Automated Speech Act Classification in Offensive German Language Tweets

Melina Plakidis[1,2] — Elena Leitner[1] — Georg Rehm[1,2]

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

ABSTRACT. *One under-researched avenue for hate speech and offensive language detection is the integration of knowledge related to speech acts. In previous work, we investigated whether the distribution of speech acts differs across offensive and non-offensive language. Our findings revealed supporting evidence. In the present article, we fine-tune several BERT models and LLMs on the German Speech Acts Dataset. Our goals are two-fold: we want to contribute relevant research results to speech act theory by developing and providing models that detect and classify speech acts in documents or other types of discourse such as tweets. We hope that detected speech acts can be used in a beneficial way as additional features in the detection of hate speech. Our best-performing model achieves a macro-averaged F_1-score of 68.68%.*

RÉSUMÉ. *En matière de détection de discours de haine et de langage offensant, l'intégration des connaissances sur les actes de langage représente une voie de recherche encore peu explorée. Dans nos précédents travaux, nous avons analysé si la répartition des actes de langage variait selon que les propos étaient injurieux ou non. Les résultats que nous avons obtenus ont confirmé cette hypothèse. Dans le présent article, pour affiner plusieurs modèles BERT et LLM, nous avons utilisé le jeu de données des actes de langage allemands. Nous poursuivons un double objectif. Nous souhaitons fournir des résultats pertinents à la théorie des actes de langage en développant et en mettant à disposition des modèles capables de mettre en œuvre la détection et la classification d'actes de langage dans des documents ou d'autres types de propos, des tweets notamment. Nous espérons que les actes de langage détectés pourront servir de caractéristiques supplémentaires et bénéficier à la détection des discours de haine. Notre modèle le plus performant atteint un score F_1 macro-moyenné de 68,68 %.*

KEYWORDS: *hate speech detection, offensive language, speech acts.*

MOTS-CLÉS : *détection de discours haineux, langage offensant, actes de langage.*

1. Introduction

Hate speech and the use of offensive language have become a pervasive phenomenon online. Wiegand *et al.* (2018) define offensive language as "hurtful, derogatory or obscene comments made by one person to another person". A study conducted by Bilewicz and Soral (2020) shows that increased exposure to hate speech can lead to desensitisation and thus decrease people's ability to identify hate speech. Furthermore, encountering derogatory language targeting immigrants and minority groups can contribute to political radicalisation (Bilewicz and Soral, 2020). The vast amount of newly created daily posts, messages and other types of content makes the manual handling of offensive language impossible. Automatic processes are needed, but even with the recent emphasis on hate speech detection (Poletto *et al.*, 2021), there are still various challenges when it comes to detecting hate speech automatically. [1]

In previous work (Plakidis and Rehm, 2022), we took a closer look at pragmatic properties of offensive language, i. e., by combining the field of *speech act theory* with *hate speech detection*, aiming to enrich text data with pragmatic characteristics and exploring possible differences between offensive and non-offensive language. We created a dataset of offensive and non-offensive German tweets and annotated them for coarse- and fine-grained speech acts. Our findings suggest a difference in the distribution of speech acts between offensive and non-offensive tweets as well as between different offensiveness categories. A similar observation made by previous studies also shows that speech acts vary depending on the discussed topic (Zhang *et al.*, 2011; Vosoughi and Roy, 2016; Laurenti *et al.*, 2022).

Building on our previous work, in this article we experiment with state-of-the-art encoder and decoder models to train speech act classifiers on our German Speech Act Dataset and investigate which models are better suited considering different levels of speech acts. For encoders, we implement various fine-tuning strategies such as default, hyperparameter search and few-shot classification to improve performance of selected models. For decoders, we focus on the parameter-efficient fine-tuning method instead of exploring different prompting approaches such as zero-shot or few-shot. This method allows to optimise performance of a model by retraining with specific data (Wang *et al.*, 2024). In addition, we provide an error analysis to identify the core issues of our best-performing classifiers.

The remainder of this article is structured as follows. Section 2 presents related work and Section 3 describes the dataset. Section 4 introduces our experiments on training a speech act classifier, providing information on the approach as well as on the evaluation. Section 5 reports on our results. Finally, Section 6 concludes the article.

1. In the wider field of research, a variety of similar terms are used such as "abusive" (Nobata *et al.*, 2016), "toxic" (Risch *et al.*, 2021) or "offensive" (Wiegand *et al.*, 2018; Zampieri *et al.*, 2019) language. We use *hate speech* and *offensive language* synonymously.

2. Related Work

The research dedicated to speech acts used in hate speech is still limited. Nevertheless, there are some works dealing with the combination of speech acts and hate speech which we will present in the following.

Oktaviani and Nur (2022) analyse a twitter account using Searle's speech act theory. They use an exploratory, qualitative approach and assign a hate speech label as well as a speech act label for each tweet. They observe the occurrence of *assertives*, *directives* and *expressives*, stating that *directives* appear most often in the data. Nevertheless, it is not clear which speech act classes occur most often in which hate speech category and how both categories relate to each other. Similarly to the study by Oktaviani and Nur (2022), Mubarok *et al.* (2024) also analyse comments of a selected social media account. They find 11 *directives*, 15 *expressives* and five *assertive* speech acts in a small sample of 31 abusive comments. In a study by Dhayef and Ali (2020), seven newspaper article extracts are selected from a Rwandan newspaper which are expected to contain racial hate speech. They examine them using Searle's five speech act classes (Searle, 1979) and additionally include Searle's distinction between direct and indirect speech acts. They present a qualitative as well as quantitative analysis and put forward three hypotheses for which their results seem to provide confirming evidence. First, they expect the excerpts to contain a high quantity of *directives*. Second, they assume that the excerpts will contain more indirect than direct speech acts and third, they estimate that direct *assertives* and indirect *expressives* are the most dominant speech acts in the excerpts. However, Dhayef and Ali (2020) provide only seven short extracts for their pragmatic analysis and they do not state what constitutes an utterance or how they intend to segment the extracts. A more recent study by Ollagnier (2024) introduces the dataset CyberAgressionAdo-V2 on cyberbullying in French multiparty chats, which, inter alia, is annotated with pragmatic aspects. These pragmatic aspects are located on the discursive level which comprises eight distinct categories such as *gaslighting*, *defend*, and *attack*. Similar to speech acts, these categories denote the intention that the user attempts to convey with his message. In addition, Ollagnier (2024) also considers the context in which these messages occur. The annotations are not restricted to aggressive messages, but comprise all messages regardless of their level of aggressiveness. The findings show that most of the time, bullies and their supporters intentionally send messages that attack the victims, while victims and their supporters predominantly issue either neutral or defensive messages.

