A Dataset of Offensive German Language Tweets Annotated for Speech Acts

Melina Plakidis, Georg Rehm
DFKI GmbH, Germany

Motivation

- Several works examine offensive language for linguistic characteristics or phenomena
- Especially difficult to detect offensive language which is expressed implicitly rather than explicitly
- However, little research on pragmatic characteristics of offensive language
- Linguistic analysis regarding pragmatic characteristics of offensive language might contribute to the improvement of hate speech detection
- Idea: Combining hate speech detection with speech act theory:
  - Austin (1962)
  - Five classes of illocutionary acts in Searle's taxonomy (1979): *Assertives*, *Directives*, *Commissives*, *Expressives*, *Declarations*

Hypotheses

1) There are more directives in offensive than in non-offensive language (excluding *address*)
2) There are more expressives of type *complain* in offensive than in non-offensive language
3) Speech acts of type *assert* occur less frequently in offensive than in non-offensive language
4) Declarative sentences are the most dominant sentence type overall

Data and Annotation

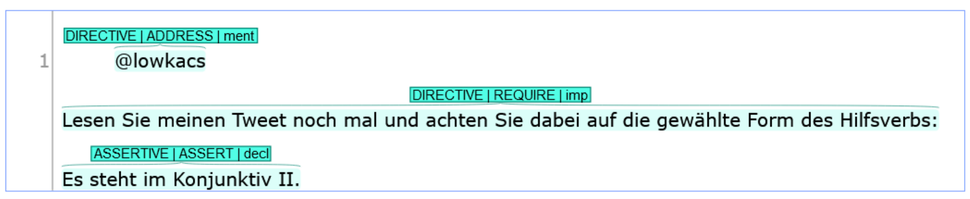
Dataset

- Created within the frame of the second edition of the GermEval Shared Task on the Identification of Offensive Language (Struß et al., 2019)
- German language Tweets
- Task 2 constitutes 3 subtasks:
  - Subtask 1: Binary (Offense, Other)
  - Subtask 2: Fine-grained classification (Profanity, Insult, Abuse, Other)
  - Subtask 3: Implicit vs. explicit offensive language

- Annotation scheme inspired by Searle (1979) and Compagno et al. (2018)
- Building upon Weisser (2018), it includes two levels:
  - Syntactical Level: Describes the sentence type of each speech act (14 sentence types)
  - Speech Act Level: Type of speech act, divided into coarse-grained (6 speech acts) and fine-grained speech act level (23 speech acts)

Speech Act Level	Speech Acts: Coarse-grained	Assertive	Expressive	Commissive	Directive	Unsure	Other
	Speech Acts: Fine-grained	Assert, Sustain, Guess, Predict, Agree, Disagree	Rejoice, Complain, Wish, Apologize, Thank, expressEmoji	Engage, Accept, Refuse, Threat	Request, Require, Suggest, Greet, Address	Unsure	Other
Syntactical Level	Sentence Types	Alternative Question, Declarative, Exclamative, Yes-/ No-Question, Fragment, Imperative, Interjection, Conjunctive, Mention, Multiple, Non-textual, W-Question					

- Annotation with INCEpTION (Klie et al., 2018)
- Final dataset consists of 600 XML files



Results

Table 1: Frequency of coarse-grained and fine-grained speech acts in offensive language categories

	Offensive		Other		Implicit		Explicit		Abuse		Profanity		Insult		Total	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Assertive	541	33.9	123	37.3	113	41.2	80	28.2	114	31.8	109	33.6	125	35.3	664	34.5
Assert	461	28.9	114	34.5	95	34.3	69	24.3	96	27.0	92	28.4	109	30.8	575	29.9
Sustain	10	0.6	2	0.6	2	0.7	0	0.0	4	1.1	1	0.3	3	0.8	12	0.6
Guess	25	1.6	1	0.3	9	3.2	2	0.7	2	0.6	7	2.2	5	1.4	26	1.4
Predict	30	1.9	2	0.6	6	2.2	7	2.5	6	1.7	4	1.2	7	2.0	32	1.7
Agree	11	0.7	2	0.6	2	0.7	1	0.4	4	1.1	4	1.2	0	0.0	13	0.7
Disagree	4	0.3	2	0.6	0	0.0	1	0.4	1	0.3	1	0.3	1	0.3	6	0.3
Expressive	345	21.6	47	14.2	44	15.9	73	25.7	76	21.4	72	22.2	80	22.6	392	20.4
Rejoice	14	0.9	3	0.9	1	0.4	6	2.1	1	0.3	4	1.2	2	0.6	17	0.9
Complain	232	14.6	17	5.2	37	13.4	52	18.3	37	10.4	45	13.9	61	17.2	249	12.9
Wish	10	0.6	1	0.3	0	0.0	3	1.1	3	0.8	4	1.2	0	0.0	11	0.6
Apologize	0	0.0	1	0.3	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.1
Thank	4	0.3	4	1.2	0	0.0	0	0.0	1	0.3	2	0.6	1	0.3	8	0.4
expressEmoji	85	5.3	21	6.4	6	2.2	12	4.2	34	9.6	17	5.2	16	4.5	106	5.5
Commissive	17	1.1	3	0.9	0	0.0	3	1.1	1	0.3	12	3.7	1	0.3	20	1.0
Engage	11	0.7	2	0.6	0	0.0	0	0.0	0	0.0	11	3.4	0	0.0	13	0.7
Accept	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0
Refuse	1	0.1	0	0.0	0	0.0	1	0.4	0	0.0	0	0.0	0	0.0	1	0.1
Threat	5	0.3	1	0.3	0	0.0	2	0.7	1	0.3	1	0.3	1	0.3	6	0.3
Directive	522	32.7	108	32.7	99	35.7	99	34.9	130	36.6	85	26.2	109	30.8	630	32.7
Request	130	8.2	33	10.0	23	8.3	23	8.1	36	10.1	24	7.4	24	6.8	163	8.5
Require	65	4.1	11	3.3	7	2.5	16	5.6	13	3.7	13	4.0	16	4.5	76	4.0
Suggest	14	0.9	1	0.3	2	0.7	1	0.4	4	1.1	3	0.9	4	1.1	15	0.8
Greet	1	0.1	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.3	1	0.1
Address	312	19.6	63	19.1	67	24.2	59	20.8	77	21.7	45	13.9	64	18.1	375	19.5
Unsure	113	7.1	37	11.2	18	6.5	15	5.3	30	8.5	35	10.8	15	4.2	150	7.8
Other	56	3.5	12	3.6	2	0.7	14	4.9	5	1.4	11	3.4	24	6.8	68	3.5
Total	1594	100.0	330	100.0	277	100.0	284	100.0	355	100.0	324	100.0	354	100.0	1924	100.0

