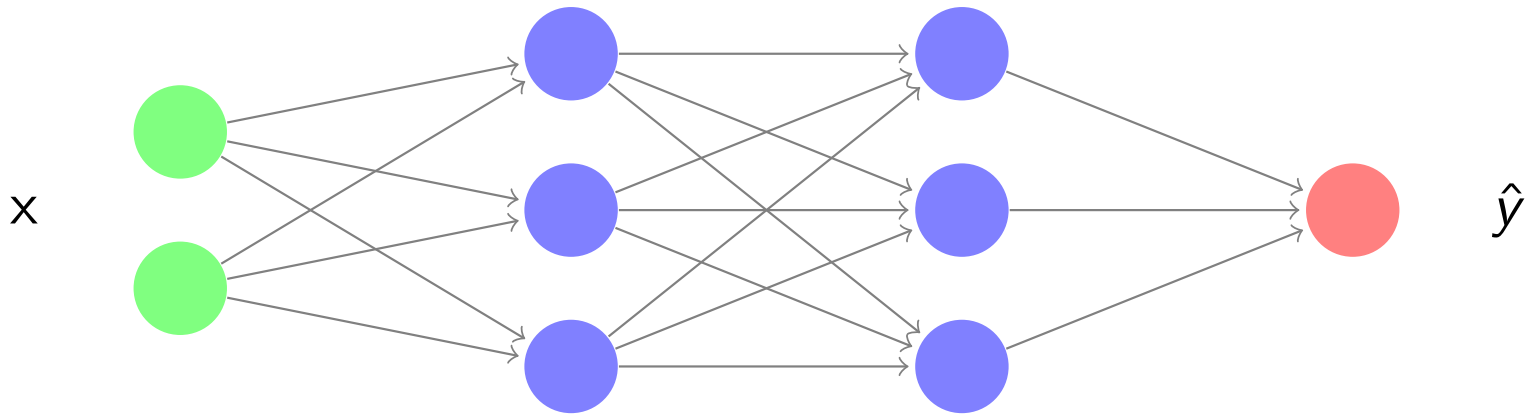
Bedarf an automatischen System zur Erkennung von bS

- Was ist beledigend - und für wen?
- Wie definiert man bS?
  $\rightarrow$ Unklar (individuelle Unterschiede), komplexes Problem
- Ansatz:
  Einschätzungen von Menschen auf realen Daten,
  Lernen einer Maschine anhand dieser Annotationen
- Experiment: bS Erkennung in Kurzbeiträgen (microposts, *Twitter*-Nachrichten, *Tweets*)

Übersicht

Übersicht

Warum neuronale Netzwerke (NNe)?



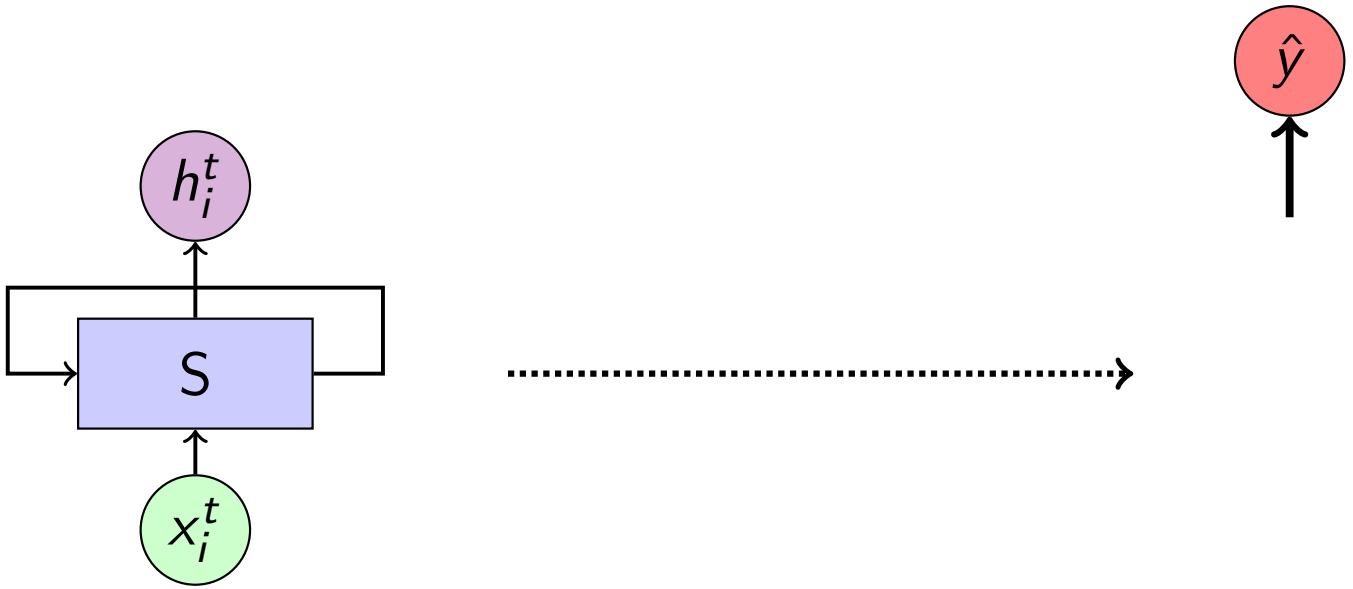
NNe lernen äußerst **komplexe** Funktionen f: $f(x) = \hat{y}$