Several attempts have been made to classify speech acts automatically and their annotation taxonomies have often been influenced to a great extent by Austin (1962) and Searle (1979). Compagno *et al.* (2018) represent one of these works. Their hierarchically structured speech act taxonomy is based on Searle's five classes which they applied to a Reddit corpus dealing with autoimmune diseases. In total, their fine-grained classification consists of 17 speech acts. Other approaches influenced by Searle (1979) and Austin (1962) include Vosoughi and Roy (2016) and Zhang *et al.* (2011). Zhang *et al.* (2011), for instance, aim to classify tweets into one of five speech act classes: *statement*, *question*, *comment*, *suggestion* and *miscellaneous*. They achieve an F_1-

score of almost 70.00% on average using a Support Vector Machine classifier with a linear kernel in addition to word- and character-based features. Similarly, Vosoughi and Roy (2016) classify tweets into one of six categories: *assertion*, *recommendation*, *expression*, *question*, *request*, and, again, *miscellaneous*. With the use of semantic and syntactic features in combination with a Logistic Regression classifier, they manage to achieve an average F_1-score of 70.00%. Another approach with the aim of annotating speech acts automatically, currently however semi-automatically, is presented by Weisser (2018). He introduces the Dialogue Annotation and Research Tool (DART), which is publicly available. Its current version 3.0 [2] classifies dialogue using various features including syntactic categories and speech acts. The proposed speech act tags in the latest version of the DART taxonomy [3] result in a total of 162 speech act tags. In a further study by Laurenti *et al.* (2022), French tweets posted during crisis events were annotated for speech acts on both tweet level and a more fine-grained segment level. Their speech act annotations on the level of tweets comprise five classes, namely *assertives*, *jussives*, *subjectives*, *interrogatives* and *other*. Additionally, their segment level annotations consist of eight speech act classes. Their findings indicate a correlation between urgent messages during crisis events and higher occurrence of *proper assertions* (assertions not relying on a third-party source). Additionally, they observe a higher occurrence of *subjective* speech acts in non-urgent tweets. Their best-performing model for tweet-level annotations with four classes is CamemBERT (Martin *et al.*, 2020) with focal loss (Lin *et al.*, 2020) and extra-features and achieves an F_1-score of 73.55%. Building on the work by Laurenti *et al.* (2022), Benamara *et al.* (2024) further extend the dataset by Laurenti *et al.* (2022) to about 13,000 French tweets. Their experiments show that FlauBERT (Le *et al.*, 2019) pre-trained on crisis domain tweets (Kozlowski *et al.*, 2020) with focal loss and additional features is the best performing model (F_1: 67.37%) for predicting the five speech act classes on tweet level. Their best-performing classifier for the eight fine-grained speech act classes is FlauBERT base with cross-entropy loss in a multi-label setting achieves an F_1 of 87.80%.

3. Dataset

Our German Speech Act Dataset (Plakidis and Rehm, 2022) comprises 600 tweets of the dataset created for task two of the 2019 GermEval Shared Task on the Identification of Offensive Language (Struß *et al.*, 2019). We chose Twitter (now: X) as the main source of data because it is the most frequently used platform in the field of hate speech detection (Poletto *et al.*, 2021). The 600 tweets were selected with the aim to analyse whether the speech act distribution differs across different offensive language classes. For each of the six offensive language classes established by Struß *et al.* (2019), i. e., *implicit*, *explicit*, *profanity*, *insult*, *abuse* and *other*, we randomly selected 100 tweets.

2. http://martinweisser.org/publications/DART_manual_v3.0.pdf.
3. http://martinweisser.org/DART_scheme.html.

According to Struß *et al.* (2019), these classes can be described as follows. In contrast to being *explicitly* offensive, 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). Tweets are labeled as *profanity* if they consist of profane words like swearwords but lack abusive language as well as insults. If they also contain abusive language or insults, they either belong to the class *insult* or *abuse*. While the class *insult* only contains offensive language targeting individuals, the class *abuse* contains tweets that target group representatives, assigning them universally negative traits.

We extended the dataset by adding speech act annotations which we included for both fine-grained as well as coarse-grained speech acts. In contrast to Struß *et al.* (2019), these annotations relate to the sentence level and not to the tweet level. Thus, the unit for a speech act is the sentence. [4] However, Twitter users often do not use punctuation properly in their tweets. During the annotation process, in order to clarify how to segment tweets into sentences, rules had to be established which are specified in our previous work (Plakidis and Rehm, 2022).

Table 1 shows the results of our speech act annotations, revealing distinct differences between offensive and non-offensive language in terms of speech act usage. Offensive language generally features more *expressives* and fewer *assertives* than non-offensive language. As tweets consisting of implicit offensive language tend to lack emotional expression, thus increasing the use of *assertives* and decreasing the use of *expressives*, this difference is most pronounced when comparing explicitly offensive with implicitly offensive tweets. The results indicate that offensive and non-offensive language differ in how speech acts are distributed.

In the following, we present our speech act annotation scheme which is based on Compagno *et al.* (2018) and Searle (1979). [5] In addition, we also provide information on the inter-annotator agreement in Section 3.3. The German Speech Act Dataset is publicly available under a CC-BY-4.0 license and can be accessed at GitHub [6].

3.1. *Coarse-Grained Speech Act Level*

The coarse-grained speech act level includes six classes: *assertive*, *directive*, *expressive*, *commissive*, *unsure* and *other*. Quoting Searle (1979), the first four speech acts can be defined as follows: "We tell people how things are (Assertives), we try

4. Exceptional cases include user mentions, hashtags and emojis. The difference between tweet and sentence level is best illustrated in examples from the dataset in Sections 3.1 and 3.2.

5. Building upon Weisser (2018), our annotation scheme contains a syntactical level and a speech act level. In this article, we focus on the speech act level exclusively.

6. The most recent Version 1.1 of the dataset contains several bugfixes: `https://github.com/MelinaPl/speech-act-analysis/blob/main/version_1-1_changes.md`.

