Content analyses of user comments in journalism: a systematic literature review spanning communication studies and computer science

Julius Reimer, Marlo Häring, Wiebke Loosen, Walid Maalej, Lisa Merten

Abstract

Different disciplines have studied the content of online user comments in various contexts, using manual qualitative/quantitative or (semi-)automated approaches. The broad spectrum and disciplinary divides make it difficult to grasp an overview of those aspects which have already been examined, e.g. to identify findings related to one’s own research, recommendable methodological approaches, and under-researched topics. We introduce a systematic literature review concerning content analyses of user comments in a journalistic context. Our review covered 192 papers identified through a systematic search focusing on communication studies and computer science. We found that research predominantly concentrates on the comment sections of Anglo-American newspaper brands and on aspects like hate speech, general incivility, or users’ opinions on specific issues, while disregarding media from other parts of the world, comments in social media, propaganda, and constructive comments. We derive a research agenda that also highlights potentials for automating analyses and for cooperation across disciplines.

Keywords

User comments, journalism, qualitative content analysis, quantitative content analysis, automated content analysis, systematic literature review, computer science, interdisciplinary research

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: this work is partly funded by the Hamburg programme Ahoi Digital (project ‘Forum 4.0’).
Content analyses of user comments in journalism: a systematic literature review spanning communication studies and computer science

Abstract

Different disciplines have studied the content of online user comments in various contexts, using manual qualitative/quantitative or (semi-)automated approaches. The broad spectrum and disciplinary divides make it difficult to grasp an overview of those aspects which have already been examined, e.g. to identify findings related to one’s own research, recommendable methodological approaches, and under-researched topics. We introduce a systematic literature review concerning content analyses of user comments in a journalistic context. Our review covered 192 papers identified through a systematic search focusing on communication studies and computer science. We found that research predominantly concentrates on the comment sections of Anglo-American newspaper brands and on aspects like hate speech, general incivility, or users’ opinions on specific issues, while disregarding media from other parts of the world, comments in social media, propaganda, and constructive comments. We derive a research agenda that also highlights potentials for automating analyses and for cooperation across disciplines.

Keywords

User comments, journalism, qualitative content analysis, quantitative content analysis, automated content analysis, systematic literature review, computer science, interdisciplinary research

Introduction

In recent years commenting on journalistic stories has become one of the most prevalent participation options that news media offer their users on their websites and social media channels (Stroud et al. 2016b). Not only are comment spaces increasingly widespread, so is the audience’s use of them, making the online discussion of news a popular way for people to engage with socio-political issues (Heise et al. 2014a, 2014b; Loosen et al. 2013a, 2013b; Reich 2011; Reimer et al. 2015; [154: Ruiz et al. 2011].
in brackets indicate the studies analysed, which are listed in Appendix 2). The sheer volume of comments flooding into a newsroom every day can be overwhelming (Braun and Gillespie 2011), especially when a large share is uncivil or ‘toxic’ in tone and content (Chen and Pain 2017; Loke 2012). Unsurprisingly, a growing number of media organisations decided to shut down comment sections on their own websites and instead refer their audience to their accounts on social media such as Facebook or Twitter, where commenting is still possible (cf. the list compiled by Cherubini 2016). In general, the experiences of incivility, hate speech, and propaganda in comments appear to have disillusioned many journalists to the extent that they no longer think of comments as furthering public debate. On the contrary, journalists tend to regard them as a threat to deliberation ([95: Ksiazek et al. 2015]), or at best, “a necessary evil […] to attract audiences” (Reich 2011: 103).

The relevance of user comments for journalism and public debate has invariably led to a wave of related research. Scholars have, for instance, investigated how journalists perceive comments and commentators (e.g., Heise et al. 2014a; Loke 2012), how media organisations (should) moderate discussions on their websites (e.g., Ziegele and Jost 2016), how comments influence users’ assessments of a news item or a media brand (e.g., Houston et al. 2011; Kümpel and Springer 2016), or what motivates audience members to voice their opinion and engage in online news discussions (e.g., Heise et al. 2014a; Springer et al. 2015).

Researchers have also looked at the comments themselves: one strand of literature leverages user comments as a proxy for measuring public opinion on various issues (Naab and Sehl 2017) from climate change ([90: Koteyko et al. 2013]) and the financial crisis ([10: Baden and Springer 2014]) to breastfeeding ([72: Grant 2016]). Another strand of studies look at user comments for their own sake, for instance, by analysing their inherent architecture and deliberative potential ([55: Freelon 2015]; [154: Ruiz et al. 2011]) or by focusing on (in-)civility in user comments ([27: Coe et al. 2014]). Moreover, comment analyses can differ in terms of the methods applied: comments can be analysed quantitatively or qualitatively in a manual process (Krippendorff 2013; Kuckartz 2014) as well as quantitatively in a (semi-)automated way (Boumans and Trilling 2016; Lewis et al. 2013).

Additionally, user comments are not only an object of investigation in the social sciences, but also in computer science and computer linguistics where they are used as a data source for gathering information on
product requirements, innovative new features or issues (Maalej et al. 2016) and as empirical data for the development of new analytical methods (e.g., [39: de-la-Peña-Sordo et al. 2015]; [53: Foster et al. 2011]).

Due to this broad thematic, methodological, and disciplinary spectrum of user comment analyses with reference to journalism, it is difficult to obtain a comprehensive overview of which specific aspects have been analysed and to identify under-researched topics or the limitations present in current work. This however appears desirable, if not necessary, in order to facilitate the advancement of knowledge in this area. This in turn is important as in the long run it can help journalists manage the flood of comments, identify, make use of, and respond to valuable and insightful statements among them, as well as generally drive back incivility, hate speech, and propaganda, and further constructive public debate.

**Aim and structure of the study**

Against this background, this study aims to build a bridge between the disciplines of communication studies and computer science and their respective stocks of content analyses of user comments referring to journalism. We focus on these two disciplines not only because they are the authors’ fields of expertise and are both particularly interested in user comments. More importantly, these two disciplines can be fruitfully brought together in this field of research: in communication studies, in which content analysis has been fundamentally developed as an empirical method, comments are often examined to understand their very content. Computer science, on the other hand, focuses more on the development of automated approaches and tools for analysing large volumes of comments. This is particularly relevant given the above-mentioned challenges which the sheer volume of comments as well as the incivility, hate speech, and propaganda found in them pose to the journalistic profession and public discourse in general. We focus on user comments that occur in a journalistic context because it is a domain of particular societal relevance and central to public debates. It is also a traditionally prominent subject in communication research and—with the emergence of online media, social networks, and computational journalism—is becoming increasingly relevant for computer scientists as well (e.g., [67: Gonçalves et al. 2016]; [86: Kabadjov et al. 2015]; [166: Sood and Churchill 2010]; [189: Zhou et al. 2014]). Additionally, journalism’s topical breadth makes user comments in this domain particularly interesting as well as challenging when it comes to the development of automated approaches, as opposed to more homogeneous comments, for example, those found in app stores or on travel websites. Finally, we had to limit our selection to studies using the same empirical method. First, including
studies with different methodological approaches would have increased the number of studies to analyse beyond a manageable amount. Second, it would have made coding too complex since different methods are characterised by different parameters. Their samples, for example, consist of different kinds of cases: comments, users, journalists, newsrooms. Not only would we have ended up with a lot of incomparable data, but also would a codebook that accounts for all this have been even more extensive than the one we used (see Appendix 3).

We compare the two disciplines in regard to their thematic and methodological foci to reveal gaps in current research and to establish the grounds for mutual inspiration and interdisciplinary cooperation that can advance the field of study as a whole. In particular, we aim to:

- provide an overview of the two disciplines’ bodies of research with respect to the different features scrutinised by user comment analyses and provide scholars with a well-informed starting point for their own studies;
- highlight qualitative work that can be used for developing quantitative, possibly automated, methods to investigate particular aspects of user comments;
- identify the potential for and limitations of automation of comment analyses, especially in communication studies, by highlighting particularly promising approaches that make use of machine learning;
- identify gaps in current research and under-investigated features of comments; and
- integrate these insights in an agenda for joint future research between communication scholars and computer science researchers.

To this end, we conducted a systematic literature review of content analyses investigating user comments in online journalism and identified what variables or constructs (e.g., opinions/sentiment, incivility, information that comments add to the original article) have, so far, been analysed in what way (qualitatively/quantitatively; manually/(semi-)automatically) and from which disciplinary perspective.

In the following section we first elaborate on the context of our research, describing the development, use, and management of user comments in journalism, placing a particular focus on findings from survey-, interview-, and observation-based studies, which we will use in the conclusion to complement and contextualise the insights gained from our systematic review of content analyses. After that, we explain the methodology of our review, including our codebook, the systematic search for studies to be analysed, and their coding. We continue with the presentation of our results, placing a special focus on the comparison
between the two disciplines. Finally, we conclude the paper with a discussion of the identified research gaps and the potentials for collaboration on user comments in order to develop the aspired agenda for joint research on user comments by computer science researchers and communication scholars.

**User comments in journalism**

In recent years, user comments in journalism have developed from being a mere sub-aspect of participatory journalism into a distinct field of research of its own, reflecting the “tension between professional control and open participation” (Lewis 2012) which journalism still struggles with: the early years of online journalism saw a growing number of news websites providing features for users to publicly comment on news stories, e.g. to not fall behind competitors who already offer comment sections, to strengthen users’ loyalty, or to obtain information about them and their preferences (Heise et al. 2014a, 2014b; Karlsson et al. 2015; Loosen et al. 2013a, 2013b; Reimer et al. 2015; Singer et al. 2011). But for a few years now, an increasing number of media outlets have shut down their comment sections, referring users who want to comment to their social media accounts, often because they fear the questionable tone of comments—the incivility or toxicity—may damage their brand and/or they lack the resources necessary to monitor and moderate the sometimes overwhelming amount of comments flooding into newsrooms every day (Bergström and Wadbring 2015; Cherubini 2016; Karlsson et al. 2015; Loke 2012). Other news organisations only delete hate speech, rude language, spam, propaganda, and other incidents of “dark participation” (Frischlich et al. 2019) as well as off-topic statements and do not pay too much attention to the rest of the comments that may be insightful ([167: Sood et al. 2012]). Correspondingly, journalists’ attitudes towards comments appear to have changed: while at first, they were viewed as “a space for a new ‘deliberative democratic potential’” ([31: Collins and Nerlich 2015]), now they are often considered as a rather shallow and aggressive form of audience participation (Bergström and Wadbring 2015; Loke 2012), a threat to deliberation ([95: Ksiazek et al. 2015]), or at best, “a necessary evil […] to attract audiences” (Reich 2011: 103).

Studies suggest that between a quarter and half of users have commented on a news story at least once, but that only a small share of them does so regularly or frequently (Heise et al. 2014b; Karlsson et al. 2015; Loosen et al. 2013a, 2013b; Purcell et al. 2010; Reimer et al. 2015; [154: Ruiz et al. 2011]; Springer et al. 2015; Ziegele et al. 2013). However, comments tend to be widely appreciated as the percentage of
‘lurkers’, i.e. users who don’t comment themselves, but read others’ comments, is usually much larger (Heise et al. 2014b; Karlsson et al. 2015; Loosen et al. 2013a, 2013b; Reimer et al. 2015). Lurkers engage with comments to gather more information on a particular story—e.g. to sense the climate of opinion—and for entertainment (Springer et al. 2015; [45: Diakopoulos and Naaman 2011]). The motives for active commenting, on the other hand, are more manifold: studies found them to relate to public debate and information (express an opinion, bring topics deemed important onto the agenda, correct errors and perceived misrepresentations in reporting, learn from dialogue with others), to identity and emotion management (present oneself, vent anger, spend time) as well as to a feeling of belonging through interaction with other users or journalists, whose participation in comment sections was also found to be a motivating factor (Aschwanden 2016; [45: Diakopoulos and Naaman 2011], Heise et al. 2014a, 2014b; Loosen et al. 2013a, 2013b; Meyer and Carey 2014; Reimer et al. 2015; Springer et al. 2015). By contrast, incivility and a lack of deliberative quality have been found to keep users from participating in comment sections (Engelke 2019; Weber 2014).

Possibly related, studies have found active users and their comments not to be representative of the general population and public opinion (Friemel and Dötsch 2015; Hölig 2018): commenting users, for example, tend to be overwhelmingly male as well as rather extroverted and narcissistic.

The amount of comments a news story receives seems to be independent of general audience interest, but depend on its topic and related news factors as most-clicked stories are not necessarily those with the most comments and commenters were found to prefer political content (Tenenboim and Cohen 2015; Ziegele et al. 2013) as well as stories characterised by proximity, impact, and frequency (Weber 2014).

While surveys suggest that discussions are an important objective for commenters (Heise et al. 2014a, 2014b; Loosen et al. 2013a, 2013b; Reimer et al. 2015; Springer et al. 2015; Ziegele et al. 2013), content analyses show that comments often lack the necessary interaction between users ([101: Len-Ríos et al. 2014]; [153: Rowe 2015]; Taddicken and Bund 2010). Instead, Ruiz et al. ([154: 2011: 8]) describe them as a “dialogue of the deaf”, finding that comments mostly contain users’ personal opinions in the form of a reaction to the article and not so much to other comments. This is problematic because, as Weber (2014: 942) states, “the potentials for quality discourse emerge only […] when there is a certain degree of interactivity among the users’ comments”. However, some studies found that, at least in moderated comment sections, substantial, albeit smaller, portions of comments refer to other users’ statements, provide alternative
perspectives or information on the commented story’s topic, give arguments for stated opinions, contain humour or point towards an error or misrepresentation in the story (Aschwanden 2016; Heise et al. 2014b; Loosen et al. 2013a, 2013b; Reimer et al. 2015; Witschge 2011).

One aspect related to the content of comments that is researched particularly often is their (in)civility. Coe et al. ([27: 2014]), for instance, found that the share of uncivil comments does not necessarily increase when the amount of incoming comments does. Instead, it increases when ‘weightier’ topics are discussed. Correspondingly, studies found that the topics more strongly related to uncivil discourse include politics, society, crime and justice, disasters and accidents, the environment, and feminism (Diakopoulos and Naaman 2011; [58: Gardiner et al. 2016]). Moreover, Coe et al. ([27: 2014]) show that article-inherent aspects, such as the sources cited, influence the level of civility in comments. Various studies also suggest that anonymity has a negative influence on civility and other indicators for comment quality ([54: Fredheim et al. 2015]; [158: Santana 2014]).

Additionally, the article author’s gender is of importance as female journalists received more comments that had to be blocked than their male colleagues ([58: Gardiner et al. 2016]). In this regard, a whole strand of research shows how much journalists—especially those of female or non-binary gender, of colour, and/or with an immigrant background—suffer from and cope with hateful comments addressed to them, which is particularly important because it has been shown that the anticipation of hateful audience feedback can impair their work and open hostility towards journalists can negatively affect the general public’s sentiment towards the profession as a whole (Obermaier et al. 2018; Post and Kepplinger 2019).

The quality and civility of comments appear to differ depending on the platform they are posted on: in a case study of the Washington Post, Rowe ([152: 2015]) found that comments on the newspaper’s website show more deliberative quality than those on the outlet’s Facebook page. This is consistent with the findings from our own previous studies in which we found that comments from three German media’s Facebook pages contain less interaction with other comments, arguments to back stated opinions, or alternative perspectives on the discussed topic, but more personal attacks than comments to the same articles on the media’s websites (Heise et al. 2014b; Loosen et al. 2013b; Reimer et al. 2015). Furthermore, in surveys the outlets’ users found the comments on the media’s Facebook pages were not as good a supplement to the commented article, less informative and, generally, of lower quality than those on their websites (Heise et al. 2014b; Loosen et al. 2013b; Reimer et al. 2015).
Because of their overwhelming volume and their often-toxic tone, monitoring and moderating comments requires considerable resources (Diakopoulos and Naaman 2011; [71: Graham and Wright 2015]; Heise et al. 2014a, 2014b; Loosen et al. 2013a, 2013b; Reimer et al. 2015). Yet, a survey by Stroud et al. (2016a) showed that most comment sections are moderated, and nearly all media organisations respond in one way or another to their audience via the comment section or on social media. While these tasks are often fulfilled by specific newsroom personnel such as social media editors, studies found that a considerable share of ‘ordinary’ journalists also respond to commenters at least occasionally or even engage in discussions in comment sections (Chen and Pain 2017; Heise et al. 2014b; Loosen et al. 2013a, 2013b; Reimer et al. 2015; Stroud et al. 2016a). In any case, the high workload and difficulty in identifying ‘good’ or insightful comments within the vast amounts of user statements keep journalists from using the information in comments for their work and from actually engaging with their audience (Braun and Gillespie 2011; [71: Graham and Wright 2015]; Heise et al. 2014a, 2014b; Loosen et al. 2013a, 2013b; Reich 2011; Reimer et al. 2015). The situation is further complicated by the multiplication of media channels—such as websites, social media, blogs—which journalists are active on and receive feedback through (Heise et al. 2014a; Kramp and Loosen 2018; Loosen et al. 2017; Neuberger et al. 2019). This is due to the different subgroups of the audience and “commenting cultures” represented on these platforms.

