Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4799–4807
Marseille, 20-25 June 2022

A Dataset of Offensive German Language Tweets Annotated for Speech Acts

Melina Plakidis[1,2], Georg Rehm[1,2]
[1] DFKI GmbH, Alt-Moabit 91c, 10559 Berlin, Germany
[2] Humboldt-Universität zu Berlin, Dorotheenstraße 24, 10117 Berlin, Germany

Abstract

We present a dataset consisting of German offensive and non-offensive tweets, annotated for speech acts. These 600 tweets are a subset of the dataset by (Struß et al., 2019) and comprises three levels of annotation, i. e., six coarse-grained speech acts, 23 fine-grained speech acts and 14 different sentence types. Furthermore, we provide an evaluation in both qualitative and quantitative terms. The dataset is made publicly available under a CC-BY-4.0 license.

Keywords: Speech acts, hate speech detection, offensive language, annotation, corpus annotation

1 Introduction

In recent years, research has invested a considerable amount of effort in offensive language and other phenomena related to online communication. Hate speech not only affects individuals or minority groups but it might also threaten social cohesion (Weber et al., 2019). Thus, the automatic detection of hate speech and offensive language[1] continues to be an important and very relevant research topic.

While there is a large body of research in this area, approaches often merely classify text using binary labels (Burnap and Williams, 2016; Risch et al., 2021). Hate speech classification is in itself a complex task because it is highly subjective what constitutes hate speech. In addition, it is even more difficult to detect hate speech if it is expressed *implicitly* rather than explicitly (Palmer et al., 2020; Struß et al., 2019). Surprisingly little research exists on the pragmatic characteristics of offensive language. Pragmatics is "the study of how utterances have meanings in situations" (Leech, 1983, p. x). To address this gap and contribute to the improvement of hate speech detection, we analysed the pragmatic characteristics of offensive language. We conducted a speech act analysis of a dataset of German tweets that contain offensive language. We use a subset of the 2019 GermEval Shared Task on the Identification of Offensive Language dataset (Struß et al., 2019).

Similar studies exist in which speech act theory (Austin, 1962; Searle, 1969) is applied to data (Jurafsky, 1997; Zhang et al., 2011; Compagno et al., 2018; Vosoughi and Roy, 2016; Weisser, 2018). However, many of these concentrate on spoken language, often confined to restricted discourse scenarios such as the SPAADIA Trainline Corpus (Leech and Weisser, 2013) which consists of phone conversations between call-centre agents and customers concerning bookings of train tickets. Other approaches apply speech act theory to data from Twitter (Zhang et al., 2011; Vosoughi and Roy, 2016) or Reddit (Compagno et al., 2018). Nevertheless, to our knowledge, only Dhayef and Ali (2020) investigate the distribution of speech acts in a hate speech dataset. Consequently, there appears to be a need for more research in this area. Our study aims to contribute to automated hate speech detection by providing a pragmatic analysis of offensive language and it also attempts to contribute to speech act theory by testing its applicability on real life data and sharing findings on frequency, syntactical realisation and common sequences of speech acts. We hypothesise the following: (i) there are more *directives* in offensive than in non-offensive language (excluding *address*); (ii) there are more *expressives* of type *complain* in offensive than in non-offensive language; (iii) speech acts of type *assert* occur less frequently in offensive than in non-offensive language; (iv) declarative sentences are the most dominant sentence types overall. Hypothesis (i) is based on Dhayef and Ali (2020); hypothesis (ii) is motivated due to the assumption that uttering a hateful comment corresponds to the speaker having a negative attitude towards the targeted person or group (Dhayef and Ali, 2020). Hypothesis (iii) is assumed because of the hypotheses (i) and (ii). If there are more *expressives* and *directives* in offensive language, then they might displace a number of *assert* speech acts. Moreover, hypothesis (iv) is expected because we regard declaratives to be the default sentence type (Weisser, 2018).

The remainder of this article is structured as follows. Section 2 reports on related work. Section 3 provides information about the dataset. Section 4 presents our annotation scheme regarding the syntactical and speech act level and in Section 5, the results are discussed. Section 6 concludes the article.

