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
Need for automatic detection in social media posts

- What is offensive - and to whom?
- How is OL/HS defined?
  → 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 → 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 → 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

1. Methods: Neural Network Systems
2. Extensions using Linguistic Features
3. Future Work: Further Features to detect HS/OL
Offensive Language Detection Task

Encoding the Input Sequences (Text)

\[ x_i: \]

@OnlineMagazin
wir
haben
Schwachmaten
und
Dumpfbacken
als
Politiker
:-)
Semantic Representation of Words
Encoding the Semantic Representations

$x_i$: @OnlineMagazin, wir, haben, Schwachmaten, und, Dumpfbacken, als, Politiker, :-) → 

$\hat{y}_i$: OL, yes/no

Input → Word Embeddings → $\hat{y}$
Learning Sequences
Recurrent Neural Network (RNN) using Long Short-Term Memory (LSTM) cells

\[ x_i \rightarrow \text{Input} \rightarrow \text{Word Embeddings} \rightarrow \text{Encoder} \rightarrow \hat{y} \]

@OnlineMagazin
wir
haben
Schwachmaten
und
Dumpfbacken
als
Politiker
:-)
Learning on N-Grams
Convolutional Neural Network (CNN)

\( x_i : \)

@OnlineMagazin
wir
haben
Schwachmaten
und
Dumpfbacken
als
Politiker
:-)

\( \hat{y}_i : \)

\begin{bmatrix}
1 & \cdots \\
4 & \cdots \\
7 & 4 & \cdots \\
5 & 9 & \cdots \\
\vdots & & \ddots \\
\end{bmatrix}

\begin{bmatrix}
\{ \\
\{ \\
\{ \\
\{ \\
to \\
to \\
to \\
end{bmatrix}

yes/no

CNN

Input \rightarrow Word Embeddings \rightarrow Encoder \rightarrow \hat{y}
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
  \[ F_{1,\text{macro-avg}} = 71.14 \% \]

⇒ Usually only part of message offensive → trigger

\(^1\)https://projects.fzai.h-da.de/iggsa/
Additional sub-networks

Overall NN architecture extended from Founta et al. 2018

Text Input (Tweet) → Text Encoder (RNN or CNN)

Metadata → Meta Encoder (Densely-connected NN)

Part-of-Speech tags → Encoder (RNN/Dense)

concatenate

\( \hat{y} \)

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 separate words assuming compositionality

Performance using compound splitting

CNN on word component embeddings: \( F_{1,\text{macro-avg}} = 73.42\% \)
Where to integrate linguistic features?

- Text Input (Tweet)
- Metadata
- Part-of-Speech tags

Text Encoder (RNN or CNN) 
Meta Encoder (Densely-connected NN) 
Encoder (RNN/Dense) 

concatenate 

\( \hat{y} \)

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

