Detecting Impoliteness and Incivility in Online Discussions

Classification Approaches for German User Comments

Anke Stoll, Marc Ziegele, Oliver Quiring

CCR 2 (1): 109–134
DOI: 10.5117/CCR2020.1.005.KATH

Abstract

Impoliteness and incivility in online discussions have recently been discussed as relevant issues in communication science. However, automatically detecting these concepts with computational methods is challenging. In our study, we build and compare supervised classification models to predict impoliteness and incivility in online discussions on German media outlets on Facebook. Using a sample of 10,000 hand-coded user comments and a theory-grounded coding scheme, we develop classifiers on different feature sets including unigram and n-gram distributions as well as various dictionary-based features. Our findings show that impoliteness and incivility can be measured to a certain extent on the word level of a comment, but the models suffer from high misclassification rates, even if lexical resources are included. This is mainly because the classifiers cannot reveal subtle forms of incivility and because comment authors often use predictive words of incivility or impoliteness in non-offensive ways or in different contexts. Still, when applying the classifiers to a comparable set of comments, we find that the machine-coded categories and the hand-coded categories reveal similar patterns regarding the distribution of and the user reactions to uncivil/impolite comments. The findings of our study therefore provide new insights into the supervised machine learning approach to the detection of different forms of offensive language.

Keywords: incivility, user comments, document classification, machine learning, automated content analysis, impoliteness, online discussions

Introduction

Discussions in comment sections often include high levels of rude, offensive, or even hateful language. Researchers in social sciences have argued that using such language can be considered a violation of democratic and social norms (Muddiman & Stroud, 2017). They have therefore used the term 'incivility' to describe different forms of disrespectful and harmful language (e.g., Coe, Kenski, & Rains, 2014; Muddiman & Stroud, 2017). Previous studies have reported various negative effects of uncivil comments on the readers of online discussions. For example, uncivil comments adversely influenced the thoughts and feelings of readers towards news media organizations (Prochazka et al., 2018) as well as towards other individuals or social groups (Hsueh et al., 2015), and towards political issues in general (Anderson, Brossard, Scheufele, Xenos, & Ladwig, 2014). Some news organizations, due to lacking capacities or an unwillingness to deal with the massive numbers of uncivil comments, have also shut down the comment sections on their websites or outsourced user comments to third-party platforms such as Facebook (Larsson, 2018).

Computational methods could support media organizations and journalists in managing user comment sections and in detecting problematic comments more efficiently. An important and popular approach is the automated classification (or categorization) of text data using *Supervised Machine Learning* (*SML*) techniques. A *Classifier* that has been trained appropriately once may be applied to automatically predict the text categories learned, such as incivility, for comparable data without much further manual coding. However, most related research is done for the English language and cannot easily be transferred to non-English text data (Gitari, Zuping, Damien, & Long, 2015; Silva, Mondal, Correa, Benevenuto, & Weber, 2016). Further, methods to automatically detect abusive or harmful user-generated content work best for 'obvious' forms that are clearly expressed through the use of specific words, such as offensive language or extreme forms of hate speech (e.g., Davidson, Warmsley, Macy, & Weber, 2017). Automatically detecting subtle forms of incivility, such as covert racism, is more challenging. Still, from a psychological viewpoint, these forms could affect the attitudes of readers even stronger than obvious forms of offensive language (Kalch & Naab, 2016; Papacharissi, 2004). Based on previous theoretical work on incivility (Muddiman & Stroud, 2017; Papacharissi, 2004), we therefore train classifiers on both *impolite* comments—postings that are offensive but not necessarily harmful to other users—and *'truly'* *uncivil* comments, which often include subtle forms of racism, extremism, and undemocratic appeals (e.g., Kalch & Naab, 2017).

For our study, we rely on a dataset of more than 10,000 manually-labeled user comments that were posted to the Facebook sites of German media outlets. We use different feature sets to build our models that predict a comment's incivility and impoliteness. These feature sets include representations of single words (*unigrams*) and word combinations (*n-grams*) as well as features based on lexical resources, such as insults, polarity, and sentiment dictionaries. We also test the contributions of *Named-Entity Recognition* (*NER*) and several standard *Natural Language Processing* (*NLP*) techniques, such as *Part-of-Speech tagging* (*POS tagging*) or *lemmatization* and *stemming*. This way, we can compare the extent to which impoliteness and incivility in user comments can be predicted on the level of words and provide insights into the features that predict each category. Finally, we apply the classifiers to another hand-coded data set of user comments. Thereby, we illustrate the applicability of classifiers that are trained on a specific dataset to other datasets and discuss their validity to answer specific research questions from social science.

In sum, the present study contributes to the literature by applying the theoretical differentiation between impoliteness and incivility, which is important in communication research, to machine learning problems. The study also offers a methodologically-focused reflection of the requirements, potentials, and limitations of the SML approach and tests the applicability of classifiers to comparable data sets. The findings can support scholars and media professionals in developing automated methods to detect and manage potentially harmful user-generated content.

Theory: Incivility and Impoliteness

Incivility is difficult to define because the decision of what is civil and uncivil is subjectively shaped (Coe et al., 2014; Herbst, 2010). Therefore, achieving consensus about where to draw the line between civil and uncivil discourse is a complex problem (Muddiman, 2017; Stryker, Conway, & Danielson, 2016). Scholars have defined incivility as the communication of disagreement combined with a dismissive, disrespectful, aggressive, or hostile tone (Coe et al., 2014; Hwang, Kim, & Kim, 2016). Impoliteness—an individual's unwillingness to minimize interpersonal conflict and adhere to the rules of etiquette (e.g., Grice, 1989)—is sometimes seen as a sub-concept of incivility (e.g., Coe et al., 2014) and sometimes as an independent concept (e.g., Papacharissi, 2004).

In user comments, various rhetorical and stylistic elements have been labeled as uncivil, including name-calling, aspersions, vulgarity, lying, and pejorative speech (Coe et al., 2014). These elements resemble the language previous work has analysed under the terms of 'flaming' (Alonzo & Aiken, 2004) or 'offensive language' (Davidson et al., 2017). Additionally, researchers have applied the concept of incivility to comments that 'threaten democracy, deny people their personal freedoms, and stereotype social groups' (Papacharissi, 2004, p. 267). Examples of such incivility include racism, sexism, attacking people for belonging to certain social or ethnic groups, or threatening democracy as a whole (Kalch & Naab, 2017; Papacharissi, 2004). Such forms of incivility have previously been analysed as 'hate speech' (Davidson et al., 2017), 'racism' (Daniels, 2009), or 'extremism' (Agarwal & Sureka, 2014).

