Neural networks at hate speech and offensive language detection with a focus on linguistic features

Johannes Schäfer
johannes.schaefer@uni-hildesheim.de

Supervisor: Professor Dr. Ulrich Heid

University of Hildesheim
Institute for Information Science
and Natural Language Processing

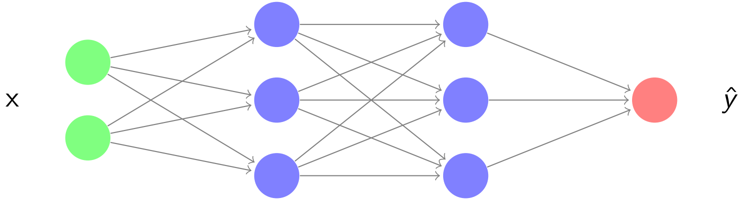
February 27th, 2019

Hate Speech (HS) and Offensive Language (OL) Detection

Need for automatic detection in social media posts

- What is offensive - and to whom?
- How is OL/HS defined?
  $\rightarrow$ Not clear (even to humans), complex problem
- Empirical approach:
  - gather (multiple) human assessments of actual data
  - learn model on this data using machine learning
  - automatically find patterns of HS/OL

Why Neural Networks (NNs)?



NNs learn highly **complex** function f: $f(x) = \hat{y}$

- Based on raw input, no predetermined features
  $\rightarrow$ can learn variety of features/combinations themselves
- Identify helpful input features for the classification task
- Complex combinations of features

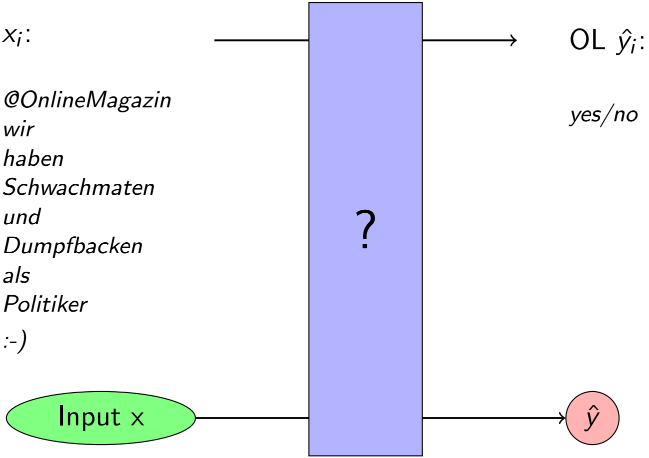
Motivation

- NN approaches: purely statistical, processing of signal data
- Linguistic utterances $\rightarrow$ contain structure
- Support the NN
  (careful: not predetermined features! only as additional input)
- Basic principle of CL:
  statistical processing with the inclusion of linguistic knowledge!

