UNCIVIL ANNOTATIONS

Annotation Scheme for Creating an Incivility-Dictionary for German Online Discussions.

About UNCIVIL ANNOTATIONS

Projektziel

Die computergestützte Erkennung von **Inzivilität in Onlinediskussionen** ist Teil des Forschungsauftrags der Nachwuchsforschungsgruppe *DEDIS – Deliberative Diskussionen im Social Web*. Hierzu gehört ein wichtiger Teilschritt: die manuelle Annotation von Nutzerkommentaren. Ziel der **manuellen Annotation** ist es, inzivile Ausdrücke wie Wörter, Emojis oder Hashtags zu sammeln, um daraus ein **Diktionär** für **Inzivilität online** zu erstellen. Mit Hilfe dieses Diktionärs aus inzivilen Ausdrücken können **Algorithmen** für das Aufspüren von Inzivilität in Nutzerkommentaren verbessert und Community Manager bei der Moderation von Onlinediskussionen unterstützt werden.

Da wir aus einer demokratietheoretischen Sicht auf öffentliche Diskussionen schauen, erhoffen wir uns auch von Onlinediskussionen einen **deliberativen** Meinungsaustausch. In der Praxis heißt das konkret, dass in einem öffentlichen Diskurs, wie z.B. auf **Facebook** oder **Twitter**, jeder gleichberechtigt seine Ansichten einbringen und Meinungen austauschen kann, ohne persönlich angegriffen, beleidigt oder bedroht zu werden – ein Grundpfeiler jedes demokratischen Systems. Wer sich jedoch in **Kommentarspalten** umsieht, findet meist keine deliberativen Diskussionen vor, sondern Beschimpfungen, Diskriminierung bis hin zu Morddrohungen. Solche Diskussionen bezeichnen wir als **inzivil**.

Was ist inzivil? Unterkategorien von Inzivilität

Inzivilität hat viele Gesichter und je nach **Situation** oder persönlichem **Empfinden** nehmen wir Aussagen anders wahr. Das macht es auch für **Algorithmen** nicht gerade leicht, Inzivilität zu erkennen. Auch in der Forschung ist man sich noch uneinig, was nun genau

unter Inzivilität zu verstehen ist. Für unser Projekt berücksichtigen wir die folgenden Formen, oder auch **Unterkategorien**, von **Inzivilität**.

Vulgarität/ Unangebrachte Sprache/ Fluchen	Die Verwendung von **obszöner**, **unflätiger** oder **rüpelhafter** Sprache, die für einen zivilisierten Diskurs unangebracht/unangemessen ist. Vulgäre Ausdrücke können, müssen aber nicht an jemanden oder etwas gerichtet sein.
Beschimpfungen/ Schimpfwörter/ Name Calling	**Schimpfwörter** und **Beleidigungen** von Personen, Personengruppen oder anderen Objekten, um diese zu verunglimpfen.
Dehumanisierung	**Herabwürdigungen** oder **Beschimpfungen** von Personen oder Personengruppen, um diese zu **entmenschlichen**, indem menschliche Eigenschaften abgesprochen werden.
Sarkasmus/ Verspottung/ Zynismus	**Herabwürdigung** eines Referenzobjektes (Person oder andere Objekte), indem jemand oder etwas ins Lächerliche gezogen, also **lächerlich dargestellt** wird.

Negative Stereotype	**Herabwürdigung** von Personen oder anderen Objekten, die sich durch die Verwendung **negativ konnotierter Stereotype, Verallgemeinerungen** äußert.
Diskriminierung	Herabsetzung einer Person oder Personengruppe, auf Grund der Zugehörigkeit einer ethnischen, religiösen, kulturellen oder sozialen Gruppe, sowie auf Grund des Geschlechts, der sexuelle Orientierung, des sozialen Status oder physischer Eigenschaften.
Androhung von Gewalt	Drohungen, Ankündigung oder Befürwortung von Gewalt. Verbale oder physische Aggressivität gegenüber einem Referenzobjekt wird ankündigt oder zu Gewalt-/Straftaten aufgerufen, gefordert oder auch nur befürwortet. Auch, wenn die Hoffnung geäußert wird, dem Referenzobjekt möge etwas zustoßen, oder Unglück widerfahren.
Rechte Absprechen	Das Absprechen von **Menschenrechten/demokratischen** Rechten.

Lügenvorwürfe	Jemand bezichtigt eine Person, Institution/Organisationen oder ein anderes Objekt der (wissentlichen) Lüge.
(Sonstige) Herabwürdigung/ Geringschätzung/ Abwertung	**Diffamierung** von Personen oder anderen Objekten, die nicht in die hier vorgestellten Kategorien einzuordnen sind, aber dennoch das mit dem erkennbare n Ziel haben, diese die Personen oder Objekte herabzuwürdigen.