Empirical research on incivility has not always differentiated between these two forms of disrespectful behaviour (e.g., Coe et al., 2014). Still, in a seminal work on civility, Papacharissi (2004) makes a strong point why there is a need to draw a line between impolite and truly uncivil behaviour: 'Polite manners are a condition necessary, but not sufficient, for civility. And yet, civility is misunderstood when reduced to interpersonal politeness, because this definition ignores the democratic merit of robust and heated discussion.' (Papacharissi, 2004, p. 260). Put differently, calling others names or using vulgar or pejorative language is inconsiderate and violates norms of interpersonal politeness, but it does not necessarily undermine democratic discourse. This view has been supported by research reporting that participants in online discussions felt they were having constructive conversations although neutral observers would rate these discussions as rude (Davidson et al., 2017). The data we use in our study offers several examples of inappropriate language that violates norms of politeness, but that is not 'harmful' in a democratic sense:

- 'Scheiße, gleich Auto verkaufen!!!'[1]
- 'Schmidt ist ein Politiker der noch EIER hat!'[2]
- 'Der machtgeile Mensch Martin Schulz allen voran.'[3]

In contrast, truly uncivil behaviour—assigning stereotypes to social groups, being racist/sexist, threatening democracy as a whole or the democratic rights of social groups—denies the 'collective traditions of democracy' (Papacharissi, 2004, p. 260) and therefore likely has more sustainable negative consequences for societies as a whole. Additionally, although these forms of incivility sometimes include offensive and aggressive language, they are often hidden behind seemingly civil language or arguments (Daniels, 2009; Kalch & Naab, 2017), for example:

- 'Muslime haben ja auch ne hohe geburtsrate.... das gleicht es wieder aus....'[4]
- 'Wir müssen uns währen, Leute geht mit knüppel auf die Straße und zeigt das es uns reicht.'[5]
- 'Es ist vielleicht ein Vorurteile aber: kaum gestohlen schon in Polen.'[6]

Following this line of research, it seems useful to differentiate between impoliteness, which is considered inappropriate but not necessarily harmful, and true incivility, which has more detrimental and lasting consequences. This view is supported by recent empirical work on incivility that has drawn similar differentiations (e.g., Rowe, 2015; Su et al., 2018). These studies typically reveal that impoliteness occurs relatively often in online discussions, whereas incivility occurs less frequently (Papacharissi, 2004; Rowe, 2015a; Su et al., 2018).

The Classification Approach to Automated Text Analysis

Classification is a *Supervised ML* (*SML*) technique, in which predictive models learn a relationship between an outcome *class* (or *category*, dependent variable) and certain *features* (or attributes, independent variables) in a set of *training data*. For example, a classifier can learn the relationship between the category *incivility* and the words used in comments. Training high-quality classifiers often requires a large amount of (manually) labeled data. This especially applies to language data (Baayen, 2002). For example, of all possible uncivil statements, many of them will *not* appear in a given sample, and therefore cannot be considered by a statistical model. It is possible to bypass this problem by using dictionary-based approaches, which rely on a limited set of words as indicators for a certain category, such as a set of words that indicate negative and positive sentiment (Denecke, 2008; Remus, Quasthoff, & Heyer, 2010) or sets of offensive words (Davidson et al., 2017). However, such methods only allow a limited and deterministic understanding of a concept – for example, lexical methods often use a predefined threshold of offensive words that separates harmful documents (comments) from harmless ones. Additionally, creating such corpora is expensive and, so far, many corpora are available in English only and therefore cannot be used to conduct automated content analyses in other languages (Burnap & Williams, 2015; Thelwall, Buckley, & Paltogou, 2011; Thelwall, Wilkinson, & Uppal, 2010b). Machine learning-based methods, unlike lexical approaches, do not necessarily require external lexical resources. Further, these methods can reveal predictive features (e.g., words) that are not 'obviously' uncivil or offensive, but still harmful in the context of or in combination with other words. A ML model, however, will only learn what the training data provides. Creating such training data often involves extensive manual coding procedures (Witten, Frank, Hall, & Pal, 2016). Consequently, creating high-quality training data quickly becomes an expensive endeavour as well. That is why supervised ML approaches often are not economically viable for answering certain research questions.

Ideally, a classifier is not only accurate in predicting a certain class but also generalizable, that is, transferable to comparable data. For example, an incivility classifier that has been trained once ideally can be applied to new user comments without further manual coding. Many classifiers, however, achieve high estimation accuracy but are not applicable to different data, for example to discussions on a different topic. The current study therefore pays special attention to the information the predictive features provide regarding the generalizability of the models. In addition, we evaluate our classifiers on a completely independent sample of user comments.

Related Work

Some concepts in texts, such as topics or sentiments, can reasonably be captured using a predefined dictionary of vocabulary (e.g., Denecke, 2008; Pang & Lee, 2008). In such cases, the use of lexical resources often improves the performance of a method, both for dictionary-based and ML techniques. In the domain of incivility research, previous studies have applied dictionary-based approaches as well (e.g., Muddiman & Stroud, 2017). Still, predefined word lists cannot comprehensively measure the concepts related to incivility. Davidson et al. (2017) found that only five percent of the Tweets that contained words of the English hate speech lexicon *Hatebase.org* were labeled as hate speech by human coders. In general, approaches to automated content analysis can perform well when the concept analysed is closely related to the word level of a statement. In contrast to the category 'hate speech', Davidson et al. (2017) found that the category 'offensive language' could satisfyingly be measured by the occurrence of offensive words. The automated detection of more abstract concepts or subtle forms of, for example, harmful content is more challenging. In their study on racism against black people on Twitter, Kwok and Wang (2013) showed that classifiers did not capture many Tweets including hate speech because these Tweets did not include any race-related words at all.

Tying in with the theoretical argumentation, a comment can be uncivil without including any words that would be considered as offensive. For example, a comment stating 'My cleaning maid performs very well although she is from Turkey' can be considered racist although it does not include any explicit offensive language. Such unobtrusive forms of incivility are more difficult to detect since there are no unambiguous word indicators. Still, a classifier may reveal the appearance of the words 'cleaning', 'performance', and 'Turkey' as indicators for racism in some online discussions. Detecting racism on Twitter, Kwok and Wang (2013) showed that words like 'black', 'white', and 'filthy' are likely used in hate speech against black people, even if they bear no racial undertones outside of the context. But using only single words (*unigrams*) as features can lead to misclassification, since certain words will be used differently in other contexts (Burnap & Williams, 2015). Pendar (2007) showed that using combinations of words (*n-grams*) instead of single words can improve the performance of a classification model. In the example above, a *bigram* structure would keep together 'cleaning maid' and a *trigram* structure would keep together 'cleaning maid performs'. These terms might be more precise predictors for incivility than the single word 'cleaning' is. Yet, such complex features will only improve the model performance if they were learned in the training data *and* if they appear again in the data on which the model is tested. This can be problematic because combinations of words are less likely than single words to appear to a statistically recognizable amount in given sample of text (Mandelbrot, 1961; Zipf, 1945). This is why the automated detection of subtle concepts, such as incivility, becomes more promising the more extensive the training data is.