Against this background of challenges that comments pose to journalism and public discourse in general, a considerable share of research has focused on automating analyses to help with the moderation of uncivil statements especially and, less frequently, to identify comments that are constructive or contain information potentially valuable to journalists or other users ([44: Diakopoulos 2015]; Loosen et al. 2017; Løvlie 2018).

In relation to uncivil comments, different moderation strategies have been found to have different effects: Ziegele and Jost (2016) show that factual responses raise other users’ willingness to participate in the discussion while sarcastic responses decrease the perceived credibility of the commented story and the news medium publishing it, but increase discussions’ entertainment value. Additionally, Stroud et al.’s ([171: 2015]) study suggests that if, instead of an unidentified staff member, a recognisable reporter responds to comments, the deliberative quality of comments increases. Ruiz et al. ([154: 2011: 482]) ascertain that a strict moderation style leads to a lower amount of comments, but also conclude that overall “the different
solutions adopted (pre-/post-moderation, in-house/outsourced) do not seem to direct the quality of the debate in a clear direction”.

**Method**

To achieve the study goals, we conducted a systematic (or structured) literature review (Greenhalgh et al. 2004). Essentially, this is “a form of content analysis whereby the unit of analysis is the article” (Massaro et al. 2015: n.p.). Such an approach appears particularly appropriate when there is a wide range of research on a subject (Petticrew and Roberts 2006). Here, systematic literature reviews give “an overview of the scope of existing research, the prevalence of the procedures used, and the identified problems” (Naab and Sehl 2017: 1257) as well as identify research gaps (Petticrew and Roberts 2006: 2). As such, systematic reviews are useful sources to consult “before embarking on any new piece of primary research” (Petticrew and Roberts 2006: 21; authors’ emphasis). The systematic procedure also reduces the risk “that seminal articles may be missed” (Massaro et al. 2015: n.p.) and “minimize[s] bias through exhaustive literature searches” (Tranfield et al. 2003: 209). This is of particular importance in areas such as user comment analyses where it is difficult to obtain a comprehensive overview of the large body of research in any other way. Consequently, the method stands and falls with how one searches for and selects the relevant literature and what aspects one analyses. We will explain both in the following.

**Literature selection and sample**

Based on our research aims, we defined three criteria for studies to be included in our sample: they have to be 1) based on a manual quantitative or qualitative, or a (semi-)automated analysis of the content of 2) user comments that 3) refer to journalistic stories.

We then created a search string that combined synonyms for each of the above-mentioned inclusion criteria (see Appendix 1 for the development and resulting search string). In December 2016, we used this search string to search the titles, keywords, and abstracts of all entries in arguably the most comprehensive literature database in communication science (EBSCO Communication and Mass Media Complete) as well as in its counterparts in computer science (ACM Digital Library and IEEE Xplore). In addition, we consulted the reading lists of ‘The Coral Project’ (2016; Diakopoulos 2016) and of our own research project ‘Systematic Content Analysis of User Comments for Journalists (SCAN4J)’\(^1\). Due to the advanced
internationalisation of both disciplines, we can assume that researchers of all nationalities (also) publish in English-language publications. To broaden our view and complement the sample with comment analyses performed in other disciplines, we also searched four extensive multidisciplinary repositories (Springer Link, ScienceDirect, Web of Science, Google Scholar; see Appendix 1 for the list of repositories and exact search procedures).

The search resulted in a list of 2,219 potentially relevant studies. Based on title, abstract and keywords, in cases of doubt on the full text, we identified the 192 studies which meet all of the three above-mentioned inclusion criteria (see Appendix 2 for more details on the inspection of the initial 2,219 search results and the final list of studies included).

**Codebook and coding procedure**

To review these studies, we developed a detailed codebook (see Table 1 for an overview and Appendix 3 for the complete codebook). At its core are categories to capture what comment-related properties the studies are concerned with. To develop these, we initially collected comment-related aspects based on the knowledge we had gained in previous research projects on audience participation in journalism and on comments mining in software engineering. Then we complemented this list based on a close examination of studies that we already knew met our three inclusion criteria and represent a variety of different perspectives on comment analyses: they refer to news media from a variety of countries (e.g., Bulgaria, Finland, Germany, Romania, the U.S., UK), focus on diverse news topics (e.g., politics, a plane crash, celebrities, data visualisations) as well as research interests (e.g., incivility, opinions, deliberation, media criticism, emotions, neo-Nazi propaganda) and come from different years (2008–2016) as well as disciplines (communication studies, computer science, linguistics, and political science). This was important to account for the diversity of perspectives on user comments as well as for potential shifts in interests over time. We stopped as soon as the three last studies examined did not investigate any new comment features, which was the case after having analysed 15 studies in total (i.e., [35: Craft et al. 2016]; [79: Hille and Bakker 2014]; [83: Hullman et al. 2015]; [95: Ksiazek et al. 2015]; [102: Loke 2012]; [107: Macovei 2013]; [129: Neagu 2015]; [142: Pond 2016]; [144: Quinn and Powers 2016]; [152: Rowe 2015]; [164: Slavtcheva-Petkova 2016]; [169: Strandberg 2008]; [171: Stroud et al. 2015]; Taddicken and Bund 2010; [182: Van den Bulck and Claessens 2014]). Finally, we added potentially relevant aspects from survey and interview
studies on comments in journalism (e.g., Heise et al. 2014a, 2014b; Loosen et al. 2013a, 2013b, 2017; Reimer et al. 2015; Springer et al. 2015). This way we are able to include variables that had not yet been the focus of content analyses, but whose investigation might be of particular interest to researchers, journalists, users, or protagonists of news stories. Examples of such variables include who is addressed in a comment (‘addressees of comments’) or whether it contains propaganda (‘kind of content: propaganda’). Through thematic clustering, we collectively and consensually grouped the identified objects of user comment analyses into seven ‘construct categories’ with similar thematic focus and added an open residual category to document any ‘unforeseen’ aspects analysed in the sample studies.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Example (sub-)categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliographical information</td>
<td>Authors, publication year, discipline, etc.</td>
</tr>
<tr>
<td>Methodology</td>
<td>Comment analysis method applied, additional methods applied, features/algorithms used in automated approaches, reliability/evaluation scores, etc.</td>
</tr>
<tr>
<td>Sampling</td>
<td>Media brands &amp; news stories comments refer to, number of analysed comments, etc.</td>
</tr>
<tr>
<td>Construct categories</td>
<td></td>
</tr>
<tr>
<td>Quantitative aspects</td>
<td>Length of comments, number of comments per story, etc.</td>
</tr>
<tr>
<td>Kinds of content</td>
<td>Personal opinion/attitude, argument for opinion, media criticism, propaganda, etc.</td>
</tr>
<tr>
<td>Incivility</td>
<td>Offensive language, personal insults, racism, sexism, etc.</td>
</tr>
<tr>
<td>Addressees of comments</td>
<td>Other users, journalist, forum moderator, etc.</td>
</tr>
<tr>
<td>Emotionality</td>
<td>Anger, hatred, fear, surprise, humour, etc.</td>
</tr>
<tr>
<td>Readability</td>
<td>Sentence length, technical/foreign terms, etc.</td>
</tr>
<tr>
<td>Facticity</td>
<td>Correctness of facts stated in comments</td>
</tr>
<tr>
<td>Other variable/construct</td>
<td>(Open category)</td>
</tr>
</tbody>
</table>

Table 1: Codebook overview.

The coding was carried out by five coders. As suggested by Lombard et al. (2002), eighteen random papers (approx. 9%) were coded by all coders. On this basis, we calculated Holsti’s coefficient for intercoder agreement ($r_H$), as proposed, for example, by Früh (2007). Of the 127 variables in our codebook 117 reached an acceptable to very good reliability score (58 variables with $r_H \geq .9$; 42 with $.9 > r_H \geq .8$; 42; 17 with $.8 > r_H \geq .7$). Of the 10 variables with reliability values of less than .7, we still included four due to their high level of relevance to our research questions. These are: ‘other methods applied in study’, ‘number of comments
analysed’, ‘kind of content: personal opinion/attitude’, ‘kind of content: other’. In these cases, we report the Holsti-score in a note.

**Results**

We first describe our sample in terms of the originating discipline of the studies, their authors, year of publication, and venue. Then we report which media, countries, languages, and platforms have been at the focus of comment analyses. Following this, we look at the methodology and sampling procedure of the studies, including what aspects of comments have been analysed automatically using machine learning techniques. Finally, we turn towards the seven construct categories, i.e. the examined quantitative aspects and kinds of comment content as well as the incivility, addressees of comments, emotionality, readability, and facticity of postings. In accordance with our research aims, we place special emphasis on comparing the two fields of communication studies and computer science throughout this section.

It is worth noting that the coding and counting were done on the level of papers and not on that of research projects. Consequently, we also use the word ‘study’ as a synonym of ‘paper’, and not of ‘research project’. The differentiation on paper level was the only one feasible because authors rarely state exactly what research project(s) the presented data originates from. It is possible, therefore, that one paper reports on the results from more than one project as well as that the findings of one project are reported on in more than one paper.

**Publication disciplines, authors, years and venues**

Nearly than half of the studies in our sample can be attributed to communication science (91 studies=47.4%). Papers from computer science and from other disciplines make up approximately a quarter each (52 studies=27.1% and 48 studies=25.0%, respectively).

The relatively small amount of studies from computer science is, in part, due to our inclusion criterion that a paper needed to present at least some results exclusively related to comments on journalistic content (see above). The most prominent other disciplines, as inferred from the publication venues in our sample, are health sciences (12 studies), language studies/linguistics (6), behavioural science/psychology (5) as well as political science, sociology, or ecology-related science (4 each).
In sum, 139 different publication venues are represented in our sample, the vast majority of them (81.3%) with one paper only. This reflects the research fragmentation resulting from the thematic breadth of user comments that makes them interesting to a variety of disciplines. In view of the significantly larger number of communication studies in the sample, it is not surprising that the ‘top three’ publication venues are: *Journalism* (8 studies), *Journalism Studies* (5), and *New Media & Society* (5). The top computer science venues are: the *ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)* (4 studies), the *International Conference on Social Informatics (SocInfo)* (4), and the *Annual ACM Web Science Conference (WebSci)* (2). The sample also comprises some journals one would not necessarily associate with analyses of user comments in online journalism, e.g. *Clinical Obesity*, the *Journal of Human Lactation*, *Crisis – The Journal of Crisis Intervention and Suicide Prevention*, *Stem Cell Reviews and Reports*, and the *Urban Water Journal*. The studies published in these and similar venues mostly investigated comments on a specific news topic that is of relevance to the respective field of research in order to ascertain users’ opinions on the topic.

The disciplinary divide becomes apparent in the publishing activities of the 433 authors in our sample: although their insights are of potential interest to more than one discipline, only eight of them (1.8%) have written for venues representing different disciplines, and in only three such instances (0.7%), these were communication and computer science publications.

Across disciplines, comment analyses have gained popularity from 2003 to 2015 as Figure 1 shows. The rather low numbers for 2016, especially for computer science, may result from proceedings for conferences in that year, which had not yet been published.
Media brands, countries, languages, and platforms covered

The 192 studies in our sample analyse comments referring to 295 different media brands. On average, each study looked at about four brands (M=3.9, n=165), the maximum being 35. However, more than half of the 165 studies that mention how many media they included analysed comments from only one brand (51.5%).

The most prevalent brands in the sample are the New York Times and the Guardian (26 studies each), supposedly due to their international reputation as well as the convenient APIs of their websites. Other media organisations of particular prominence are the BBC (17 studies), the Washington Post (15), the Daily Mail (10), the British Daily Telegraph (9), and the Wall Street Journal (8). Especially the ‘top group’ of the 21 brands that were included in studies four or more times shows a strong tendency towards broadsheet newspapers, whereas broadcasters (TV: 4, radio: 1), digital natives (3), and tabloids (2) are rarely the focus of research.

The media brands come from 43 different countries and all continents except Antarctica. However, comment analyses strongly focus on Anglo-American countries. Nearly half the studies are concerned with
US or UK media (30.1%, 18.6%, n=183). The next large category contains studies that look at media from more than one country (12.6%), including another five papers combining UK and US media only. Another recurring country combination is Qatar and Saudi Arabia/United Arab Emirates for comparing comments related to Al Jazeera and Al Arabiya (3 studies). The least attention has been devoted to African media, with only five studies.

The focus on Anglo-American media is reflected in the languages of the comments analysed. In total, the studies examine comments in twenty-six different languages. Nearly two thirds of papers, however, are concerned with English comments (62.4%, n=186). A distant second is the group of studies that investigate comments in more than one language (5.9%), with only one of these eleven works not also including English comments. Comments in other languages than English are examined in less than five percent of the cases each.

A striking difference between the disciplines is that more widely spoken languages are analysed in an automated manner more often (Arabic: 4 of 6 studies involve automated analyses; Spanish: 4 of 9; Chinese: 3 of 5; English: 33 of 116). This suggests that analysis software and lexica for less common languages might still be rare or immature.

In ninety percent of cases, the comments analysed are gathered from the comment sections of the respective media’s websites, while comments from Facebook, Twitter and other social platforms play a much smaller role despite the attention they receive in the public discourse about online discussions (see Table 2). Comparing disciplines, we find that Facebook comments are examined even less in computer science than in communication studies, while for tweets it is the other way around. Similarly, only 6.6 percent of studies look at comments from more than one platform and only two studies include three platforms.

<table>
<thead>
<tr>
<th></th>
<th>Total (n=182)</th>
<th>Communication studies (n=85)</th>
<th>Computer science (n=51)</th>
<th>Other discipline (n=46)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment section or forum on news website</td>
<td>90.1</td>
<td>87.1</td>
<td>92.2</td>
<td>93.5</td>
</tr>
</tbody>
</table>
Methodological aspects

Looking at the analysis methods applied in the studies, we realised an expected methodological divide between communication science and other disciplines, which predominantly use manual methods, and computer science in which automated analyses are by far the most prevalent (see Table 3). However, computational methods appear to be slowly becoming more prominent among communication scholars with the share of communications studies including automated approaches rising from ten percent in 2013 (n=10) to 22.2 percent in 2016 (n=18).

<table>
<thead>
<tr>
<th></th>
<th>Total (n=191)</th>
<th>Communication studies (n=90)</th>
<th>Computer science (n=52)</th>
<th>Other discipline (n=49)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automated quantitative content analysis</td>
<td>28.8</td>
<td>8.9</td>
<td>84.6</td>
<td>6.1</td>
</tr>
<tr>
<td>Manual content analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quantitative</td>
<td>42.9</td>
<td>57.8</td>
<td>30.8</td>
<td>28.6</td>
</tr>
<tr>
<td>Qualitative</td>
<td>48.3</td>
<td>54.4</td>
<td>11.5</td>
<td>77.6</td>
</tr>
</tbody>
</table>

Table 3: Comment analysis methods applied (multiple coding possible).

Studies in which two or even all three different content analysis methods were applied, for instance when a qualitative pre-study is conducted to prepare a manual or automated quantitative analysis, make up less than a fifth of our sample (17.2 and 1.6%, respectively). Multi-method designs are prevalent, though, since nearly half of the studies (47.4%), in addition to analysing comments, also investigate the corresponding journalistic stories or include interviews or surveys with users, journalists or comment moderators. Such approaches are significantly less common in computer science than in communication studies or other disciplines (26.9%, n=52 vs. 52.7, n=91 vs. 59.2, n=49; χ²=12.52, df=2, p<.01).

On the other hand, procedures and programmes for automated analyses appear to be created and tested almost exclusively in computer science. In that discipline, nearly all studies involving automated methods also elaborate on the development, implementation, and/or evaluation of software programmed
especially for the analysis (90.9%, n=44). In communication studies, this kind of work is virtually non-existent (14.3%, n=7).

In our corpus, two thirds of automated analyses employ machine learning approaches (supervised: 46.3%, unsupervised: 14.8%, both: 5.6%, n=54). Because one of our research objectives is to identify potentials for automating comment analyses, we further inspected these studies qualitatively. First, we grouped the different aspects of comments that have been examined in a (semi-)automated manner into six larger categories. (Note that the categories often overlap because aspects can be analysed on their own as well as as one part of another category. Categorising comments as on- or off-topic, for instance, can also be one step towards identifying troll comments.) We then identified particularly promising (semi-)automated approaches within each of the six categories that we briefly present now:

- **Sentiment**: Fifteen studies apply different procedures to determine the sentiment users express in their comment, i.e. their positive, neutral, or negative attitude towards a topic, usually that of the news story. Sentiment analysis is often used as one means among others (e.g., named entity recognition, part of speech tagging, vector space models) in a larger framework designed to achieve a higher goal. The respective sentiment analysis performance is not assessed individually. The approaches are also developed for different languages. For instance, Tumitan and Becker ([181: 2014]) test different procedures to predict Brazilian election results based on sentiments expressed in comments (in Brazilian Portuguese) before the polls. Kabadjov et al. ([86: 2015]) describe different approaches to summarising forum discussions (in English and Italian) including argumentation mining and sentiment analysis.

