International Journal on Digital Libraries (2019) 20:391–415
https://doi.org/10.1007/s00799-018-0261-y

Automated identification of media bias in news articles: an interdisciplinary literature review

Felix Hamborg[1] · **Karsten Donnay**[1] · **Bela Gipp**[1]

Received: 21 June 2018 / Revised: 8 October 2018 / Accepted: 29 October 2018 / Published online: 16 November 2018

Abstract
Media bias, i.e., slanted news coverage, can strongly impact the public perception of the reported topics. In the social sciences, research over the past decades has developed comprehensive models to describe media bias and effective, yet often manual and thus cumbersome, methods for analysis. In contrast, in computer science fast, automated, and scalable methods are available, but few approaches systematically analyze media bias. The models used to analyze media bias in computer science tend to be simpler compared to models established in the social sciences, and do not necessarily address the most pressing substantial questions, despite technically superior approaches. Computer science research on media bias thus stands to profit from a closer integration of models for the study of media bias developed in the social sciences with automated methods from computer science. This article first establishes a shared conceptual understanding by mapping the state of the art from the social sciences to a framework, which can be targeted by approaches from computer science. Next, we investigate different forms of media bias and review how each form is analyzed in the social sciences. For each form, we then discuss methods from computer science suitable to (semi-)automate the corresponding analysis. Our review suggests that suitable, automated methods from computer science, primarily in the realm of natural language processing, are already available for each of the discussed forms of media bias, opening multiple directions for promising further research in computer science in this area.

Keywords News bias · News slant · Natural language processing (NLP)

1 Introduction

The Internet has increased the degree of self-determination in how people gather knowledge, shape their own views, and engage with topics of societal relevance [1]. Unrestricted access to unbiased information is crucial for forming a well-balanced understanding of current events. For many individuals, news articles are the primary source to attain such information. News articles thus play a central role in shaping personal and public opinion. Furthermore, news consumers rate news articles as having the highest quality and trustworthiness compared to other media formats, such as TV or radio broadcasts, or more recently, social media [1–3]. However, media coverage often exhibits an internal bias, reflected in news articles and commonly referred to as *media bias*. Factors influencing this bias can include ownership or source of income of the media outlet, or a specific political or ideological stance of the outlet and its audience [4].

The literature identifies numerous ways in which media coverage can manifest bias. For instance, journalists *select events, sources*, and from these sources the *information* they want to publish in a news article. This initial selection process introduces bias to the resulting news story. Journalists can also affect the reader's perception of a topic through *word choice*, e.g., if the author uses a word with a positive or a negative connotation to refer to an entity [5], or by varying the credibility ascribed to the source [6–8]. Finally, the *placement* and *size* of an article within a newspaper or on a website determine how much attention the article will receive [9].

The impact of media bias on shaping public opinion has been studied by numerous scholars [10]. Historically, major outlets exerted a strong influence on public opinion, e.g., in elections [11, 12], or the social acceptance of tobacco

✉ Felix Hamborg
felix.hamborg@uni-konstanz.de

Karsten Donnay
karsten.donnay@uni-konstanz.de

Bela Gipp
bela.gipp@uni-konstanz.de

[1] University of Konstanz, Constance, Germany

consumption [13, 14]. The influence of media corporations has increased significantly in the past decades. Today, six corporations control 90% of the media in the USA [15]. This naturally increases the risk of media coverage being intentionally biased [16, 17]. Also on *social media*, which typically reflects a broader range of opinions, people may still be subject to media bias [18–20], despite social media being characterized by more direct and frequent interaction between users, and hence presumably more exposure to different perspectives. Some argue that social media users are more likely to actively or passively isolate themselves in a "filter bubble" or "echo chamber" [21], i.e., only be surrounded by news and opinions close to their own. However, this isolation is not necessarily as absolute as often assumed, e.g., Barberá et al. [22] find noticeable isolation for political issues but not for others, such as reporting on accidents and disasters. Recent technological developments are another reason for topical isolation of social media consumers, which might lead to a general decrease in the diversity of news consumption. For instance, Facebook, the world's largest social network with more than one billion active users [23], recently introduced *Trending Topics*, a news overview feature. Users can now discover current events by exclusively relying on Facebook. However, the consumption of news from only a single distributor amplifies the previously mentioned level of influence further: Only a single company controls what is shown to news consumers.

The automated identification of media bias, and the analysis of news articles in general, have recently gained attention in computer science. A popular example are news aggregators, such as *Google News*, which give news readers a quick overview of a broad news landscape. Yet, established systems currently provide no support for showing the different perspectives contained in articles reporting on the same news event. Thus, most news aggregators ultimately tend to facilitate media bias [24, 25]. Recent research efforts aim to fill this gap and reduce the effects of such biases. However, the approaches suffer from practical limitations, such as being fine-tuned to only one news category, or relying heavily on user input [26–28]. As we show in this article, an important reason for the comparably poor performance of the technically superior computer science methods for automatic identification of instances of media bias is that such approaches currently tend to not make full use of the knowledge and expertise on this topic from the social sciences.

This article is motivated by the question of how computer science approaches can contribute to identifying and mitigating media bias by ultimately making available a more balanced coverage of events and societal issues to news consumers. We address this question by comparing and contrasting established research on the topic of media bias in the social sciences with the state-of-the-art technical approaches from computer science. This comparative review thus serves as a guide for computer scientists to better benefit from already more established media bias research in the social sciences. Similarly, social scientists seeking to apply powerful, state-of-the-art approaches from computer science to their own media bias research will also benefit from this review.

The remainder of this article is structured as follows: In Sect. 2, we introduce the term media bias, and describe its effects (Sect. 2.1), develop a conceptual understanding of how media bias arises in the process of news production (Sect. 2.2), and briefly introduce the most important approaches from the social sciences to analyze bias in the media (Sect. 2.3). Each of the subsections in Sect. 3 focuses on a specific *form* of media bias, describes studies from the social sciences that analyze the specified form of media bias, and discusses methods from computer science that either have been used, or could be used, to automatically identify the specified form of bias. In Sect. 4, we discuss the reliability and generalizability of the manual approaches from the social sciences and point out key issues to be considered when evaluating interdisciplinary research on media bias. In Sect. 5, we conclude the article with a discussion of the main findings of our literature review.