Was wird annotiert? Stichprobe und Untersuchungsmaterial

Als Untersuchungsmaterial/die Stichprobe für die manuelle Annotation erhältst du eine Zufallsauswahl von **Nutzerkommentaren** auf **Facebook**, die aus Diskussionen zu unterschiedlichen Themen und über einen längeren Zeitraum von verschiedenen Nachrichtenseiten stammen, z.B. *Hart aber Fair*, *Tagesschau* und *Spiegel Online*. Nutzerkommentare können **Initialkommentare** (Level 1 in einem Diskussionsbaum/Thread) oder **Antwortkommentare** sein (Level 2 in einem Diskussionsbaum/Thread).

Für die Annotation wirst du einzelne Initialkommentare und Antwortkommentare lesen, deren Reihenfolge **zufällig** ausgewählt ist. Das Diskussionsthema wird daher oft nicht klar sein. Auch kann es passieren, dass Bezug auf einen anderen Kommentar genommen wird, der nicht zurückzuverfolgen ist. Für die Annotation ist das jedoch kein Problem. Ganz im

Gegenteil: Für unser Wörterbuch eignen sich besonders gut Wörter und Phrasen, die auch **ohne Kontext** inzivil sind.

Wie wird annotiert? Vorgehen und Annotationsregeln

Für das Incivility-Diktionär sollen alle Unterkategorien von Inzivilität berücksichtigt werden. Jedes Mal, wenn du in einem Kommentar oder Tweet auf ein Wort, eine Emoji oder einen Hastag stößt, der sich eine der Kategorien zuordnen lässt, wird dieses/r **markiert**. Dabei spielt es keine Rolle, um welche der Unterkategorien genau es sich handelt. Lies jeden Kommentar einmal sorgfältig durch. Manche Kommentare enthalten nur wenige Wörter, andere mehrere Sätze. Es kann also durchaus vorkommen, dass du **mehr als eine inzivile Textstelle** entdeckst. Alle inzivilen Ausdrücke werden **separat** markiert.

Für die Annotation benutzen wir die Software **Excel**. Leider ist Excel etwas umständlich für die Annotation, weil das Programm nicht direkt auf das Markieren und Sammeln von Textstellen in großem Stil gedacht ist. Markieren heißt hier also, dass du das inzivile Wort aus dem Kommentar in der Spalte *Text* **kopieren** und in eine der Spalten *Wort1, Wort2, ..., Wort10* **einfügen** musst. Pro Kommentar kannst du maximal 10 verschiedene Ausdrücke markieren.

Annotationsregeln Was musst du beachten?

Ziel der Annotation ist es, ein **Wörterbuch** inziviler Ausdrücke zu erstellen. Gesammelt werden **inzivile *Unigramme***. Mit Unigrammen sind zusammenhängende Zeichenketten aus Buchstaben und/oder anderen Zeichen gemeint, die durch Leerzeichen voneinander getrennt sind. Dazu gehören in erster Linie einfach einzelne Wörter. Da wir uns im Bereich

Social Media bewegen, haben wir es darüber hinaus auch mit besonderen Unigrammen zu tun, nämlich Emojis und Hashtag oder bewusst oder unabsichtlich „falsch“ geschriebenen Wörtern. Die folgenden Regeln erklären dir, was und wie genau du annotierst.