- **Trolling, hate speech, and spam**: Five papers are concerned with identifying destructive user contributions including hate speech, spam, or troll comments. Supervised machine learning approaches reach high accuracy scores, but they also depend on pre-labelled training data. Contrarily, de-la-Peña-Sordo et al. ([39: 2015]; [40: 2014a]; [41: 2014b]) employ semi-supervised machine learning and compression models (for comments in Spanish). Their procedures perform nearly as well as supervised approaches, while depending on less labelled data and being incrementally updateable.

- **On/off-topic**: In seven studies, machine learning is used to detect whether or not a comment is related to the original article or fits into the discussion context. This is important for tasks such as discourse and argumentation analysis, troll detection, or forum moderation. Here, de-la-Peña-Sordo et al.’s ([40: 2014a]; [41: 2014b]) unsupervised approach of comparing the comments’ vector space model with that of the
news story’s lead seems promising. This approach, developed for the Spanish language, was only part of a larger framework and its performance was not tested individually.

- **Discussion structure:** Three studies try to determine the structure of the discussion between users in a forum thread. For instance, Schuth et al. ([160: 2007]) propose a method for the precise detection of (Dutch-language) comments referring to other comments. de-la-Peña-Sordo et al. ([39: 2015]) show that a Random Forest algorithm performs best in identifying whether a (Spanish-language) comment refers to the news story or to another comment.

- **Topics:** Three studies seek to cluster topics discussed in comments or identify ‘hot topics’, i.e. themes that spark considerable discussion. Both supervised ([189: Zhou et al. 2014]) and unsupervised methods ([3: Aker et al. 2016]; [115: Meguebli et al. 2014]) have been used for these tasks, with the former, more labour-intensive approaches performing (expectably and) considerably better.

- **Diversity and anomaly:** Six studies deal with assessing the diversity of comments or with detecting unusual user contributions. For instance, Giannopoulos et al. ([63: 2012]) seek to identify within an English language discussion thread a subset of comments that are most heterogeneous with regard to the aspects of the news article they refer to and the sentiments expressed. These diverse comments, they argue, could be highlighted for other users in order to counter the risk of filter bubbles and echo chambers.

In the case of manual quantitative content analyses, researchers usually calculate inter- or intracoder reliability scores to account for the robustness of their method (Lombard et al. 2002). However, although most studies in our sample were published in high-quality academic journals, only half of the papers including a manual quantitative analysis provide such metrics (52.6%, n=78). With less than a fifth, the share is even smaller for studies employing a qualitative method (16.1%, n=87), which might be due to the fact that reliability scores are not undisputed in qualitative research. Similarly, only about a fifth of studies with a manual quantitative approach and a quarter of those including qualitative comment analyses report to have exercised consensual or peer coding (22.2%, n=81; 24.7%, n=93) which is a common measure for strengthening validity and reliability (Kuckartz 2014; Kurtanović and Maalej 2017; Schmidt 2004).

*Number of stories and comments*
The average and maximum number of news stories a study is concerned with seem extremely high: $M=6,569.3$; $\text{Max}=200,000$. However, this is due to some extreme outliers, the median is ‘only’ 50 stories.

Comparing methods reveals that the high mean value is connected with automated analyses that on average deal with significantly more news stories than studies only using other methods ($M_{\text{automated}}=26,338.6$, $n=34$ vs. $M_{\text{other}}=824.3$, $n=117$; Welch-test: $t=2.53$, $p<.05$). Unsurprisingly, the result is similar when comparing computer science and communication research ($M_{\text{comp}}=26,797.6$, $n=32$ vs. $M_{\text{comm}}=398.4$, $n=77$; Welch-test: $t=2.52$, $p<.05$). Entirely qualitative analyses on average examine the comments of a much smaller number of news stories than other manual analyses with a purely quantitative approach; however, the difference while large is not statistically significant ($M_{\text{qual}}=71.5$, $n=60$ vs. $M_{\text{quant}}=2,118.0$, $n=41$).

Concerning the actual number of comments analysed, the high mean value ($M=11,946,993.3$, $n=165$) is largely due to seven studies that analyse extremely large amounts of comments from approximately 2.5 million up to 1.8 billion individual examples. The median is ‘only’ 1,795 comments, and the minimum number of comments analysed is only 25.

**Construct categories: aspects of comments analysed**

We organised the aspects of comments that studies may investigate in categories of similar constructs. The first category is *quantitative aspects* which were investigated in nearly half of the studies (95 studies=49.5%). Here, the most prevalent aspect was the number of comments per individual news story, media brand, platform, etc. (see Table 4). Other aspects repeatedly examined were the number of individual commentators engaged, how many comments each of them posted, and how long these comments were. We found only minor differences between the disciplines in regard to the percentage of studies including quantitative aspects.

<table>
<thead>
<tr>
<th>%</th>
<th>Total (n=95; 49.5% of corpus)</th>
<th>Communication studies (n=52)</th>
<th>Computer science (n=25)</th>
<th>Other discipline (n=18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of comments per individual news story, media brand, platform, etc.</td>
<td>62.1</td>
<td>65.4</td>
<td>48.0</td>
<td>72.2</td>
</tr>
<tr>
<td>Number of individual commentators</td>
<td>44.2</td>
<td>42.3</td>
<td>44.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Length of comments</td>
<td>38.9</td>
<td>30.8</td>
<td>64.0</td>
<td>27.8</td>
</tr>
<tr>
<td>Number of comments per commentator</td>
<td>34.7</td>
<td>32.7</td>
<td>36.0</td>
<td>38.9</td>
</tr>
</tbody>
</table>
The category most frequently researched comprises *particular kinds of content* the occurrence and/or nature of which was examined in nearly all studies in our sample (170 studies=88.5%). As expected, personal opinions, e.g. on the topic of the related news story, are the most frequently researched aspect of this category (see Table 5). This includes sentiment analyses that categorise comments into positive, neutral, and negative opinions. User comments often seem to be used as a proxy to determine ‘public opinion’ although commenting users are not representative of the general population (Naab and Sehl 2017).

Other kinds of content that studies often investigated include instances where users provide an argument for the opinion they state, what frames or perspectives on a topic comments contain, and whether users react to each other’s comments or merely post their own thoughts. The prevalence of these aspects might be due to their assumed relation to the (deliberative) quality of user discussions which is a frequently researched topic (e.g., [55: Freelon 2015]; [154: Ruiz et al. 2011]).

Interestingly, computer science studies look at ‘constructive’ content that might also be of interest for other users or journalists considerably less frequently than communication research (or other disciplines for that matter). This applies to media criticism that could help improve reporting as well as aspects that could result or be included in future stories, such as personal experiences, additional pro/contra arguments, new information and leads, or untouched frames and perspectives (Loosen et al. 2017). Similarly, studies involving automated content analysis rarely focus on these kinds of content.

<table>
<thead>
<tr>
<th></th>
<th>Total (n=170; 88.5% of corpus)</th>
<th>Communication studies (n=82)</th>
<th>Computer science (n=41)</th>
<th>Other discipline (n=47)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal opinion, attitude, evaluation, judgement, verdict</td>
<td>70.6</td>
<td>68.3</td>
<td>68.3</td>
<td>76.6</td>
</tr>
<tr>
<td>Argument for opinion</td>
<td>34.1</td>
<td>37.8</td>
<td>4.9</td>
<td>53.2</td>
</tr>
<tr>
<td>Frame, perspective, etc.</td>
<td>34.1</td>
<td>37.8</td>
<td>7.3</td>
<td>51.1</td>
</tr>
<tr>
<td>Reaction to other comment</td>
<td>25.3</td>
<td>31.7</td>
<td>24.4</td>
<td>14.9</td>
</tr>
<tr>
<td>On-/off-topic</td>
<td>23.5</td>
<td>25.6</td>
<td>29.3</td>
<td>14.9</td>
</tr>
</tbody>
</table>
REIMER ET AL. (2021): CONTENT ANALYSES OF USER COMMENTS IN JOURNALISM.

<table>
<thead>
<tr>
<th>Personal experience</th>
<th>22.4</th>
<th>30.5</th>
<th>4.9</th>
<th>23.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Additional information, leads, material, etc.</td>
<td>16.5</td>
<td>24.4</td>
<td>7.3</td>
<td>10.6</td>
</tr>
<tr>
<td>Reference or link to external source</td>
<td>14.1</td>
<td>19.5</td>
<td>9.8</td>
<td>8.5</td>
</tr>
<tr>
<td>Media criticism</td>
<td>12.4</td>
<td>15.9</td>
<td>2.4</td>
<td>14.9</td>
</tr>
<tr>
<td>Mentioning of specific persons</td>
<td>11.8</td>
<td>11.0</td>
<td>9.8</td>
<td>14.9</td>
</tr>
<tr>
<td>Additional frame, perspective, etc.</td>
<td>10.6</td>
<td>14.6</td>
<td>2.4</td>
<td>10.6</td>
</tr>
<tr>
<td>Additional argument</td>
<td>7.6</td>
<td>11.0</td>
<td>4.9</td>
<td>4.3</td>
</tr>
<tr>
<td>Propaganda</td>
<td>0.6</td>
<td>1.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other kind of content, e.g. questions, personal information about user, style/structure (rhetorical, interactional, etc.), meta-discourse on comments</td>
<td>38.8</td>
<td>39.0</td>
<td>34.1</td>
<td>42.6</td>
</tr>
</tbody>
</table>

Table 5: Kinds of content researched in comment analyses.

More than a third of papers in the sample examined the incivility of comments (68 studies=35.4%). This may be because it is negatively related to the (deliberative) quality of user debates and a much-discussed topic among practitioners (Loosen et al. 2013a, 2013b, 2017; Reimer et al. 2015; Ziegele and Jost 2016). By far the most researched forms of incivility are general hostility and personal insults (57.4%) as well as profanity, such as the use of swear words (55.9%). Hostility and personal insults are examined more often when automated methods are involved (61.1% vs. 42.9%), while more specific forms of hate speech are rarely a topic of automated analyses: sexism, racism, religious or political intolerance were all investigated in only one such study which was conducted by an interdisciplinary team of journalism practitioners including a communication scholar, a data scientist, and graphic editors ([58: Gardiner et al. 2016]). The picture regarding hate speech is similar for manual quantitative analyses. By contrast, if a study involves a qualitative approach, it is more likely to also look at these forms of hate speech (sexism: 16.1 vs. 2.8%; racism: 32.3 vs. 13.9%; religious: 16.1 vs 2.8%; political: 12.9 vs. 8.3%; n=31 vs n=36).

In our sample, forty-five studies, or 23.4%, dealt with emotions expressed in comments or their overall emotionality. The most frequently studied forms of emotion are irony, sarcasm and cynicism, which are central to similar extents in both communication and computer science (36.8% vs. 33.3%). This is interesting because automated analyses are considered error-prone when recognising these particular variants of human emotion and language (e.g., [121: Moreo et al. 2012]; [175: Thelwall et al. 2012]), which might explain computer scientists’ attention to this topic. However, the absolute numbers of studies concerned with them are low (7 vs. 3). Positive emotions (pity/sympathy, surprise, curiosity/interest, love, happiness/joy,
enthusiasm, humour: 43.2%) are researched nearly as often as those feelings with a negative connotation (anger, hatred, contempt/disgust/nausea, fear, sadness, shame/guilt: 52.3%).

A quarter of studies (48 studies=25.0%) look at variables indicating who is addressed in a comment (Häring et al. 2018). Research examining whether specific users are addressed occurs most often (see Table 6). Much less frequently, scholars were interested in finding out if protagonists of the story or other people affected—e.g., obese people in studies on emotions towards obesity or weight loss surgery—were addressed. This also applies to individual journalists, the newsroom in general, or the general public. Interestingly, no computer science study looked at whether newsrooms, individual journalists, or community managers were addressed.

<table>
<thead>
<tr>
<th>%</th>
<th>Total (n=48; 25.0% of corpus)</th>
<th>Communication studies (n=33)</th>
<th>Computer science (n=5)</th>
<th>Other discipline (n=10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific user(s)</td>
<td>70.8</td>
<td>66.7</td>
<td>80.0</td>
<td>70.0</td>
</tr>
<tr>
<td>Protagonists of journalistic story or other external figures</td>
<td>31.3</td>
<td>30.3</td>
<td>40.0</td>
<td>30.0</td>
</tr>
<tr>
<td>Individual journalist(s)</td>
<td>25.0</td>
<td>24.2</td>
<td>-</td>
<td>40.0</td>
</tr>
<tr>
<td>Newsroom</td>
<td>20.8</td>
<td>21.2</td>
<td>-</td>
<td>30.0</td>
</tr>
<tr>
<td>Audience as a whole / general public</td>
<td>18.8</td>
<td>15.2</td>
<td>40.0</td>
<td>20.0</td>
</tr>
<tr>
<td>Community manager(s) / forum moderator(s)</td>
<td>4.2</td>
<td>6.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other addressees</td>
<td>16.7</td>
<td>24.2</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 6: Addressees of comments investigated.

In nearly one tenth of our sample (18 studies=9.4%), scholars also determine aspects of readability or comprehensibility of comments. In these studies, researchers identify if comments contain technical, foreign, or other terms that users might not understand (38.9 %), how complex and long the sentences in it are (38.9% and 33.3%, respectively), and if there are typos or other errors (22.2%).

Only twelve studies (6.3%) are concerned with checking the facticity of comments, i.e. they somehow check if comments contained factual statements or not. Interestingly, of these studies only one comes from computer science.
Finally, around eleven to thirteen percent of studies across disciplines also examined comments in relation to a variable we could not attribute to one of our construct categories, for instance the location attached to a post.

**Conclusion: an agenda for future (interdisciplinary) research**

This paper presents a systematic literature review of 192 analyses of user comments with reference to journalism identified in a systematic database search. It shows that user comments referring to journalistic stories on various topics—from elections, to climate change, to breastfeeding—are analysed in order to answer a wide range of research questions—for example, to determine users’ opinions on specific topics or the deliberative quality of online discussions—through the employment of manual quantitative or qualitative as well as (semi-)automated methods. In some cases, comments are the empirical data required for developing and testing automated approaches.

What all these studies have in common is that ultimately, they aim to **understand who communicates about what and how in user comments**. In view of the inundation of user comments posted every day and their diverse nature, there is an evident urge to (partially) automate their analysis. This is true for both academia and the journalistic field, and also for all other domains in which users share comments, such as streaming services, shopping websites, or app stores (e.g., Maalej et al. 2016). Whether automated analyses are created to support comment moderation or for scientific research, we consider it useful that they are designed collaboratively by the fields of social science and computer science. For this reason, rather than just summarising our results in this conclusion we draw on them to develop a **joint agenda for future research** that we believe should be addressed in close collaboration by communication studies and computer science.

Our findings show that interdisciplinary cooperation is still rare: for instance, only three of the 433 authors represented in the sample (0.7%) have published in both communication and computer science venues. However, our results show that computer scientists and communication scholars obviously share many research interests as nearly all aspects of user comments have been investigated separately in both disciplines.

This suggests considerable potential for cooperation in theoretical, methodological, and research-practical terms: even though the use of computational comment analysis is slowly gaining importance in
communication science, the tools for these analyses are predominantly developed by the field of computer science. However, particularly precise procedures can probably be best produced if they are guided by a clearly contoured research question and a profound understanding of the associated phenomena. Here, communication science offers a rich stock of elaborate theories and empirical findings which can help identify and formulate pressing research questions as well as operationalise the relevant aspects methodologically: communication scholars can contribute their expert knowledge of journalism, public debate, opinion formation, and other communicative phenomena to precisely tailor methodological approaches to the complexity of the aspect under consideration as well as to appropriately interpret the sometimes unclear results of, for instance, topic models. Similarly, the many existing communications studies investigating certain comment aspects qualitatively can be consulted to assess if and how their analysis processes can be automated or, at least, how particular aspects of interest may be operationalised for computational approaches. The automated detection of hate speech, for instance, could be improved to differentiate between its different forms—racism, sexism, religious and political intolerance, etc.—like qualitative analyses do. Another research desideratum would be to develop software that can identify comments with ‘constructive’ content that is likely to be of interest to other users or even of use to journalists: media criticism that could help improve reporting as well as aspects that could result or be included in future stories on the topic at hand, such as users’ personal experiences related to the topic, additional pro/contra arguments, new information and leads, or untouched frames and perspectives (Loosen et al. 2017).

Second, we found that it is significantly less common in computer science studies to combine the analysis of comments with other methods and data, such as surveys of users or journalists. In other words, computer scientists often look at comments in isolation. By contrast, as we show in the overview on existing research, communication scholars are often also interested in the relationships between comments and other factors, such as the article commented on or the attitudes and motivations of commenting users. This may represent an important research gap within computer science because including such possibly independent or moderating variables factors could prove effective in refining automated approaches for the analysis of comments. An example could be that computer science researchers have so far only looked at whether a comment addresses other users or protagonists of the news story, but have not taken into account that
comments may also be intended for the journalist who authored the piece, the newsroom as a whole or the forum moderators, which is likely to make a difference in terms of the comment’s content and moderation.