7. The class *accept* has been constructed and does not represent a real instance because the category did not occur in the data at all.

Table 1. *Frequency of coarse-grained and fine-grained speech acts in offensive language categories. Note that speech acts were annotated on sentence level, while offensive language categories were annotated on tweet level.*

	Offensive		**Other**		**Implicit**		**Explicit**		**Abuse**		**Profanity**		**Insult**		**Total**	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Assertive	557	34.3	126	37.7	116	41.6	85	28.9	118	32.2	111	33.8	127	35.5	683	34.9
Assert	473	29.1	117	35.0	97	34.8	73	24.8	99	27.0	93	28.4	111	31.0	590	30.1
Sustain	11	0.7	2	0.6	2	0.7	0	0.0	5	1.4	1	0.3	3	0.8	13	0.7
Guess	26	1.6	1	0.3	9	3.2	2	0.7	3	0.8	7	2.1	5	1.4	27	1.4
Predict	32	2.0	2	0.6	6	2.2	8	2.7	6	1.6	5	1.5	7	2.0	34	1.7
Agree	11	0.7	2	0.6	2	0.7	1	0.3	4	1.1	4	1.2	0	0.0	13	0.7
Disagree	4	0.2	2	0.6	0	0.0	1	0.3	1	0.3	1	0.3	1	0.3	6	0.3
Expressive	353	21.7	47	14.1	44	15.8	76	25.9	78	21.3	73	22.3	82	22.9	400	20.4
Rejoice	14	0.9	3	0.9	1	0.4	6	2.0	1	0.3	4	1.2	2	0.6	17	0.9
Complain	240	14.8	17	5.1	37	13.3	55	18.7	39	10.7	46	14.0	63	17.6	257	13.1
Wish	10	0.6	1	0.4	0	0.0	3	1.0	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	0.0
Thank	4	0.2	4	1.2	0	0.0	0	0.0	1	0.3	2	0.6	1	0.3	8	0.4
expressEmoji	85	5.2	21	6.3	6	2.2	12	4.1	34	9.3	17	5.2	16	4.5	106	5.4
Commissive	17	1.0	3	0.9	0	0.0	3	1.0	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[7]	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	0.0	0	0.0	0	0.0	1	0.3	0	0.0	0	0.0	0	0.0	1	0.0
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	524	32.2	109	32.6	99	35.5	100	34.0	131	35.8	85	25.9	109	30.4	633	32.3
Request	130	8.0	33	9.9	23	8.2	23	7.8	36	9.8	24	7.3	24	6.7	163	8.3
Require	66	4.1	12	3.6	7	2.5	17	5.8	13	3.6	13	4.0	16	4.5	78	4.0
Suggest	15	0.9	1	0.3	2	0.7	1	0.3	5	1.4	3	0.9	4	1.1	16	0.8
Greet	1	0.0	0	0.0	0	0.0	0	0.0	0	0.0	0	0.0	1	0.3	1	0.0
Address	312	19.2	63	18.9	67	24.0	59	20.1	77	21.0	45	13.7	64	17.9	375	19.1
Unsure	113	7.0	37	11.1	18	6.5	15	5.1	30	8.2	35	10.7	15	4.2	150	7.7
Other	61	3.8	12	3.6	2	0.7	15	5.1	8	2.2	12	3.7	24	6.7	73	3.7
Total	1,625	100.0	334	100.0	279	100.0	294	100.0	366	100.0	328	100.0	358	100.0	1,959	100.0

to get them to do things (Directives), we commit ourselves to doing things (Commissives), [and] we express our feelings and attitudes (Expressives)" (p. viii). The categories *assertive*, *directive* and *expressive* are shown in Examples (1, 2) and *commissive* in (3). The category *unsure* is used in cases where an utterance in a tweet cannot be classified due to missing or insufficient context as in Example (4). Finally, the category *other* in (2) is used for all speech acts not represented in this annotation scheme. The examples below reflect coarse- and fine-grained labels which are separated using "|".

(1) *[@Alexplantsatree @griechenwoos2 @Die_Gruenen]*$_{directive|address}$ *[Schon die*
@Alexplantsatree @griechenwoos2 @Die_Gruenen already the
Worter "schmutzige Technologien" implizieren, dass der
words dirty technologies imply that the
Automobilbau eine Technologie ist, die entsorgt werden
automobile.manufacturing a technology is that disposed.of be
musse.]$_{assertive|assert}$ *[Nur leider ist die Elektromobilitat keine adaquate*
must now unfortunately is the electric.mobility no adequate

Alternative zum Auto mit Verbrennungsmotor und Kernenergie wurde aus
alternative to cars with combustion.engines and nuclear.energy was of
reinem Opportunismus aufgegeben.]$_{expressive|complain}$
pure opportunism up.give
'@Alexplantsatree @griechenwoos2 @Die_Gruenen Already the words "dirty technologies" imply that automobile manufacturing is a technology that must be disposed of. Unfortunately, electric mobility is not an adequate alternative to cars with combustion engines and nuclear energy has been abandoned out of pure opportunism.'

(2) *[2/2]*$_{other}$ *[sollten wir nicht in Berlin und Brüssel stehen und die Banditen*
2/2 should we not in Berlin and Brussels stand and the bandits
aus ihren Ämtern jagen?]$_{directive|request}$ *[Mit Schimp und Schande, geteert*
from their offices chase with disgrace and shame tarred
und gefedert?]$_{directive|request}$ *[Suizid?]*$_{directive|request}$ *[Unfassbar.]*$_{expressive|complain}$
and feathered suicide unbelievable.
2/2 Shouldn't we stand in Berlin and Brussels and chase the bandits from their offices? With disgrace and shame, tarred and feathered? Suicide? Unbelievable.'

(3) *[@_denk_mal_]*$_{directive|address}$ *[ES WIRD ZEIT, DIESE KRANKE ZU*
@_denk_mal_ it will time this sick.person to
WARNEN!]$_{commissive|threat}$
warn
'@_denk_mal_ It will be time to warn this sick person!'

(4) *[@Snakecleaver @Metalwilli]*$_{directive|address}$ *[OK.....!]*$_{unsure}$
@Snakecleaver @Metalwilli okay
'@Snakecleaver @Metalwilli OK....!'