In sum, automatically classifying documents into abstract concepts based on text patterns requires an appropriate training sample (Grimmer & Stewart, 2017). Creating such training sets is costly and some studies therefore rely on small sample sizes (Su et al., 2018). Further, Ross et al. (2017) have shown that the manual coding of hate speech requires clear definitions and guidelines to produce reliable annotations. Since the labelling of large data sets is expensive, many studies have used non-experts, such as crowd workers, to label their data (Davidson et al., 2017; Hsueh, Melville, & Sindhwani, 2009). Additionally, for the German language, data sets that can be used for training text classification models using supervised machine learning are still rare.

Data

Supervised ML requires categorized instances on which a predictive model can learn a generalizable relationship between data attributes (features) and a certain category (class). The required sample size depends on several factors, such as model complexity and heterogeneity of the data. Especially for text classification, large data sets are needed since language data is particularly heterogeneous (Manning & Schütze, 2000; Zipf, 1945). It is therefore important to obtain training data, in which a concept has been (a) validly measured in (b) a sufficiently large sample.

Sample

To train our impoliteness and incivility models, we used a training data set of 10,114 hand-coded German user comments. This subsample was drawn from a corpus of more than 1,000,000 comments from the Facebook pages of nine German news media that were collected in 2016 via Facebook Graph API. The hand-coded subsample included top-level comments (TLCs)[7] and reply comments that were posted to the different news media outlets as well as to different topics and at various stages of the discussions. The data therefore offers a solid basis for identifying online impoliteness and incivility in a broad variety of contexts.

Operationalization of impoliteness and incivility

Six student assistants were extensively trained to code the comments in the data set regarding their impoliteness and incivility. The coding scheme included the definitions of the two categories as well as various example comments. Overall, three intercoder reliability tests were conducted in which each coder annotated the same comments with regard to their incivility and impoliteness.

The coders read all message in their entirety and then coded the presence of impoliteness and incivility using two comprehensive measures. While these measures differentiated between impoliteness and incivility, they did not further classify the exact sub-type of impoliteness (e.g., name-calling) or incivility (e.g., racism; see next section). Although this approach is commonplace in communication research (e.g., Papacharissi, 2004; Ziegele et al., 2014), future research could consider differentiating between these different subtypes to generate test data for more precise and sophisticated algorithms.

Impoliteness. For every comment, the coders rated the level of impoliteness on a three-point scale (0 = *not impolite*, 1 = *slightly impolite*, 2 = *predominantly impolite*). The scale was adapted from the impoliteness measure used by Papacharissi (2004). That is, a comment was coded as impolite when it included name-calling (e.g., 'weirdo', 'idiot'), aspersions (e.g., 'politicians are corrupt'), synonyms for liar (e.g., 'hoax'), vulgarity, or pejorative speech.[8] A comment was rated as impolite when it included at least one of the impoliteness-related concepts. Intercoder reliability was tested on a sample of 100 comments and reached a satisfactory level of *Krippendorff's* α = .83. For the current analysis, we recoded the impoliteness measure dichotomously (0 = *no impoliteness present*, 1 = *impoliteness present*). Within the sample of hand-coded comments, the category 'no impoliteness' was assigned to 7,419 comments (73.24 %), and the category 'impoliteness' was assigned to 2,707 comments (26.76 %).

Incivility. Following Papacharissi (2004), comments were coded as uncivil when they assigned negative stereotypes to individuals or groups (e.g., 'women are less intelligent than men'), when they included political extremism (e.g., 'we should evict every single refugee'), or when they threatened individual's democratic rights (e.g., 'you have no right to speak') or the integrity of democratic norms and values (e.g., 'we need to overthrow this government'). It is important to note that uncivil comments often also included impoliteness (e.g., 'we need to overthrow this f*cking government'). Coders used a dichotomous measure (0 = *no extreme incivility present*, 1 = *extreme incivility present*) to assign each comment the respective degree of incivility. Intercoder reliability reached an acceptable level of *Krippendorff's* α = .73. In the hand-coded sample, the category 'no incivility' was assigned to 8,454 comments (83.5%) and the category 'incivility' was assigned to 1,676 comments (16.6%).

Feature Sets

For the present study, we used different feature sets to examine and compare the predictability of both concepts and to draw conclusions about which information lead to the classification of the categories. These features include single words (*unigrams*), word combinations of *bigrams* and *trigrams* in a *Bag-of-Words* (*BoW*) representation, and features that were created from dictionaries and word lists of sentiments, polarity, and offensive words for the German language. We also included a feature that depicts the occurrence of *Named Entities* (NE) in comments, since incivility often includes stigmatization of individuals or social and ethnic groups.

Bag-of-Words Features

As a first step, we included *Bag-of-Words* (*BoW*) features of unigrams (single words), bigrams (groups of two words), and trigrams (groups of three words) in our models. A BoW representation displays the distribution of words in a given document. Instead of absolute word counts (absolute *term frequency*, *TF*), we used *TF-IDF* weighted terms frequency, which assigns high weight to terms that occur often but only in few documents (Manning, Schütze, & Raghavan, 2008; Sebastiani, 2002). To reduce (unnecessary) variance in the text data, we applied several *pre-processing*[9] techniques in advance that are commonly used in many *NLP* (*Natural Language Processing*) tasks and applications (e.g., Bird, Klein & Loper, 2009; Manning & Schütze, 2000). We applied *Snowball Stemming* by Porter (2001)[10] and *lemmatization* [11] to reduce different word forms to their mutual stem, respectively lemma. Additionally, we included *Part-of-Speech* (*POS*) *tags*[12] as features, meaning the information whether a word is a noun, a verb, or an adjective, for example. Further, we removed *Stop Words*[13], that is, very frequent, 'non-informative' words (Bird et al., 2009; Jurafsky & Martin, 2014) and excluded non-alphanumeric characters as well as all comments that include only two or less word tokens.

Dictionary-based Features

Previous work on offensive language in user comments suggests that some obvious forms of impoliteness can be detected by the occurrence of insults or offensive words. For our study, we tested features based on lexical resources that are available for the German language and that quantify the appearance of insults, anger, and swearing, but also negative emotions and the polarity of a comment.

insult.wiki Word Count. Due to the lack of academic open source dictionaries of offensive words for the German language, we created a look-up feature from the collection of German offensive words from i*nsult.wiki*[14]. The collection provides a non-weighted list of 1,800 insult words that includes common insults such as 'Idiot' or 'Depp'[15] and a variety of surprising word compositions that are probably not used very often. The list also contains insults that are clearly discriminatory against social groups, such as 'Schlampe', 'Hure', or 'Schwuchtel', 'Transe'[16]. We therefore assume that this feature might improve both the impoliteness and incivility classifiers. Nevertheless, the collection of insults also contains words that might not be offensive in many contexts, such as ‘Brot’, ‘Bürokrat’, ‘Currywurst’, or ‘Dennis’[17].