Third, content analysis is considered one of the core methods for communication science and has been developed decisively in this field of study (Krippendorf 2013). The acquired methodological expertise could help generate better training data and truth sets for supervised machine learning, i.e. manually labelled data used to train an automated approach and to evaluate its performance. Computer science papers seldomly make transparent how the variables under consideration were theoretically derived and operationalised in order to create this training and benchmark data. Additionally, the task of coding is regularly assigned to untrained annotators of services like Amazon’s Mechanical Turk (e.g., [22: Cheng et al. 2015]; [61: Ghorbel 2012]; [167: Sood et al. 2012]) with no reliability scores reported to determine coding accuracy (a notable exception being, for example, the study by Carvalho et al. [19: 2011]). More rigorous procedures could generate more valid and reliable training and benchmark data and help improve an automated approach.

Fourth, we found a slight difference in regard to the disciplines’ predominant epistemological interests, which also speaks for their complementarity. Researchers from the field of computer science often seek to detect if a comment contains a certain aspect: does it, for example, include racism, or criticism of the news story? Communication scholars on the other hand usually want to analyse how they do so: what kind of racist statements are made (and which potential others are not)? What exactly is being criticised: the selection of the topic (and disregard of others), the presentation, alleged bias, etc.? Consequently, in interdisciplinary projects, computer scientists may determine which comments to include in a sample in order to study a particular phenomenon while communication scholars provide insights into the very nature of that phenomenon.

In terms of more concrete research gaps that should be addressed in the future, we observe that research tends to concentrate on comments referring to Anglo-American newspaper brands while simultaneously disregarding others, especially African media. While this may, in part, be due to our own focus on English-language studies, there certainly is a need for further investigation, as commentary cultures, the diversity of opinions expressed in comments, and the (in-)civility of discourse, etc. are likely to vary in different countries with different political and media systems. In addition, many automated methods have been developed for the English language so that there are still considerable gaps with respect to other languages that are less common.
Comment analyses are seldomly concerned with the positive or useful aspects of user contributions. This is especially true for studies in computer science and research involving automated analyses in particular. There is more work needed, therefore, on automatically identifying ‘constructive content’, e.g. to help journalists “mak[e] sense of user comments” (Loosen et al. 2017) and use them for newswork. The highlighted valuable comments could serve as positive examples for other users to follow and consequently improve the overall discussion. Moreover, as shown in the research overview, this would likely enhance users’ perception of and loyalty to the media brand. Additionally, more constructive comments could encourage journalists to join the debate, which, as other studies have found, can in turn motivate lurkers to participate actively as well.

Nevertheless, the strong focus on negative aspects—most notably on incivility and hate speech—is justified by their exceptional relevance for journalism as illustrated in the research overview. In this regard, automated analyses developed jointly by communication scholars and computer scientists have great potential: first, they can help overburdened moderators manage the inundation of comments, for example by identifying comments that likely have to be moderated or deleted. Existing research suggests that this can have a positive effect on the perception of the media brand and also raise the overall quality of the comment section. This in turn could grant moderators the time to also check comments on their medium’s profiles on Facebook and other social media and improve debate there. Second, automated approaches can help further test two aspects which, as shown in the research overview, have been found to affect the civility and deliberative quality of user comments: different moderation strategies and diverse designs of comment section interfaces on the journalists’ and moderators’ as well as on the users’ side (Loosen et al. 2017; Løvlie 2018; Schneider and Meter 2019). With the help of automated approaches, both moderation strategies and interface designs could be tested in field experiments in real comment sections with large amounts of comments and evaluations carried out in real-time. Computer science researchers often have the know-how when it comes to developing the interfaces to test. Communication scholars on the other hand cannot only contribute their domain expertise on how moderation strategies and interfaces that improve user discussions might look, but also point out the typical problems of automated procedures and examine the results with relation to these: their difficulties related to “accounting for context, subtlety, sarcasm, and subcultural meaning” as well as their tendency to “lay[] the burden of error on underserved, disenfranchised, and minority groups” (Gillespie 2020: 3).
All of the above has the potential to reduce incivility and strengthen the constructive voices in comment sections and, thereby, motivate more journalists and users to participate, which is crucial as “the potentials for quality discourse emerge only when a substantial amount of users participate in commenting on a news item” (Weber 2014: 942). This way, the deliberative potential theoretically associated with user discussions could be promoted in practice.

Despite the focus on incivility, some negative aspects of user comments apparently receive too little attention as well. There are, for instance, hardly any comment analyses focusing on identifying propaganda or determining if comments are based on facts. In view of the current debates about these kinds of ‘dark participation’ (Frischlich et al. 2019), these represent fundamental research desiderata.

Instead we see a pervasiveness of studies examining the occurrence or nature of personal attitudes or sentiments expressed by users. This shows that researchers—just like lurkers in comment sections—often use comments as a ‘proxy’ to fathom ‘public opinion’ on a particular issue. Based on our sample, this is the case especially in disciplines other than communication or computer science. This may, of course, be problematic due to the lack of representativeness of active users and their comments in relation to the general population (e.g., Friemel and Dötsch 2015; Hölig 2018)—a fact that communication scholars can point out when working across disciplinary boundaries.

A particular shortcoming of comment analyses, so far, is the near-complete focus on user posts in comment sections and the small number of studies investigating comments made on social media, not to mention comparative analyses of comments from different platforms and media brands which also attract different groups of users. This is unsatisfying because of the outstanding real-world relevance of these points: first, as outlined in the research overview, most media also maintain accounts in social networks, especially Facebook, or have even ‘outsourced’ comments to them entirely. Second, there is much debate among practitioners about how much comments on platforms differ in tone and ‘quality’ (Loosen et al. 2017; Reimer et al. 2015), something that is very much worth scrutinising empirically; and third and most importantly, social networks have been discussed as a main distribution channel for disinformation, and it is highly relevant that we empirically determine what role comments play in framing and (re)interpreting ‘real and true’ as well as ‘fake and false’ news, for example when someone shares a journalistic post with their own introductory comment. Automated approaches that process far more content appear to have the potential to scale up comment analyses and, in so doing, also advance cross-platform comparisons.
Through our literature review we identified particularly promising approaches that can be adopted and adapted in the contexts discussed above in order to detect trolling or spam, ‘hot topics’, and exceptional statements as well as determine the sentiments, relatedness to the story’s topic, discussion structure, and diversity of comments.

Our study, of course, has some limitations. For instance, since we collected our data at the end of 2016, other relevant studies have been published; however, due to the considerable effort required, we were unable to research and code these as well. Furthermore, this review does not consider studies that also analyse comments to news stories (besides other content), but do not present the journalism-related results separately. Additionally, we have not included papers that solely report on the development and testing of a software tool for automated analyses. As a consequence, computer science papers and those studies that make use of automated methods may be generally underrepresented in our sample. As a result, some relevant tools and procedures may not be included in this review. Of equal importance, as this paper is in itself a manual quantitative content analysis, it provides only little insight into what the studies in our sample actually discovered about the comment properties they focus on.

Despite these points, our review produced valuable insights. We identified under-researched aspects of user comments, hopefully point researchers towards studies from either discipline upon which they can build their own research, and we were able to develop a research agenda of particularly pressing research gaps and how these could be addressed jointly by communication studies and computer science. We would like to see these efforts be considered as the foundations of a bridge that seeks to overcome the disciplinary divide in a bid to meaningfully and productively advance the analysis of user comments in journalism.

**Funding**

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: this work is partly funded by the Hamburg programme Ahoi Digital (project ‘Forum 4.0’).

**Acknowledgments**
The authors would like to thank the peer reviewers for their valuable feedback on previous drafts of this manuscript. Our special thanks go to Volodymyr Biryuk, Hannah Immler, Anne Schmitz, and Louise Sprengelmeyer for their great coding efforts.

1. The literature database and further information can be found on the project website: https://scan.informatik.uni-hamburg.de

2. This study meets our inclusion criteria but was excluded from the sample because it was not found through our systematic research and is in German.

3. There are two ahead-of-print articles in our sample ([4: Al-Rawi 2016]; [93: Ksiazek 2016]), for which we coded the year of online publication as the year of publication.

4. In the reliability test, this variable only had a Holsti-score of $r_H=0.62$.

5. In the reliability test, this variable only had a Holsti score of $r_H=0.56$.

6. In the reliability test, this variable only had a Holsti-score of $r_H=0.53$.

7. In the reliability test, this variable only had a Holsti-score of $r_H=0.69$. 
References


Hölig, Sascha. 2018. “Eine meinungsstarke Minderheit als Stimmungsbarometer?! Über die


APPENDIX

List of supplemental material

1. 
   Title: Appendix 1: Search strings and repositories searched
   Description: Overview of the development of the search strings and search procedure applied in the systematic literature review, including a list of the repositories searched; supplemental material for Content Analyses of User Comments in Journalism: a Systematic Literature Review Spanning Communication Studies and Computer Science by Reimer et al. (2021)
   File name: Content_analyses_of_user_comments_APPENDIX_1-Search_strings.docx

2. 
   Title: Appendix 2: Inspection of initial search results and final list of studies analysed
   Description: Explanation of the inspection of the initial search results and final list of studies analysed in the systematic literature review; supplemental material for Content Analyses of User Comments in Journalism: a Systematic Literature Review Spanning Communication Studies and Computer Science by Reimer et al. (2021)
   File name: Content_analyses_of_user_comments_APPENDIX_2-Inspection_and_list_of_studies.docx

3. 
   Title: Appendix 3: Codebook
   Description: Codebook used for the systematic literature review; supplemental material for Content Analyses of User Comments in Journalism: a Systematic Literature Review Spanning Communication Studies and Computer Science by Reimer et al. (2021)
   File name: Content_analyses_of_user_comments_APPENDIX_3-Codebook.docx
## Appendix 1:
### Search strings and repositories searched

1. **Complete search string**

<table>
<thead>
<tr>
<th>Inclusion criterion</th>
<th>Part of search string representing criterion/synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. (Content) Analysis...</td>
<td>(“content analysis” OR “textual analysis” OR “discourse analysis” OR “sentiment analysis” OR “text analysis” OR “network analysis” OR “thematic analysis” OR “corpus analysis” OR “corpus linguistic” OR “corpus linguistics” OR “linguistic analysis” OR “text mining” OR “sentiment mining” OR “opinion mining” OR “aspect-based mining” OR “aspect based mining” OR “argumentation analysis” OR “argumentation mining”)</td>
</tr>
<tr>
<td>AND</td>
<td></td>
</tr>
<tr>
<td>2. ...of user comments...</td>
<td>(comment OR comments OR commenting OR discussion OR discussions OR debate OR debates OR conversation OR conversations OR &quot;online communication&quot; OR post OR posts OR posting OR postings OR tweet OR tweets OR “asynchronous collaboration” OR discourse OR discourses OR forum OR fora)</td>
</tr>
<tr>
<td>AND</td>
<td></td>
</tr>
<tr>
<td>3. ...referring to online journalism</td>
<td>(news OR newspaper OR newscast OR journalism OR journalistic OR coverage OR press)</td>
</tr>
</tbody>
</table>

*Note.* Terms joined by an ‘AND’ as well as at least one of the terms in parentheses connected by an ‘OR’ must appear; phrases in quotes must appear verbatim. We did not use wildcards (e.g. “for*” for “forum” and “fora”) because they caused problems in some literature repositories.

We designed the search string in a way that it includes synonyms of the three criteria for studies to be included in our sample: studies have to be 1) based on a manual quantitative or qualitative, or a (semi-) automated content analysis of 2) user comments that 3) refer to journalism. First, we collected synonyms through a review of the keywords and abstracts of sixteen communication and computer science studies we had already identified as meeting the three inclusion criteria. Then our research team which consists of computer as well as communication scientists with extensive experience in comment analyses complemented the list with additional commonly used terms from both disciplines. The final search string combines eighteen synonyms for ‘content analysis’ with twenty-one synonyms for ‘comment(s)’ and seven synonyms for ‘journalism’. We tested and adjusted the search string several times. For instance, the term “media” (as a synonym for “journalism”) had to be eliminated because it is also part of the term “social media” which produced a large number of false positives without reference to journalism (e.g., analysis of comments in personal blogs, on Amazon, ebay, AirBnB, Instagram).
2. Literature repositories searched

<table>
<thead>
<tr>
<th>EBSCO Communication and Mass Media Complete (EBSCO CMMC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discipline: Communication science</td>
</tr>
<tr>
<td>Search string: Complete search string</td>
</tr>
<tr>
<td>Search in: Titles, abstracts, keywords</td>
</tr>
<tr>
<td>Search results: 1,065 potentially relevant papers</td>
</tr>
<tr>
<td>Retrieval of search results: Import of all search results into literature database software (Zotero) via plugin for browser (Mozilla Firefox), then export as CSV-file and import into Excel spreadsheet</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ACM Digital Library (&quot;The ACM Guide to Computing Literature&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discipline: Computer science</td>
</tr>
<tr>
<td>Volume of database: “407,367 Full-text articles; 2.0+ Million Pages of full-text articles; […] 44+ High Impact Journals […]; 2,000+ Proceedings Volumes; 8 Magazines […]; 37 Technical Newsletters from ACM's Special Interest Groups (SIGs); 6,500+ Video files; 594 Audio files” (<a href="http://libraries.acm.org/digital-library">http://libraries.acm.org/digital-library</a>; accessed 20 Jan 2017)</td>
</tr>
<tr>
<td>Search string: Complete search string</td>
</tr>
<tr>
<td>Search in: Titles, abstracts, keywords</td>
</tr>
<tr>
<td>Search results: 695 potentially relevant papers</td>
</tr>
<tr>
<td>Retrieval of search results: Direct export of all search results as CSV-file and import into Excel spreadsheet</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IEEE Xplore Digital Library</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discipline: Computer science</td>
</tr>
<tr>
<td>Volume of database: “170+ journals; 1,400+ conference proceedings; 5,100+ technical standards; approx. 2,000 eBooks; 400+ educational courses” (<a href="http://ieeexplore.ieee.org/xpl/aboutUs.jsp">http://ieeexplore.ieee.org/xpl/aboutUs.jsp</a>; accessed 20 Jan 2017)</td>
</tr>
<tr>
<td>Search string: Complete search string: A script was programmed that sent multiple simpler queries of all possible combinations of the terms included in the complete search string.</td>
</tr>
<tr>
<td>Search in: Titles, abstracts, keywords</td>
</tr>
<tr>
<td>Search results: 107 potentially relevant papers (excluding duplicates from the multiple queries)</td>
</tr>
<tr>
<td>Retrieval of search results</td>
</tr>
<tr>
<td>----------------------------</td>
</tr>
</tbody>
</table>

**Springer Link**

| Search string | This repository does not support complex search strings and was, therefore, searched with combinations of multiple, simpler queries, more precisely with twenty-six less complex search strings with combinations of those terms from the complete search string that, according to the studies already known to be relevant, are used most often to represent one of the three inclusion criteria: (analy* OR mining) AND (“reader comments” OR “user comments”) AND (news* OR coverage OR journalis* OR press). |
| Search in | Full-text (search cannot be restricted to titles, abstracts and keywords) |
| Intermediate search results | 2,281 (excl. duplicates from the multiple queries) |
| Additional steps | Search only in titles of results for the terms “comment*”, “forum”, “fora” |
| Search results | 40 potentially relevant papers |

**Web of Science**

| Search string | Complete search string |
| Search in | Titles, abstracts and keywords |
| Intermediate search results | 2,633 |
| Additional steps | Search for the terms “comment*”, “forum”, “fora” in titles, abstracts and keywords of results |
| Search results | 71 potentially relevant papers |

**ScienceDirect**

<p>| Search string | Complete search string: A script was programmed that sent multiple less complex queries of all possible combinations of the terms included in the complete search string. |
| Search in | Titles, abstracts, and keywords |</p>
<table>
<thead>
<tr>
<th>Search results</th>
<th>203 potentially relevant papers</th>
</tr>
</thead>
</table>

**Google Scholar**

<table>
<thead>
<tr>
<th>Search string</th>
<th>This repository does not support complex search strings, so that a more focused query was used (“journalism + comments + analysis”).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search in</td>
<td>Full-text (search cannot be restricted to titles, abstracts, and keywords)</td>
</tr>
<tr>
<td>Intermediate search results</td>
<td>220,000</td>
</tr>
</tbody>
</table>
| Additional steps | Inspection of title and abstract of the first 200 search results  
(As inspecting all roughly 500,000 search results would not have been feasible and this was the last list we checked, we decided to stop the inspection as soon as the last 20 results inspected had not contained a previously unknown and actually relevant study. This was the case after exactly 200 search results.) |
| Search results | 15 potentially and actually relevant studies |

**Reading lists of “The Coral Project”**

| Articles | 78 |
| Manual inspection of… | Titles, abstracts, keywords, and, in case of doubt, the article itself |
| Search results | 7 potentially and actually relevant studies |
Appendix 2:
Inspection of initial search results and final list of studies analysed

1. Inspection of initial search results and identification of actually relevant studies

The first two columns of Table 1 show the numbers of initial search results by the sources searched, totalling 2,219 potentially relevant studies (see Appendix 1 for details on the search procedure). Of these we first excluded:

- forty-nine publications from before 1999 as they are highly unlikely to be concerned with online comments;
- eight publications in languages other than English due to a lack of language proficiency;
- eleven publications that could not be found or obtained from databases or local libraries;
- eleven monographs and theses because these publication types are not provided systematically by all repositories and too voluminous to be analysed in full.

