Offence in Dialogues: A Corpus-Based Study

Johannes Schäfer and Ben Burtenshaw

Institute for Information Science and Natural Language Processing
University of Hildesheim, Universitätsplatz 1, Hildesheim, Germany
johannes.schaefer@uni-hildesheim.de

Computational Linguistics & Psycholinguistics Research Center
The University of Antwerp, Lange Winkelstraat 40-42, Antwerp, Belgium
benjamin.burtenshaw@uantwerpen.be

Research on **Offensive Language** and **Hate Speech**

- Mostly detection based on isolated instances (e. g. Tweets)
- System accuracy of max. 70-80% (*Germeval 2018*, *Offenseval 2019*)
- Suggested course of action: Deletion!?
  → Should a system act based on that?

Our vision:

Counter offence! (automatically) - before conversations turn illegal

This study:

- How do humans react and use tactics?
- Data acquisition and first steps of analysis

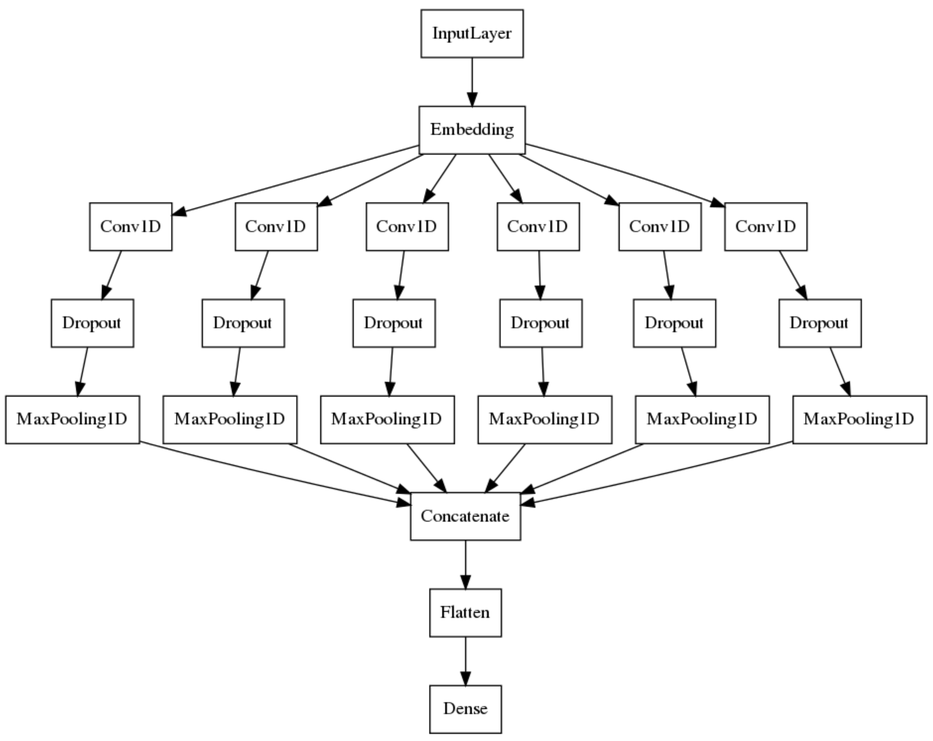
Corpus data

Target: English text,
Large set of long conversations (empirical, statistical analysis)

Source: *Europe-Subreddit (reddit.com)*,
Comments in forum-like tree-structured threads

Corpus creation process[a]**:**

* Download of submissions and comments: *Python 3 psaw* module using the *pushshift.io reddit* API
  ⇒ over 11M posts, 357k submissions
* Reformating into a tree-structured XML-corpus (comments nested recursively)
* Annotation (Offensive Language Detection):
  - Classification of comments in isolation
  - OLID (Zampieri et al., 2019b) as training data
  - CNN on word embeddings