2 Media bias

The study of biased news reporting has a long tradition in the social sciences going back at least to the 1950s [29]. In the classical definition of Williams, media bias must both be intentional, i.e., reflect a conscious act or choice, and it must be sustained, i.e., represent a systematic tendency rather than an isolated incident [30]. In this article, we thus focus on *intentional media bias,* which journalists and other involved parties implement purposely to achieve a specific goal [13]. This definition sets the media bias that we consider apart from other sources of *unintentional* bias in news coverage. Source of unintentional bias include the influence of *news values* [31] throughout the production of news [27], and later the news consumption by readers with different backgrounds [7]. Examples for news values include the geographic vicinity of a newsworthy event to the location of the news outlet and consumers, or the effects of the general visibility or societal relevance of a specific topic [32].

Various definitions of media bias and its specific forms exist, each depending on the particular context and research questions studied. Mullainathan and Shleifer define two high-level types of media bias concerned with the intention of news outlets when writing articles: *ideology* and *spin* [33]. Ideological bias is present if an outlet biases articles to promote a specific opinion on a topic. Spin bias is present if the outlet attempts to create a memorable story. A second definition of media bias that is commonly used distinguishes between three types: *coverage*, *gatekeeping*, and *statement* (cf. [34]).

Coverage bias is concerned with the visibility of topics or entities, such as a person or country, in media coverage. Gatekeeping bias, also called selection bias or agenda bias, relates to which stories media outlets select or reject for reporting. Statement bias, also called presentation bias, is concerned with how articles choose to report on concepts. For example, in the US elections, a well-observed bias arises from *editorial slant* [35], in which the editorial position on a given presidential candidate affects the quantity and tone of a newspaper's coverage. Further forms of media bias can be found in the extensive discussion by D'Alessio and Allen [34].

2.1 Effects of biased news consumption

Media bias has a strong impact on both individual and public perception of news events, and thus impacts political decisions [10, 36–41]. Despite the rise of social media, news articles published by well-established media outlets remain the primary source of information on current events (cf. [1–3]). Thus, if the reporting of a news outlet is biased, readers are prone to adopting similarly biased views. Today, the effects of biased coverage are amplified by social media, in which readers tend to "follow" only the news that conforms with their established views and beliefs [42–46]. On social media, news readers thus encounter an "echo chamber," where their internal biases are only reinforced. Furthermore, most news readers only consult a small subset of available news outlets [47], as a result of information overload, language barriers, or their specific interests or habits.

Nearly all news consumers are affected by media bias [11, 12, 48–50], which may, for example, influence voters and, in turn, influence election outcomes [11, 12, 35, 51, 52]. Another effect of media bias is the polarization of public opinion [21], which complicates agreements on contentious topics. These strong negative effects have led some researchers to believe that media bias challenges the pillars of our democracy [40, 41]: If media outlets influence public opinion, is the observed public opinion really the "true" public opinion? For instance, a 2003 survey showed that there were significant differences in the presentation of information on US television channels. Fox News viewers were most misinformed about the Iraq war. Over 40% of viewers believed that weapons of mass destruction were actually found in Iraq [50], which is the reason used by the US government to justify the war.

According to social science research, the three key ways in which media bias affects the perception of news are *priming*, *agenda setting*, and *framing* [35], [53]. Priming theory states that how news consumers tend to evaluate a topic is influenced by their (prior) perception of the specific issues that were portrayed in news on that topic. Agenda setting refers to the ability of news publishers to influence which topics are considered relevant by selectively reporting on topics of their choosing. News consumers' evaluation of topics is furthermore based on the perspectives portrayed in news articles, which are also known as *frames*, [54]. Journalists use framing to present a topic from their perspective to "promote a particular interpretation" [55].

We illustrate the effect of framing using an example provided by Kahneman and Tversky [41]: Assume a scenario in which a population of 600 people is endangered by an outbreak of a virus. In a first survey, Kahneman and Tversky asked participants which option they would choose:

A. 200 people be will be saved.
B. 66% chance that 600 people will be saved. 33% chance that no one will be saved.

In the first survey, 72% of the participants chose A, and 26% chose B. Afterward, a second survey was conducted that objectively represents the exact same choices, but here the options to choose from were framed in terms of likely deaths rather than lives saved.

C. 400 people will die.
D. 66% chance that no one will die. 33% chance that 600 people will die.

In this case, the preference of participants was reversed. 22% of the participants chose C, and 72% chose D. The results of the survey thus demonstrated that framing alone, that is, the way in which information is presented, has the ability to draw attention to either the negative or the positive aspects of an issue [41].

In summary, the effects of media bias are manifold and especially dangerous when individuals are unaware of the occurrence of bias. The recent concentration of the majority of mass media in the hands of a few corporations amplifies the potential impact of media bias of individual news outlets even further.

2.2 Understanding media bias

Understanding not only various forms of media bias but also at which stage in the news production process they can arise [27] is beneficial to devise methods and systems that help to reduce the impact of media bias on readers. We focus on a specific conceptualization of the news production process, depicted in Fig. 1, which models how media outlets turn events into news stories and how then readers consume the stories (cf. [6, 27, 38, 56–58]). The stages in the process map to the forms of bias described by Baker et al. [6]. Since each stage of the process is distinctively defined, we find this conceptualization of the news production process and the included bias forms to be the most comprehensive model of media bias for the purpose of devising future research in