Each of the remaining studies was assigned to one of six researchers who, based on title, abstract and keywords, decided whether or not the study met the three inclusion criteria. In cases of doubt, we checked the full text and decided consensually.

The search string was designed to be as inclusive as possible so that we did not overlook any relevant papers. This, however, also means that it produced a large number of false-positive search results that we needed to identify and exclude to build the final sample: for instance, articles whose abstracts contain the three terms ‘journalism’, ‘content analysis’, and ‘comments’ often do not analyse user comments but journalistic commentary.

We also excluded analyses of mixed material that mingle comments on journalistic stories with other web content if they do not report results for journalism-related comments separately. For example, Twitter studies that examine tweets with certain hashtags often include tweets related to journalistic stories, but not exclusively.

The third column of Table 1 shows the results of the inspection, totalling 192 studies actually relevant for the review. A list of the 192 studies can be found below.

<table>
<thead>
<tr>
<th>Database, journal, reading list searched</th>
<th>Search results/potentially relevant studies (without duplicates)</th>
<th>Final corpus/relevant studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBSCO</td>
<td>1,065</td>
<td>66</td>
</tr>
<tr>
<td>ACM</td>
<td>695</td>
<td>18</td>
</tr>
<tr>
<td>IEEE</td>
<td>107</td>
<td>13</td>
</tr>
<tr>
<td>Springer Link</td>
<td>40</td>
<td>21</td>
</tr>
<tr>
<td>Web of Science</td>
<td>71</td>
<td>41</td>
</tr>
<tr>
<td>ScienceDirect</td>
<td>203</td>
<td>8</td>
</tr>
<tr>
<td>Google Scholar (first 200 search results)</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>The Coral Project reading list (78 entries)</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Literature database of research project ‘SCAN4J’ (280 entries)</td>
<td>16</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>2,219</td>
<td>192</td>
</tr>
</tbody>
</table>

Table 1: Overview of studies analysed by sources searched.

2. List of studies

Case ID


3. Aker, Ahmet; Kurtic, Emina; Balamurali, A.R.; Paramita, Monica; Barker, Emma; Hepple, Mark; Gaizauskas, Rob (2016) A graph-based approach to topic clustering for online comments to news. *European Conference on Information Retrieval*: 15–29. DOI:10.1007/978-3-319-30671-1_2


19. Carvalho, Paula; Sarmento, Luís; Teixeira, Jorge; Silva, Mario J. (2011) Liars and saviors in a sentiment annotated corpus of comments to political debates. In *49th Annual Meeting of the Association for Computational Linguistics*, pp. 564–568.


40. de-la-Peña-Sordo, Jorge; Pastor-Lopez, Iker; Santos, Igor; Bringas, Pablo G. 2014 Social news website moderation through semi-supervised troll user filtering. In *International Joint Conference SOCO’13-CISIS’13-ICEUTE’13*, pp. 577–587. DOI:10.1007/978-3-319-01854-6_59
41. de-la-Peña-Sordo, Jorge; Pastor-López, Iker; Ugarte-Pedrero, Xabier; Santos, Igor; Bringas, Pablo G. (2014) Anomalous user comment detection in social news websites. In Gaviria de la Puerta, José; García Ferreira, Iván; García Bringas, Pablo; Klett, Fanny; Abraham, Ajith; de Carvalho, André C.P.L.F.; Herrero, Álvaro; Baruque, Bruno; Quintián, Héctor; Corchado, Emilio (eds) *International Joint Conference SOCO’14-CISIS’14-ICEUTE’14*, pp: 1–10. DOI:10.1007/978-3-319-07995-0_51
50. Fan, Wen; Sun, Shutao; Song, Guohui (2011) Sentiment classification for Chinese netnews comments based on multiple classifiers integration. In 2011 Fourth International Joint Conference on Computational Sciences and Optimization, pp. 829–834. DOI:10.1109/CSO.2011.239


73. Grant, Aimee (2016) 'I...don't want to see you flashing your bits around': exhibitionism, othering and good motherhood in perceptions of public breastfeeding. *Geoforum* 71(1): 52–61. DOI:10.1016/j.geoforum.2016.03.004


77. Hepp, Andreas; Elsler, Monika; Lingenberg, Swantje; Mollen, Anne; Möller, Johanna; Offerhaus, Anke (2016) Citizens' online engagement: the Euro crisis in online forums. In Hepp, Andreas; Elsler, Monika; Lingenberg, Swantje; Mollen, Anne; Möller, Johanna; Offerhaus, Anke; Sword, Keith; Pospielovsky, Dimitry V. (eds) *The Communicative Construction of Europe*, pp. 109–140. DOI:10.1057/9781137453136_5


Quality Sociology 35(2): 163–181. DOI:10.1007/s11133-012-9224-6


100. Lei, Yang; Pereira, Jennifer A.; Quach, Susan; Bettinger, Julie A.; Kwong, Jeffrey C.; Corace, Kimberly; Garber, Gary; Feinberg, Yael; Guay, Maryse (2015) Examining perceptions about mandatory influenza vaccination of healthcare workers through online comments on news stories. *PLoS one* 10(6). DOI:10.1371/journal.pone.0129993


106. Luo, Zhunchen; Tang, Jintao; Wang, Ting (2013) Improving keyphrase extraction from web news by exploiting comments information. In *Web Technologies and Applications*, pp. 140–150. DOI:10.1007/978-3-642-37401-2_16


126. Musat, Claudiu-Cristiart; Faltings, Boi; Rousille, Philippe (2013) Negative opinions in online discussions. In 2013 International Conference on Social Computing, pp. 219–226. DOI:10.1109/SocialCom.2013.28


147. Regan, Áine; Shan, Liran; McConnon, Áine; Marcu, Afrodita; Raats, Monique; Wall, Patrick; Barnett, Julie (2014) Strategies for dismissing dietary risks: insights from user-generated comments online. *Health, Risk & Society* 16(4): 308–322. DOI:10.1080/13698575.2014.919993


181. Tumitan, Diego; Becker, Karin (2014) Sentiment-based features for predicting election polls: a case study on the Brazilian scenario. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pp. 126–133. DOI:10.1109/WI-IAT.2014.89


186. Xie, Sihong; Wang, Jing; Amin, Mohammad S.; Yan, Baoshi; Bhasin, Anmol; Yu, Clement; Yu, Philip S. (2015) A context-aware approach to detection of short irrelevant texts. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. DOI:10.1109/DSAA.2015.7344831


Appendix 3: Codebook

Procedures, preliminary remarks and inclusion criteria

1. First of all, coding for a literature review, as for any content analysis, is not trivial. In fact, it’s a double balancing act: firstly, between reducing as much complexity as necessary and preserving as much information as possible; and secondly, between the terminology and labelling as used in the coded material and the coders’ own terminology. The remarks provided here as well as the coding guide in general provide guidelines for you to solve the first problem. For the latter problem, there is a clear rule: “Translate” the terminology used in the papers you code into our own terminology. For an example, see the remarks regarding variables “Em02-Em16 – Emotionality subcategories” and “Kc05 – Personal opinion, …”.

2. Look into the list of papers assigned to you and open the PDF of the next paper listed there, which you will find in the folder “Literature”.

3. Of all studies, read at least the following sections: abstract, method(ology), sample/dataset, results/evaluation. The labelling in the studies may differ from the terms listed here. Look also for synonymously named sections.

4. One line of the coding sheet represents one study.

5. We only code variables which refer to comments on journalistic content as their unit of analysis. If, for instance, the journalistic content itself is analysed as well, we do not consider the variables used to do so.

6. Highlight all codings you are unsure about in a different colour. Note the reason why you are unsure in the “coding problem”-column at the end of the coding sheet.

7. At the end of the coding sheet you also find a column where you can note anything noteworthy, e.g. if the paper mentions an algorithm or tool we could use for our own developments, or if

8. Coding language is English. This means terms in other languages (e.g., the names of variables in non-English studies) must be translated into English. Note that in English, a full stop (a.k.a. dot, period) is used to separate the integer part of a number from its fractional part: “1.28” (instead of “1,28” in German).

9. Do not use any thousands separator (comma or full stop/dot/period) when writing numbers: “1000” (instead of “1.000” or “1,000”).

n.a. or empty cell = not applicable, i.e. the variable cannot be applied to the study at hand or does not make sense with regard to it. Example: The study has been identified as a book section. Consequently, the variable “Volume” does not apply to it since it only makes sense with regard to journal articles.

n.i. = not indicated, i.e. the variable is (in principal) applicable to the study at hand but the study does not contain (sufficient) information to decide which code to assign. Example: In the study there is no indication as to what time span the analysed comments cover. Consequently, with regard to the variable “Time span covered by sample” we have to assign “n.i.”.

0 = no / not included / not true, i.e. the variable is not true for the study at hand although theoretically it could have been. Example: The authors of a quantitative content analysis do not report whether they conducted a reliability check. Consequently, with regard to the variable “Reliability check” we have to assign “0” for “no reliability check (mentioned)”. Note the difference between “0” and an empty cell/“n.a.”.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Codes &amp; coding style</th>
<th>Definition &amp; Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study info</td>
<td>(automatically filled in cells)</td>
<td></td>
</tr>
<tr>
<td>1. PI01 – Case ID</td>
<td>Case ID of study at hand</td>
<td>Check case ID of the study at hand</td>
</tr>
<tr>
<td></td>
<td>Check category</td>
<td></td>
</tr>
<tr>
<td>2. PI02 – Authors</td>
<td>Authors of study at hand</td>
<td>Check if author names are complete, written out and spelled correctly.</td>
</tr>
<tr>
<td></td>
<td>Check category</td>
<td>Mode: surname, first name; surname, first name; ...</td>
</tr>
<tr>
<td>3. PI02a – No. of authors</td>
<td>Number of authors of study at hand</td>
<td>Check the number of authors of the study at hand which will be calculated automatically here.</td>
</tr>
<tr>
<td></td>
<td>Calculated category</td>
<td></td>
</tr>
<tr>
<td>4. PI03 – Year of publication</td>
<td>Year of publication</td>
<td>Check if year of publication is correct.</td>
</tr>
<tr>
<td></td>
<td>Check category</td>
<td>Mode: yyyy</td>
</tr>
<tr>
<td>5. PI04 – Title of study</td>
<td>Title of study at hand</td>
<td>Check if title of study is complete (including subtitles) and spelled correctly.</td>
</tr>
<tr>
<td></td>
<td>Check category</td>
<td>Mode: title : or: ? or: – or: : subtitle.</td>
</tr>
<tr>
<td>6. PI06 – Type of publication</td>
<td>Type of publication</td>
<td>Check if type of publication is indicated.</td>
</tr>
<tr>
<td></td>
<td>Check category</td>
<td>In case of “Other type”, note the type of publication.</td>
</tr>
<tr>
<td></td>
<td>• Journal paper</td>
<td>Note that a blogpost presenting the results of or on-going research is coded as a &quot;research report&quot;.</td>
</tr>
<tr>
<td></td>
<td>• Conference/proceedings paper</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Book section</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Research report (incl. certain blogposts)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Other type of publication: ______</td>
<td></td>
</tr>
<tr>
<td>7. PI07 – Title of publication</td>
<td>Title of publication</td>
<td>Check if title of journal, proceedings or edited volume the study was published in is complete (including subtitles) and spelled correctly.</td>
</tr>
<tr>
<td></td>
<td>Check category</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PI08 – Editor(s)</td>
<td>Editor(s) of parent publication</td>
</tr>
<tr>
<td>---</td>
<td>-----------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>9.</td>
<td>PI09 – Volume</td>
<td>Volume of journal</td>
</tr>
<tr>
<td>10.</td>
<td>PI10 – Issue</td>
<td>Issue of journal</td>
</tr>
<tr>
<td>11.</td>
<td>PI11 – Starting page</td>
<td>Starting page of paper</td>
</tr>
<tr>
<td>12.</td>
<td>PI12 – Ending page</td>
<td>Ending page of paper</td>
</tr>
<tr>
<td>13.</td>
<td>PI13 – Database</td>
<td>Database(s) the study was found in</td>
</tr>
<tr>
<td>14.</td>
<td>PI14 – Keywords</td>
<td>Keywords / tags</td>
</tr>
<tr>
<td>15. PI15 – Coder</td>
<td>Coder ID</td>
<td>Note your coder ID, i.e. the initials of your name. Example: “CC” for Cody Coder</td>
</tr>
<tr>
<td>------------------</td>
<td>----------</td>
<td>------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Open category Mode: Initials</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Me – Methodology</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>16. Me03-Me05 – Methods of comment-content analysis applied</th>
<th>Methods of comment-content analysis applied in the study Multiple coding category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0 = no</td>
</tr>
<tr>
<td></td>
<td>1 = yes</td>
</tr>
</tbody>
</table>

Indicate all method(s) of comment-content analysis that have been applied in the study at hand. If the authors of a study do not explicitly state if their analysis (or parts of it) are quantitatively or qualitatively oriented, use the explication of these two types below.

In case of multi-method studies indicate only those methods for which results are presented in the paper. Mind that the authors may use other terms for “content analysis”, e.g.: discourse analysis, text(ual) analysis, sentiment analysis, opinion mining, argument mining, etc.

This variable refers only to the analysis of comments. Neither content analyses of other forms of content (e.g., the journalistic stories commented on) nor the procedures of retrieving the comments to be analysed are to be coded here. The latter remark means that the automated retrieval of comments (e.g., via a crawler or API) alone does not qualify the analysis of these comments to be coded as “automated” since the actual coding of the comments could still have been accomplished manually.

If you have already completed some coding sheet lines for a certain paper, and then realise that another method has been applied, don’t forget to change your coding also in the previous lines.

Me03 – Automated comment-content analysis ➔ Filter variable: If “0”, skip Me06 and Me07

Me04 – Manual quantitative comment-content analysis – i.e., the researchers count how many comments belong to different, pre-defined categories (just like we do it with their papers here). E.g., a study on whom users address in their comments may group comments into three categories: “other users addressed”, “author of the news story addressed”, “forum moderator addressed”.

Note that despite its focus on pre-defined categories, a quantitative study may very well include residual categories in case the a priori developed typology is not exhaustive of the different forms a variable can take. In the abovementioned example, some users might direct their comment towards other than those three groups, e.g. the protagonist of the news story (“Dear Mrs Merkel,…”). These “unforeseen” categories are often coded in an “other”-category (just like we often do in this codebook, as well), which might be differentiated into the different newly found categories afterwards (“protagonists”, “general public”, etc.). However, the focus of a quantitative analysis lies on pre-defined categories and counting how often they appear in a sample of comments.
Me05 – Qualitative comment-content analysis – i.e., in contrast to quantitative analyses, in qualitative studies the categories used to cluster and describe the comments usually are not pre-defined. Instead, they are the result of the analysis. E.g., a study on whom users address in their comments would analyse the given sample and afterwards, as a result, state that comments are directed towards: the general public, other users of the discussion forum, the author of the news story, etc.

Note that often the share of comments that falls into the different categories is mentioned, too. However, this does not necessarily mean that the study has to be coded as a (manual or automated) quantitative study, as well. Only if the typology is developed qualitatively in a pre-study on the basis of a rather small sample, and afterwards another, often larger sample, is coded quantitatively (in a manual or automated manner) in a subsequent study, the paper is coded as containing a qualitative as well as a quantitative (manual or automated) comment-content analysis. Two examples for qualitative studies that also present percentage values are: Glenn et al. 2012: 127–128 and Reader 2012: 502–505.

17. **Me03-05a – No. of different comment analysis methods**

<table>
<thead>
<tr>
<th>Number of different methods of comment analysis applied</th>
<th>Calculated category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Here, the number of different methods of comment analysis applied in the study will be calculated automatically.</td>
<td></td>
</tr>
</tbody>
</table>

18. **Me06 – Tool development**

<table>
<thead>
<tr>
<th>Discussion of the development of a tool or its parts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = no (phase of) tool development discussed</td>
</tr>
<tr>
<td>1 = tool development</td>
</tr>
<tr>
<td>2 = tool evaluation</td>
</tr>
</tbody>
</table>

Only if “Automated content analysis” = “1” = “yes”:

Indicate whether the paper reports on phases of the development of a tool or its parts:

0 = no phase of tool development discussed (e.g., as in Chmiel et al. 2011).
1 = tool development – i.e., the paper discusses/reports on the development one or more parts of a tool (e.g., single algorithms, a mock-up, a prototype, a user interface, etc.) or a proof of concept. Examples: Du et al. 2015: 52; Neunerdt et al. 2013.
2 = tool evaluation – i.e., the paper discusses/reports on the evaluation of a complete version of a tool, i.e. more than just a single algorithm, but a tool with different features and/or a user interface, etc. The evaluation is usually connected with terms like “performance”, “accuracy”, “precision”, “usability”, etc. Examples: Park et al. 2016; Neri et al. 2011.