Addressing the Hypotheses

- Hypothesis 1 refuted:
  - 16.4% directives in offensive and 16.9% directives (excluding *address*) in non-offensive tweets
- Hypothesis 2 confirmed:
  - 14.6% of *complain* in offensive tweets and 5.2% in non-offensive tweets
- Hypothesis 3 confirmed:
  - 28.9% *of assert* in offensive tweets, 34.5% in non-offensive tweets
- Hypothesis 4 confirmed:
  - *Declarative* most frequent sentence type (27.2%)

Table 2: Frequency of sentence types in offensive language categories

	Offensive		Other		Implicit		Explicit		Abuse		Profanity		Insult		Total	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Alt-f	3	0.2	2	0.6	0	0.0	0	0.0	1	0.3	2	0.6	0	0.0	5	0.3
Decl	434	27.2	90	27.3	92	33.2	70	24.6	73	20.6	93	28.7	106	29.9	524	27.2
Excl	133	8.3	10	3.0	16	5.8	47	16.5	28	7.9	23	7.1	19	5.4	143	7.4
F	63	4.0	19	5.8	17	6.1	7	2.5	16	4.5	13	4.0	10	2.8	82	4.3
Frag	270	16.9	59	17.9	43	15.5	31	10.9	59	16.6	74	22.8	63	17.8	329	17.1
Hashtag	57	3.6	14	4.2	4	1.4	15	5.3	7	2.0	12	3.7	19	5.4	71	3.7
Imp	45	2.8	6	1.8	2	0.7	10	3.5	10	2.8	11	3.4	12	3.4	51	2.7
Intj	5	0.3	2	0.6	0	0.0	3	1.1	2	0.6	0	0.0	0	0.0	7	0.4
Kon	37	2.3	8	2.4	13	4.7	9	3.2	6	1.7	2	0.6	7	2.0	45	2.3
Ment	310	19.4	63	19.1	66	23.8	59	20.8	78	22.0	43	13.3	64	18.1	373	19.4
Mult	7	0.4	2	0.6	0	0.0	2	0.7	1	0.3	3	0.9	1	0.3	9	0.5
Non-txt	85	5.3	21	6.4	6	2.2	12	4.2	33	9.3	18	5.6	16	4.5	106	5.5
Other	86	5.4	22	6.7	10	3.6	5	1.8	23	6.5	26	8.0	22	6.2	108	5.6
W-f	59	3.7	12	3.6	8	2.9	14	4.9	18	5.1	4	1.2	15	4.2	71	3.7
Total	1594	100.0	330	100.0	277	100.0	284	100.0	355	100.0	324	100.0	354	100.0	1924	100.0

Conclusions

- Offensive language mainly differs from non-offensive language in the respect that offensive language contains more expressives and less assertives than non-offensive language
- Biggest difference when comparing tweets containing implicit offensive language with tweets containing explicit offensive language
  - Implicit: Seem to lack the tendency to overtly express emotions
    - Have the lowest frequency of expressives (excluding non-offensive tweets) and the highest frequency of assertives
  - Explicit: Show the opposite
    - Lowest frequency of assertives and highest frequency of expressives
- Results suggest that differences exist regarding the distribution of speech acts in offensive language and non-offensive language
- Remains to be seen if an accurate speech act classifier can be developed as one additional component in larger hate speech detection system

References: https://github.com/MelinaPl/speech-act-analysis#references
Dataset, Examples, Code: https://github.com/MelinaPl/speech-act-analysis
Corresponding author: Melina Plakidis, melina.plakidis@dfki.de

LREC 2022, 20th – 25th June 2022