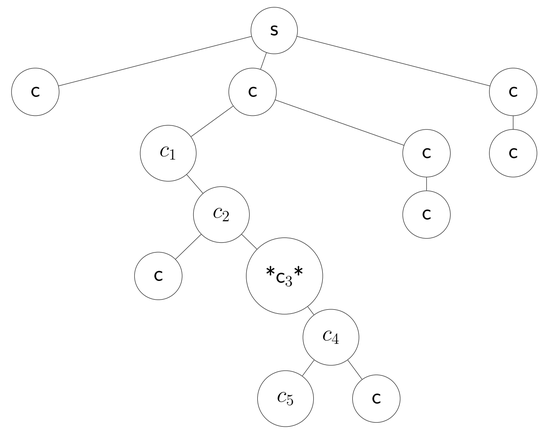


- Annotation: Binary and probabilistic output
  ⇒ Measuring the level of offensiveness of comments

[a] Code to build and annotate corpus available at https://github.com/Johannes-Schaefer/oid_ranlp19

Extraction of linear dialogues

Target: multiple turns in linear (order) dialogues from tree-structured corpus threads

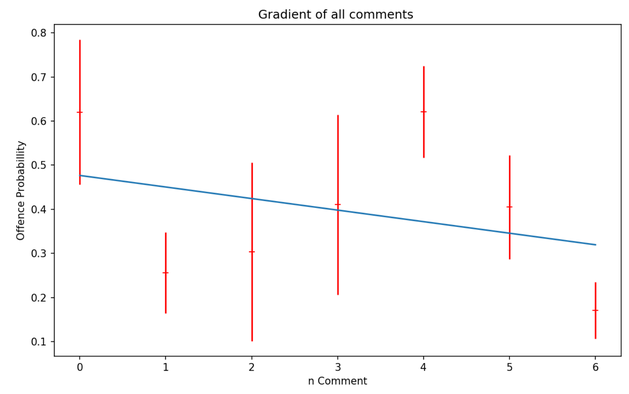


- Fixed structure:
  Comment with off_score >0.5 as trigger;
  window size 3 leads to 67k linear dialogues
- Branching? Rather infrequent

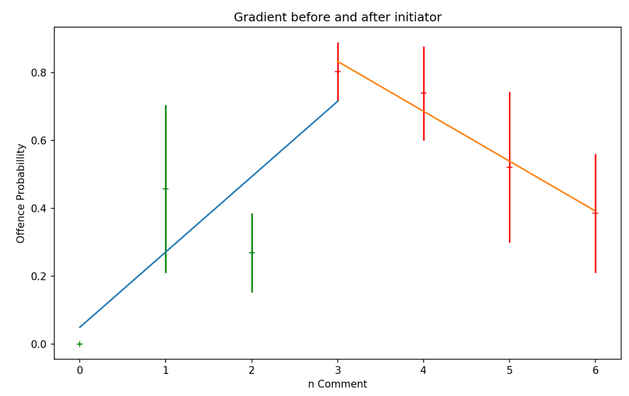
Analysis of **linear dialogues** using **decoupling functions**

→ Decoupling - reduction of sequential values of a variable

a) Extraction by declining **gradient** of entire linear conversation:



b) Split dialogue in half; analysis of 2 gradients:



Future work:

- Detailed analysis of decoupling methods and **evaluation** (human annotation of extracted linear dialogues); revised detection system (Burtenshaw and Schäfer, submitted)
- Towards **automatically extracting tactics** from social media data; frequencies; measure success of tactics?
- **Goal**: Automatic generation of tactics
- Outlook:
  EU-project *DeTACT (Detect Then ACT: Taking Direct Action against Online Hate Speech by Turning Bystanders into Upstanders)*

REFERENCES:
[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.] [Marcos Zampieri, Shervin Malmasi, Preslav Nakov,Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). arXiv preprint arXiv:1903.08983.] [Marcos Zampieri, Shervin Malmasi, Preslav Nakov,Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Predicting the Type and Target of OffensivePosts in Social Media. In Proceedings of NAACL.] [Ben Burtenshaw and Johannes Schäfer. Detecting Declining Offence in Conversations. submitted.]