Annotation Scheme to Create an Incivility-Dictionary for German-Language Online Discussions.

The goal of the **manual annotation** is to collect uncivil expressions such as words, emojis, and hashtags to create a **dictionary** for **incivility in German-language online discussions**. This collection of uncivil expressions can be utilized to improve **algorithms** for the automated content analysis of user comment sections.

What is uncivil? Subcategories of Incivility

Incivility has many facets and can be perceived differently depending on the **situation** or **personal** background. We consider the following forms (**subcategories**) of incivility.

Vulgarity **Inappropriate Language** **Swearing**	The use of **swear words**, **obscene**, **vulgar**, or **boorish language** that is inappropriate/incongruous to civilized discourse. Vulgar language and swear words may or may not be directed at someone or something.

Insults **Name Calling** **Profanity**	**Name Calling** and **insults** of persons, groups of persons, or objects to **degrade** and **abuse** them.
Dehumanization	**Degrading** or **insulting** persons or groups of persons by denying human qualities.
Sarcasm **Mockery** **Cynicism**	**Degradation** of a reference object (person or other objects) by **ridiculing** someone or something, i.e., depicting it in a **ridiculous manner**.
Negative Stereotypes	**Degradation** of persons or other objects, expressed through the use of **negatively connoted stereotypes** and **generalizations.**
Discrimination	Disparagement of a person or group of persons, based on membership of an **ethnic**, **religious**, **cultural** or **social group**, as well as on the basis of **gender**, **sexual orientation**, social **status** or **physical** characteristics.

Threats of Violence	Threats, **announcement** or **advocacy** of and **incitations** to violence. **Verbal** or **physical aggressiveness** towards a reference object is **announced** or called for, demanded or even advocated violent/criminal acts. Also, when the hope is expressed that something bad my happen to the reference object, or **misfortune** befalls it.
Denail of Rights	The denial of **human/democratic** rights of a person or group of persons.
Accusations of Lying	Someone accuses a person, institution/organization or other object of (knowingly) lying.
(Other) Degradation **Disrespect** **Devaluation**	Any form of **Defamation** of persons or other objects that have the recognizable goal of disparaging them.

How to annotate? Procedure and Annotation Rules

For the Incivility dictionary, all subcategories of incivility should be considered. Every time you come across a word, emoji, or hashtag in a comment or tweet that can be assigned to one or more of the categories, it should be **flagged.** It doesn't matter which of the subcategories it is exactly. Read each comment once carefully. Some comments contain only a few words, others several sentences. It is further possible that you will find **more than one uncivil passage** in one comment. Note that all uncivil expressions should be marked separately**.**

For the annotation we use the software **Excel**. Unfortunately, Excel is a bit cumbersome for annotation because the program is not directly intended for flagging and collecting text passages on a large scale. Therefore, you have to **copy** the uncivil from the comment and **paste** it into one of the columns *Word1, Word2, ..., Word10*. You can flag a maximum of 10 different expressions per comment.

The goal of the annotation is to create a dictionary of uncivil expressions here, uncivil *unigrams*. Unigrams refer to contiguous strings of letters and/or other characters separated by spaces. Primarily, this simply includes single words. In our case, we are also dealing with special social media-unigrams, namely **emojis**, **hashtags**, or intentionally or unintentionally "**misspelled**" words. The following rules explain what and how exactly to annotate.

Annotation Rules

1. The goal of the annotation is to collect entries for an incivility dictionary. Entries for the dictionary should therefore be as **unambiguously uncivil**. That means, they should be considered as uncivil in other online discussion contexts, too.
2. The classic case for uncivil unigrams are **single words**, such as insults or swear words, e.g. *Turd*, *this shit, I could puke,* worst *dick comparison*, *Son of a bitch*, *fuck*!, *load of shit*. These will probably appear most often. Here, the entire word is simply copied and pasted.
    a. Tip: **Double-clicking** on the word will highlight the whole word between the spaces/punctuation marks.
3. In addition to explicit insults, supposedly more harmless disparagement also counts as incivility, e.g., *witch*, *Uschi*, *blabbering*, *blablablaaa*
4. Adjectives can also be uncivil, e.g., *braindead*, *stupid*, *dumb*, *retarded*, *fat*
5. **"Camouflaged"**, intentionally misspelled or differently spelled words are also marked, e.g., *fukc*, *sh..t*., *fuuuuuuck!!*. This does not apply if words have been "accidentally" written together, e.g., *thisshit*.
    a. **Attention:** Words that are accidentally written apart are annotated as one word, e.g. *Gut menschentum*, and corrected in the column "correct", e.g. to *Gutmenschentum*.
6. **English words** are also marked: *bullshit*, *Loser*.
7. **Uncivil hashtags** are also flagged, including the hashtag #: *#whiteprides*.
8. **Uncivil emojis/emoticons** are flagged as well:*:pile_of_poo:* 😬 👊 🔥. If a statement is only uncivil due to the combination of emojis, the emojis are coded together as one uncivil expression, e.g., *:woman_with_headscarf:: angry_face_with_horns:* For emoticons that do not appear as a graphic but as a label, the special characters are encoded as well, i.e. colons : at the beginning and end and underscores _ in the middle.
9. **Numbers** are marked if they are part of an uncivil expression, e.g. *0815minister*.
10. Generally **omitted** during annotation are:

   a. **Whitespace**
   b. **Links** (websites, people): *@petertauber @ARDde*
   c. Names of **private** and **public persons**, e.g., *Anja, Angela Merkel*
   d. **Punctuation marks, special characters, numbers** and **smileys:** *!!! ;) ? … @ $ % +*
      i. Exception: hashtags are coded with if the hashtag is uncivil, e.g., *#left-greenstuffed*
      ii. Exception: If related words, emojis, or hashtags are separated by special characters but belong together, they are marked as one expression, e.g., *#fuck_ni**er*
      iii. Exception: If an uncivil word is only uncivil due to the special characters/punctuation marks, e.g., *„refugees“*. Further, special characters in misspelled or "camouflaged" words are flagged, e.g. *f**k*.
      iv. Exception: If emojis are not displayed as graphics, but the name is written out, the colons at the beginning and at the end are also coded, e.g., *:pile_of_poo:*
      v. Exception: If numbers are part of the uncivil expression, they are coded as well, e.g., *0815minister*

11. For compound uncivil expressions in which both individual words are uncivil, the individual parts are annotated separately, e.g., *stupid* *motherfucker*
12. **Attention:** contextual incivility is not annotated, i.e., if the incivility cannot be attached to individual uncivil expressions, it is not considered. These include:
   a. Uncivil words or emojis that are not uncivil per se, only in context, e.g., *Mouse, bird,* 🤣, 😨, *out*
   b. Ironic words and emojis, which can only be identified as sarcasm in context, e.g., *super*, *:thumbs_up:*
   c. Words or hashtags that are not uncivil themselves, but only vouch for uncivil potential, e.g., *#metoo*, *Todesstrafe*

d. **Attention:** In case of doubt, it is better to mark than to lose! If you are not 100% sure, annotate the expression anyway.