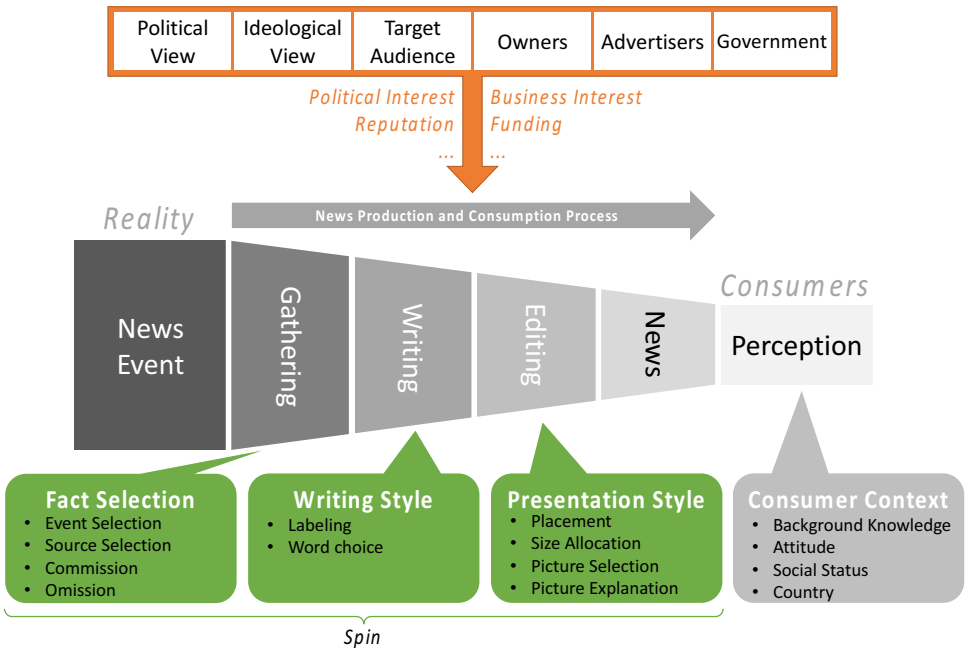

Fig. 1 Motives underlying media bias and forms of media bias introduced in the news production process. The "consumer context" label (far right) additionally shows factors influencing the perception of the described news event that are not related to media bias. Adapted from [27]

computer science. In the following paragraphs, we exemplarily demonstrate the different forms of media bias within the news production and consumption process. In Sect. 3, we discuss each form in more detail. Note that while the process focuses on news articles, most of our discussion in Sect. 3 can be adapted to other media types, such as social media, blogs, or transcripts of newscasts.

Various parties can directly or indirectly, intentionally or structurally influence the news production process (refer to the motives underlying media bias shown in the orange rectangle in Fig. 1). News producers have their own *political* and *ideological views* [59]. These views extend through all levels of a news company, e.g., news outlets and their journalists typically have a slant toward a certain political direction [42]. Journalists might also introduce bias in a story if the change is supportive of their career [60]. In addition to these internal forces, external factors may also influence the news production cycle. News stories are often tailored for a current *target audience* of the news outlet [42, 59, 61], e.g., because readers switch to other news outlets if their current news source too often contradicts their own beliefs and views [43–46, 61]. News producers may tailor news stories for their *advertisers and owners*, e.g., they might not report on a negative event involving one of their main advertisers or partnered companies [38, 59, 62]. Similarly, producers may bias news in favor of *governments* since they rely on them as a source of information [58, 63, 64].

In addition to these external factors, business reasons can also affect the resulting news story, e.g., investigative journalism is more expensive than copyediting prepared press releases. Ultimately, most news producers are profit-oriented companies that may not claim the provision of bias-free information to their news consumers as their main goal [65]; in fact, news consumers expect commentators to take positions on important issues, and filter important from unimportant information (cf. [66, 67]).

All these factors influence the news production process at various stages (gray). In the first stage, *gathering*, journalists *select facts* from all the news events that happened. This stage starts with the *selection* of *events,* also named story selection. Naturally, not all events are relevant to a new outlet's target audience, or sensational stories might yield more sales [42]. Next, journalists need to *select sources*, e.g., press releases, other news articles, or studies, to be used when writing an article. Ultimately, the journalists must decide which information from the sources to be included and which to be excluded from the article to be written. This step is called *commission* or *omission*, and likewise affects which perspective is taken on the event.

In the next phase, *writing*, journalists may use different *writings styles* to bias news. For instance, two forms defined in the production process are *labeling* (an event, action, attribute, etc., is labeled positively, e.g., "independent politician," whereas for the other party no label or a negative label is used), and *word choice* (how the article refers to an entity, e.g., "coalition forces" versus "invading forces").

The last stage, *editing*, is concerned with the *presentation style* of the story. This includes, for instance, the *placement* of the story and the *size allocation* (a large cover story receives more attention than a brief comment on page three), the *picture selection* (e.g., usage of emotional pictures or their size influence attention and perception of an event), and the *picture explanation* (i.e., placing the picture in context using a caption).

Table 1 Overview: forms of media bias

Name of form	Citations	Medium	Target object	Stage	Explanation/example
Event selection	[71–73]	News outlet	News article	Gathering	The news outlet rarely reports on events criticizing the government
Source selection	[42, 74, 75]	News article	Text, picture	Gathering	Inclusion of more sources that report on a certain perspective
Commission and omission	[27, 61, 76]	News article	Text	Gathering	Facts that support or question a specific perspective are added to or omitted from the article
Labeling and word choice	[5, 77, 78]	Text	Entity, action, attribute, etc.	Writing	Liberal versus conservative, expert, independent; intervene versus invade, clever versus sneaky, refugee versus immigrant
Story placement	[79]	News outlet	News article	Editing	Cover story receives more attention than a 3rd page story.
Size allocation	[73, 80]	News outlet	News article	Editing	A large story is likely to receive more attention than a small story
Picture selection	[81–83]	News article	Picture	Editing	What does the picture show? For example, fighting versus a peace flag
Picture explanation	[84]	Text	Picture caption	Editing	The caption puts the picture into context and may either support or criticize what is pictured
Spin	[57, 85–87]	News article and outlet	One or more news articles	All phases	The overall *slant* of the article, i.e., the result when the various types of news bias are combined

The second column contains for each form of bias references to an exemplary study from the social sciences, and the most relevant publications from computer science, if any

Lastly, *spin bias* is a form of media bias that represents the overall bias of a news article. An article's spin is essentially a combination of all previously mentioned forms of bias and other minor forms (see Sect. 3.8).

In summary, the resulting news story has potentially been subject to various sources of media bias at different stages of the story's genesis before it is finally consumed by the reader. The *consumer context*, in turn, affects how readers actually perceive the described information (cf. [68, 69]). The perception of any event will differ, depending on the readers' *background knowledge*, their preexisting *attitude* toward the described event (sometimes called *hostile media perception*) [70], their *social status* (how readers are affected by the event), and *country* (news reporting negatively about a reader's country might lead to refusal of the discussed topic), and a range of other factors. Note, however, that "consumer context" is not a form of media bias, and thus will be excluded from analysis in the remainder of this article.