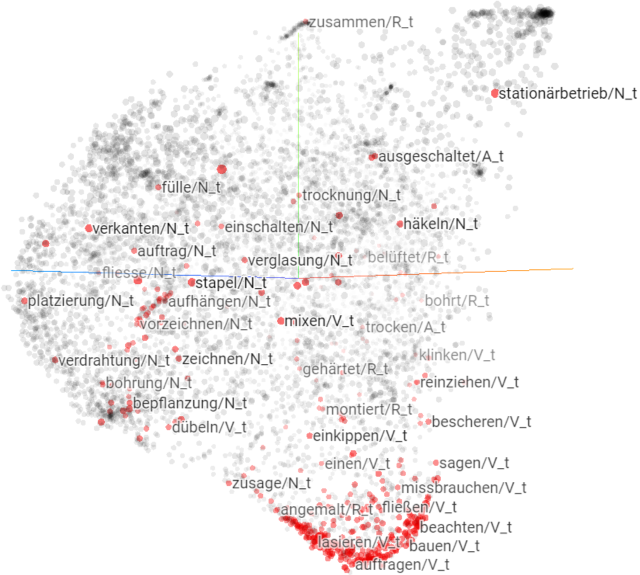
Overview

Offensive Language Detection Task

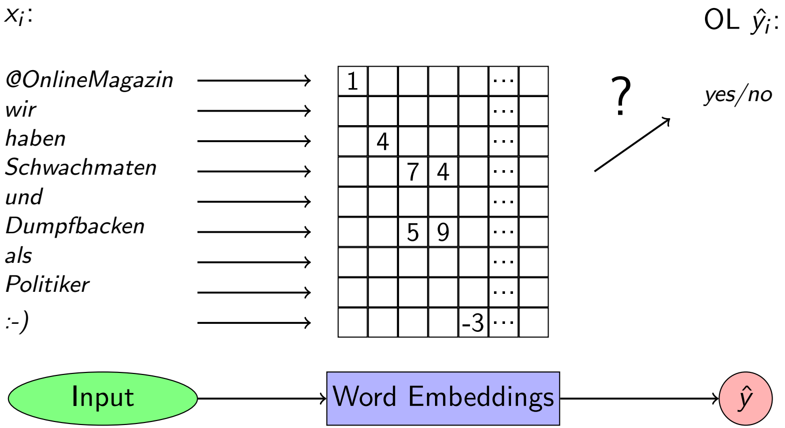
Encoding the Input Sequences (Text)



Semantic Representation of Words

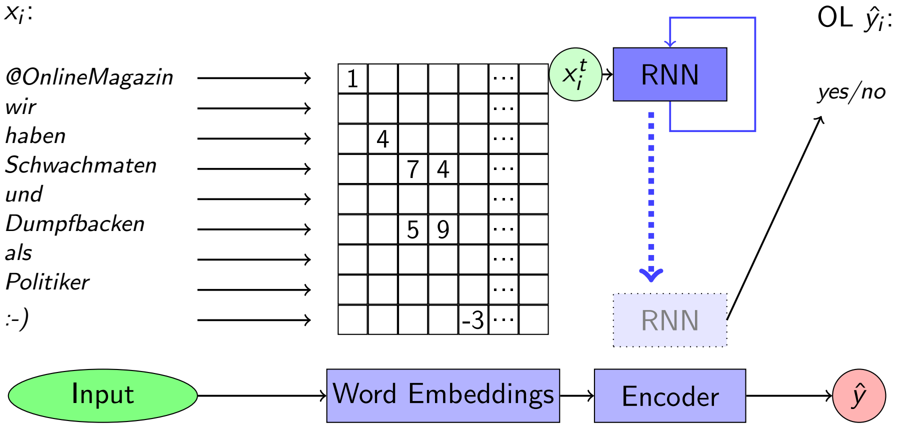


Encoding the Semantic Representations



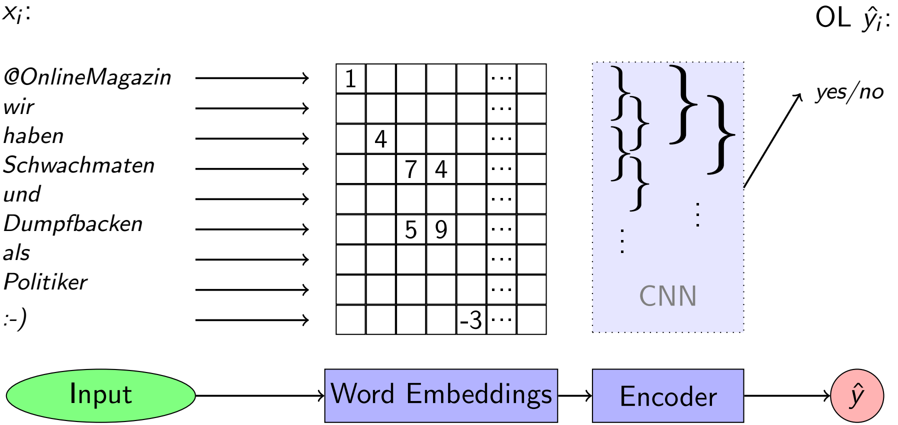
Learning Sequences

Recurrent Neural Network (RNN) using Long Short-Term Memory (LSTM) cells



Learning on N-Grams

Convolutional Neural Network (CNN)