3.2. *Fine-Grained Speech Act Level*

The fine-grained speech act level consists of 23 speech acts. We modified the taxonomy by Compagno *et al.* (2018) by adding the categories *predict*, *expressEmoji*, *threat*, *address* and *unsure* as well as by moving *greet* to *directives* and maintaining the distinction between the two classes *request* and *require*. Each fine-grained speech act has a corresponding coarse-grained speech act. However, the categories *unsure* and *other* remain the same on both levels. Several examples are shown in (5-9) and in the previous subsection in (1-4). [8]

(5)

8. Additional examples can be found in our repository: `https://github.com/MelinaPl/speech-act-analysis/blob/main/README.md`.

[Er geht mir ziemlich auf den Keks, aber wegen Vorstehendem habe
he goes me quite on the biscuit but because.of before.standing have
ich ihn noch nicht einfach geblockt!]$_{sustain}$
I him yet not simply blocked
'*He really gets on my nerves but because of the preceding I haven't blocked him yet.*'

(6) *[ich kotzt das so an,]*$_{complain}$ *[fragt die deutschen Staatsbürger,]*$_{require}$
me throw.up that so of ask the German citizens
[schätze 80% sind gegen den Migrationspackt]$_{guess}$ *[#Maischberger]*$_{other}$
estimate 80% are against the migration.pact #Maischberger
'I'm so sick of it, ask the German citizens, I estimate that 80% are against the migration pact.'

(7) *... [ich werde ihnen auch in den Hintern Kriechen so bald ich bei der*
... I will you also in the butt creep as soon.as I by the
Merkel raus bin.]$_{predict}$ *[Ich biete Ihnen gute Zusammenarbeit an..]*$_{engage}$ *...*
Merkel out am I offer you good cooperation on..
'I will kiss their asses as soon as I leave Merkel. I offer you good cooperation.'

Assertive speech acts comprise statements that *assert* something, statements sustained with arguments (*sustain* in (5)) as well as weaker forms of assertions (*guess* in (6) or *predict* in (7)). In addition, assertive speech acts can also be used to signal agreement (*agree*) or disagreement (*disagree*) with something or someone.

Expressive speech acts include statements about positive (*rejoice*) or negative (*complain* in (1, 2, 4, 6)) attitude towards someone or something, serve by wishing for something (*wish* in (8)), apologising to someone for something (*apologize*) or thanking someone (*thank*). Additionally, *expressEmoji* is used for an emoji or series of emojis.

Directive speech acts either *require* or *request* someone to do something, provide a suggestion about something (*suggest* in (9)) or are used to greet (*greet*) or address someone (*address* in (1, 2, 4)).

Commissive speech acts are utterances that either *engage* oneself to do something (7), *accept* or *refuse* something based on a previous utterance or are used to threaten someone (*threat* as in (3)).

(8) *[Schönen Freitag.]*$_{wish}$
beautiful Friday
'Have a nice Friday.'

(9) *[Die linke, deutsch/islamische #Bundesregierung kann den #korantreuen*
the left German/Islamic #federal.government can the #Koran.faithful
#Moslems #IS #Hamas doch gleich den Schlüssel zu Deutschland
#Muslims #IS #Hamas still immediately the key to Germany
überreichen.]$_{suggest}$
over.give
'The leftist, German/Islamic #federalgoverment may as well hand the #Koranfaithful #Muslims #IS #Hamas the key to Germany.'

3.3. *Inter-Annotator Agreement*

As our original dataset had only been annotated by one annotator, we decided to extend it by including two more annotators. Two of the annotators are authors of this paper, while the third annotator is a Master student with a background in linguistics. Currently, 200 of the 600 tweets have been annotated by two annotators (100 tweets by each of the two additional annotators). We pre-segment the tweets so that the annotators only have to choose the correct speech act labels. To compute the agreement between two annotators, we choose Cohen's κ (Cohen, 1968), resulting in an average κ score of 0.69 for coarse-grained speech acts and a κ of 0.66 for fine-grained speech acts. The values for both granularities indicate a substantial agreement. Similar values were achieved by Laurenti *et al.* (2022), who report a Cohen's κ of 0.62. Furthermore, Compagno *et al.* (2018) report a moderate to substantial agreement for all annotators with values between 0.57 and 0.87 for five coarse-grained speech act classes and between 0.48 and 0.73 for 18 fine-grained speech act classes. However, it should be noted that the authors computed the inter-annotator agreement using Fleiss' κ (Fleiss, 1971).

One of the greatest difficulties during the annotation process was distinguishing between *assertives* and *expressives* as it is often challenging to specify whether an utterance merely describes reality (= *assertive*) or expresses the speaker's feelings or attitude towards something (= *expressive*). Sometimes, both can be true at the same time, resulting in a rather subjective choice by the annotator. Similar observations were also made by Laurenti *et al.* (2022), who report issues distinguishing between *assertives* and *subjectives*, the latter class being comparable to *expressives*, and by Compagno *et al.* (2018), who report that their annotation results point to a continuity between *assertives* and *expressives*.

4. Experiments

The following section presents the experiments.[9] First, we describe which annotations are used for the training process and how we modify the classes with sparse

9. In addition to training several speech act classifiers, we also conducted an initial experiment to fine-tune an offensive language classifier with and without speech acts. These results have

data. Second, we provide information on the evaluation method and metric. Third, we present the selected models and fine-tuning strategies.

4.1. *Training Data*

For the experiments, we use both sets of annotations in the German Speech Act Dataset. The first version consists of data annotated for six coarse-grained speech act classes. The second version consists of fine-grained classes that have been modified. Due to rather sparse occurrences of a few fine-grained classes, only classes with ten or more instances are included. *Disagree*, *apologize*, *thank* and *greet* were combined in the new *excluded* class. As for the coarse-grained speech act *commissive*, we decided not to divide it into fine-grained classes due to sparse occurrences of its fine-grained classes. Thus, the number of fine-grained speech acts was reduced from 23 to 17.

4.2. *Evaluation Method and Metric*

Due to the dataset size and label distribution, we apply 5-fold cross-validation, a best-practice evaluation method. Sentences were shuffled and stratified in order to preserve the percentage of samples for each class in each fold and split. A train split contains 1,567 sentences (80%) and a validation split 392 sentences (20%). For the coarse- and fine-grained labels, we created individual splits. The mean number of instances across the 5-fold splits can be found in Table 2. As evaluation metrics, we use precision, recall and macro F_1, which calculate the unweighted mean between all labels.