[1]The terms "hate speech" and "offensive language" are used synonymously in this paper. It should be noted, however, that these terms are not always used synonymously in the literature and that there are various differing definitions. Poletto et al. (2020) provide a corresponding overview.

2 Related Work

Starting out with the publication of the first hate speech dataset (Spertus, 1997), a considerable body of research has been carried out in this field of research

(Poletto et al., 2020), including a number of survey papers (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018; Mishra et al., 2020; Poletto et al., 2020). While English is the dominant language in current research, datasets in and for other languages also exist, including German (Ross et al., 2017; Wiegand et al., 2018; Struß et al., 2019). With regard to the annotation schemes, we can distinguish three main approaches (Poletto et al., 2020), i. e., binary classification (Risch et al., 2021; Burnap and Williams, 2016), non-binary classification consisting of more than two labels (Kumar et al., 2018; Struß et al., 2019) and more complex schemes that consist of multiple levels (Zampieri et al., 2019).

Several authors examine hate speech for specific linguistic phenomena. Neutral adjectives, for example, can acquire a pejorative meaning if they are nominalized (Palmer et al., 2017; Palmer et al., 2020), using the definite plural instead of the bare plural can be used to indicate non-membership (Palmer et al., 2020) and the use of both including and excluding pronouns may contribute to construct a dichotomy between the in- and the out-group, i. e., *distancing* or *othering* (Palmer et al., 2020). Those findings have contributed to the development of more complex annotation schemes. In addition, Palmer et al. (2020) create an annotation scheme based on four questions which include the presence or absence of offensiveness, slurs, adjectival nominalizations and distancing to establish a dataset.

Regarding speech act theory, nowadays, various studies attempt to classify speech acts automatically. To enable automatic classification, they build speech act taxonomies which are often based on Austin (1962) and Searle (1979). Compagno et al. (2018), for example, develop a taxonomy based on Searle's five speech act classes (*assertives*, *directives*, *expressives*, *commissives*, *declarations*) which they validate on a Reddit corpus with threads on autoimmune diseases. They also provide sub-classes, resulting in 17 speech acts in total. Similar approaches with taxonomies based on Searle's classes include Zhang et al. (2011) and Vosoughi and Roy (2016). Weisser (2018) introduce a more complex scheme with the Dialogue Annotation and Research Tool (DART).[2] The current version (version 3.0[3]) of DART classifies dialogue by syntactic categories (ten in total) and speech acts (162 classes), among others.

3 Dataset

The annotation scheme was applied to a subset of the dataset created for task two of the 2019 GermEval Shared Task on the Identification of Offensive Language (Struß et al., 2019).It consists of offensive and non-offensive German language tweets that do not include any surrounding context in the form of other tweets. Evidently, this is not ideal for the application of speech act theory which highly depends on context. This is why we established the category *unsure*.

Task two of the GermEval shared task included three subtasks. The first subtask uses binary classification (*offense*, *other*) and the second subtask uses a more fine-grained classification (*profanity*, *insult*, *abuse*, *other*). Tweets labeled as *profanity* only contain profane words such as swearwords but do not contain insults or abusive language. If they do, they are labeled as *insult* or *abuse*. The category *abuse* differs from *insult* insofar as in abusive language, the target functions as a representative of a group and "is ascribed negative qualities that are taken to be universal, omnipresent and unchangeable characteristics of the group" (Struß et al., 2019, p. 356) while tweets labeled *insult* are only targeted at individuals without being associated with a group. In addition, tweets containing instances of dehumanization are labeled as *abuse* as well. The third subtask uses binary classification by distinguishing between implicit and explicit offensive language. According to Struß et al. (2019), offensive language counts as being implicit when the reader needs to infer that the tweet is offensive, as the offense is only implied. Moreover, implicit offensive language also entails using figurative language (e. g., sarcasm or irony). The first and second subtask are based on the 2019 data which includes 7,025 tweets while the third subtask is based on the 2018 data only including tweets categorised as abuse or insult. The data for the third subtask consists of 8,541 tweets of which 2,888 are categorised as offensive (393 of these are instances of implicit offensive language) (Struß et al., 2019). The class *profanity* was excluded from the data used in subtask three as the authors argue that it is "by definition explicit offensive language" (Struß et al., 2019, p. 358). In both the 2018 and 2019 data, tweets were excluded if they did not exclusively contain German language, were retweets, contained less than five tokens or contained a URL (Wiegand et al., 2018; Struß et al., 2019). For our analysis we selected 600 tweets from this dataset. From each of the six classes (*implicit*, *explicit*, *profanity*, *insult*, *abuse*, *other*), 100 tweets were picked randomly. For the classes *implicit* and *explicit*, we used the 2019 gold standard files of the test data of subtask 3 (Struß et al., 2019b) and for the other four classes, we used the 2019 gold standard files of the test data from subtask 1 and 2 (Struß et al., 2019a). We randomly shuffled both test datasets using Python and for each class, the first 100 occurrences were selected; every tweet was saved as a text file.