Performance of the Architectures

Results on the GermEval-2018[1] test dataset

- Recurrent NN (RNN) using Long short-term memory (LSTM) units: Learning representations on sequences $F_{1,\text{macro-avg}} = 70.66\ \%$
- Convolutional Neural Network: Learning representations as combination of n-grams $F_{1,\text{macro-avg}} = 71.14\ \%$

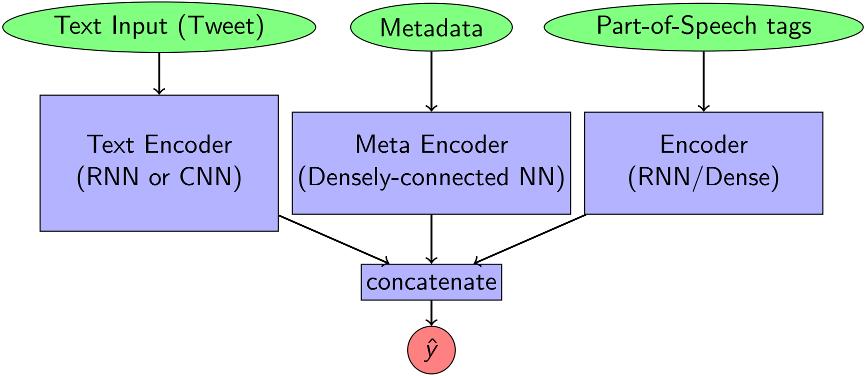
$\Rightarrow$ Usually only part of message offensive $\rightarrow$ *trigger*

[1] `https://projects.fzai.h-da.de/iggsa/`

Additional sub-networks

Overall NN architecture

extended from Founta et al. 2018



Results:

Metadata sub-network - improvements; minimal with POS tags

Considering Word Components

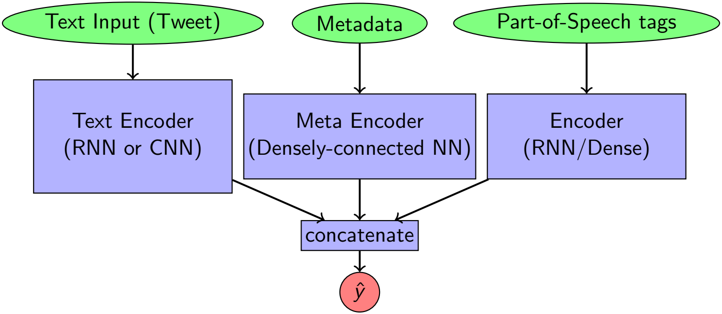
Motivation

- Pre-trained word embeddings (initial weights)
- OOV words
  (*Politidioten, Oberdummzicke, Sozialschmarotzer, Migrantenpack*)
- First implementation:
  Handle compounds as seperate words assuming compositionality

Performance using compound splitting

CNN on word component embeddings: $F_{1,\text{macro-avg}} = 73.42\ \%$

Where to integrate linguistic features?



Effect of additional features in parallel sub-networks is low

→ Linguistic features directly in the text encoding!

Conclusion: *"Digital Methods in Political Science"*

Sophisticated analysis necessary
for automatic offensive language and hate speech detection

- Offensive language hidden in words or multi-word constructions
- What NN approaches and linguistic features can be discussed to analyze political discussions in particular?
- Possibilities to include the target/victims (detection of typical groups)

References

- 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.
- Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. *Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language*. 14th Conference on Natural Language Processing KONVENS 2018. 2018. `https://projects.fzai.h-da.de/iggsa/`, Results: `https://github.com/uds-lsv/GermEval-2018-Data`.
- Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. *A unified deep learning architecture for abuse detection.* CoRR, abs/1802.00385. 2018.