19. **Me07 – Content-focused paper**

<table>
<thead>
<tr>
<th>Does the paper also report on the actual content of comments?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = no, not at all</td>
</tr>
<tr>
<td>1 = yes, to a smaller extent (less than a third of the space dedicated to describing content of comments)</td>
</tr>
</tbody>
</table>

Only if “Me06 – Tool development” = “1 = tool development or implementation” or “2 = tool evaluation”:

Papers about the automated analysis of comments often are not (so much) concerned with describing what the actual content of the comments analysed looks like, i.e. what portion of the comments contained the variables of the content analysis, e.g. how many were positive/neu/tal/neutral/negative, how many referred to a certain topic, how many contained a particular argument, how many included insults or other forms of incivility, etc. Instead, these papers only or predominantly deal with the development of an algorithm,
2 = yes, to a substantial extent (a third and more of the space dedicated to describing content of comments)

Formula, prototype, mock-up, tool, etc. and/or with the testing and evaluation of their performance or usability. Hence, an indicator of such non-content-focused papers can be the extensive reporting of accuracy or precision values or measures for usability. Example: in a non-content-focused paper about the sentiment of comments, the authors would report extensively on how accurately an algorithm assesses the sentiment of a set of comments (as compared to a manual coding of said comments). It would, however, not elaborate on how many comments actually were positive or negative or what that means for the discussion in the comment section, for the news topic commented on, and so on.

In contrast, a content-focused paper is (also) concerned with describing in depth the content of the comments.

An example for code “1” = “yes, to a smaller extent” is Foster et al. (2011) who, in table 1, also present information on the “content” they analysed, which in this case is the length of sentences in the user comments. The rest of the paper, however, is concerned with describing how accurately differently trained versions of their tool assess the sentence length in comments without ever saying how long the sentences in the analysed comments actually are. (Thus, if it were not for table 1, Foster et al. (2011) would be coded as “0” = “no, not at all”.)

<table>
<thead>
<tr>
<th>Me08 – Reliability value reported</th>
<th>Reliability value(s) of content analysis reported in paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = no reliability value reported</td>
<td></td>
</tr>
<tr>
<td>1 = yes, inter-coder reliability value reported</td>
<td></td>
</tr>
<tr>
<td>2 = yes, intra-coder reliability value reported</td>
<td></td>
</tr>
</tbody>
</table>

Only if “Me03 – Automated content analysis” and/or “Me04 – Manual quantitative content analysis” = “1 = yes”:

Do the authors of the study report that they have performed a reliability check and what the resulting value was?

<table>
<thead>
<tr>
<th>Me09 – Peer coding procedure</th>
<th>Application of peer coding procedure mentioned in paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = no</td>
<td></td>
</tr>
<tr>
<td>1 = yes</td>
<td></td>
</tr>
</tbody>
</table>

Do the authors of the study report that the coders engaged in a form of peer coding to enhance inter-coder reliability? That is, did coders discuss codings they disagreed or were unsure about, in order to collectively reach a shared decision?

“1 = yes” has to be coded also when this procedure was only used in the building of a truth set.

<table>
<thead>
<tr>
<th>Me10 – Additional methods applied</th>
<th>Additional methods applied in the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = no</td>
<td></td>
</tr>
<tr>
<td>1 = yes</td>
<td></td>
</tr>
</tbody>
</table>

Indicate if additional methods have been applied in the study at hand. Examples of such additional methods are:

- quantitative or qualitative content analyses of other content – e.g., of the journalistic stories that the
- quantitative surveys, qualitative interviews or focus groups – e.g., among commenters to find out why they comment, etc.
REIMER ET AL. (2021): CONTENT ANALYSES OF USER COMMENTS IN ONLINE JOURNALISM.

Quantitative or qualitative observations – e.g., of journalists and forum moderators to find out how they handle comments, etc.

Feature analyses – i.e., analyses of the features and/or functions provided by a website, social network, discussion forum, etc.

In case of such multi-method studies, count only those methods for which results are presented in the paper. → content analysis of other content. In case of such multi-method studies, count only those methods for which results are presented in the paper.

Note that the authors might not explicitly name what kind of method they applied (e.g., “We also analysed the sentiment of the news stories the comments refer to.” → content analysis of other content) or not even call it a method or analysis (e.g., Anstead/O’Loughlin (2011) analyse tweets that refer to a political talk show. They do not explicitly state that they analysed the content of the talk show, as well. However, this becomes clear when they explain which statements from the talk show panelists triggered how many and what kind of tweets from the viewers: “[D]escriptions of social identities by the Question Time panel drive descriptions among the viewertariat. […] Recall that the point of the broadcast that elicited most comments occurred at 22:20, following Bonnie Greer’s extended attack on Griffin’s use of history to justify his policies […]” (BBC Question Time Transcript 2009: 17)” (ibid.: 456; emphasis in original). This can definitely be considered a qualitative content analysis.)

<table>
<thead>
<tr>
<th>I01 – Informatics / automated analyses</th>
<th>Machine learning</th>
<th>Filter variable: Indicate whether a machine learning approach was applied:</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = no ML</td>
<td>proceed with I17</td>
<td>0 = no ML: No machine learning approach was used.</td>
</tr>
<tr>
<td>1 = supervised</td>
<td>proceed with I02</td>
<td>1 = supervised: Supervised learning is the machine learning task of inferring a function from labelled training data. Example: “Through investigation of supervised learning techniques, we show that content-based […]”</td>
</tr>
<tr>
<td>2 = unsupervised</td>
<td>proceed with I03</td>
<td>2 = unsupervised: Unsupervised learning is the machine learning task of inferring a function to describe hidden structure from unlabelled data. Example: “This paper presents a simple unsupervised learning algorithm for classifying reviews as recommended […]”</td>
</tr>
<tr>
<td>-99 = n.i.</td>
<td>proceed with I17</td>
<td>-99 = n.i.: It is not indicated if a machine learning approach was used.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I02 – Truth set</th>
<th>Size of truth set</th>
<th>Only if “Machine learning” = “1” = “supervised”:</th>
</tr>
</thead>
</table>
In case a supervised machine learning approach is used \((I01 = \text{Supervised})\), note the size of the training set / labelled data used.

Example: “Annotators were asked to label each sentence as ‘controversial’ or ‘non-controversial’ based on the content of the news article, resulting in 6,218 labelled sentences: […]”

Only if “Machine learning” = “1” = “supervised” or “2” = “unsupervised”:

### I03 – Readability
Feature that measures how readable a text is according to a metric (e.g. SMOG, Flesh-Reading-Ease).

**Example:** “We assessed readability using the Flesh-Reading-Ease Test scoring method.”

### I04 – Length
Feature that measures how long a text is based on number of words or characters.

**Example:** “In this work, we utilize the validated analytic operationalization of several of these criteria to score comments, including readability, personal experience, length, and relevance (i.e., to the article and conversation).”

### I05 – Word frequencies (e.g. bag-of-word, personal experience)
Indicate whether some features are based on word frequencies of specific words (e.g. to identify personal experience within a comment).

**Example:** “The word-frequency of the author’s name”

**Example:** “The word-frequency of the candidate (candidate freq) is measured equal to the feature above.”

**Example:** “Personal experience: Measures the rate of use of words in Linguistic Inquiry and Word Count (LIWC) categories “I”, “We”, “Family”, and “Friends:””

### I06 – Relevance
Feature that measures how relevant topic is with regard to the article commented on or other comments.

**Example:** “How relevant a comment is with respect to the article, based on word feature vector similarity”

### I07 – Informativeness
Feature that measures the informativeness quantified and extracted as a feature?

**Example:** “They propose to use entropy rate as a way to measure informativeness of a comment.”

### I08 – Ratings / recommendation votes
Feature that measures the rating of comments (e.g. likes, ratings, recommendations, …)?

**Example:** “How many recommendations/votes a comment has received”

### I09 – Activity (user interaction / user history)
Feature that measures the degree to which users participate in an online discussion.

**Example:** “We incorporate understanding from other literature in the domain of online reviews which suggests that a measure of user activity level will be usefully correlated to quality”
### I10 – Linguistic feature
(e.g. sentence structure) – Feature concerned with linguistic aspects of comments.

*Example:* “In addition, we also order the words in an article based on decreasing *perplexity values* and the average *perplexity* of the top 10, 20 and 30 words in this list are added as features.”

### I11 – Other feature: ______ – List the other features comma separated that were used.

---

### 26. I12–16 – Metrics

<table>
<thead>
<tr>
<th>Evaluation metrics</th>
<th>Only if &quot;Machine learning&quot; = &quot;1&quot; = &quot;supervised&quot; or &quot;2&quot; = &quot;unsupervised&quot;: Describe the accuracy of the machine learning approach. If just a single value x is mentioned, indicate this with the range x-x (see examples). Often multiple values or ranges are mentioned.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open multiple coding category Note range of numeric values in % or: -99 = not indicated</td>
<td><strong>Example:</strong> “The overall performance measures used on Sentiment Analysis used <strong>Recall</strong> and <strong>Precision</strong>: in our tests, they are normally above 87% and 93% respectively.” =&gt; 87-93 If multiple ranges are mentioned per metric, create a range from the smallest and the largest value reported. <strong>Example:</strong> If the ranges 50-70 and 60-80 are reported, note 50-80.</td>
</tr>
</tbody>
</table>

**I12 – Accuracy**
*Example:* “Overall the accuracy of classification was 70%” => 70-70
*Example:* “The accuracy was usually around 80%” => 80-80

**I13 – Precision**
*Example:* “Average precision score using 5-fold cross validation result was 0.13±0.07 with 95 percent confidence interval using linear SVM classifier” (transform decimals to ones and tens) => 6-20

**I14 – Recall**
*Example:* “The three methods all have a fairly low recall of 0.39 to 0.66” (transform decimals to ones and tens) => 39-66

**I15 – F-Score (F1 and F2):** If both are mentioned <F1 range>,<F2 range>. Transform decimals to ones and tens.

**I16 – Other evaluation metric: ______ – Note other evaluation metric.

*Example:* Receiver-operating characteristic (ROC)

---

### 27. I17 – External tools

<table>
<thead>
<tr>
<th>Tools / frameworks / libraries external software names</th>
<th>Note common software used for data preparation, analysis, evaluation, etc. (separated by commas): Common tools are: NLTK, scikit, Google parser, Weka, word2vec, Stanford, Mallet, etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open category</td>
<td><strong>Example:</strong> “[…], which is implemented as part of the Natural Language Toolkit (NLTK) (Bird, Loper, &amp; Klein, 2009), and which translates any word into its root stem allowing us to look it up in the dictionary.”</td>
</tr>
<tr>
<td>Column</td>
<td>Code</td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
</tr>
</tbody>
</table>
| 28.   | I18  | Self-developed tools | Note self-developed tools (e.g., scripts or spreadsheets) (separated by commas).  
Example: "In particular, we created a Python script to categorize a large dataset; used spreadsheet and statistical software to organize the data and identify the objects of our analysis."
| 29.   | I19  | Algorithms | Shortly note the algorithms used (separated by commas).  
These are mostly optimization, sorting, heuristics, machine learning etc. algorithms. Common algorithms are:
- for Machine learning: SVM, Naïve Bayes, kNN, etc.
- for other purposes: SMOG, Flesh-Reading-Ease, LIWC, Tf-idf, LDA, tokenizer, POS-tagger/Part-of-speech (tagger), tree-learner, rule-learner, etc.
Example: "Since the final score for the comment is a weighted sum of weights and scores from each criteria, we tested a linear support vector machine (SVM) and a logistic regression."
Example: "The experiment uses three kinds of single classifiers (Naïve Bayes, kNN, SVM) and two multiple classifiers integration algorithms (Bagging, Boosting) for three groups of corpus."
| 30.   | I20  | Cross-fold | Indicate whether a cross-fold validation is conducted.  
Example: "To evaluate the performance of the proposed algorithm for sentiment classification of Chinese netnews comments, we adopted 10-fold cross validation."
Example: "The prediction experiments described in the next section were developed using sentiment classified using the SVM approach with cross-fold validation."
| 31.   | I21  | Evaluation | Name the validation method for assessing how an automatic statistical analysis performs on unseen data.  
Example: "Thus we have collected total 10 sentences from starting of the document. Each of these sentences were then classified by OntoCat, and the answers were checked manually." \(\rightarrow\) qualitative validation
| 32.   | I22  | Limitations | Note restrictions for the application of the mentioned approach and/or preconditions for the data in order to be analyzed as mentioned by the authors of the study.  
Example: "Yet we must also consider the limitations of such an automated approach, such as the possible censorship (or at least de-emphasis) of comments that may in fact be related but not detected by the algorithm as relevant due to differences in language use."
### REIMER ET AL. (2021): CONTENT ANALYSES OF USER COMMENTS IN ONLINE JOURNALISM.

| 33. | Mo01 – Names of media outlets | Names of media outlets the comments refer to | Note the names of the media outlets the comments analysed refer to (separated by commas). Do *not* distinguish between online and offline editions of media outlets. Also, do *not* mention media outlets from which no user comments were collected and analysed.  
- In case of print media, do not use common abbreviations, but write out the title, e.g.: *New York Times* (instead of: *NYT*), *Washington Post* (instead of: *WP*), etc.  
- The same rule applies to TV or radio programmes (newscasts, political talk shows, etc.), e.g.: *Tagesschau, heute, Anne Will, Günther Jauch, Hart aber fair*, etc.  
- In case of online editions of print, TV or radio media, add “online” to the full name of the original medium, e.g.: *Der Spiegel online* (instead of: *Spiegel Online* or *SpOn*), *Bild online* (instead of: *Bild.de*), *Frankfurter Allgemeine Zeitung online* (instead of: *FAZ.net*), *Süddeutsche Zeitung online* (instead of: *SZ.de* or *Sueddeutsche.de*).  
- The full-name rule also applies to native online media, e.g. write: “*Huffington Post*” (instead of: “*HuffPo*”), *BuzzFeed, Watson, etc.*  
In case of no names mentioned, note “-99” = “n.i.”.  
In case the names of the media outlets are not mentioned but their number is mentioned, make sure to add the number in variable “Mo01a – No. of media outlets” (because the number cannot be calculated automatically if no names have been entered in this cell). |
| 34. | Mo01a – No. of media outlets | Number of media outlets | Calculated category |
| 35. | Mo02 – Countries of origin of media outlets | Countries of origin of media outlets | Note the countries of origin of the media outlets the comments analysed refer to (separated by commas):  
- Australia = “AU”  
- Austria = “AT”  
- Belgium = “BE”  
- Brazil = “BR”  
- Canada = “CA”  
- China = “CN”  
- Czech Republic = “CZ”  
- Denmark = “DK”  
- Finland = “FI”  
- France = “FR”  
- Germany = “DE”  
- Greece = “GR”  
- India = “IN”  
- Italy = “IT”  
- Japan = “JP”  
- Netherlands = “NL”  
- Norway = “NO”  
- Poland = “PL”  
- Portugal = “PT”  
- Romania = “RO”  
- Russia = “RU”  
- South Korea = KR  
- Spain = “ES”  
- Sweden = “SE”  
- Switzerland = “CH”  
- Turkey = “TR”  
- USA = “US”  
Note other countries of origin fully.  
In case no country of origin is reported by the authors, note “-99” = “n.i.”. |
### Language of the comments analysed

<table>
<thead>
<tr>
<th>Code</th>
<th>Language Code</th>
<th>Language</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>zh</td>
<td>Chinese (any kind)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fi</td>
<td>Finnish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ko</td>
<td>Korean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ru</td>
<td>Russian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cs</td>
<td>Czech</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fr</td>
<td>French</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nb</td>
<td>Norwegian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>Spanish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>da</td>
<td>Danish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>German</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pl</td>
<td>Polish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sv</td>
<td>Swedish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nl</td>
<td>Dutch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>el</td>
<td>Greek</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pt</td>
<td>Portuguese</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tr</td>
<td>Turkish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>it</td>
<td>Italian</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ro</td>
<td>Romanian</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note other languages fully.

In case of no indication of the comments' languages, note “-99” = “n.i.”

### Platform or channel the comments analysed were posted on/through

<table>
<thead>
<tr>
<th>Code</th>
<th>Platform or Channel</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>C02</td>
<td>Comment section or forum on news site – i.e. the comment sections beneath a news story or blog on a media outlet’s website or the forum on a media outlet’s website in which comments can be posted. Blogs of individual journalists are counted as a medium website, too.</td>
<td></td>
</tr>
<tr>
<td>C03</td>
<td>Facebook – i.e. a media outlet’s Facebook profile or Facebook posts referring to a media outlet’s content, e.g. by linking to it</td>
<td></td>
</tr>
<tr>
<td>C04</td>
<td>Twitter – i.e. a media outlet’s Twitter profile or tweets referring to a media outlet’s content, e.g. by linking to it</td>
<td></td>
</tr>
<tr>
<td>C05</td>
<td>Other platform: _____ – Note other platform.</td>
<td></td>
</tr>
</tbody>
</table>

### Important remarks on the following multiple coding variables:

**Sc – Story and comment selection**

- **C01** – Open category
  - Language of the comments analysed
  - Note the languages of the comments analysed as mentioned by the authors or indicated by the media outlet the comments refer to (separated by commas):
  - Chinese (any kind) = “zh”
  - Finnish = “fi”
  - Korean = “ko”
  - Russian = “ru”
  - Czech = “cs”
  - French = “fr”
  - Norwegian = “nb”
  - Spanish = “es”
  - Danish = “da”
  - German = “de”
  - Polish = “pl”
  - Swedish = “sv”
  - Dutch = “nl”
  - Greek = “el”
  - Portuguese = “pt”
  - Turkish = “tr”
  - English = “en”
  - Italian = “it”
  - Romanian = “ro”
  - Note other languages fully.
  - In case of no indication of the comments’ languages, note “-99” = “n.i.”.