4.3. *Models*

For training, we use state-of-the-art encoder and decoder models developed for German. As encoders, we selected three pre-trained BERT (Devlin *et al.*, 2018) models as many previous text classification experiments make use of Transformer-based architectures (Risch *et al.*, 2021) which have been shown to be more effective than other approaches (Struß *et al.*, 2019). We use the base versions of the cased [10] and uncased [11] pre-trained Digitale Bibliothek Münchener Digitalisierungszentrum (DBMDZ) BERT models. The two models were trained on Wikipedia, the EU Bookshop

not shown improvements when including speech act features: F_1 of 67.99% without speech acts and F_1 of 65.31% with speech acts. However, as offensive language classification is conducted on the tweet level and not on the sentence level, the experiments were carried out on a much smaller dataset. We plan to replicate this experiment on a significantly increased dataset in the future.

10. `https://huggingface.co/dbmdz/bert-base-german-cased`.

11. `https://huggingface.co/dbmdz/bert-base-german-uncased`.

	Train	Val	#
Assertive	546	137	683
Expressive	320	80	400
Commissive	16	4	20
Directive	506	127	633
Unsure	120	30	150
Other	59	14	73
Total	1,567	392	1,959

(a) *Coarse-grained annotations.*

	Train	Val	#
Assert	472	118	590
Sustain	10	3	13
Guess	22	5	27
Predict	27	7	34
Agree	10	3	13
Rejoice	14	3	17
Complain	206	51	257
Wish	9	2	11
Expressemoji	85	21	106
Commissive	16	4	20
Request	130	33	163
Require	62	16	78
Suggest	13	3	16
Address	300	75	375
Unsure	120	30	150
Other	58	15	73
Excluded	13	3	16
Total	1,567	392	1,959

(b) *Fine-grained annotations.*

Table 2. *Mean number of speech acts in 5-fold splits.*

corpus, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl. Additionally, we use Deepset's base version of the German BERT model called GBERT [12] (Chan *et al.*, 2020).

As decoders, we utilise Gemini 1.5 Flash [13] from Google AI, multilingual Llama 3.2 [14] from Meta (3B) as well as German Llama3 [15] (8B). Gemini is a transformer decoder model with 2M+ context and multimodal capabilities trained on a variety of multimodal and multilingual data (Gemini Team *et al.*, 2024). This model achieved good results as a text classifier (Wang *et al.*, 2024). Llama 3.2 (Llama Team *et al.*, 2024) is an auto-regressive language model trained on up to 9 trillion tokens from publicly available online data. German Llama3 is developed by the open-source research collective Disco Research that concentrates on the German language. This model is based on Meta's Llama3-8B and was pretrained on 65 billion tokens.

12. `https://huggingface.co/deepset/gbert-base`.
13. `https://ai.google.dev/gemini-api/docs/models/gemini`.
14. `https://huggingface.co/meta-llama/Llama-3.2-3B`.
15. `https://huggingface.co/DiscoResearch/Llama3-German-8B`.

4.4. *Fine-Tuning Strategy*

For each encoder model and granularity level (coarse-grained and fine-grained), we apply three methods: (i) default, (ii) bestrun with hyperparameter search, (iii) few-shot classification. The default models were fine-tuned with default hyperparameters which were the same for each model and granularity. We also performed a hyperparameter search on the first train and validation split using Ray Tune [16] (Liaw *et al.*, 2018). The goal was to maximise the macro F_1-score during 30 trials. After finding the best hyperparameters, we trained and evaluated a bestrun model on 5-folds. For few-shot classification, we used Fastfit. [17] This method utilises an approach that integrates batch contrastive learning and a token-level similarity score which provides accurate classification of semantically similar classes (Yehudai and Bandel, 2024).

For the decoder models, instead of prompting strategies such as few-shot prompting, we apply a supervised fine-tuning strategy to improve the model's performance as a speech act classifier. Each decoder model was fine-tuned in the same manner as encoder models, each on a full train set from 5-folds and evaluated on a corresponding validation set. For Gemini, we leverage the fine-tuning procedure with suggested hyperparameters described in its model tuning card. For both pretrained base Llama models, we performed parameter efficient fine-tuning (PEFT) (Mangrulkar *et al.*, 2022) using quantized low-rank adaptation (QLoRA) (Dettmers *et al.*, 2023).

Detailed information on the hyperparameters and results for each model, granularity and fine-tuning strategy (including results per class) as well as training scripts can be found in our GitHub Repository. [18]

5. Results

The following section presents the results of the experiments as well as a brief error analysis.

5.1. *Performance*

Table 3 presents the mean results during 5-fold cross-validation for the encoder models GBERT, $\text{BERT}^{\text{cased}}_{\text{german}}$ and $\text{BERT}^{\text{uncased}}_{\text{german}}$ as well as for the decoder models Gemini 1.5 Flash, Llama 3.2 and German Llama 3 across the two granularities and the different fine-tuning strategies. Regarding the encoder models, we can see that the results are improving based on the fine-tuning strategy, i. e., hyperparameter search is better than the default, and few-shot classification is better than hyperparameter search. The only exceptions are GBERT and $\text{BERT}^{\text{cased}}_{\text{german}}$ trained on coarse-grained labels; here,

16. `https://docs.ray.io/en/latest/tune/index.html`.
17. `https://github.com/IBM/fastfit`.
18. `https://github.com/elenanereiss/German-Speech-Act-Classification`.

	coarse-grained labels			*fine-grained labels*		
encoder	**GBERT**	**BERT**$^{\text{cased}}_{\text{german}}$	**BERT**$^{\text{uncased}}_{\text{german}}$	**GBERT**	**BERT**$^{\text{cased}}_{\text{german}}$	**BERT**$^{\text{uncased}}_{\text{german}}$
	Default					
precision	68.81	67.04	70.44	55.80	57.82	55.91
recall	65.62	64.33	66.21	48.39	50.70	49.88
F_1*-score*	66.51	65.05	<u>67.76</u>	50.14	<u>52.55</u>	51.84
	Hyperparameter search – Bestrun					
precision	69.44	70.19	65.80	63.18	58.74	57.20
recall	68.76	67.11	64.27	54.15	51.68	51.39
F_1*-score*	**68.68**	67.96	64.47	<u>56.37</u>	53.48	52.72
	Few-shot classification – Fastfit					
precision	73.97	70.07	72.39	67.68	63.58	62.46
recall	66.25	65.03	66.06	53.11	52.48	53.13
F_1*-score*	<u>68.45</u>	66.39	68.15	**57.04**	55.29	55.80
decoder	**Gemini**$^{\text{1.5 Flash}}$	**Llama3.2**$^{\text{3B}}$	**Llama3**$^{\text{8B}}_{\text{german}}$	**Gemini**$^{\text{1.5 Flash}}$	**Llama3.2**$^{\text{3B}}$	**Llama3**$^{\text{8B}}_{\text{german}}$
precision	33.42	64.97	62.69	45.06	39.68	40.62
recall	31.07	64.05	62.59	32.93	41.80	42.97
F_1*-score*	28.96	<u>62.56</u>	61.41	34.26	39.51	<u>39.88</u>