For the annotation proper we used the open source tool INCEpTION (Klie et al., 2018).[4] The tool supports span annotations as well as the creation of new tagsets and annotation layers. The annotator first segmented the tweet (Section 4.1) and then decided on the speech act and sentence type labels. If uncertain which label to choose, the issue with the tweet was documented

[2] http://martinweisser.org/DART_scheme.html

[3] http://martinweisser.org/publications/DART_manual_v3.0.pdf

[4] https://inception-project.github.io

so that a list of challenges could be assembled (Section 5.2). After the initial pass through all tweets, all annotations were checked and revised once again. The final dataset consists of 600 XML files.

4 Annotation Scheme

Our annotation scheme is mainly inspired by Searle (1979) and Compagno et al. (2018). Additionally, building upon Weisser (2018), it includes two levels: the *syntactical level* describes the sentence type of each speech act and the *speech act level* (consisting of a coarse-grained and a fine-grained level) encodes the type of speech act.

4.1 Tweet Segmentation

Annotation guidelines have been established to help annotators in their decisions. Tweets were segmented according to the following rules (square brackets indicate segmentation boundaries):

- If two adjacent main clauses are connected using a comma or without a conjunction, they are split into two sentences. Exceptions are enumerations and sentences where one main clause has the V2 form due to colloquial language use but should actually be a subordinate clause (Vfin) as in example b) (*Fahrer ist Fluchthelfer*).

  Ex. a) *Wir brauchen #Gelbwesten in der ganzen #EU.. |LBR| Die Völker müssen zeigen das sie diese Scheiß Armutspolitik nicht mehr mit machen.weg mit #Macron weg mit #Merkel und Merkel 2.0 #AKK*[5]

  → *[Wir brauchen #Gelbwesten in der ganzen #EU..] [|LBR| Die Völker müssen zeigen das sie diese Scheiß Armutspolitik nicht mehr mit machen.] [weg mit #Macron] [weg mit #Merkel und Merkel 2.0 #AKK]*

  Ex. b) *Wenn das bislang nicht gefunden wurde, kann man zunächst davon ausgehen, Fahrer ist Fluchthelfer.*[6]

  → *[Wenn das bislang nicht gefunden wurde, kann man zunächst davon ausgehen, Fahrer ist Fluchthelfer.]*

- If unsure where to draw the line between two adjacent fragments, take punctuation into account. If there is none, treat the fragments as one unit.

  Ex. c) *Deutschland das Land der drei Geschlechter der Duckmäuser und ja Sager der Toleranz Fanatiker. Das Land der Verblödung.*[7]

  → *[Deutschland das Land der drei Geschlechter der Duckmäuser und ja Sager der Toleranz Fanatiker.] [Das Land der Verblödung.]*

- Sentence-ending punctuation (".", "*?*", "*!*") is given priority. Exceptions are cases where main clauses have been split for stylistic reasons (ex. d))[8].