LIWC Anger, Swear and Negative Emotions. The *Linguistic Inquiry and Word Count* (*LIWC*) by Pennebaker, Francis, and Booth (2001) is a lexical resource and a program for automated text analysis to count words in a series of psychologically relevant categories (Tausczik & Pennebaker, 2010). For our models, we used the categories *Anger*, *Swearing,* and *Negative Emotions* of the German translation of the LIWC2015.[18] The category *Swear* counts the number of vulgar or pejorative words in a given text. The category *Anger* is a subcategory of *Affect* and counts words that include a negative and aggressive valence. *Negative Emotions* include words that are indicative of various negative feelings (Tausczik & Pennebaker, 2010). Like most of the LIWC categories, *Anger*, *Swear*, and *Negative Emotions* are quantified as relative frequencies proportional to the absolute word count of the document and we implemented the corresponding features that way as well.

Polarity. The polarity of a document can be identified by assigning weighted negativity and positivity to each of the words it contains. Doing so, the polarity of a statement can be quantified from -1 to 1, meaning from completely positive to completely negative. We measured the polarity of a comment by applying a polarity dictionary provided by the Institute of Computational Linguistics, University of Zurich, Switzerland[19] and scaled the scores in a range from 0 to 1, since not all ML models can be applied to negative values.

Named Entities

We applied a Named Entity Recognition (NER) system trained on the TIGER[20] and the WikiNER[21] corpus that supports the automated identification of PER (persons), ORG (organizations), and LOC (locations) entities[22]. Following the definition of ‘true’ incivility discussed above, including information about NEs in a comment could help detect impoliteness or incivility directed against certain persons or (democratic) institutions. We implemented the NE feature as a TF-IDF weighted ‘Bag-of-NE’ distribution of single NEs (unigram entities) and combinations of two NEs (bigram entities).

Classification Models

Several ML algorithms can be applied to classification problems. Not all algorithms, however, fit all data structures or characteristics, such as language data (Aggarwal & Zhai, 2012). Further, the algorithms differ in terms of pragmatic circumstances, such as required computing power, which can actually become problematic the more complex a model and the larger the sample is (Kotsiantis, Zaharakis, & Pintelas, 2007; Robert, 2014). For the current classification problem, we compared several models that have been successfully applied to comparable tasks, including *Logistic Regression* (*LogReg*), *Naive Bayes*, *Decision Trees,* and *Support Vector Machines* (*SVM*) (Wiegand, Siegel, & Ruppenhofer, 2018). We further compared models based on different feature sets of unigrams, bigrams, and trigrams (referred to as feature set *BOW(1-3)*), dictionary-based features (*DICT*), and named entity distributions (*NE*). In sum, models based on the relatively simple *Naive Bayes* (*NB*) algorithm outperformed systems using LogReg, Decision Trees, and SVM. SVM systems with nonlinear kernel performed worst for predicting impoliteness and incivility.

Naive Bayes classifiers are statistical models that are based on *Bayes' theorem* and that calculate the probability P for a class C_k given the features $x_1, \dots x_n$.

$$P(C_k \mid x) = \frac{P(C_k) \times P(x \mid C_k)}{P(x)}$$

$P(C_k)$ is the a-priori probability for the occurrence of a class C_k that is based on the frequency of C_k in the training set. For example, as the proportion of polite and impolite comments is roughly 70:30 in the training data, $P(C_{\text{impolite}}) = 0.30$. $P(x)$ is the a priori probability for a feature x. $P(x \mid C_k)$ is the conditional probability for x given C_k, that is, the probability that a certain feature x is classified as C_k. It is based on the common occurrence of x with the C_k in the training data. Following the Bayes' theorem, the probability $P(C_k \mid x)$, which is the conditional probability for a class C_k, such as impoliteness given a feature x, such as a certain word, can be calculated. Classification models based on *Naive Bayes* (*NB*)[23] are among the most efficient classifiers in terms of implementation and computational effort (McCallum & Nigam, 1998; Sebastiani, 2002; Aggarwal & Zhai, 2012). NB models are very fast regarding both their training and testing. In addition, NB classifiers can be used for data with many thousands to millions of features. In practice, NB classifiers are often as powerful or even superior compared to more complex classification algorithms (Pedregosa et al., 2018).

Model building process

All models were trained on a *train set* and tested on a *test set*, that is, a data set of unseen instances. For our analysis, we applied *k-Fold Cross-Validation* (*CV*) that splits the data *k* times into a train set and a test set to overcome a strong dependency of the model performance to the distribution of data in the test set and train set (Garreta & Moncecchi, 2013, Raschka & Mirjalili, 2017). Hence, all performance measures reported are the mean of *k* model runs.

Classifiers can be very sensitive to *unbalanced* training data. When in doubt, a model will predict the predominant class without actually 'learning' a relationship and still will achieve satisfying results (Larose & Larose, 2015). In our sample, impoliteness and incivility are distributed unequally, that is, impolite (n = 2,707) and uncivil comments (n = 1,676) appeared less frequently than polite and civil ones. Therefore, we additionally calculated models on a resampled, balanced training data basis. For impoliteness, we created a random set of 5,400 user comments that included each 2,700 polite and impolite comments. For incivility, we created a set of 3,320 user comments that included 1,660 comments for each class.

In practice, the data points of the various classes cannot always be perfectly separated by a (linear) classification boundary (e.g., due to noisy data or missing informative features). In these cases, models often achieve better results by ignoring some data points instead of getting mislead (Coelho & Richert, 2015; Han & Kamber, 2011). To reduce such *overfitting*, a model can be forced to learn a less flexible but more generalizable classification boundary. In general, ML models provide different *hyper parameters* to control for overfitting or under fitting (Han & Kamber, 2011; Robert, 2014). For NB models, we optimized the hyper parameter *alpha* (Coelho & Richert, 2015)[24].

Results and Evaluation

We trained and tested different classification models to predict impoliteness and incivility in German user comments based on unigram, bigram, and trigram features in a TF-IDF weighted BoW representations (BOW(1-3))[25], dictionary-based features (DICT) and named entity distributions for each comment (NE). Overall, the best model performances were achieved by NB systems. These models outperformed linear SVM and LogReg and Decision tree models by, on average, at least three percent points. For all models, lemmatization worked better than stemming. Further, removing stop words did not strongly affect the model performance but changed the impact of the single unigram, respectively n-gram features for the estimation. Part-of-Speech information did not improve model performances significantly.

Incivility Models

For predicting incivility, the best results in terms of F1, precision, and accuracy scores were achieved by a NB[26] model that used a combination of unigram, bigram, and trigram lemmas (BOW(1-3), stop words removed), the dictionary feature union (DICT) and the named entity distribution feature (NE) on a balanced data set of 3,320 instances. Table 1 shows an overview of the results.

Table 1. Model Overview for Incivility Classification of a NB system

Features	F_1	Recall	Precision	Accuracy
BOW(1)	0.68	0.83	0.58	0.62
BOW(1-2)	0.69	0.83	0.58	0.63
BOW(1-3)	0.69	0.83	0.59	0.62
BOW(1-3) + DICT	0.69	0.82	0.59	0.63
BOW(1-3) + DICT + NE	0.69	0.80	0.61	0.64

Notes. Recall, Precision and F_1 for the positive class (incivility); *alpha = 1.0;* $N_{comments}$=3,320; Cross-validation: *k* = 7.