Table 3. *Mean precision, recall and macro F_1-score during 5-fold cross-validation. The best F_1-score in each setting is underlined, the best overall F_1-score is typeset in bold.*

hyperparameter search provides the best results. Overall, the best performing model on coarse-grained labels is bestrun GBERT with 68.68 macro F_1-score after hyperparameter search. The results for few-shot classification are almost similar and differ by 0.23 points. Regarding the fine-grained labels, few-shot classification with Fastfit has a clear advantage regarding macro F_1-score. Compared to the default, Fastfit achieves 3-7 points more on macro F_1-score; compared to bestrun with hyperparameter search, it achieves 0.7-3 points more. The best results are achieved by GBERT with 57.04 macro F_1-score.

Concerning the results of the decoder-based models, we observe that Gemini 1.5 Flash achieves exceptionally low scores in both settings (macro F_1-score of 28.96 and 34.26, respectively). While Llama 3.2 achieves the best F_1-score with 62.56 in the coarse-grained setting, and Llama 3 achieves the best F_1-score with 39.88 in the fine-

grained setting. All encoder-based models still outperform the two Llama models. We observe a particularly large difference concerning the results on the fine-grained speech acts: even the default models with default hyperparameters achieve macro F_1-scores that are at least 10 points better.

	precision	*recall*	*F_1-score*	*support*
Assertive	73.65	82.77	77.81	137
Expressive	71.25	63.25	66.79	80
Commissive	57.33	60.00	58.10	4
Directive	93.83	89.29	91.45	127
Unsure	31.46	28.67	29.31	30
Other	89.10	88.57	88.60	14
macro F_1-score	69.44	68.76	68.68	392

Table 4. *Mean results per class during 5-fold cross-validation for the best performing model GBERT on the coarse-grained labels after hyperparameter search.*

Table 4 presents the mean results of the best-performing model GBERT after hyperparameter search for each coarse-grained speech act class. The class *directive* achieves the best macro F_1-score (91.45) while the class *unsure* achieves the lowest macro F_1-score (29.31). This indicates that the class either was not well defined during the annotation or that it is difficult to predict whether the surrounding context is sufficient enough for a valid interpretation. An interesting finding is that the *commissive* class does not achieve the lowest macro F_1-score, although it is the class with the lowest number of instances during training and validation.

Finally, Table 5 shows the mean results of GBERT (in the few-shot classification setting with Fastfit) for each fine-grained speech act class. The best-performing classes are *address* and *expressEmoji* with a macro F_1-score of 99.60 and 98.56, respectively. This should come as no surprise as these are the most well-defined classes that are almost exclusively used whenever emojis or mentions are involved. The class *request* has the third best macro F_1-score (87.84), closely followed by *other* with an F_1-score of 86.13. As *request* is mostly used for questions, the presence of a question mark is most probably the best indicator of the class, leading to the good F_1-score. Similarly, the class *other* is most often used for uses of hashtags. Thus, the presence of a hashtag is a very likely sign to classify the utterance as an instance of *other*. The two worst-performing classes are *rejoice* and *other* with a macro F_1-score of 20.00 and 21.46, respectively.

5.2. *Error Analysis*

We conducted an error analysis with the aim of improving our understanding of the models' performance. For our analysis, we choose the best-performing models for each data version. As we evaluated each model in a 5-fold cross-validation manner,

	precision	*recall*	*F_1-score*	*support*
Assert	67.97	76.95	72.16	118
Sustain	50.00	20.00	28.00	3
Guess	55.71	36.00	40.00	5
Predict	60.00	42.86	48.21	7
Agree	63.33	33.33	41.33	3
Rejoice	40.00	13.33	20.00	3
Complain	50.37	56.47	53.17	51
Wish	60.00	40.00	46.67	2
expressEmoji	99.09	98.10	98.56	21
Commissive	83.33	65.00	71.90	4
Request	93.15	83.64	87.84	33
Require	62.53	50.00	54.61	16
Suggest	76.67	33.33	44.67	3
Address	100.00	99.20	99.60	75
Unsure	23.71	20.00	21.46	30
Other	84.63	88.00	86.13	15
Excluded	80.00	46.67	55.43	3
macro F_1-score	67.68	53.11	57.04	392

Table 5. *Mean results per class during 5-fold cross-validation for best performing model GBERT on the fine-grained labels in few-shot classification with Fastfit.*

we selected a fold where the results of the model were closest to the mean results on all folds. We create confusion matrices for the two models to illustrate common errors. Figure 1 shows the results of GBERT which was fine-tuned on the coarse-grained data version consisting of six classes and Figure 2 shows the confusion matrix for GBERT which was fine-tuned on the fine-grained data version. It should be noted that the two confusion matrices only show incorrectly predicted labels, while all correctly predicted labels were removed.