  Ex. d) *Ein Egomane. Nutzt jede Gelegenheit, um im Mittelpunkt zu stehen.*[9]

  → *[Ein Egomane. Nutzt jede Gelegenheit, um im Mittelpunkt zu stehen.]*

4.2 Syntactical Level

The main clause of a sentence determines the sentence type (Dudenredaktion, 2016, p. 899). We adapted and slightly modified the annotation scheme of the Georgetown University Multilayer Corpus (GUM) (Zeldes, 2017)[10] for the annotation of sentence types. GUM uses a total of 11 sentence types based on Leech et al. (2003), which explains the similarity between the sentence types used by Zeldes (2017) and Weisser (2018). Table 1 describes the 14 sentence types we use. Our public repository contains additional examples.[11]

4.3 Speech Act Level

For the speech act annotation, we modify the taxonomy by Compagno et al. (2018). Our hierarchically structured scheme consists of 23 fine-grained and six coarse-grained speech acts. The six coarse-grained classes are: *assertives*, *expressives*, *directives*, *commissives*, *unsure* and *other*. Table 2 shows the 23 fine-grained speech acts. Examples can be found in our repository.[12].

4.4 Examples

Figure 1 shows an example tweet from the dataset that was labeled as non-offensive and includes an *expressive* of the type *complain* that has the syntactical form of a *fragment*. Figure 2 shows a tweet labeled as *abusive* and contains a *directive* of the type *require* in the syntactical form of an imperative.

[5]Transl.: *'We need #yellowvests in the whole #EU.. |LBR| The people have to show that they won't take this crappy poverty policy any longer.away with #Macron away with #Merkel and Merkel 2.0 #AKK'*

[6]Transl.: *'If that hasn't been found yet, one can assume that the driver is a getaway driver.'*

[7]Transl.: *'Germany, the land of three sexes, of cowards and yes-men, the land of tolerance fanatics. The land of stupidity.'*

[8]This example is disputable.

[9]Transl.: *'An egomaniac uses every opportunity to be the center of attention'*

[10]https://corpling.uis.georgetown.edu/wiki/doku.php?id=gum:tokenization_segmentation

[11]https://github.com/MelinaPl/speech-act-analysis#sentence-types

[12]https://github.com/MelinaPl/speech-act-analysis#speech-acts

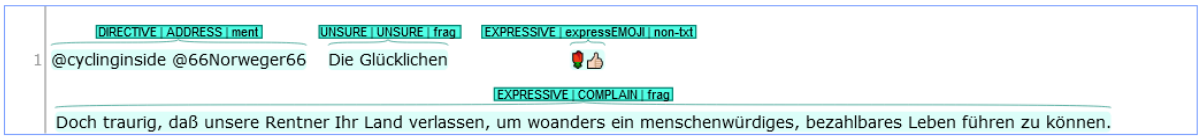

Figure 1: Example for a tweet labeled "*other*" (*s_1–2_Tweet_45_805_other.txt*). Transl.: *'The lucky ones. But it's sad that our pensioners leave their country to live a decent, affordable life elsewhere.'*

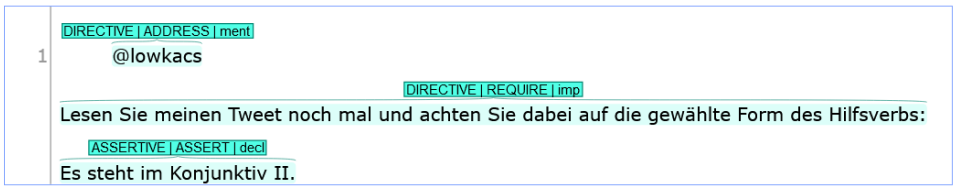

Figure 2: Example for a tweet labeled "*abuse*" (*s_1–2_Tweet_441_2404_abuse.txt*). Transl.: *'Read my tweet again and pay attention to the chosen form of the auxiliary verb: It is in the subjunctive II.'*

5 Results and Discussion

Section 5.1 presents our quantitative results, generated using Python.[13] A qualitative evaluation follows in Section 5.2. Section 5.3 discusses our main findings.

5.1 Quantitative Evaluation

Tables 1, 2 and 3 show the results of the statistical analysis of the dataset. Table 1 presents the absolute frequencies of sentence types (Section 4.2) for each offensive language category. Table 2, illustrates the frequencies of coarse-grained and fine-grained speech acts (Section 4.3), for each offensive language category including binary categorization (*offensive/ other*). Table 3 shows the frequencies of sentence types for each offensive language category, again including binary categorization. Additional details regarding the syntactical form of fine-grained speech act types are shown in Table 6 in the Appendix.