All models generally identified uncivil comments well (Recall 80 to 83 percent) but, at the same time, they were imprecise because they often classified a comment as uncivil even it was actually civil (Precision 58 to 61 percent). Adding the DICT feature to BOW(1-3) (line 4) did not significantly change the model performance. In other words, the information whether a comment matches offensive words from the dictionaries is not very important for the classification. In contrast, the NE feature slightly improved the precision of the model for the uncivil class (Prec. = 0.61, line 5). That means, information about persons, locations, or organizations in a comment support the model in retrieving uncivil comments. The absolute numbers of the true and predicted labels are shown in the confusion matrix (Table 2). In sum, the best-performing model (BOW(1-3) + DICT + NE) correctly classified 2,100 instances but also made 1,220 mistakes, mostly when identifying civil comments.

Table 2. Model Results Confusion Matrix – True and Predicted Incivility by NB-Classifier

True Label	*uncivil*	367	1303
	civil	797	853
		civil	*uncivil*
		Predicted Label	

Notes. Correctly classified instances shaded in dark grey; $N_{comments}$ =3,320; alpha=1.0; $N_{features}$ =88,306; CV =7.

For predicting incivility, adding bigrams (BOW(1-2)) and trigrams (BOW(1-3)) to the unigram features did not significantly affect the model performance in any direction. Accordingly, the most informative features of the model are mainly unigrams.

The importance of single features can be quantified as the estimated probability P of the feature X given the class C. The most informative features with $P(C = civil|X) = 0.001$ for the best model classifying *incivility* include 'deutschland', 'menschen', 'euro', 'griechenland', 'geld', 'welt', 'merkel', 'land', 'volk', 'schuld', 'politiker', 'regierung'[27]. However, the model also included features that obviously could lead to misclassification on the test set, including 'mal', 'machen', 'werden', 'mehr', 'sehen', 'können'[28]. Similarly, *civility* was classified based on features that are seemingly suitable to predict the class, such as 'gut', 'geben', 'endlich', 'lieb', 'respekt'[29]; $P(C = uncivil|X) = 0.001$). However, the classifier also identified features that were rather ambiguous and possibly led to misclassification, including 'immer', 'zeit', and 'viel'.[30] Consistent with the models' performance indicators, no features of the offensive words dictionaries are included in the most informative features.

Impoliteness Models

Table 3 provides an overview of the performances of the models predicting impoliteness using different features. A NB system with a feature combination of BOW(1-3) lemmas (stop words removed) and the DICT feature union achieved the best model performance on a balanced data set (Table 3, line 2).

Table 3. Model Overview for Impoliteness Classification of a NB system

Features	F_1	Recall	Precision	Accuracy
DICT (all)	0.54	0.43	0.70	0.62
BOW(1-3) + DICT	0.70	0.80	0.62	0.65
BOW(1-3) + DICT + NE	0.69	0.81	0.61	0.64
BOW(1-3)	0.69	0.83	0.59	0.63

Notes. Recall, Precision and F_1 for positive class (impoliteness); *alpha = 1.0;* $N_{comments}$=5,400, CV=7.

When only using the DICT feature union (line 1), the model achieved better precision than any of the other models (Prec. = 70). That is, in 70 percent of the cases, this model correctly classified comments as impolite. At the same time, using only the DICT features made the model miss most of the impolite comments (Rec.= 0.43). In other words, only using lexical resources

(DICT) to classify impoliteness may lead to more precise but less thorough predictions and many impolite comments would not be retrieved.

Similar to the incivility models, using BOW features pushed the recall of the impoliteness classifier (i.e., more impolite comments were found), but reduced the model's precision (line 4). In other words, the model found up to 83 percent of the impolite comments, but, additionally, many of the comments that were classified as impolite were actually polite. Including the DICT features closed that gap to some extent (line 2). In contrast to the incivility models, including the NE feature did not improve the model performance. Apparently, information about persons, locations, or organizations are more important to classify incivility than impoliteness.

The best-performing impoliteness model (BOW(1-3) + DICT) mainly used unigrams for predictions, such as insults 'dumm', 'scheiß', 'schwachsinn', or topic-related words, such as 'volk', 'griechen', 'merkel'[31] ($P(C = \mathit{polite}|X)$ = 0.001). However, it also used ambiguous words that probably cause misclassifications, such as 'machen', 'geben', 'einfach', 'kommen', 'gehen'[32] ($P(C = \mathit{polite}|X)$ = 0.001). For predicting politeness, the model used unigrams such as 'endlich', 'menschen', 'warum', 'immer', 'kinder', or 'machen'[33] ($P(C = \mathit{impolite}|X)$ = 0.001). Table 4 shows the absolute numbers of the true and predicted class labels.

Table 4. Model Results Confusion Matrix – True and Predicted Impoliteness by NB-Classifier

True Label	impolite	545	2153
	polite	1338	1314
		polite	impolite
		Predicted Label	

Notes. Correctly classified instances dark shaded; $N_{comments}$ =5,400; alpha=1.0; $N_{features}$ =130,216; CV=7.

In sum, both the impoliteness and incivility models performed better when they were trained on balanced data. The models mainly based their predictions on unigrams, which increased their recall but decreased their precision. In other words, the models correctly retrieved many uncivil and impolite comments, but, at the same time, many comments that were classified as uncivil/impolite were actually civil/polite. Including only dictionary-based features increased the precision of the impoliteness model, but, at the same time, decreased its recall. For incivility, the DICT feature did not significantly improve the model performance, but the NE features helped to increase its precision. Overall, these results show that the information to which extent a comment includes offensive words only improves the precision of classifying impolite comments.

Additional Model Validation and Application

The model evaluation process in SML already provides important information regarding the extent to which a model can be transferred to other data: All evaluation measures describe the performance of the model on unseen data (test set) and the most predictive features provide some ideas about how good the predictions will be on new data that includes, for example, comments posted to different topics).

Still, few studies have actually tested classifiers that have been trained on a specific data set on new data sets. To overcome this gap, we applied the best-performing models for predicting impoliteness and incivility to a new data set of 3,500 user comments that had been posted to the Facebook sites of different news outlets in 2018.[34] The comments had also been hand-coded regarding their civility and politeness by six student assistants. As our classifiers tend to label many comments as uncivil/impolite that are actually civil/polite, we adjusted the probability threshold of the classifiers to predict a positive category from .50 to .75. This forces the model to be more 'certain' about its prediction for uncivil/impolite comments. At the same time, we can achieve better results on the new data, since the models predict civility/politeness in the case of uncertainty. As most comments of the new data set were also civil/polite, this adjustment led to better model results. The accuracy of the models on the new data set was 0.84 for incivility/civility (F1 = 0.85, Recall = 0.85, Precision = 0.86). For impoliteness/politeness, the accuracy was 0.67 (F1 = 0.66, Recall= 0.67, Precision= 0.65)[35].