The confusion matrix for the coarse-grained model shows that most instances of errors consist either of *assertives* that were classified as *expressives* or *expressives* classified as *assertives*. This finding corroborates our observations made during the annotation of the data, where the greatest challenge was to distinguish between *assertives* an *expressives*. An example can be seen in Table 6, ex. (1). While the declarative sentence structure might appear like a mere statement on the surface, the utterance is actually used to express a negative attitude which is not recognised by the classifier. Furthermore, the confusion matrix illustrates that the class *unsure* often leads to misclassifications with *assertives*. This class was originally created to address uncertainties due to missing or unclear contexts. As the nature of this class heavily relies on the surrounding context of the utterance which is not available to the classifier during prediction, the classifier cannot accurately predict the class *unsure* which is reflected in the high error rate. Thus, the classifier tends to incorrectly classify instances which

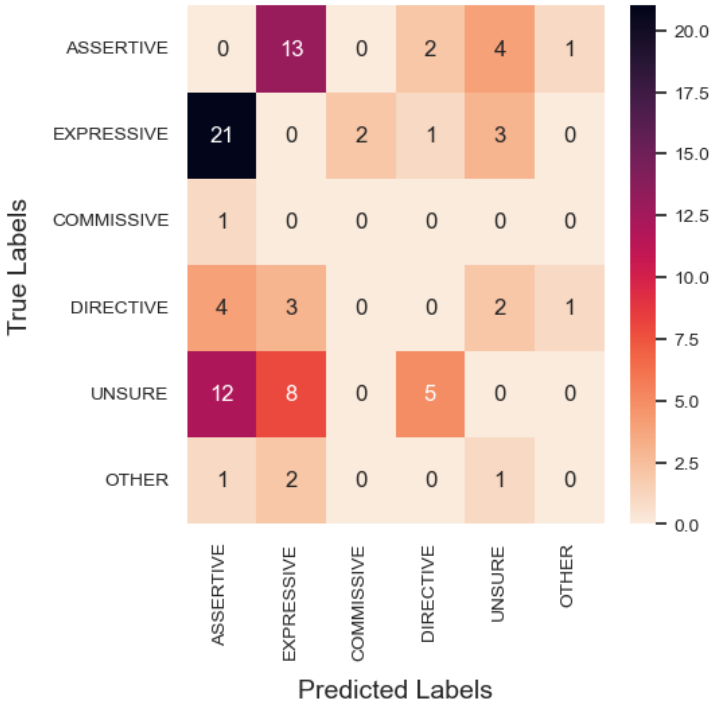

Figure 1. *Confusion matrix for all incorrectly classified instances of the best performing coarse-grained classifier GBERT after hyperparameter search to illustrate common errors.*

are usually shorter and do not provide much context by themselves, as can be seen in Examples (2), (3) and (4) in Table 6.

Figure 2 shows similar observations. The class *unsure* leads to several errors, repeatedly, while the classes *assert* (a subclass of *assertives*) and *complain* (a subclass of an *expressive*) are frequently confused with each other. In example (5), for instance, one could argue that the utterance is both describing the world and expressing an attitude, simultaneously, leading to misclassifications. In example (6) in Table 6, the expression is clearly used to express a negative feeling of the speaker. Nevertheless, the form of a declarative sentence, which is the sentence structure most often used for *assertives* (Plakidis and Rehm, 2022), might have led the classifier to incorrectly classify it as an instance of the *assert* class.

	Text	*Correct label*	*Predicted label*
	coarse-grained labels		
1	*Wir leben in einem Irrenhaus.* we live in a madhouse 'We live in a madhouse.'	Expressive	Assertive
2	*Ein Witz.* a joke 'A joke.'	Expressive	Unsure
3	*Wie kannst du!* how can you 'How could you!'	Unsure	Directive
	fine-grained labels		
4	*Ja!* yes 'Yes!'	Agree	Unsure
5	*Kein Wunder, dass Bewegungen wie z.B. AfD usw. so viel Zulauf haben.* no wonder that movements like e.g. AfD etc. so much support have 'No wonder movements like the AfD, etc., have so much support.'	Assert	Complain
6	*Ich habe eine Scheißangst.* I have a shit.fear 'I'm scared as hell.'	Complain	Assert

Table 6. *Examples of misclassifications for coarse- and fine-grained speech act classification.*

6. Conclusion

Our results demonstrate that encoder-based models outperform decoder-based models in the task of speech act classification. The best performing classifier is GBERT in both settings, achieving a macro F_1-score of 68.68 for coarse-grained classification and a macro F_1-score of 57.04 for fine-grained classification.

Our results show that there is still room for improvement regarding the automated detection of speech acts, which could involve a new annotation scheme with more precise guidelines in order to diminish potentially overlapping classes and vagueness concerning definitions. During the annotation process, we observed that some examples in our data fit multiple classes at the same time, especially with regard to the distinction between *assertives* and *expressives*, which renders the task of speech act annotation rather subjective. This is also reflected in the error analysis of this paper which shows that the distinction between *assertives* and *expressives* is a frequent error.

For future work, we thus intend to revise our annotation scheme and annotate a larger, more balanced dataset, enabling us to train improved speech act classifiers as well as an offensive language classifier to investigate whether the inclusion of speech acts improves the detection of offensive language on a larger dataset. In addition, we plan to release a curated version of the annotations.