Table 2 demonstrates that there are 16.4% *directives* in offensive and 16.9% *directives* in non-offensive tweets (excluding the class *address*). Therefore, hypothesis (i) is refuted. Moreover, there are 14.6% of the class *complain* in offensive tweets while there are 5.2% in non-offensive tweets, confirming hypothesis (ii). Furthermore, *assert* occurs with a frequency of 28.9% in offensive tweets and with a frequency of 34.5% in non-offensive tweets, thus confirming hypothesis (iii).

With regard to the overall sentence type distribution, Table 3 shows that declarative sentences are the most frequently occurring sentence type by far, followed by mentions (19.4%), fragments (17.1%), which probably entail many declarative sentences as well that contain an ellipsis, and exclamatives (7.4%). The types that occur the least are alternative questions (0.3%), interjections (0.4%) and *multiple* (0.5%). Thus, the numbers show evidence in favor of hypothesis (iv). There is also a higher number of exclamatives in offensive tweets (8.3%) compared to non-offensive tweets (3.0%). The biggest difference can be viewed when comparing the number of exclamatives in tweets containing explicit language (16.5%) with tweets containing implicit offensive language (5.8%).

5.2 Qualitative Evaluation

The annotation process has raised a number of challenges regarding the decision which label to choose.

One issue encountered during the annotation phase was the distinction between *assert* and *rejoice* or *complain* speech acts. Insults, for example, could be viewed as expressing a negative attitude towards someone and therefore could be annotated as *complain* speech acts. Nevertheless, if someone were to say "He is a devious man" or "He is not the brightest man" and these statements were actually true, could they still be considered as insults? Or do they rather describe reality, therefore fitting more into the speech act class *assert*? There are various cases where both speech acts occur simultaneously, leading to the difficult task of deciding on the most dominant speech act.[14]

Taking a closer look at the data, a rather unexpected finding was the frequent use of *rejoice* speech acts in tweets classified as containing *explicit* offensive language. After examination of these tweets, this can be explained with two ways in which *rejoice* speech acts have been used in general. The first use can be seen in example e). Here, "Hahaha!" has been annotated as an instance of *rejoice*, expressed through an interjection. In this example, the author uses an apparently joyful expression to comment on the previous question for the purpose of ridiculing it.

- **Ex. e)** *@ObenausThomas @Jung_us_Koelle @anna_IIna Der syrische Sozialtourist und eine Haftplichtversicherung? Hahaha!*[15]

[13] Due to class imbalance and the limited size of the dataset, no classification experiments have been carried out yet.

[14] We assign one speech act class per segment only.

[15] Transl.: *'The Syrian social tourist and a liability insur-*

Table 1: Frequency of sentence types in coarse-grained speech acts

Sentence Types		Coarse-Grained Speech Acts						
		Assertive	**Commissive**	**Expressive**	**Directive**	**Unsure**	**Other**	**Total**
Alt-f	Questions asking the addressee to decide for one option	0	0	0	5	0	0	5
Decl	Declarative sentence (only indicative)	359	6	110	13	36	0	524
Excl	Exclamative sentence	56	2	59	11	15	0	143
F	A question which can be answered with "yes" or "no"	2	0	0	80	0	0	82
Frag	Containing an ellipsis/ no subject predicate structure/ finite verb	152	10	76	12	78	1	329
Hashtag	Initial #, only annotated if not part of a sentence	1	0	5	3	0	62	71
Imp	Finite verb needs to be in imperative mood	0	1	2	47	1	0	51
Intj	Short exclamations, annotated if they form a sentence on their own	1	0	5	0	1	0	7
Kon	Finite verb is in conjunctive mood	29	0	6	6	4	0	45
Ment	Mentioning a person/ another twitter account, only annotated if not part of a sentence	0	0	1	372	0	0	373
Mult	Combination of two or more types due to the conjunction of two main clauses	5	0	2	1	1	0	9
Non-txt	Non-textual units such as symbols and emojis	0	0	105	1	0	0	106
Other	Sentence types not fitting in other categories (e.g. using English phrases/ sentences, constructions with ":")	59	1	21	9	13	5	108
W-f	Questions formed with w-phrases	0	0	0	70	1	0	71
Total		664	20	392	630	150	68	1924