Other models exist of how media bias arises, but their components can effectively be mapped to the news production and consumption process detailed previously. For instance, Entman defines a *communication process* that essentially mirrors all the same steps discussed in Fig. 1: (1) Communicators make intentional or unintentional decisions about the content of a text. (2) The text inherently contains different forms of media bias. (3) Receivers, i.e., news readers, draw conclusions based on the information and style presented in the text (which, however, may or may not reflect the text's perspective). (4) Receivers of a social group are additionally subject to *culture*, also known as a common set of perspectives [54].

Table 1 gives an overview of the previously described forms of media bias, where the "Medium" column shows the medium that is the source of the specific form of bias, and the column "Target Object" shows the items within the target medium that are affected.

2.3 Approaches in the social sciences to analyze media bias

Researchers from the social sciences primarily conduct so-called *content analyses* to identify and quantify media bias in news coverage [34], or to, more generally, study patterns

in communication. First, we briefly describe the concept and workflow of content analysis. Next, we describe the concept of *frame analysis*, which is a specialized form of content analysis commonly used to study the presence of frames in news coverage [88]. Lastly, we introduce *meta-analysis*, in which researchers combine the findings from other studies and analyze general patterns across these studies [89].

2.3.1 Content analysis

Content analysis quantifies media bias by identifying and characterizing its instances within news texts. In a content analysis, researchers first define one or more analysis questions or hypotheses. Researchers then gather the relevant news data, and coders systematically read the news texts, annotating parts of the texts that indicate instances of media bias relevant to the analysis being performed. Afterward, the researchers use the annotated findings to accept or reject their hypotheses [90, 91].

In a *deductive* content analysis, researchers devise a *codebook* before coders read and annotate the texts [92, 93]. The codebook contains definitions, detailed rules, and examples of what should be annotated and in which way. Sometimes, researchers reuse existing codebooks, e.g., Papacharissi and Oliveira used annotation definitions from a previous study by Cappella and Jamieson [94] to create their codebook, then they performed a deductive content analysis comparing news coverage on terrorism in the USA and the UK [77]. In an *inductive* content analysis, coders read the texts without specified instructions on how to code the text, only knowing the research question [42]. Since statistically sound conclusions can only be derived from the results of deductive content analyses [95], researchers conduct inductive content analyses mainly in early phases of their research, e.g., to verify the line of research, or to find patterns in the data and devise a codebook [88, 95]. Usually, creating and refining the codebook is a time-intensive process, during which multiple analyses or tests using different iterations of a codebook are performed. Achieving a sufficiently high inter-coder reliability (ICR), e.g., when the individual coders annotate the same parts of the documents with same codes from the codebook, is a common criterion that must be satisfied before the final deductive analysis can be conducted [96].

Social scientists distinguish between two types of content analyses: quantitative and qualitative [97]. A qualitative analysis seeks to find "all" instances of media bias, including subtle instances that require human interpretation of the text. In a quantitative analysis, researchers in the social sciences determine the frequency of specific words or phrases (usually as specified in a codebook). Quantitative content analyses may also measure other, non-textual features of news articles, such as the number of articles published by a news outlet on a certain event, or the size and placement of a story in a printed newspaper. These measurements are also called *volumetric measurements* [34].

Thus far, the majority of studies on media bias performed in the social sciences conduct qualitative content analyses because the findings tend to be more comprehensive. Quantitative analyses can be performed faster and can by partially automated, but are more likely to miss subtle forms of bias [98].

Content analysis software, generally also called *computer-assisted qualitative data analysis software (CAQDAS)*, supports analysts when performing content analyses [99]. Most tools support the manual annotation of findings for the analyzed news data or for other types of reports, such as police reports [90]. To reduce the large amount of texts that need to be reviewed, the software helps users find relevant text passages, e.g., by finding documents or text segments containing the words specified in the codebook or from a keyword list [100] so that the coder must review less texts manually. In addition, most software helps users find patterns in the documents, e.g., by analyzing the frequencies of terms, topic, or word co-occurrences [99].

2.3.2 Frame analysis

Frame analysis investigates how readers perceive the information in a news article [54]. This is done by broadly asking two questions: (1) *what* information is conveyed in the article? (2) *How* is that information conveyed? Both questions together define a "frame." As described in Sect. 2.1, a frame is a selection of and emphasis on specific parts of an event.

Not all frame analyses focus on the text of news articles. For instance, DellaVigna and Kaplan analyzed the gradual adoption of cable TV of Fox News between 1996 and 2000 to show that Fox News had a "significant impact" on the presidential elections [51]. Essentially, the study analyzed whether a district had already adopted the Fox News channel, and what the election result was. The results revealed that the Republican party had an increased vote share in those towns that had adopted Fox News.

2.3.3 Meta-analysis

In a *meta-analysis*, researchers combine the results of multiple studies to derive further findings from them [89]. For example, in the analysis of event selection bias, a common question is which factors influence whether media organizations will choose to report on an event or not. McCarthy et al. [32] performed a meta-analysis of the results of prior work suggesting that the main factors for media to report on a demonstration are the demonstration size and the previous media attention on the demonstration's topic.

2.4 Summary

News coverage has a strong impact on public opinion, i.e., what people think about (*agenda setting*), the context in which news is perceived (*priming*), or how topics are communicated (*framing*). Researchers from the social sciences have extensively studied such forms of media bias, i.e., the intentional, non-objective coverage of news events. The extensive research has resulted in a broad literature on different forms and possible sources of media bias and their impact on (political) communication or opinion formation. In tandem, various well-established research methodologies, such as content analysis, frame analysis, and meta-analysis, have emerged in the social sciences.

The three forms of analysis discussed in Sect. 2.3 require significant manual effort and expertise [27], since those analyses require human interpretation of the texts and cannot be fully automated. For example, a quantitative content analysis might (semi-)automatically count words that have previously been manually defined in a codebook but they would be unable to read for "meaning between the lines," which is why such methods continue to be considered less comprehensive than a qualitative analysis. However, the recent methodological progress in natural language processing in computer science promises to help alleviate many of these concerns.