- **C02-C05** – Multiple coding category
  - Platform or channel the comments analysed were posted on/through
  - Note other platform.

- **C02** – Comment section or forum on news site
  - Journalistic content can be commented on on different platforms or through different channels. Usually, media outlets offer a comment section underneath every news story published as well as on every blog hosted on their website and/or a central forum in which users can discuss and comment on news items. Other platforms on which users often comment on media outlets’ profiles on social media or social media in general.
  - Indicate all channels or platforms comments were retrieved and analysed from.

- **C03** – Facebook
  - i.e. a media outlet’s Facebook profile or Facebook posts referring to a media outlet’s content, e.g. by linking to it

- **C04** – Twitter
  - i.e. a media outlet’s Twitter profile or tweets referring to a media outlet’s content, e.g. by linking to it

- **C05** – Other platform
  - Note other platform.

**Important remarks on the following multiple coding variables:**

- **Sc – Story and comment selection**
  - Usually, a study does not consider all the comments ever posted on a platform. Instead, selecting comments for an analysis usually follows a two- or even three-step approach:
  - In the first step, researchers usually aren’t directly concerned with the comments they want to analyse. Instead, they select certain journalistic stories from all the stories the investigated media outlets publish.
on their website, post about on Facebook or tweet about on Twitter. These stories can be selected using a traditional sampling technique (e.g., random sampling or a "constructed week") or by defining specific criteria they have to match or features they have to contain (e.g. they have to cover a specific topic, originate from a specific beat or website section, be covered in a certain issue of the TV programme or print outlet at hand or on a certain day on the website in question, have to be most liked or clicked, etc.). These selection criteria are coded in the first of the following categories “Selection of journalistic stories” (Sc01–09).

However, it is also possible that researchers do not limit their study to a certain set of stories but would include (comments to) all stories the media outlets have ever published on the platform in question. (The actual sample of comments to be analysed can still be limited and refined in the following step(s).)

It is only in the second step that researchers select the actual comments they want to analyse. Here again, they can choose to include all comments to the stories defined in the first step, to use a traditional sampling technique (e.g., random sampling) to limit the number of comments, or to define selection criteria on the comments level, as well. These selection criteria are coded in the second of the following categories “Selection of comments” (Sc11–19). The thus compiled sample comprises those comments that are analysed vicariously for all the comments belonging to the population. However, here it is also possible that all the comments to the before defined stories are included in a “comprehensive sample”.

In some cases, studies may comprise more than one sample of comments. For instance, researchers could compare the comments to two different kinds of journalistic stories (differentiation in the first step) or two different kinds of comments to the same kind of story (differentiation in the second step). In such a case, code all criteria for selecting journalistic stories as well as for selecting the actual comments.

| 38. | Sc01-Sc10 – Journalistic story selection | Criteria for the selection of journalistic stories
|     |     | Open multiple coding category
|     |     | 0 = no
|     |     | 1 = yes
|     |     | -99 = n.i.

Shorty note all criteria for the selection of journalistic stories, the comments to which shall be analysed. If a criterion is definitely not used to define the population, note “0”.

If it remains unclear whether a criterion is used to define the population, note “-99” = “n.i.”.

Sc01 – Specific beats – i.e., all stories from one or more specific beats or newspaper/website sections (e.g., politics, sports, business) are considered. If this is the case, shortly note the beat(s) (separated by commas).
### Sc02 – Specific authors

- **Definition:** Only stories by specific authors. If this is the case, *shortly note the authors’ names or defining characteristics* (e.g., only columnists, only journalists of a particular political orientation, only female journalists, etc.).
- **Example:** If the selection criterion is based on specific authors, note their names or defining characteristics, e.g., “only articles by John Smith and Jane Doe.”

### Sc03 – Specific topic or event

- **Definition:** Only stories referring to a specific topic or event (e.g., comments to stories on the Olympic Games). If this is the case, *note the topic, event, etc.* in an understandable way, e.g., “only articles that refer to the US election 2016.”
- **Example:** If the selection criterion is based on specific topics or events, clearly define them, e.g., “comments to stories on the US election 2016.”

### Sc04 – Single specific story/ies, episode(s), issue(s)

- **Definition:** The paper is concerned with only one specific story, one episode of a TV or radio programme, one issue of a print medium (with multiple stories), stories on a website from one specific day. This code also applies to cases in which single episodes/issues/stories of one or different media are compared, e.g., if the paper analyses the comments regarding one day’s episodes of the *Tagesschau* and *heute* or two specific episodes of *Tagesschau* from different days.
- **Example:** If the selection criterion is based on a single story, episode, or issue, clearly define it, e.g., “comments to the US election 2016.”

### Sc05 – Specific features

- **Definition:** Only comments referring to stories with specific features (e.g., a data visualisation, a listicle, a cliffhanger headline). If this is the case, *shortly note said feature.*
- **Example:** If the selection criterion is based on specific features, note them, e.g., “only articles that feature data visualisations.”

### Sc06 – Specific time period (including constructed/artificial week)

- **Definition:** Only stories published in one or more specific time periods (e.g., between 2012 and 2014; in 2010 and in 2016) were considered. Note: A “constructed week” or “artificial week” consists of all seven weekdays (Monday to Sunday), but those aren’t chosen from one single week but randomly from a longer time period (e.g., a year).
- **Example:** If the selection criterion is based on a specific time period, clearly define it, e.g., “between 2012 and 2014.”

### Sc07 – User metrics

- **Definition:** Only stories that produced particular user metrics are considered (e.g., the X most clicked/read, most liked, most commented, most shared, best rated, etc. stories). If this is the case, *shortly note said the metrics in question.*
- **Example:** If the selection criterion is based on user metrics, note them, e.g., “only stories with more than 10,000 clicks.”

### Sc08 – Other story selection criteria

- **Definition:** Shorty note the other criteria that journalistic stories need to meet in order for the comments referring to them to be considered for analysis.
- **Example:** If there are other criteria not covered by the above codes, briefly mention them, e.g., “only stories that were published on Fridays.”
| Sc09 – All stories of the media outlet(s) – i.e., explicitly all stories ever published on the platform(s) in question are considered for further selection of at least one sample of comments in the second step (Sc11−19). Only if this is the case, you can code “1” (= “yes”) here. |
| Sc10 – Story selection process/criteria applied but not explicated – i.e., from the paper it becomes clear that the authors have limited the number of journalistic stories the comments to which they want to analyse, but they do not explicate the selection process/criteria they used. Only if this is the case, you can code “1” (= “yes”) here. |

<table>
<thead>
<tr>
<th>Sc11 – No. of news stories</th>
<th>Number of news stories the comment sample refers to</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sc12-Sc20 – Selection of comments</td>
<td>Criteria for the selection of comments</td>
</tr>
</tbody>
</table>

### Sc11 – No. of news stories
- **Number of news stories the comment sample refers to**
  - Open category
  - or: -99 = n.i.

### Sc12-Sc20 – Selection of comments
- **Criteria for the selection of comments**
  - Open multiple coding category
  - 0 = no
  - Open coding / 1 = yes
  - -99 = n.i.

### Note
- Note the number of news stories the comments in the sample refer to in total.
  - Example: “We collected 300 comments on 15 different news articles on the website.” → Note: “15”.
  - If no number is given by the authors of the study, note “-99” = “n.i.”.

### Criteria for the selection of comments
- Indicate/note all selection criteria/procedures applied to compile the sample(s) of comments actually analysed from all the comments to the above defined journalistic stories. This means, if some variables are analysed using a comprehensive sample while others are analysed using a random sample drawn from the comprehensive sample, indicate that both “Comprehensive sampling” and “Random sampling” apply.

### Sc12 – Comprehensive sampling
- i.e., no further sampling but all comments to the above defined journalistic stories are subjected to analysis. If this is the case, note “1” (= “yes”).

### Sc13 – Random sampling
- i.e., a number of comments are selected randomly (not: arbitrarily). If this is the case, note “1” (= “yes”).

### Sc14 – Specific time period
- i.e., only comments posted in one or more specific time periods (e.g., in the first two hours after publication of the above defined stories) were considered for further sampling. If this is the case, shortly note the time period(s) as mentioned in the paper.
  - Or: If it is stated in the paper that comments were collected from a particular time period but the time period itself is not specified, note: “Specific time period, but not explicated”.
  - Also note if other than pragmatic reasons are given for the choice of the specific position, e.g. “The first comments are likely to be most original, while the last ones are likely to reflect on all aforementioned viewpoints.”
  - Examples: “comments from first five minutes after publication of story”; “tweets posted 12 hours before and after the airing of the programme XY”; etc.
| Sc15 – Specific position in thread | i.e., comments are selected based on their position in the overall comment thread (e.g., the first five and last ten comments to the above defined articles). If this is the case, shortly note said position. Also note if other than pragmatic reasons are given for the choice of the specific position, e.g. “The first comments are likely to be most original, while the last ones are likely to reflect on all aforementioned viewpoints.” |
| Sc16 – Specific users | i.e., only comments posted by users with specific characteristics (e.g., highly or rarely active users, female users, were considered for further sampling. If this is the case, shortly note said characteristics defining the users whose comments are considered. |
| Sc17 – Specific content | i.e., only comments that contain a certain link, the name of another user, particular terms or phrases, etc. If this is the case, shortly note said content. |
| Sc18 – Comment metrics | i.e., only comments that produced particular metrics are selected (e.g., only the X comments that were liked most, replied to most often, rated best, etc.). If this is the case, shortly note the metrics in question. |
| Sc19 – Other selection criteria/procedure | Note the other selection criteria or method applied. Example: “Snowball sample: 1st comment under story and all comments referring to this comment”, etc. |
| Sc20 – Comment selection process/criteria applied but not explicated | i.e., from the paper it becomes clear that the authors have limited the number of comments actually analysed, but they do not explicate the selection process/criteria they used. Only if this is the case, you can code “1” (= “yes”) here. |

| Sc21 – Number of samples | Number of different comment samples
Open coding category
or: -99 = n.i. |
| Sc22 – Group comparisons | Group comparisons
0 = no
1 = yes |

Some researchers use more than one sample:
Either not all variables are analysed using the same sample, e.g. some variables can be analysed in an automated manner using a comprehensive sample while others are analysed manually using a smaller random sample from that comprehensive sample. Or variables are analysed for different samples (or: groups of comments), e.g. the hostility of comments on Facebook can be compared to that of comments on the medium’s website; the sentiment of comments from different media outlets can be compared; etc. Note the number of different samples used in the paper at hand. If only one sample was used, note: “1”. If from the paper it becomes clear that more than one sample was used but the exact number is not specified by the authors, note: “-99” (= “n.i.”).
comments on Facebook can be compared to that of comments on the medium’s website; the sentiment of comments from different media outlets can be compared; etc.

<table>
<thead>
<tr>
<th>43.</th>
<th>Sc23-Sc28 – Data collection</th>
<th>Method of retrieving comments from comment platform</th>
<th>Indicate all the techniques used in the study to retrieve comments from the comment platform.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Multiple coding category</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 = no</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-99 = n.i.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sc23 – Manual copy &amp; paste: The comments were copied manually from the platform and pasted into a database.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sc24 – Crawled from web page: The data was gathered with programs (e.g. a script) from a web page. Example: “For our dataset, we crawled the most-Dugg stories of the past 365 days in November 2008, […]”</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sc25 – API: An application program interface (API) is used for gathering the data (e.g. via a REST interface). Some media sites provide APIs dedicated for developers to access their data. Example: “a dataset collected via the Times Community API (<a href="https://developer.nytimes.com)%E2%80%9D">https://developer.nytimes.com)”</a></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sc26 – Existing dataset: The dataset is already aggregated and publicly available on the internet. Example: “The complete data repository is available for download, thus allowing the extraction of the articles and their features.”</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sc27 – Other: _____ – Note the data collection method if it does not match any of the above mentioned.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sc28 – Method of data collection not indicated – Note</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>44.</th>
<th>Sc29 Open data</th>
<th>Openness / accessibility of the data set</th>
<th>Indicate whether the data set is made publicly available.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 = closed / not available</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = open on request</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = open / published</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-99 = availability n.i.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 = Closed: The authors clearly state that the data is not available. (Cf. “-99 = Not indicated”!)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 = Open on request: The authors of the study promise to provide the data set on request.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 = Published: The data is publicly accessible in the internet.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>-99 = Not indicated: The authors do not state whether the data is published or can be requested.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>45.</th>
<th>Co03 – Number of variables in the study</th>
<th>Number of variables in the study</th>
<th>Note the number of variables the constructs or dimensions investigated in the study consists of or is are operationalised with. Here, by “variables” we mean all nominal, ordinal and metric variables (incl. dichotomous variables and open ending categories). Example: If the superordinate construct “readability” is calculated by 1) the number of words of a comment, 2) the length of these words and 3) the length of the sentences, you note: “3”.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Open category</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>or: -99 = n.i.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>46.</th>
<th>Co04-Co08 – Level of measurement</th>
<th>Level of measurement</th>
<th>Indicate the level(s) of measurement (also called scale of measure(ment) or scale type) the different variables are measured on. That means it is possible to indicate more than one level of measurement.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Multiple coding category</td>
<td></td>
</tr>
</tbody>
</table>
Note that variables in qualitative studies can often be coded as being "measured" in a dichotomous manner – either on the nominal level (e.g., if comments belong to a particular category or not) or on the ordinal level (e.g., if comments contain a particular element or not).


"The nominal type differentiates between items or subjects based only on their names or (meta-)categories and other qualitative classifications they belong to [...]. Examples of these classifications include gender, nationality, ethnicity, language, genre [...]. The numbers in nominal measurement are assigned as labels and have no specific numerical value or meaning. No form of mathematical computation (+, - x etc.) may be performed on Nominal measures. [...] The ordinal type allows for rank order (1st, 2nd, 3rd, etc.) by which data can be sorted, but still does not allow for relative degree of difference between them. Examples include, on one hand, dichotomous data with dichotomous (or dichotomized) values such as ‘sick’ vs. ‘healthy’ when measuring health, ‘guilty’ vs. ‘innocent’ when making judgments in courts, ‘wrong/false’ vs. ‘right/true’ when measuring truth value, and, on the other hand, non-dichotomous data consisting of a spectrum of values, such as ‘completely agree’, ‘mostly agree’, ‘mostly disagree’, ‘completely disagree’ when measuring opinion. [...] Metric levels of measurement, i.e. interval and ratio scales, consist of items with always the same degree of distance between two neighbouring items: ‘1’ is ‘as far away from ‘2’ as ‘2’ is from ‘3’, and so on.] Examples for interval scales include temperature with the Celsius scale, [...] date when measured from an arbitrary epoch (such as AD) [...]. Ratios are not meaningful since 20 °C cannot be said to be ‘twice as hot’ as 10 °C, nor can multiplication/division be carried out between any two dates directly. [...] Examples for ratio scales include mass, length, [...] [number of items]. In contrast to interval scales, ratios are now meaningful because having a non-arbitrary zero point makes it meaningful to say, for example, that one object has ‘twice the length’ of another (= is ‘twice as long’). Very informally, many ratio scales can be described as specifying ‘how much’ of something (i.e. an amount or magnitude) or ‘how many’ (a count)."

Co04 – Dichotomous (no matter which level of measurement) – Note that sometimes variables are described as measuring if a comment includes “aspect A or aspect B”. In this case, the variable is only dichotomous if, firstly, aspects A and B are mutually exclusive and, secondly, it is impossible that a comment contains neither A nor B. Otherwise the variable is nominal or ordinal with more than two categories since the comments might contain: “A” or “B” or “A and B” or “neither A nor B”.

Co05 – Nominal with more than two categories
Co06 – Ordinal with more than two ranks
### Co07 – Metric (interval & ratio)

### Co08 – Qualitative variable with principally unlimited number of categories/forms it can take – i.e., the (number of) categories were not pre-defined, but the categories were the result of a qualitative analysis. E.g., Yahav et al. (2014: 76) in the qualitative part of their study classified the way in which information can be unintentionally leaked through comments “into 9 a-posteriori sub categories according to severity level of information exposure”.