Alt-f: alternative question; Decl: declarative; Excl: exclamative; F: yes-/ no-question; Frag: fragment; Imp: imperative; Intj: interjection; Kon: conjunctive; Ment: mention; Mult: multiple; Non-txt: non-textual; W-f: w-question

The second use of *rejoice* speech acts is illustrated in example f). Here, *rejoice* speech acts are used to express a positive attitude over something that can be viewed as offensive, thereby praising it. The segment "Das einizg Gute an #Jamaika @HeikoMaas ist endlich weg!" has been annotated as an instance of a *rejoice* speech act expressed through an exclamative sentence. The author of this tweet is expressing his positive attitude towards the assertion that "@HeikoMaas ist endlich weg!", an utterance which can be regarded as being rather offensive.

- **Ex. f)** *@Beatrix_vStorch @HeikoMaas @RegSprecher @BMJV_Bund Das einizg Gute an #Jamaika @HeikoMaas ist endlich weg!*[16]

ance? Hahaha!'

5.3 Discussion

With respect to the results on the frequency of speech acts in the data, it is striking that some of the fine-grained speech acts do not occur at all or only very rarely. This concerns the class *accept* (no occurrences) and the classes *greet* (0.1%), *apologize* (0.1%), *refuse* (0.1%), *threat* (0.3%), *disagree* (0.3%) and *thank* (0.4%). In general, commissives are very rare in the data (1.0%). The rare occurrence of *greet* is probably

[16]Transl.: *'The only good thing about #Jamaika, @HeikoMaas is finally gone!'*

Table 2: 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

Table 3: 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

partly owed to the type of dataset (single tweets without any context) and partly owed to the way how tweets were segmented. As Compagno et al. (2018) show in their annotation scheme, it is in the nature of *accept* and *refuse* speech acts that they depend on previous utterances. With the lack of conversational context, it is sometimes impossible to either distinguish *accept* and *refuse* from *agree* and *disagree* or to be entirely certain that the labels *accept* or *refuse* are the correct ones. Hence, for an updated version of the annotation scheme that deals with the same type of dataset, rarely occurring categories should be excluded or subsumed by the class *other*; we could also reconsider the segmentation approach and foresee smaller segments that we assign speech acts to, as to reduce the possibility of multiple speech acts occurring simultaneously in one sentence. This approach, however, could lead to a demand for more speech act categories and, hence, annotations.

Another finding is that tweets with implicitly offensive language seem to be more made up of statements (consisting of 41.2% *assertives*) and containing less *expressives* overall (meaning *rejoice* and *complain*). This seems reasonable, as the author does not appear to express their feelings or attitudes in an obvious, explicit way. A further discovery that supports this explanation is that the category *expressEmoji* was used least frequently in tweets with *implicit* offensive language. It should be noted that the distinction between *expressives* and *assertives* (excluding the category *expressEmoji*) was the most prominent issue faced during annotation. It seems that tweets with *implicit* offensive language contained more speech acts that were phrased more like statements than *expressives* so that, when in doubt, the annotator chose the *assert* label instead of *complain* or *rejoice*.

Finally, the findings concerning the most common syntactic realisation of each speech act type (Table 6) correspond to the view in the literature regarding sentence types and their most frequently used speech acts. König and Siemund (2007) state that declaratives are used most frequently for speech acts such as asserting something, interrogatives are usually used for requests and imperatives are commonly used for orders. The observation that exclamatives mostly realize *complain* speech acts in this data seems logical as the definition in the Duden states that exclamatives are sentences that are uttered with emphasis (Dudenredaktion, 2016) and expressives such as *complain* probably transmit a lot of strong emotions that are best expressed by using an exclamative sentence.

6 Conclusions and Future Work

The results show that offensive language mainly differs from non-offensive language in the respect that offensive language contains more expressives and less assertives than non-offensive language. The biggest difference is most prominent when we compare tweets with implicit offensive language tweets with explicit offensive language. Tweets with implicit offensive language seem to lack the tendency to overtly express emotions, hence they have the lowest frequency of expressives (excluding non-offensive tweets) and the highest frequency of assertives. In contrast, tweets with explicit offensive language show the opposite: they have the lowest frequency of assertives and the highest frequency of expressives.