In the remainder of this article, we discuss different forms of media bias defined by the news production and consumption process. The process we have laid out in detail previously is in our view the most suitable conceptual framework to map analysis workflows from the social sciences to computer science, and thus helps us to discuss where and how computer scientists can make unique contributions to the study of media bias.

3 Manual and automated approaches to identify media bias

This section is structured into eight subsections discussing all of the forms of media bias depicted in Table 1. In each subsection, we first introduce each form of bias and then provide an overview of the studies and techniques from the social sciences used to analyze that particular form. Subsequently, we describe methods and systems that have been proposed by computer science researchers to identify or analyze that specific form of media bias. Since media bias analysis is a rather young topic in computer science, often no or few methods have been specifically designed for that specific form of media bias, in which case, we describe the methods that could best be used to study the form of bias. Each subsection concludes with a summary of the main findings highlighting where and how computer science research can make a unique contribution to the study of media bias.

3.1 Event selection

From the countless stream of events happening each day, only a small fraction can make it into the news. Event selection is a necessary task, yet it is also the first step to bias news coverage. The analysis of this form of media bias requires both an event-specific and a long-term observation of multiple news outlets. The main question guiding such an analysis is whether an outlet's coverage shows topical patterns, i.e., some topics are reported more or less in one as compared to another outlet, or which factors influence whether an outlet reports on an event or not.

To analyze event selection bias, at least two datasets are required. The first dataset consists of news articles from one or more outlets; the second is used as a ground truth or baseline, which ideally contains "all" events relevant to the analysis question. For the baseline dataset, researchers from the social sciences typically rely on sources that are considered to be the most objective, such as police reports [71]. After linking events across the datasets, a comparison enables researchers to deduce factors that influence whether a specific news outlet reports on a given event. For instance, several studies compare demonstrations mentioned in police reports with news coverage on those demonstrations [32, 90, 91]. During the manual content analyses, the researchers extracted the type of event, i.e., whether it was a rally, march, or protest, the issue the demonstration was about, and the number of participants. Two studies found that the number of participants and the issue of the event, e.g., protests against the legislative body [90], had a high impact on the frequency in news coverage [71].

Meta-analyses have also been used to analyze event selection bias, mainly by summarizing findings from other studies. For instance, D'Alessio and Allen [34] found that the main factors influencing media reporting on demonstration are the demonstration size and the previous media attention on the demonstration's topic.

To our knowledge, in computer science, only few approaches have been proposed that specifically aim to analyze event selection bias. Other than in social sciences studies, none of them compare news coverage with a baseline that is considered objective, but they compare the coverage of multiple outlets or other online news sources [72, 73]. First, we describe these approaches in more detail, then we also describe state-of-the-art methods and systems that can support the analysis of this form of bias.

Bourgeois et al. span a matrix over news sources and events extracted from GDELT [101], where the value of each cell in the matrix describes whether the source (row) reported on the event (column) [99]. They use matrix factorization (MF) to extract "latent factors," which influence whether a source reports on an event. Main factors found were the affiliation, ownership, and geographic proximity of two sources.

Saez-Trumper et al. analyze relations between news sources and events [79]. By analyzing the overlap between news sources' content, they find, for example, that news agencies, such as AP, publish most non-exclusive content—i.e., if news agencies report on an event, other news sources will likely also report on the event—and that news agencies are more likely to report on international events than other sources. Media type was also a relevant event selection factor. For example, magazine-type media, such as The Economist, are more likely to publish on events with high prominence, i.e., events that receive a lot of attention in the media.

Similar to the manual analyses performed in the social sciences, automated approaches need to (1) find articles relevant to the question being analyzed (we describe relevant techniques later in this subsection, see the paragraphs on news aggregation), (2) link articles to baseline data or other articles, and (3) compute statistics on the linked data. In addition to automating analysis relevant for researchers studying media bias, we believe that news aggregators could thus also be used to reveal event selection bias to news consumers, e.g., by providing visual cues on how selective reporting is.

In task (2), we have to distinguish whether one wants to compare articles to a baseline, or technically said, across different media, or to other articles. Linking events from different media, e.g., news articles and tweets on the same events, has recently gained attention in computer science [73, 102]. However, to our knowledge, there are currently no *generic* methods to extract the required information from police reports or other, non-media databases, since the information that needs to be extracted strongly depends on the particular question studied and the information structure and format differs greatly between these documents, e.g., police reports from different countries or states usually do not share common formats (cf. [103, 104]).

To link news articles reporting on the same event, various techniques can be used. *Event detection* extracts events from text documents. Since news articles are usually concerned with events, event detection is commonly used in news related analyses. For instance, in order to group related articles, i.e., those reporting on the same event [105], one needs to first find events described in these articles. *Topic modeling* extracts semantic concepts, or topics, from a set of text documents where topics are typically extracted as lists of weighted terms. A commonly employed implementation is Latent Dirichlet Allocation (LDA) [106], which is, for instance, used in the Europe Media Monitor (EMM) news aggregator [107].

Related articles can also be grouped with the help of *document clustering* methods. Hierarchical agglomerative clustering (HAC) [108] computes pair-wise document similarity on text-features using measures such as the cosine distance on TF-IDF vectors [109] or word embeddings [110]. This way HAC creates a hierarchy of the most similar documents and document-groups [111]. HAC has been used successfully in several research projects [27, 112]. Another state-of-the-art clustering method is affinity propagation [113]. Other methods to group related articles exploit news-specific characteristics, such as the *five journalistic W questions* (5Ws). The 5Ws describe the main event of a news article, i.e., who did what, when, where, and why. Journalists usually answer the 5W questions within the first few sentences of a news article [85]. 5Ws extraction approaches automatically extract phrases that answer the 5Ws [114, 115]. These phrases can then be used to group articles that report on the same main event (cf. [114]).