- Automatische Auswahl der für die Erkennung hilfreichen Merkmale
- Komplexe Kombinationen der Merkmale und Abhängigkeiten

Recurrent NN (RNN) mit Long short-term memory (LSTM) Einheiten

Lernen Representationen anhand von Sequenzen

LSTM Einheit: *Eingabe*-, *Merk/Vergess*- und *Ausgabe*-Tor



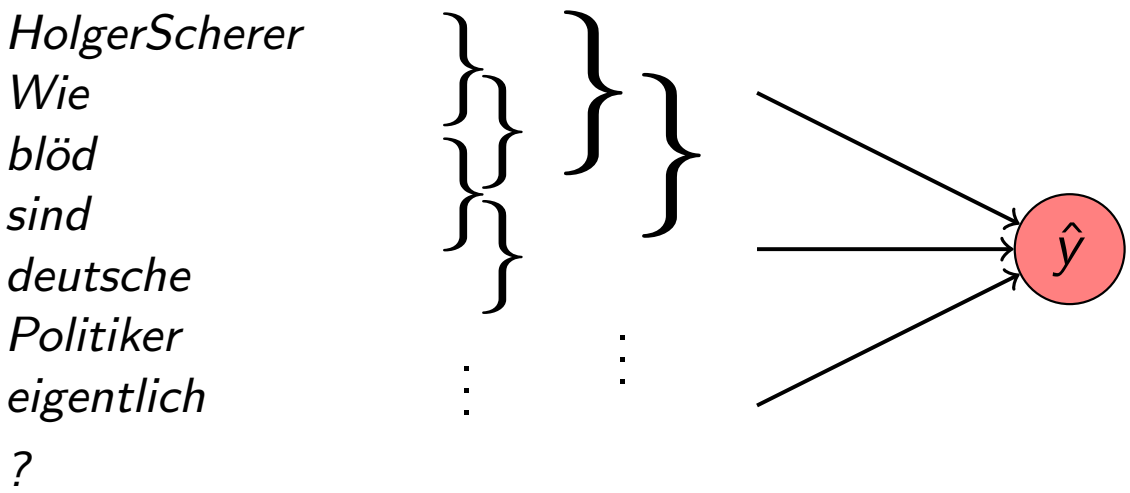
x_i: *@HolgerScherer Wie blöd sind deutsche Politiker eigentlich???*

Convolutional Neural Network (CNN)

Lernen Representation aus Kombinationen von n-Grammen

Convolution und *Max-pooling*

x_i:



Segmentation von Kurzbeiträgen

Repräsentation von Sätzen als Matrix: *Word embeddings* -
auf *social-media*-Daten? → Häufig out-of-vocabulary (OOV) Tokens

- Buchstaben/Zeichen n-Gramme?
- Wortkomponenten embeddings (mittels Kompositazerlegung)
- Ähnlichste Worte (Ähnlichkeit nach Zeichenfolge/*Levenshtein*)

Übersicht

Ergebnisse auf dem GermEval-2018 Testdatensatz

Sequenzen oder Wort n-Gramme?

CNN (Wort n-Gramm-Kombinationen) besser als
LSTM (Sequenz-lerner)

Performanz bei der Textsegmentierung

- Wort-Komponent *embeddings* besser als Wort *embeddings*
- CNN auf Zeichen/Buchstaben-n-Grammen? Noise!? $\rightarrow$ Future Work

Übersicht

Beleidigende Sprache in Microposts

Beobachtungen

- Normalerweise ist nur ein Teil der Nachricht beleidigend → *Auslöser*
- Meist direkter Gebrauch von beleidigendem Wort/Wörtern
  → leicht Erkennbar (Wörterbuch)
  ⇒ Allerdings: Variierende Wortformen; Normalisierungsmethoden
- Komplexe Ausdrücke:
  - Vergleiche/Dehumanisierungen (`"... ist ..."`)
  - Unterstellungen (`"... hat ..."`)
  - Ironie/Sarkasmus

Referenzen

- Germeval 2018: `https://projects.fzai.h-da.de/iggsa/`, Ergebnisse: `https://github.com/uds-lsv/GermEval-2018-Data`.
- Johannes Schäfer. *HIIwiStJS at GermEval-2018: Integrating Linguistic Features in a Neural Network for the Identification of Offensive Language in Micropost*, In Proceedings of the Workshop Germeval 2018 – Shared Task on the Identification of Offensive Language. Vienna, Austria. September 21, 2018.