Researchers and (media) experts are also often interested in using classified data to answer specific research questions. To investigate whether the answers to such research questions depend on whether we use machine-coded or hand-coded data, we considered the new data set that included both the hand-coded and the machine-coded measures of impoliteness and incivility. Using these measures, we investigated whether impolite and uncivil user comments receive fewer or more 'Likes' than polite and civil comments. Investigating this question is important, because many algorithms used in comment systems display comments more prominently that received a high number of likes (or recommendations or up-votes, respectively). Previous research found that impolite comments received fewer recommendations than polite comments (Muddiman & Stroud, 2017). Uncivil comments, in contrast, received significantly more recommendations (Muddiman & Stroud, 2017) or upvotes (Coe et al., 2014) than civil comments.

Using the new data set, we computed two generalized linear models. The number of Likes a comment received was entered as the dependent variable. This variable had non-negative responses (a range from 0 to 977), a high initial peak, a rapid drop, and a long right tail and, therefore, was estimated using a negative-binomial distribution (Dunteman & Ho, 2006). Regarding the independent variables, we added the hand-coded impoliteness and incivility measures (model 1) or the machine-coded measures (model 2) to the models. Additionally, we entered the different media outlets and news articles to which the comments were posted as control variables. Using these models, we were partly able to replicate the findings from previous research: In the model that included the hand-coded predictors (model 1), impolite comments did not receive more Likes than polite comments (B = 0.012, p = .802). Uncivil comments, in contrast, received more Likes than civil comments (B = 0.264, p < .001). In the model that included the machine-coded labels (model 2), impolite comments received fewer Likes than polite comments (B = -0.243, p < .001) and uncivil comments received more Likes than civil comments (B = 0.664, p < .001). Although we cannot answer the question which classification procedure yielded the 'true' results, the findings that are based on the automated classification are closer to the findings from previous research (Coe et al., 2014; Muddiman & Stroud, 2017).

A second test pertained to the question whether there are different shares of impoliteness and incivility on the sites of different news outlets. This question is important as well, because previous research has linked the different numbers of impolite/uncivil comments to the varying efforts of news organizations to moderate comment sections (e.g., Su et al., 2018). Using the hand-coded measure of impoliteness, the share of impolite comments on the ten sites we examined ranged from 14 to 29 percent. The machine-coded impoliteness measure detected a roughly comparable range of the share of impolite comments (15 to 25 percent). Similar results were obtained when comparing the hand-coded and machine-coded incivility measures. In sum, these findings suggest that, to some extent, the classifiers reveal comparable patterns and correlations when applied to specific research questions.

Discussion

Social sciences are increasingly interested in the automated detection and analysis of text documents (e.g., Mahrt & Scharkow, 2013; Muddiman

& Stroud, 2017; Ross et al., 2017). One promising approach is text classification using supervised machine learning techniques. These techniques train predictive models to estimate document categories based on language patterns. The current study built and compared classification models to predict impoliteness and incivility in user comments on German news media outlets on Facebook. Such models can support media professionals in managing user comment sections. They also provide insights into communication behaviour in online discussions and into the potentials and limitations of the automated detection of different concepts related to incivility. Following previous research, we differentiated between impoliteness and incivility in user comments: impoliteness violates interpersonal norms but is not necessarily detrimental on a societal level, whereas incivility is considered more harmful but often does not include the use of abusive words. To predict both categories on the word level of a comment, we tested different feature sets including single words (unigrams), word combinations (bigrams and trigrams), and dictionary-based features. We also applied the classifiers to a new but comparable data set of 3,500 (hand-coded) user comments to examine the transferability of the models. Finally, we used the classifiers to answer two research questions from social sciences. Finally, we compared the results with the results obtained from answering these questions using hand-coded data.

Overall, our findings suggest that impoliteness and incivility in user comments can be measured to certain extent based on the words that appear in a comment. Nevertheless, even the best-performing models showed a high misclassification rate, although they were trained on a data basis of several thousand manually-coded instances. For impoliteness, the best model performance was achieved by a NB system using a combination of Bag-of-Words (BoW) unigrams, bigrams, and trigrams, and dictionary-based features that quantify the amount of insults, polarity, and positive, respectively negative sentiment in a comment. For incivility, which was expected to occur more subtly than impoliteness, the best-performing model used a combination of Bag-of-Words (BoW) unigrams, bigrams, and trigrams, dictionary-based features, and a named entity (NE) feature, which represents the distribution of person, location, and organization entities in a comment. A closer look at the predictive features revealed that the incivility model used both ambiguous words for classification and words that are closely related to a discussion topic or a debate, such as ‘merkel’ or ‘griechen’ (‘Greeks’). This aligns with the results of Davidson et al. (2017) and Kwok and Wang (2013) who reported that words that are related to the subject of an uncivil discussion are often used as predictive features.

Furthermore, our results show that the dictionary-based features, which represent the (combined) appearance of insults, anger, and swear words, and the polarity of a statement, led to more precise predictions of impoliteness but not of incivility in user comments. These results are consistent with our theoretical rationale and with the related work on the automated detection of harmful language stating that impoliteness often includes the use of obvious insults and pejorative language while truly uncivil communication often is more difficult to detect. Nevertheless, even for the impoliteness classifier, the predictive power of these lexical resources was quite low, although one would expect that some of these resources, such as the list of insults, would perform strongly in predicting impoliteness. This is probably because many of the included words are not always used in a negative, insulting, or harmful way. For example, comment writers talk about the impolite behaviour of other users and cite passages of these impolite comments, such as the following one: 'Hat hier überhaupt irgendjemand derjenigen, die schon den Artikel absolut lächerlich finden, ihn überhaupt gelesen? Wirkt nicht so. (...) Der Artikel ist jedenfalls alles andere als dumm und unwichtig.'[36] The comment includes derogatory words (such as 'stupid') but the coders did not label it as impolite because the author of the comment tries to make other commenters think about how they expressed themselves. Such ambiguous comments make it more difficult for automated classifiers to identify the relative importance of predictive features such as insults.

Our results also revealed that including information about named entities in a comment helps to predict incivility more precisely, even if the improvement is small. This is possibly because public figures ('merkel') or governmental organizations are often targets of incivility in online discussion of news media outlets. Unfortunately, established NER systems for the German language do not provide the same possibilities to predict different entity types as the systems for the English language (e.g., ethnic and religious groups). Including this information might have further improved the performance of the incivility classifier.

However, all models also used several ambiguous 'neutral' words as predictive features, which apparently are no reasonable indicators for any form of harmful content, such as 'kommen', 'warum', 'endlich'[37]. Such words probably have led to inaccurate predictions. Since user-generated content on many social media platforms has already been filtered for the use of offensive language to some amount, it is possible that obvious forms of impoliteness and incivility is underrepresented in the training data. Possible options to solve this issue include training the models on a sample of comments that include both filtered and unfiltered comments or weighing the instances of the underrepresented class in the training data.