News aggregation[1] is one of the most popular approaches to enable users to get an overview of the large amounts of news that is published nowadays. Established news aggregators, such as Google News and Yahoo News, show related articles by different outlets reporting on the same event. Hence, the approach is feasible to reveal instances of bias by source selection, e.g., if one outlet does not report on an important event. News aggregators rely on methods from computer science, particularly methods from natural language processing (NLP). The analysis pipeline of most news aggregators aims to find the most important news topics, and present them in a compressed form to users. The analysis pipeline typically involves the following tasks [116, 117]:

1. *Data gathering*, i.e., crawling articles from news websites.
2. *Article extraction* from website data, which is typically HTML or RSS.
3. *Grouping*, i.e., finding and grouping related articles reporting on the same topic or event.
4. *Summarization* of related articles.
5. *Visualization*, e.g., presenting the most important topics to users.

For the first two tasks, data gathering and article extraction, established and reliable methods exist, e.g., in the form of web crawling frameworks [118]. Articles can be extracted with naive approaches, such as website-specific wrappers [119], or more generic methods based on content heuristics [120]. Combined approaches perform both crawling and extracting, and offer other functionality tailored to news analysis. For example, *news-please*, a web crawler and extractor for news articles, can extract information from all news articles on a website, given only the root URL of the news outlet to be crawled [121].

The objective of grouping is to identify topics and group articles on the same topic, e.g., using LDA or other topic modeling techniques, as described previously. Articles are then

[1] The paragraphs about news aggregation have been adapted partially from [118].

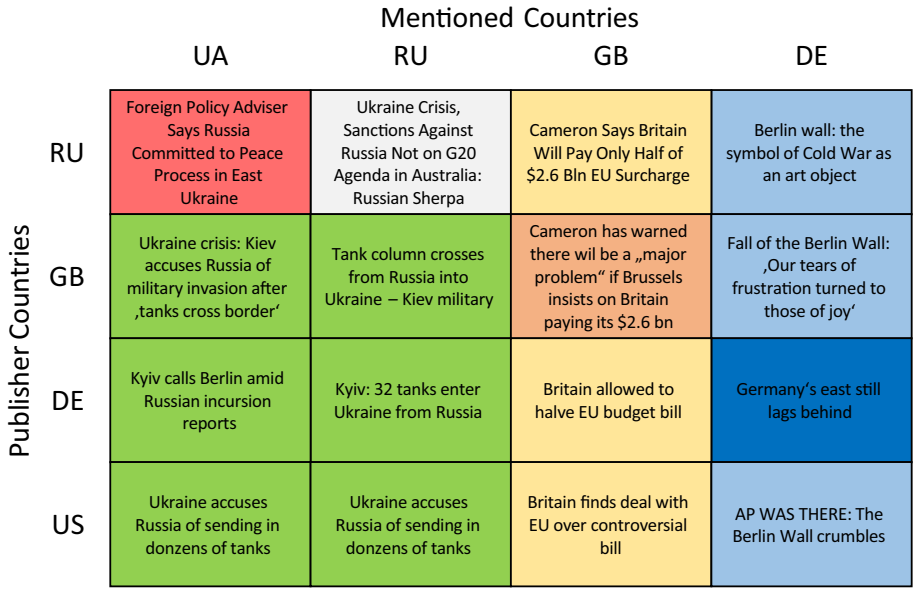

Fig. 2 Matrix news overview to enable comparative news analysis in MNA. The color of each cell refers to its main topic. *Source*: [123]

summarized using methods such as simple TF-IDF-based scores or complex approaches considering redundancy and order of appearance [122]. By performing the five tasks of the news aggregation pipeline in an automatized fashion, news aggregators can cope with the large amount of information produced by news outlets every day.

However, no established news aggregator reveals event selection bias of news outlets to their users. Incorporating this functionality for short-term or event-oriented analysis of event selection bias, news aggregators could show the publishers that did *not* publish an article on a selected event. For long-term or trend-oriented analysis, news aggregators could visualize a news outlet's coverage frequency of specific topics, e.g., to show whether the issues of a specific politician or party, or an oil company's accident is either promoted or demoted.

In addition to traditional news aggregators, which show topics and related topics in a list, recent news aggregators use different analysis approaches and visualizations to promote differences in news coverage caused by biased event selection. *Matrix-based news aggregation* (MNA) is an approach that follows the analysis workflow of established news aggregators while organizing and visualizing articles into rows and columns of a two-dimensional matrix [87], [117]. The exemplary matrix depicted in Fig. 2 reveals what is primarily stated by media in one country (rows) about another country (columns). For instance, the cell of the publisher country Russia and the mentioned country Ukraine, denoted with RU-UA, contains all articles that have been published in Russia and mention Ukraine. Each cell shows the title of the most representative article, determined through a TF-IDF-based summarization score among all cell articles [87]. Users either select rows and columns from a list of given configurations for common use cases, e.g., to analyze only major Western countries, or define own rows and columns from which the matrix shall be generated.

To analyze event selection bias, users can use MNA to explore main topics in different countries as in Fig. 2, or span the matrix over publishers and topics in a country.

Research in the social sciences concerned with bias by event selection requires significant effort due to the time-consuming manual linking of events from news articles to a second "baseline" dataset. Many established studies use event data from a source that is considered "objective," for example, police reports (cf. [90, 104, 124]). However, the automated extraction of relevant information from such non-news sources requires the development and maintenance of specialized tools for each of the sources. Reasons for the increased extraction effort include the diversity or unavailability of such sources, e.g., police reports are structured differently in different countries or may not be published at all. Linking events from different sources in an automated fashion poses another challenge because of the different ways in which the same event may be described by each of the sources. This places a limit on the possible contributions of automated approaches for comparison across sources or articles.

In our view, the automated analysis of events within news articles, however, is a very promising line of inquiry for computer science research. Sophisticated tools can already gather and extract relevant data from online news sources. Methods to link events in news articles are already available or are the subject of active research [27, 105–107, 109, 111, 112, 114]. Of course, news articles must originate from a carefully selected set of news publishers, which not only represent mainstream media but also alternative and independent pub-

lishers, such as Wikinews.[2] Finally, revealing differences in the selection of top news stories between publishers, or even the mass media of different countries has shown promising results [117], and could eventually be integrated into regular news consumption using news aggregators demonstrating the potential for computer science approaches to make a unique contribution to the study event selection.