Our results suggest that differences exist with regard to the distribution of speech acts in offensive language and non-offensive language. It remains to be seen if an accurate speech act classifier can be developed as one additional component in larger hate speech detection systems. Also in terms of future work, we plan to annotate part of the dataset with two more annotators so that we can further improve our annotation approach and also to assess the quality of the dataset and annotations with regard to inter-annotator agreement.

Acknowledgements

The research presented in this article was partially funded by the German Federal Ministry of Education and Research (BMBF) through the projects QURATOR (Unternehmen Region, Wachstumskern, no. 03WKDA1A) and PANQURA (no. 03COV03E). We would like to thank the anonymous reviewers for their helpful comments.

Bibliographical References

Austin, J. L. (1962). *How to Do Things with Words*. Oxford University Press.

Burnap, P. and Williams, M. L. (2016). Us and Them: Identifying Cyber Hate on Twitter Across Multiple Protected Characteristics. *EPJ Data Science*, 5(11):1–15.

Compagno, D., Epure, E., Deneckere, R., and Salinesi, C. (2018). Exploring Digital Conversation Corpora with Process Mining. *Corpus Pragmatics*, 2:193–215.

Dhayef, Q. and Ali, A. (2020). A Pragmatic Study of Racial Hate Speech. *Journal of Tikrit University for Humanities*, 27(8):1–24.

Dudenredaktion. (2016). *Duden - Die Grammatik: Unentbehrlich für richtiges Deutsch*, volume 4. Bibliographisches Institut GmbH.

Fortuna, P. and Nunes, S. (2018). A Survey on Automatic Detection of Hate Speech in Text. *ACM Computing Surveys*, 51(4):1–30.

Jurafsky, D. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13. *Institute of Cognitive Science Technical Report*.

Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico, Juni. Association for Computational Linguistics.

König, E. and Siemund, P. (2007). Speech Act Distinctions in Grammar. In Timothy Shopen, editor, *Language Typology and Syntactic Description*, pages 276–324. Cambridge University Press, Cambridge.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Leech, G. and Weisser, M. (2013). The SPAADIA Annotation Scheme. *Retrieved from http://martinweisser.org/publications/SPAADIA_Annotation_Scheme.pdf (last accessed January 2022)*.

Leech, G., McEnery, T., and Weisser, M. (2003). SPAAC Speech-Act Annotation Scheme. *University of Lancaster, Technical Report*.

Leech, G. (1983). *Principles of Pragmatics*. Longman, London.

Mishra, P., Yannakoudakis, H., and Shutova, E. (2020). Tackling Online Abuse: A Survey of Automated Abuse Detection Methods. *arXiv e-prints*.

Palmer, A., Robinson, M., and Phillips, K. K. (2017). Illegal is not a Noun: Linguistic Form for Detection of Pejorative Nominalizations. In *Proceedings of the First Workshop on Abusive Language Online*, pages 91–100, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Palmer, A., Carr, C., Robinson, M., and Sanders, J. (2020). COLD: Annotation Scheme and Evaluation Data Set for Complex Offensive Language in English. *Journal for Language Technology and Computational Linguistics*, 34(1):1–28.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2020). Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review. *Language Resources and Evaluation*, pages 1–47.

Risch, J., Stoll, A., Wilms, L., and Wiegand, M. (2021). Overview of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of the 3rd Workshop on Natural Language Processing for Computer-Mediated Communication (NLP4CMC III)*, pages 6–9.

Schmidt, A. and Wiegand, M. (2017). A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Searle, J. R. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, Cambridge.

Spertus, E. (1997). Smokey: Automatic Recognition of Hostile Messages. In *Proceedings of the Innovative Applications of Artificial Intelligence (IAAI)*, pages 1058–1065.

Struß, J. M., Siegel, M., Ruppenhofer, J., Wiegand, M., and Klenner, M. (2019). Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *German Society for Computational Linguistics. Proceedings of the 15th Conference on Natural Language Processing (KONVENS) 2019*, pages 354–365, Nürnberg/Erlangen.