Relying on a relatively large database of over 10,000 hand-coded user comments, we also hoped to reveal features that are indicative for 'subtle' forms of incivility, such as combinations of words. A descriptive post analysis of the data showed that the comments indeed contained subtle and probably better n-gram indicators for incivility, such as 'an wand stellen', 'raus euro schnell', or 'verhungern lass'[38]. A reason why our models still preferred unigram features may be that the given sample is still too heterogeneous to identify such complex expressions as discriminative features. In other words, some complex expressions did not appear frequently enough to be considered as discriminative features for incivility/impoliteness.

To additionally test our models, we applied them to a new but comparable data set of hand-coded user comments. Certainly, the observed misclassification patterns appeared again in the new data set. Since we knew about the tendency of the models to classify polite/civil comments as impolite/uncivil, we could adjust the classification boundary afterwards, which led to overall satisfactory results on the new data set. Nevertheless, the models will not identify new and unknown forms of incivility/impoliteness that have not been learned during training. Keeping this in mind, the results show that classifiers can be applied to new data, but with certain limitations that can already be derived from model evaluation process. Therefore, it is no surprise that manual and machine-based labelling sometimes leads to different results.

Limitations, Implications, and Conclusion

Differentiating between impolite and uncivil comments can help news media professionals evaluate different levels of inappropriate or harmful communication behaviour and respond to these different comments appropriately. The findings of the current study suggest that it remains challenging to automatically detect abstract and elusive concepts such as impoliteness and incivility in text data. Overall, there is no simple rule for labelling a text document as uncivil or impolite with both high reliability and external validity. This, of course, applies both to ML systems and to the manual coding process, since even human coders often rate the same comments differently. One main reason for this is that civility and politeness are subjective concepts to some extent (Herbst, 2010). Some words are common triggers of impoliteness or incivility, but others only work in a specific context, and some statements are only labelled as uncivil or impolite by 'insiders' or by coders with specific knowledge. Oftentimes, it is hardly generalizable which indicators or triggers lead human coders to label a comment as civil or uncivil. However, there are several approaches to address the problem, which come with different advantages and disadvantages.

First, it is possible to teach the human coders to assign categories according to rules that a statistical model will reveal more easily, for example, by only taking the appearance of unambiguous insult words into account. This procedure, however, possibly decreases the validity of the measure in favour of its reliability. Therefore, many instances of incivility would remain undiscovered. These should be kept in mind when compiling training data for SML. Second, variance in the data could be reduced by statistically controlling for the topics of discussions, for example (Blei, Ng, & Jordan, 2003), or for the media outlets that host the discussions (Sue et al., 2017). In our data, we had sampled several media outlets, topics, and times of the discussions. This might have increased the variance in the training data to a level that is inappropriate for a text-based SML model. Third, recent approaches to automated NLP achieved promising results in detecting abstractive concepts from texts using more complex classification models, namely *Deep Neural Networks* (*NNs*). But these models require very large amounts of labeled data and would overfit our sample of a few thousands of user comments quickly (Pennington, Socher, & Manning, 2014). Another way to improve the model performances integrating recent NN methods is to use word vectors (word embeddings) instead of BoW to represent words in a statement (Mikolov, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 2014). For the German language, such models are currently trained on *Common Crawl* and *Wikipedia* (Bojanowski, Grave, Joulin, & Mikolov, 2016). This approach could improve the model performance on small samples given that the corpora the word embeddings are trained on are comparable to the data set analysed.

Despite the limitations of the current study, our findings provide new insights into when and why the SML approach can be applied to the automated detection of impoliteness and incivility in German user comment sections. We hope that this study can help other researchers apply and improve SML methods for comparable problems.

Notes

[1] 'Fuck, need to sell my car at once!!!'

[2] 'Schmidt is a politician who still has BALLS!'

[3] 'Greedy human Martin Schulz leads the way.'

[4] 'Muslims also have a high birth rate.... that compensates for it...'

[5] 'We have to fight back, folks, take your truncheons to the streets and show that we've had enough.'

[6] 'It is perhaps a prejudice, but: just stolen and the car is already in Poland.'

[7] On Facebook, top-level comments are postings that appear on the first 'level' of a discussion. Users can reply to top-level comments and these replies are then displayed in chronological order under each top-level comment.

[8] These categories were similar to the categories used by Coe et al. (2014) – however, these authors labeled the use of these instances in comments as examples of incivility.

[9] For preprocessing and vectorization, we used the Python libraries *NLTK* (https://www.nltk.org/), *scikit-learn* (https://scikit-learn.org/stable/) and SpaCy (https://spacy.io/).

[10] Documentation on http://snowballstem.org/.

[11] We used the lemmatizer for German, implemented in SpaCy.

[12] We used the POS tagger for German, trained on the TIGER and WikiNER corpus (implemented in SpaCy).

[13] Retrieved from the NLTK Stopwords Corpus for German.

[14] http://www.insult.wiki/wiki/Schimpfwort-Liste

[15] 'idiot', 'douchebag'.

[16] 'bitch', 'whore', 'faggot', 'tranny'.

[17] 'bread', 'bureaucrat', 'currywurst', 'Dennis'.

[18] https://liwc.wpengine.com/

[19] Version:1.1 2010/08/01; http://bics.sentimental.li/files/8614/2462/8150/german.lex (For Python: https://pypi.org/project/textblob-de/).

[20] https://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html

[21] Documentation: https://spacy.io/models/de

[22] NE Models trained in OntoNotes 5 (https://catalog.ldc.upenn.edu/LDC2013T19), implemented in SpaCy (https://spacy.io/api/annotation#named-entities).

[23] Our models are built using the class *MultinomialNB* from the Python library scikit-learn.

[24] For parameter optimization we applied the sklearn class *GridSearchCV*(*cv=5*).

[25] BOW(1-3) mean feature of unigram, bigrams and trigrams. BOW(1) refers to only unigrams, BOW(1-2) to unigrams and bigrams, etc.

[26] Multinomial Naïve Bayes-Classifier: *class* sklearn.naive_bayes.MultinomialNB (*alpha=1.0*, *fit_prior=True*, *class_prior=None*).

[27] 'germany', 'people', 'euro', 'greece', 'money', 'world', 'merkel', 'country', 'people', 'guilt', 'politician', 'government'.

[28] 'times', 'do', 'will', 'more', 'see', 'can'.

[29] 'good', 'give', 'finally', 'sweet', 'respect'.

[30] 'always', 'time', 'a lot'.

[31] 'stupid', 'shit', 'bullshit', 'people', 'greeks', 'merkel'

[32] 'do', 'give', 'just', 'come', 'go'

[33] 'finally', 'people', 'why', 'always', 'children', 'make'

[34] While both the 'new' and the 'old' data set were collected on Facebook, the specific news outlets examined and (potentially) the topics of the articles examined differed between the two data sets.

[35] Here, the evaluation metrics report overall prediction results. Macro *F*1 is weighted on class distribution. The overall agreement between the manually-assigned labels and the prediction the classifiers made is 84.8 percent for incivility. For impoliteness, it is 67.4 percent.