3.2 Source selection

Journalists must decide on the trustworthiness of information sources and the actuality of information for a selected event. While source selection is a necessary task to avoid information overload, it may lead to biased coverage, e.g., if journalists mainly consult sources supporting one perspective when writing the article. The choice of sources used by a journalist or an outlet as a whole can reveal patterns of media bias. However, journalistic writing standards do not require journalists to list sources [81], which make the identification of original sources difficult or even impossible. One can only find hints in an article, such as the use of quotes, references to studies, phrases such as "according to [name of other news outlet]" [5], or the dateline, which indicates whether and from which press agency the article was copyedited. One can also analyze whether the content and the argumentation structure match those of an earlier article [92].

The effects of source selection bias are similar to the effects of commission and omission (Sect. 3.3), because using only sources supporting one side of the event when writing an article (source selection) is similar to omitting all information supporting the other side (omission). Because many studies in the social sciences are concerned with the *effects* of media bias, e.g., [10–12, 36, 38–41, 48–50, 61], and the effects of these three bias forms are similar, bias by source selection and bias by commission and omission are often analyzed together.

Few analyses in the social sciences aim to find the selected sources to derive insights on the source selection bias of an article or an outlet. However, there are notable exceptions, for example, one study counts how often news outlets and politicians cite phrases originating in think tanks and other political organizations. The researchers had previously assigned the organizations to a political spectrum [42]. The frequencies of specific phrases used in articles, such as "We are initiating this boycott, because we believe that it is racist to fly the Confederate Flag on the state capitol" [42], which originated in the civil rights organization NACCP, are then aggregated to estimate the bias of news outlets. In another study of media content, Papacharissi and Oliveira annotate indications of source selection in news articles, such as whether an article refers to a study conducted by the government or independent scientists [77]. One of their key findings is that UK news outlets often referred to other news articles, whereas US news outlets did that less often but referred to governments, opinions, and analyses.

On *social media*, people can be subject to their *own* source selection bias, as discussed in Sect. 1. For instance, on Facebook people tend to be friends with likewise minded people, e.g., who share similar believes or political orientations [18]. People who use social media platforms as their primary news source are subject to selection bias not only by the operating company [16, 23], but also by their friends [18].

To our knowledge, there are currently no approaches in computer science that aim to specifically identify bias by source selection. However, several automated techniques are well suited to address this form of bias. *Plagiarism detection* (PD) is a field in computer science with the broad aim of identifying instances of unauthorized information reuse in documents. Methods from PD may be used to identify the potential *sources* of information for a given article beyond identifying actual "news plagiarism" (cf. [125]). While there are some approaches focused on detecting instances of plagiarism in news, e.g., using simple text-matching methods to find 1:1 duplicates [74], research on news plagiarism is not as active as research on academic plagiarism. This is most likely a consequence of the fact that authorized copyediting is a fundamental component in the production of news. Another relevant field that we describe in this section is *semantic text similarity* (STS), which measures the semantic equivalence of two (usually short) texts [75].

The vast majority of *plagiarism detection* techniques analyzes text [126], and thus could also be adapted and subsequently applied to news texts. State-of-the-art methods can reliably detect *copy and paste* plagiarism, the most common form of plagiarism [126, 127]. *Ranking* methods use, for instance, TF-IDF and other information retrieval techniques to estimate the relevance of other documents as plagiarism candidates [128]. *Fingerprinting* methods generate hashes of phrases or documents. Documents with similar hashes indicate plagiarism candidates [128, 129].

Today's plagiarism detection methods thus already provide most of the functionality to identify the potential sources of news articles. Copyedited articles are often shortened or slightly modified, and in some cases, are a 1:1 duplicate of a press agency release. These types of slight modifications, however, can be reliably detected with ranking or fingerprinting methods (cf. [74, 126]). Current methods only continue to struggle with heavily paraphrased texts [126], but research is extending also to other non-textual data types such as analyzing links [130], an approach that can be used for the analysis of online news texts as well. Another text-independent approach to plagiarism detection is *citation-based* plagiarism detection algorithms, which achieve good results by comparing patterns of citations between two

[2] https://en.wikinews.org/wiki/Main_Page.

scientific documents [76]. Due to their text-independence, these algorithms also allow a cross-lingual detection of information reuse [76]. News articles typically do not contain citations, but the *patterns* of quotes, hyperlinks, or other named entities can also be used as a suitable marker to measure the semantic similarity of news articles (cf. [42, 130, 131]). Some articles also contain explicit referral phrases, such as "according to the New York Times." The *dateline* of an article can also state whether and from where an article was copyedited [132]. Text search and rule-based methods can be used to identify referral phrases and to extract the resources being referenced. In our view, future research should focus on identifying the span of information that was taken from the referred resource (see also Sect. 3.3).

Semantic textual similarity (STS) methods measure the semantic equivalence of two (usually short) texts [75]. State-of-the-art STS methods use basic measures, such as n-gram overlap, WordNet node-to-node distance, and syntax-based, e.g., compare whether the predicate is the same in two sentences [133]. More advanced methods combine various techniques and use deep learning networks, achieving a Pearson correlation to human coders of 0.78 [134]. Recently, these methods have also focused on *cross-lingual* STS [75], and use, for example, machine translation before employing regular mono-lingual STS methods [135]. Machine translation has proven useful also for other cross-lingual tasks, such as event analysis [87].

Graph analysis is concerned with the analysis of relations between nodes in a graph. The relation between news articles can be used to construct a dependency graph. Spitz and Gertz analyzed how information propagates in online news coverage using hyperlinks linking to other websites [136]. They identified four types of hyperlinks: *navigational* (menu structure to navigate the website), *advertisement*, *references* (links within the article pointing to semantically related sites), and *internal links* (further articles published by the same news outlet). They only used reference links to build a network, since the other link types contain too many unrelated sites (internal) or irrelevant information (advertisement and navigational). One finding by Spitz and Gertz is that networks of news articles can be analyzed with methods of citation network analysis. Another method extracts quotes attributed to individuals in news articles to follow how information propagates over time in a news landscape [131]. One finding is that quotes undergo variation over time but remain recognizable with automated methods [131].