<table>
<thead>
<tr>
<th>Cs – Comment sample</th>
<th>47. Cs01 – Sample size</th>
<th>Sample size(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Open category</td>
<td>Note the sample size(s) (n) of the study. Make sure you do not only note the number of comments containing the variable in question. E.g., if Slavtcheva-Petkova (2016: 77) writes: “The words ‘journalism,’ ‘journalist(s),’ and/or ‘journalistic’ were used in 178 comments—about 11 percent of the sample.”, the sample size is not 178 (11 %) but all the comments in which the researchers searched for the words (100 %; in this case: 1,583, a number mentioned earlier in the paper (ibid.: 76)). If the variables are all coded for the same sample but the sample sizes (i.e., the numbers of valid values) differ due to differing numbers of missing or invalid values, calculate the average sample size (round numbers to ones): sample “A” for var1: n = 219 sample “A” for var2: n = 218 → average sample size n → 218,5 → 219. In case of more than one sample – e.g., for comments from different platforms –, not all sample sizes, separated by semicolon: sample “Facebook” for var1: n = 220 sample “Twitter” for var1: n = 300 → “220; 300” In case there is no indication of the sample size, note “-99” = “n.i.”.</td>
</tr>
<tr>
<td></td>
<td>Mode: X</td>
<td>or: -99 = n.i.</td>
</tr>
</tbody>
</table>

### Important remarks:

When deciding whether a variable is used as an indicator to measure a particular aspect of comments, what is pivotal are the remarks made in the study at hand: e.g., Fredheim et al. (2015: n.p.) state they use the “[n]umber of words in all capitals” as “an indicator of shouting”. Hence, although “shouting” might also
be an indicator of “anger”, it would not be correct to code it as such, since this is not what the study’s authors themselves state.

Single terms used by the authors of the study might, however, be translated into our terminology if it is clearly appropriate: e.g., Chmiel et al. (2011) write of “emotions” and “sentiments”, which at first glance, one would subsume under “Em01 – Emotionality”. Throughout the paper, however, it becomes clear that what they mean by these terms actually is what we call “personal opinion” or “attitude”. Hence, we code it under “Kc05 – Personal opinion, ...”

| Qa01 – Quantitative aspects | Quantitative aspects | Filter variable:
| 0 = no | skip Qa02-Qa07 |
| 1 = yes | proceed with Qa02-Qa07 |

| Qa02-Qa07 – Quantitative aspects – subcategories | Quantitative aspects of comments | Only if “Quantitative aspects” = “1” = “yes”:
| Multiple coding category |
| 0 = no |
| 1 = yes |
| or: Note other quantitative aspect |

- **Qa02 – Development of amount of comments over time (“Kommentarverlauf”)** – i.e., the researchers present how many comments have been posted per minute/hour/etc. over a certain time period (e.g., the first 24 hours after the publication of the news story commented on). (Note that in this case, the unit of analysis can also be considered to be the news story. However, this alone does not justify to count the mentioned analysis as an additional method applied in the study (Me10)!

- **Qa03 – Length of comments** – in characters, words, sentences, paragraphs, etc.

- **Qa04 – Number of comments per individual news story, media outlet, platform, etc.** – i.e., researchers list the different news stories they considered and the number of comments each of them received or compare the average number of comments posted on different platforms/in different media outlets, etc. (Note that in this case, the unit of analysis can also be considered to be the news story. However, this alone does not justify to count the mentioned analysis as an additional method applied in the study (Me10)!

- **Qa05 – Number of individual commentators** – i.e., the researchers state how many commentators have produced the comments in the sample.

- **Qa06 – Number of comments per commentator** – i.e., the researchers state how many comments the individual commentators in the sample wrote (in total or on average). (Note that in this case, the unit of analysis can also be considered to be the commentator. However, this alone does not justify to count the mentioned analysis as an additional method applied in the study (Me10)!)
<table>
<thead>
<tr>
<th>50.</th>
<th>Ad01 – Addressees</th>
<th>Addressees of the comments</th>
<th>Filter variable:</th>
<th>Addressees of the comments</th>
<th>Filter variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 = no</td>
<td>skip Ad02-Ad08</td>
<td>1 = yes</td>
<td>proceed with Ad02-Ad08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

51. Ad02-Ad08 – Addressees – subcategories

<table>
<thead>
<tr>
<th>51.</th>
<th>Ad02-Ad08 – Addressees – subcategories</th>
<th>Addressees in comments</th>
<th>Filter variable:</th>
<th>Addressees in comments</th>
<th>Filter variable:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Multiple coding category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 = no</td>
<td>skip Ic02-Ic08</td>
<td>1 = yes</td>
<td>proceed with Ic02-Ic08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

52. | Ic01 – (In-)Civility | (In-)Civility | Filter variable: | (In-)Civility | Filter variable: |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0 = no</td>
<td>skip Ic02-Ic08</td>
<td>1 = yes</td>
<td>proceed with Ic02-Ic08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

53. | Ic02-Ic08 – (In-)Civility – subcategories | (In-)Civility of comments | Filter variable: | (In-)Civility of comments | Filter variable: |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Multiple coding category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 = no</td>
<td>skip Ic02-Ic08</td>
<td>1 = yes</td>
<td>proceed with Ic02-Ic08</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Qa07 – Other quantitative aspect: _____ – Note the other quantitative aspect of comments the variables measure.

50. Ad01 – Addressees

Filter variable:
Indicate whether the one or more variables measure who is addressed in the comments analysed, e.g. the media outlet/newsroom, individual journalists like the author of the news story commented on, comment moderators/forum managers, other users, the general public, etc (an aspect of deliberative quality as well as of inter-reactivity). Example: The findings regarding the “intended audience” of comments in Glenn et al. (2012: 127–128).

51. Ad02-Ad08 – Addressees – subcategories

Only if “Addressees” = “1” = “yes”:
Indicate all types of addressees the addressing of whom the variables measure.

Ad02 – Newsroom
Ad03 – Individual journalist(s)
Ad04 – Community manager(s) / forum moderator(s)
Ad05 – Audience as a whole / general public
Ad06 – Specific user(s)
Ad07 – Protagonists of journalistic story or other external figures – i.e., politicians, business representatives, NGO members etc. who are portrayed or featured in the news story commented on or connected to its topic.
Ad08 – Other addressees: _____ – Note the other addressees.

52. Ic01 – (In-)Civility

Filter variable:
Indicate whether the one or more variables are concerned with hostility, hate speech, profanity (use of swear words), but also politeness or other forms of (in-)civility in comments.

53. Ic02-Ic08 – (In-)Civility – subcategories

Only if “(In-)Civility” = “1” = “yes”:
Indicate all types of (in-)civility the variables measure.

Ic02 – Profanity / use of swear words / offensive language – e.g.: “To measure offensive language we used the offensive word list, compiled by Luis von Ahn of Carnegie Mellon University which contains in excess of 1,300 words […] The list includes many words that usually are inoffensive, but which in certain
circumstances may be antagonistic. For instance, ‘kill’ need not express hostile intent.” (Fredheim et al. 2015: n.p.).

Note that such measures may include personal insults, as well. In that case, the following code “Ic03” applies, too.

**Ic03 – Hostility and personal insults** – Here, “hostility” comprises all kinds of hostile behaviour in which a target person or smaller group is clearly addressed, e.g. insults directed towards/referring to other users, the journalists of the media outlet commented on, etc.

In contrast to hostility, “hate speech” comprises all kinds of hostile behaviour directed at larger groups of people who usually are characterised by only one or a few features (e.g., sex, colour of skin, occupation). The members of this group do not necessarily know each other personally as is the case with women, politicians in general, etc.

Note that this definition may differ from the use of the term “hate speech” in the study at hand.

**Ic04 – Hate speech – sexism**

**Ic05 – Hate speech – racism**

**Ic06 – Hate speech – religious intolerance** – i.e., hate speech directed towards members of a specific faith.

**Ic07 – Hate speech – political intolerance** – i.e., hate speech directed towards members of a specific party or followers of a certain ideology, e.g. communists.

**Ic08 – Other form of (in-)civility: _____** – Note the other form of (in-)civility measured by the variables in question.

**Em01 – Emotionality**

<table>
<thead>
<tr>
<th>Emotionality</th>
<th>Filter variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = no</td>
<td>skip Em02-Em16</td>
</tr>
<tr>
<td>1 = yes</td>
<td>proceed with Em02-Em16</td>
</tr>
</tbody>
</table>

Indicate whether the one or more variables measure the inclusion and/or intensity of emotions in comments, e.g.: anger, hatred, contempt/disgust/nausea, fear, sadness, pity/sympathy, shame/guilt, surprise, curiosity/interest, love, happiness/joy, enthusiasm, but also humour, irony/sarcasm/cynicism or other kinds of emotions.

Note that some variables might be named as if they refer to emotionality while in fact they refer only to the question whether the comment contains an opinion/verdict/evaluation/judgement on a particular matter (which is coded with category Kc05). E.g., “sentiment” can mean “mood” or “spirit” but is often used in the meaning of “opinion” or “pro/contra verdict”. The same applies to “negativity”.

54.

57.
## Emotionality of comments

**Multiple coding category**

0 = no  
1 = yes  

or: Note other form of emotionality.

Only if "Emotionality" = "1" = "yes":

Indicate all emotions/emotionality aspects the variables measure. (Measuring the emotional “mood”, “spirit” or “tone” of comments is sometimes called “sentiment analysis”. However, the term “sentiment analysis” can also be used for “only” measuring whether a comment represents a positive or negative / pro or contra opinion. Make sure to use the following categories only for the former kind of sentiment analysis. The latter kind is coded with category Kc05.)

**Basic emotions:**

- Em02 – Anger
- Em03 – Hatred
- Em04 – Contempt / disgust / nausea
- Em05 – Fear
- Em06 – Sadness
- Em07 – Pity / sympathy
- Em08 – Shame / guilt
- Em09 – Surprise
- Em10 – Curiosity / interest
- Em11 – Love
- Em12 – Happiness / joy
- Em13 – Enthusiasm

**Humour and irony etc.:**

- Em14 – Humour
- Em15 – Irony, sarcasm, cynicism

**Em16 – Other form of emotionality:** ______ – Note the other form of emotionality of comments the variables measure.

## Kinds of content

**Filter variable:**

Indicate whether the one of more variables are concerned with measuring if / how much of a particular kind of content is included in comments, e.g.: whether the common is on-topic or off-topic, contains media criticism, a personal opinion, judgment or evaluation, an argument, etc.

Note: Some kinds of content have already been mentioned above. E.g., the inclusion of a joke might better be coded as “humour” in the emotionality-section; personal insults falls in the category “hostility” in the (in-)civility-section; etc.

0 = no -> skip Kc02-Kc15  
1 = yes -> proceed with Kc02-Kc15
### Kinds of content of comments

<table>
<thead>
<tr>
<th>Multiple coding category</th>
<th>Indicate all kinds of content of comments one or more variables are concerned with.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 = no</td>
<td></td>
</tr>
<tr>
<td>1 = yes</td>
<td></td>
</tr>
</tbody>
</table>

- **Kc02 – On/off-topic** – i.e., one or more variables measure to what extent a comment refers to the topic of the article commented on.
- **Kc03 – Media criticism** – i.e., one or more variables measure to what extent a comment contains praise or criticism of the quality of the news story commented on or journalism in general.
- **Kc04 – Personal experience** – i.e., one or more variables measure to what extent a comment contains a user’s account of a personal experience.
- **Kc05 – Personal opinion, attitude, evaluation, judgement, verdict** – i.e., one or more variables measure to what extent a comment contains the commenting user’s personal opinion/judgment about or evaluation of a specific, non-media criticism related aspect, e.g. a political decision or idea described in the news story commented on. Pro/contra analyses fall in this category. (This is the other kind of sentiment analysis mentioned above under Em01 and Em02-Em16.)
- **Kc06 – Argument for opinion** – i.e., one or more variables measure to what extent the commenting user backs his opinion or judgment with an argument/explained how he is come to it. (This an aspect of rationality/deliberative quality.)
- **Kc07 – Additional argument** – i.e., one or more variables measure to what extent a comment contains additional arguments for a specific opinion on the topic of the news story commented on, i.e. arguments which have not been mentioned in the news story. (This is an aspect of deliberative quality.)
- **Kc08 – Additional information, leads, material, etc.** – i.e. one or more variables measure to what extent a comment contains additional information on or leads or material referring to the topic of the news story commented on, e.g. if a user directs the authors of the news story towards a scientific study on the topic or links to other news articles on the topic. (This is an aspect of deliberative quality.)
- **Kc09 – Reference or link to external source** – i.e., one or more variables measure to what extent a comment contains a reference or link to an external source, e.g. another news story, a scientific study, etc. (This is an aspect of deliberative quality.)
- **Kc10 – Frame, perspective, etc.** – i.e., one or more variables measure to what extent a comment contains one or more frames/ framings of a news story or its topic or perspectives on it. (This is an aspect of deliberative quality.)

These aspects can, but need not necessarily go along with a personal opinion, attitude, etc. on the news story or its topic (Kc05). For clarification: “Different frames highlight different aspects of a situation, construct different interpretations” (Baden/Springer 2014: 530; own emphasis). Similar to this, a perspective on a topic represents what you measure (e.g., the economic consequences of a political decision) and by which criteria you do so (e.g., the expected effects on the GDP), while an opinion...
represents the concrete *result* of that measurement (e.g., if you think these economic consequences are favourable or not).

However, a frame can “imply different evaluative judgments and suitable courses of action” (ibid.) and a perspective can be blended with an opinion on the story or its topic. In that case, Kc05 applies to this/these variable(s), too.

Also, make sure to note the different frames measured in the name of the variable.

**Kc11 – Additional frame, perspective, etc.** – i.e., one or more variables measure to what extent a comment provides an *additional* framing of or perspective on the news story, i.e. a frame/perspective which has not already been mentioned in the news story commented on. (This is an aspect of deliberative quality.) E.g., user comments might *complement* a news story about a political decision with ideas about what this decision might mean in *economical* terms.

**Kc12 – Propaganda** – i.e., one or more variables measure to what extent a comment contains propaganda. The coding depends on what the authors of the study at hand define as “propaganda”, e.g. whether the mere expression of an (extremist) opinion is counted as such.

**Kc13 – Reaction to other comment** – i.e., one or more variables measure to what extent a user’s comment is the reaction to a previous comment by another user or refers to it. Note that this does not necessarily mean that the previous comment’s author is also addressed in the comment (“Ad06 – Specific user”).

**Kc14 – Mentioning of specific persons** – i.e., one or more variables measure to what extent specific persons are mentioned in a comment, e.g. the protagonists of the news story commented on. This doesn’t necessarily mean that these persons are also addressees of the comments.

**Kc15 – Other kind of content: ______** – Note the other kind of content the variable(s) is/are concerned with.

| 58. | Re01 – Readability / comprehensibility | Readability / comprehensibility  
| 0 = no | skip Re02-Re06  
| 1 = yes | proceed with Re02-Re06 | Filter variable: 
Indicate whether one or more variables measure how comprehensible or readable a comment is, how many typos or errors it contains, etc. |

| 59. | Re02-Re06 – Readability / comprehensibility – subcategories | Readability / comprehensibility of comments  
| Multiple coding category  
| 0 = no  
| 1 = yes  
| or: Note other form of readability / comprehensibility. | Only if “Readability / comprehensibility” = “1” = “yes”  
Indicate all aspects of readability / comprehensibility one or more variables measure.  

**Re02 – Length of sentences**  
**Re03 – (Complexity of) sentence structure**  
**Re04 – Technical terms / foreign words** – i.e., one or more variables measure to what extent a comment contains technical terms
<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Re05 – Typos / errors</strong> – i.e., one or more variables measure to what extent a comment contains spelling, grammatical, punctuation errors, etc.</td>
<td></td>
</tr>
<tr>
<td><strong>Re06 – Other aspect of readability / comprehensibility:</strong> ______ – Note the other aspect of readability / comprehensibility of comments measured by the variable(s).</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Note:</strong> The aspect “length of comments” falls in the category of “Quantitative aspects of comments”.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th><strong>60. Fa01 – Facticity</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Facticity</strong></td>
<td>Coding category</td>
<td></td>
</tr>
<tr>
<td>0 = no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 = yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indicate whether one or more variables measure to what extent the facts stated by the commenting user are correct, e.g. by comparing the user’s statement on climate change with the current scientific knowledge about it. (This is an aspect of deliberative quality.)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th><strong>61. Ot01 – Other focus of variable(s)</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable(s) (also) measure(s) other aspects</strong></td>
<td>Open coding category</td>
<td></td>
</tr>
<tr>
<td>0 = no</td>
<td></td>
<td></td>
</tr>
<tr>
<td>or: Note other aspect variable measures.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indicate/note whether one or more variables (additionally or solely) measure one or more aspects other than those of the aforementioned variable categories (Qa01-Fa01).</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th><strong>Appendix</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>62. Ir – Interesting or exemplary results</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Interesting or exemplary results</strong></td>
<td>Open category</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Note here if the study produced results that are either exemplary for a variable of this category (in that they are in line with other studies’ results on this variable) or particularly surprising, counterintuitive, strange, funny, graphic or noteworthy in any other form. Include page numbers.</td>
<td></td>
</tr>
</tbody>
</table>

|   | **63. Na – Notes and additional info** |   |
| **Notes and additional info** | Open category |   |
|   | Note here, if, for instance, ...   |
| 1. the study analyses comments on a highly unusual topic; |   |
| 2. the study investigates an unusual research question; |   |
| 3. the study compares different groups or samples of comments with respect to certain variables; |   |
| 4. an unusual or innovative methodology is employed; |   |
| 5. the authors mention limitations of their study; |   |
| 6. the authors apply an algorithm or tool which could be of use for us; |   |
| 7. etc. |   |
| You can also provide a short assessment of the study's substance and methodological quality. |   |
| 64. | Cp – Coding problems | Coding problems Open category | Note any problems you experienced during the coding of the study. |