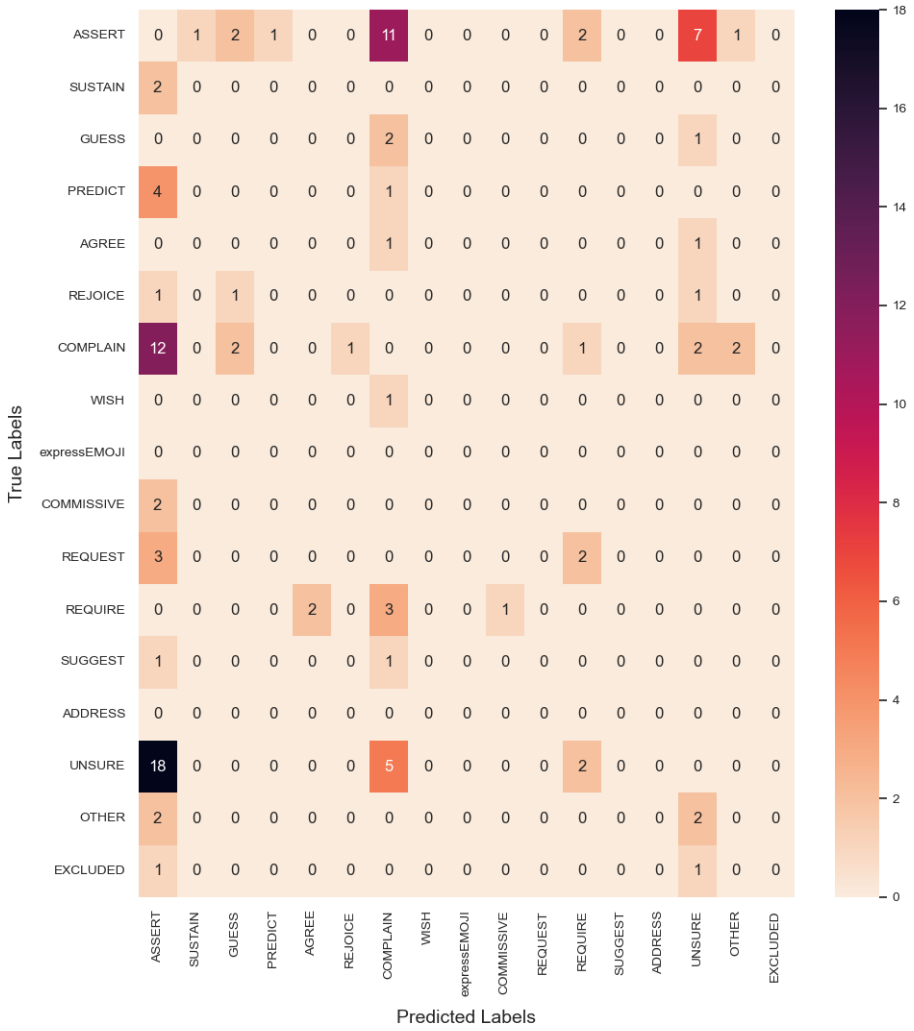

Figure 2. *Confusion matrix for all incorrectly classified instances of the best performing fine-grained classifier GBERT to illustrate common errors.*

7. References

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

Benamara F., Mari A., Meunier R., Moriceau V., Moudjari L., Tinarrage V., "Digging Communicative Intentions: The Case of Crises Events", *Dialogue Discourse*, 2024.

Bilewicz M., Soral W., "Hate speech epidemic. The dynamic effects of derogatory language on intergroup relations and political radicalization", *Political Psychology*, vol. 41, p. 3-33, 2020.

Chan B., Schweter S., Möller T., "German's Next Language Model", *arXiv preprint arXiv:2010.10906*, 2020.

Cohen J., "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.", *Psychological Bulletin*, vol. 70, p. 213-220, 1968.

Compagno D., Epure E., Deneckere R., Salinesi C., "Exploring Digital Conversation Corpora with Process Mining", *Corpus Pragmatics*, vol. 2, p. 193-215, 2018.

Dettmers T., Pagnoni A., Holtzman A., Zettlemoyer L., "QLORA: efficient finetuning of quantized LLMs", *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

Devlin J., Chang M.-W., Lee K., Toutanova K., "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding", *arXiv preprint arXiv:1810.04805*, 2018.

Dhayef Q., Ali A., "A Pragmatic Study of Racial Hate Speech", *Journal of Tikrit University for Humanities*, vol. 27, nº 8, p. 24-1, 2020.

Fleiss J. L., "Measuring nominal scale agreement among many raters.", *Psychological bulletin*, vol. 76, nº 5, p. 378, 1971.

Gemini Team, Georgiev P., Lei V. I., Burnell R., Bai L., Gulati A., Tanzer G., Vincent D., Pan Z., Wang S., Mariooryad S., et al., "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context", *arXiv*, 2024.

Kozlowski D., Lannelongue E., Saudemont F., Benamara F., Mari A., Moriceau V., Boumadane A., "A three-level classification of French tweets in ecological crises", *Information Processing & Management*, vol. 57, nº 5, p. 102284, 2020.

Laurenti E., Bourgon N., Benamara F., Mari A., Moriceau V., Courgeon C., "Give me your Intentions, I'll Predict Our Actions: A Two-level Classification of Speech Acts for Crisis Management in Social Media", *13th Conference on Language Resources and Evaluation (LREC 2022)*, p. 4333-4343, 2022.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L., Schwab D., "Flaubert: Unsupervised language model pre-training for french", *arXiv preprint arXiv:1912.05372*, 2019.

Liaw R., Liang E., Nishihara R., Moritz P., Gonzalez J. E., Stoica I., "Tune: A Research Platform for Distributed Model Selection and Training", *arXiv preprint arXiv:1807.05118*, 2018.

Lin T.-Y., Goyal P., Girshick R., He K., Dollár P., "Focal Loss for Dense Object Detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, nº 2, p. 318-327, 2020.

Llama Team, Grattafiori A., Dubey A., Jauhri A., Pandey A., Kadian A., Al-Dahle A., Letman A., Mathur A., Schelten A., Vaughan A., et al., "The Llama 3 Herd of Models", *arXiv*, 2024.

Mangrulkar S., Gugger S., Debut L., Belkada Y., Paul S., Bossan B., "PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods", 2022.

Martin L., Muller B., Ortiz Suárez P. J., Dupont Y., Romary L., de la Clergerie É., Seddah D., Sagot B., "CamemBERT: a Tasty French Language Model", *in* D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (eds), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, p. 7203-7219, July, 2020.

Mubarok Y., Sudana D., Yanti D., Aisyah A. D., Af'idah A. N. *et al.*, "Abusive Comments (Hate Speech) on Indonesian Social Media: A Forensic Linguistics Approach", *Theory and Practice in Language Studies*, vol. 14, n° 5, p. 1440-1449, 2024.

Nobata C., Tetreault J., Thomas A., Mehdad Y., Chang Y., "Abusive Language Detection in Online User Content", *Proceedings of the 25th International Conference on World Wide Web*, p. 145-153, 2016.

Oktaviani A. D., Nur O. S., "Illocutionary Speech Acts and Types of Hate Speech in Comments on @Indraakenz's Twitter Account", *International Journal of Science and Applied Science: Conference Series*, 2022.

Ollagnier A., "CyberAgressionAdo-v2: Leveraging Pragmatic-Level Information to Decipher Online Hate in French Multiparty Chats", *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024.

Plakidis M., Rehm G., "A Dataset of Offensive German Language Tweets Annotated for Speech Acts", *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, p. 4799-4807, 2022.

Poletto F., Basile V., Sanguinetti M., Bosco C., Patti V., "Resources and benchmark corpora for hate speech detection: a systematic review", *Language Resources and Evaluation*, vol. 55, p. 477-523, 2021.

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

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

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

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

Wang Z., Pang Y., Lin Y., Zhu X., "Adaptable and Reliable Text Classification using Large Language Models", *arXiv*, 2024.

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

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

Yehudai A., Bandel E., "When LLMs are Unfit Use FastFit: Fast and Effective Text Classification with Many Classes", *ArXiv*, 2024.

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

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