[36] 'Has anyone here who finds the article absolutely ridiculous read it at all? It doesn't seem so. (...) The article is anything but stupid and unimportant.'

[37] 'come', 'why', 'finally'.

[38] 'put on the wall', 'get out euro fast', or 'starve to death'.

References

Alonzo, M., & Aiken, M. (2004). Flaming in electronic communication. *Decision Support Systems*, *36*(3), 205-213.

Anderson, A.A., Brossard, D., Scheufele, D.A., Xenos, M.A., & Ladwig, P. (2014). The nasty effect: Online incivility and risk perceptions of emerging technologies. *Journal of Computer-Mediated Communication*, *19*(3), 373-387.

Agarwal, S., & Sureka, A. (2014). A focused crawler for mining hate and extremism promoting videos on YouTube. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pp. 294-296.

Aggarwal, C.C., & Zhai, C. (2012). A survey of text classification algorithms. In *Mining text data* (pp. 163-222). Boston: Springer.

Baayen, R.H. (2002). *Word frequency distributions*. Berlin: Springer Science & Business Media.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. London: O'Reilly Media.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *arXiv:1607.04606*.

Burnap, P., & Williams, M.L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223-242.

Coe, K., Kenski, K., & Rains, S.A. (2014). Online and uncivil? Patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, *64*(4), 658–679. doi:10.1111/jcom.12104

Coelho, L.P., & Richert, W. (2015). *Building machine learning systems with Python*. Birmingham: Packt Publishing Ltd.

Daniels, J. (2009). Cloaked websites: propaganda, cyber-racism and epistemology in the digital era. *New Media & Society*, *11*(5), 659-683.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv:1703.04009*.

Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. In *Data Engineering Workshop, 2008. ICDEW 2008*, pp. 507-512.

Dunteman, G.H., & Ho, M.-H.R. (2006). *An introduction to generalized linear models*. Thousand Oaks, CA: Sage.

Garreta, R., & Moncecchi, G. (2013). *Learning scikit-learn: machine learning in python*. Birmingham: Packt Publishing Ltd.

Gitari, N.D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215-230.

Grimmer, J., & Stewart, B.M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21*(3), 267-297.

Grice, P. (1989) *Studies in the Way of Words*. Cambridge: Harvard University Press.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Herbst, S. (2010). *Rude democracy: Civility and incivility in American politics*. Philadelphia: Temple University Press.

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave your comment below": Can biased online comments influence our own prejudicial attitudes and behaviors? *Human Communication Research*, *41*(4), 557–576. doi:10.1111/hcre.12059

Hsueh, P.Y., Melville, P., & Sindhwani, V. (2009, June). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009, workshop on active learning for natural language processing*, pp. 27-35.

Jurafsky, D., & Martin, J.H. (2014). *Speech and language processing* (Vol. 3). London: Pearson.

Kalch, A., & Naab, T.K. (2018). Replying, disliking, flagging: How users engage with uncivil and impolite comments on news sites. *SCM Studies in Communication and Media*, *6*(4), 395–419.

Kotsiantis, S.B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, *160*, 3-24.

Kwok, I., & Wang, Y. (2013, July). Locate the Hate: Detecting Tweets against Blacks. In *Proceedings of the 27th National Conference on Artificial Intelligence* (*AAAI*), pp. 1621-1624.

Larsson, A.O. (2018). Assessing "The Regulars"—and Beyond: A study of comments on Norwegian and Swedish newspaper Facebook pages. *Journalism Practice*, *12*(5), 605-623.

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20-33.

Mandelbrot, B. (1961). On the theory of word frequencies and on related Markovian models of discourse. *Structure of language and its mathematical aspects*, *12*, 190-219.

Manning, C.D., Schütze, H., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press.

Manning, C.D., & Schütze, H. (2000). *Foundations of statistical natural language processing* (3. print.). Cambridge: MIT Press.

McCallum, A., & Nigam, K. (1998, July). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, No. 1, pp. 41-48).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*.

Muddiman, A., & Stroud, N.J. (2017). News values, cognitive biases, and partisan incivility in comment sections. *Journal of Communication*, *67*(4), 586–609. doi:10.1111/jcom.12312

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1-135.

Papacharissi, Z. (2004). Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media & Society*, *6*(6), 259–283. doi:10.1177/1461444804041444

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*, 2825-2830.

Pendar, N. (2007). Toward spotting the pedophile telling victim from predator in text chats. In *Semantic Computing, 2007* (*ICSC 2007*), pp. 235-241.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (*EMNLP*), pp. 1532-1543.

Prochazka, F., Weber, P., & Schweiger, W. (2018). Effects of civility and reasoning in user comments on perceived journalistic quality. *Journalism Studies*, *19*(1), 62–78.

Raschka, S., & Mirjalili, V. (2017). *Python machine learning*. Packt Publishing Ltd.

Remus, R., Quasthoff, U., & Heyer, G. (2010). SentiWS – a public available german-language resource for sentiment analysis. In *Proceedings of the 7th International Language Resources and Evaluation* (*LREC'10*), pp. 1168-1171.

Robert, C. (2014). Machine Learning, a Probabilistic Perspective. *CHANCE*, *27*(2), 62-63.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., & Wojatzki, M. (2017). Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Rowe, I. (2015). Civility 2.0: A comparative analysis of incivility in online political discussion. *Information, Communication & Society*, *18*(2), 121-138.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys* (*CSUR*), *34*(1), 1-47.

Silva, L.A., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016, March). Analyzing the Targets of Hate in Online Social Media. In *ICWSM* (pp. 687-690).

Su, L.Y.F., Xenos, M.A., Rose, K.M., Wirz, C., Scheufele, D.A., & Brossard, D. (2018). Uncivil and personal? Comparing patterns of incivility in comments on the Facebook pages of news outlets. *New Media & Society*, 1461444818757205.

Tausczik, Y.R., & Pennebaker, J.W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Thelwall, M., Buckley, K., & Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, *62*(2), 406-418.

Thelwall, M., Wilkinson, D., & Uppal, S. (2010). Data mining emotion in social network communication: Gender differences in MySpace. *Journal of the American Society for Information Science and Technology*, *61*(1), 190-199.

Wiegand, M., Siegel, M., & Ruppenhofer, J. (2018). Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing* (*KONVENS 2018*) (pp. 1-10). Vienna, Austria.

Witten, I.H., Frank, E., Hall, M.A., & Pal, C.J. (2016). *Data Mining: Practical machine learning tools and techniques*. Burlington: Morgan Kaufmann.

Ziegele, M., Breiner, T., & Quiring, O. (2014). What Creates Interactivity in Online News Discussions?. *Journal of Communication*, *64*, 1111–1138. doi: 10.1111/jcom.12123

Zipf, G.K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology*, *33*(2), 251-256.

About the author

Anke Stoll and **Marc Ziegele** work at the Heinrich-Heine University Düsseldorf, Department of Communication and Media Studies.
Oliver Quiring works at the University Mainz, Institut für Publizistik.