Vosoughi, S. and Roy, D. (2016). Tweet Acts: A Speech Act Classifier for Twitter. In *Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*, Cologne, Germany.

Weber, M., Viehmann, C., Ziegele, M., and Schemer, C. (2019). Online Hate Does Not Stay Online – How Implicit and Explicit Attitudes Mediate the Effect of Civil Negativity and Hate in User Comments on Prosocial Behavior. *Computers in Human Behavior*, 104.

Weisser, M. (2018). *How to Do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Wiegand, M., Siegel, M., and Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018 Workshop (GermEval)*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Zeldes, A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zhang, R., Gao, D., and Li, W. (2011). What Are Tweeters Doing: Recognizing Speech Acts in Twitter. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Language Resource References

Struß, Julia Maria and Siegel, Melanie and Ruppenhofer, Josef and Wiegand, Michael and Klenner, Manfred. (2019a). *Germeval Task 2, 2019 – Shared Task on the Identification of Offensive Language. Gold file of GermEval 2019 (Subtask I & II)*. https://projects.fzai.h-da.de/iggsa/wp-content/uploads/2019/08/germeval2019GoldLabelsSubtask1_2.txt.

Struß, Julia Maria and Siegel, Melanie and Ruppenhofer, Josef and Wiegand, Michael and Klenner, Manfred. (2019b). *Germeval Task 2, 2019 – Shared Task on the Identification of Offensive Language. Gold file of GermEval 2019 (Subtask III)*. https://projects.fzai.h-da.de/iggsa/wp-content/uploads/2019/08/germeval2019GoldLabelsSubtask3.txt.

Appendix

Table 4: Absolute frequencies of fine-grained speech acts in sentence types

	Sentence Type														
Speech Act	**Alt-f**	**Decl**	**Excl**	**F**	**Frag**	**Hashtag**	**Imp**	**Intj**	**Kon**	**Ment**	**Mult**	**Non-txt**	**Other**	**W-f**	**Total**
Accept	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Address	0	0	1	0	0	0	0	0	0	**371**	0	1	0	2	375
Agree	0	4	3	0	**5**	0	0	1	0	0	0	0	0	0	13
Apologize	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	1
Assert	0	**305**	43	0	140	1	0	0	24	0	4	0	58	0	575
Complain	0	**100**	55	0	65	5	2	3	4	0	2	0	13	0	249
Disagree	0	**6**	0	0	0	0	0	0	0	0	0	0	0	0	6
Engage	0	3	0	0	**10**	0	0	0	0	0	0	0	0	0	13
Greet	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0	1
Guess	0	**11**	1	2	6	0	0	0	5	0	1	0	0	0	26
Other	0	0	0	0	1	**62**	0	0	0	0	0	0	5	0	68
Predict	0	**24**	7	0	0	0	0	0	0	0	0	0	1	0	32
Refuse	0	0	0	0	0	0	**1**	0	0	0	0	0	0	0	1
Rejoice	0	**7**	2	0	5	0	0	2	0	0	0	0	1	0	17
Request	5	2	2	**79**	1	0	2	0	2	1	0	0	1	68	163
Require	0	7	7	1	6	3	**45**	0	1	0	0	0	6	0	76
Suggest	0	4	1	0	**5**	0	0	0	3	0	1	0	1	0	15
Sustain	0	**9**	2	0	1	0	0	0	0	0	0	0	0	0	12
Thank	0	1	1	0	2	0	0	0	0	0	0	0	**4**	0	8
Threat	0	**3**	2	0	0	0	0	0	0	0	0	0	1	0	6
Unsure	0	36	15	0	**78**	0	1	1	4	0	1	0	13	1	150
Wish	0	2	1	0	**4**	0	0	0	2	0	0	0	2	0	11
expressEmoji	0	0	0	0	0	0	0	0	0	1	0	**105**	0	0	106
Total	5	**524**	143	82	329	71	51	7	45	373	9	106	108	71	1924

Alt-f: alternative question; Decl: declarative; Excl: exclamative; F: yes-/ no-question; Frag: fragment; Imp: imperative; Intj: interjection; Kon: conjunctive; Ment: mention; Mult: multiple; Non-txt: non-textual; W-f: w-question