Annotationsregeln

1. Ziel der Annotation ist es, Einträge für ein Inzivilitätswörterbuch zu sammeln, dass z.B. Algorithmen bei der automatischen Erkennung von inzivilen Kommentaren helfen kann. Einträge für das Wörterbuch sollten also möglichst **eindeutig inzivil** sein und im besten Fall auch in anderen Onlinediskussionen als inzivil gelten.
2. Der klassische Fall für **inzivile** Unigramme sind **Einzelwörter**, wie Beleidigungen, Beschimpfungen oder obszöne Wörter, z.B. *Kackhaufen*, *diese Scheiße*, *ich könnte kotzen*, *Brechreiz*, übelster *Schwanzvergleich*, *Hurensohn*, *Wutbürger*, zum *kotzen*!, *Scheißverein*. Diese werden vermutlich am häufigsten auftauchen. Hier wird einfach das gesamte Wort kopiert und in eingefügt.
    a. Tipp: Mit einem **Doppelklick** auf das Wort wird direkt das ganze Wort zwischen den Leerzeichen/Satzzeichen markiert.
3. Neben expliziten Beschimpfungen zählen auch vermeintlich harmlosere Herabwürdigungen zu Inzivilität, z.B. *Sitzenbleiber*, *Uschi*, *Tante*, *rumlabern*, *blablablaaa*
4. Auch Adjektiven können inzivil sein, z.B. *hirnlos*, *dämlich*, *super ätzend*, *beknackt*, *gestört*, *behindert*
5. Unkenntlich gemachte, „**getarnte**“, absichtlich falsch oder anders geschriebene Wörter werden ebenfalls markiert, z.B. *fukc*, *gesch..ssen.*, *fuuuuuuck!!*. Dies gilt nicht, wenn Wörter „versehentlich“ zusammengeschrieben wurden, z.B. *diesescheiße ey*.
    a. Achtung: Wörter, die „versehentlich“ auseinandergeschrieben sind, werden als ein Wort annotiert, z.B. *Gut menschentum*, und nach dem Einfügen korrigiert z.B. zu *Gutmenschentum*.
6. Auch **englische Wörter** werden markiert: *fuck*, *shit*, *Loser*.

7. **Inzivile Hashtags** werden codiert, inklusive des Hashtags #: *#whitepridesl*.
8. **Inzivile Emojis/Emoticons** werden mitcodiert: *:pile_of_poo:* 😫 👊 🔥. Ist eine Aussage erst durch die Kombination von Emojis inzivil, werden die Emojis zusammen als inzivil codiert, z.B. *:woman_with_headscarf:: angry_face_with_horns:* Bei Emoticons, die nicht als Grafik sondern als Bezeichnung auftauchen, werden die Sonderzeichen mitcodiert, also Doppelpunkte : am Anfang und Ende und Unterstriche _ in der Mitte.
9. **Zahlen** werden mitcodiert, wenn sie Teil eines inzivilen Ausdrucks sind, Z.B. *365Tages-Schulz*.
10. Bei der Annotation werden generell **ausgespart**:
    a. **Leerzeichen**
    b. **Verlinkungen** (Websites, Personen): *@petertauber @ARDde*
    c. Namen von **privaten** und **öffentlichen Personen**, z.B. *Anja, Angela Merkel*
    d. **Satzzeichen**, **Sonderzeichen, Zahlen** und **Smileys**: *!!! ;) ? … @ $ % +*
        i. Ausnahme**:** Hashtags werden mitcodiert, wenn der Hashtag inzivil ist, z.B. *#linksgrünversifft*
        ii. Ausnahme: Sind zusammenhängende Wörter, Emojis oder Hashtags, durch Sonderzeichen getrennt, gehören aber zusammen, werden sie im Ganzen markiert, z.B. *#fuck_nigger*
        iii. Ausnahme: Ist ein inziviles Wort erst durch die Sonderzeichen/Satzzeichen inzivil, z.B. *„flüchtlinge“* Auch bei falschgeschriebenen („getarnten“) Wörtern werden die Sonderzeichen mitcodiert, z.B. *f**k*.
        iv. Ausnahme: Sind Emojis nicht als Graphik angezeigt, sondern die Bezeichnung ausgeschrieben, werden die Doppelpunkte am Anfang und am Ende mitcodiert, z.B. *:pile_of_poo:*
        v. Ausnahme: Sind Zahlen Teil des inzivilen Ausdrucks, werden sie mitcodiert, z.B. *365TagesSchulz*

11. Bei zusammengesetzten Inzivilen Ausdrücken, in denen beide Einzelwörter inzivil sind, werden die Einzelteile separat annotiert, z.B. *dumme* *Sau*
12. **Achtung**: Kontextabhängige Inzivilität wird *nicht* annotiert, das heißt, wenn die Inzivilität nicht an einzelnen inzivilen Ausdrücken festzumachen ist, wird sie nicht berücksichtig. Dazu gehören:
    a. Inzivile Wörter oder Emojis, die nicht per se inzivil sind, nur im Kontext, z.B. *Maus, Vogel,* 🤣, 😨, *raus*
    b. Ironische Wörter und Emojis, die sich jedoch (wirklich) nur im Kontext als Sarkasmus enttarnen lassen, z.B. *super, :thumbs_up:*
    c. Wörter oder Hashtags, die selbst nicht inzivil sind, sondern nur inziviles Potenzial bürgen, z.B. *#metoo, Todesstrafe*
    d. **Achtung**: Im Zweifel gilt jedoch: Lieber markieren als verlieren! Wenn du dir nicht 100% sicher bist, annotiere den Ausdruck trotzdem.