In our view, computer science research could therefore provide promising solutions to long-standing technical problems in the systematic study of source selection by combining methods from PD and graph analysis. If two articles are strongly similar, the later published article will most likely contain reused information from the former published article. This is a typical case in news coverage, e.g., many news outlets copyedit articles from press agencies or other major news outlets [137]. Using PD, such as fingerprinting and pattern-based analysis as previously described, to measure the likelihood of information reuse between all possible pairs of articles in a set of related articles, implicitly constructs a directed dependency graph. The nodes represent single articles, the directed edges represent the flow of information reuse, and the weights of the edges represent the degree of information reuse. The graph can be analyzed with the help of methods from graph analysis, e.g., to estimate importance or slant of news outlets or to identify clusters of articles or outlets that frame an event in a similar manner (cf. [136]). For instance, if many news outlets reuse information from a specific news outlet, the higher we can rate its importance. The detection of semantic (near-)duplicates would also help lower the number of articles that researchers from the social sciences need to manually investigate to analyze other forms of media bias in content analyses.

In summary, the analysis of bias by source selection is challenging, since the sources of information are mostly not documented in news articles. Hence, in both the social sciences and in computer science research, only few studies have analyzed this form of bias. Notable exceptions are the studies discussed previously that analyzed quotes used by politicians originating from think tanks. Methods from computer science can in principle provide the required techniques for the (semi-)automated analysis of this form of bias and thus make a very valuable contribution. The methods, most importantly those from plagiarism detection research, could be (and partially already have been [74]) adapted and extended from academic plagiarism detection and other domains, where extensive and reliable methods already exist.

3.3 Commission and omission of information

Analyses of bias by commission and omission compare the information contained in a news article with those in other news articles or sources, such as police reports and other official reports. The "implementation" and effects of commission and omission overlap with those of source selection, i.e., when information supporting or opposing a perspective is either included or left out of an article. Analyses in the social sciences aim to determine which frames the information included in such articles support. For instance, frame analyses typically compare the frequencies of frame-attributing phrases in a set of news articles [61, 138]. More generally, content analysis compares which facts are presented in news articles and other sources [139]. In the following, we describe exemplary studies of each of the two forms.

A frame analysis by Gentzkow and Shapiro analyzed the frequency of phrases that may sway readers to one or the other side of a political issue [61]. For this analysis, the researchers first examined which phrases were used signif-

icantly more often by politicians of one party over another, and vice versa. Afterward, they counted the occurrence of phrases in news outlets to estimate the outlet's bias toward one side of the political spectrum. The results of the study showed that news producers have economic motives to bias their coverage toward the ideological views of their readers. Similarly, another method, briefly mentioned in Sect. 3.2, counts how often US congressmen use the phrases coined by think tanks, which the researchers previously associated with political parties [42]. One finding is that Fox News coverage was significantly slanted toward the US Republican party.

A content analysis conducted by Smith et al. investigated whether the aims of protesters corresponded to the way in which news reported one demonstrations [139]. One of their key hypotheses was that news outlets will tend to favor the positions of the government over the positions of protesters. In the analysis, Smith et al. extracted relevant textual information from news articles, transcripts of TV broadcasts, and police reports. They then asked analysts to annotate the data, and statistically tested the hypotheses using the annotated data.

Bias by commission and omission of information has not specifically been addressed by approaches in computer science despite the existence of various methods that we consider beneficial for the analysis of both forms of bias in a (semi-)automated manner. Researchers from the social sciences are already using text search to find relevant documents and phrases within documents [100]. However, search terms need to be constructed manually, and the final analysis still requires a human interpretation of the text to answer coding tasks, such as "assess the spin of the coverage of the event" [139]. Another challenge is that content analyses comparing news articles with other sources require the development of scrapers and information extractors tailored specifically to these sources. While there are mature generic tools to crawl and extract information from news articles (cf. [121]), there are no established or publicly available generic extractors for police reports.

An approach that partially addresses commission and omission of information is *aspect-level browsing* as implemented in the news aggregator *NewsCube* [27]. Park et al. define an "aspect" as the semantic proposition of a news topic. The aspect-level browsing enables users to view different perspectives on political topics. The approach follows the news aggregation workflow described in Sect. 3.1, but with a novel grouping phase: NewsCube extracts aspects from each article using keywords and syntactic rules, and weighs these aspects according to their position in the article (motivated by the *inverted pyramid* concept: the earlier the information appears in the article, the more important it is [85]). Afterward, NewsCube performs HAC to group related articles. The visualization is similar to the topic list shown in established news aggregators, but additionally shows different aspects of a selected topic. A user study found that users of NewsCube became aware of the different perspectives, and subsequently read more original articles containing perspective-attributing aspects. However, the approach cannot reliably assess the diversity of the aspects. NewsCube shows all aspects, even though many of them are similar, which decreases the efficiency of using the visualization to get an overview of the different perspectives in news coverage. Word and phrase embeddings might be used to recognize the similarity of aspects (cf. [110, 140]). The NewsCube visualization also does not highlight which information is subject to commission and omission bias, i.e., what information is contained in a particular article and left out by another article.

Methods from plagiarism detection (see Sect. 3.2) open a promising research direction for the automated detection of commission and omission of information in news. More than 80% of related news articles add no new information and only reuse information contained in previously published articles [137]. Comparing the presented facts of one article with the facts presented in previously published articles would help identify commission and omission of information. Methods from PD can detect and visualize which segments of a text may have been taken from other texts [76]. The relatedness of bias by source selection and bias by commission and omission suggests that an analysis workflow may ideally integrate methods from PD to address both issues (also see Sect. 3.2).

Centering resonance analysis (CRA) aims to find how influential terms are within a text by constructing a graph with each node representing a term that is contained in the noun phrases (NP) of a given text [141]. Two nodes are connected if their terms are in the same NP or boundary terms of two adjacent NPs. The idea of the approach is that the more edges a node has, the more influential its term is to the text's meaning. To compare two documents, methods from graph analysis can be used to analyze both CRA graphs (Sect. 3.2 gave a brief introduction to methods from graph analysis). Researchers from the social sciences have successfully employed CRA to extract influential words from articles, and then manually compare the information contained in the articles [77]. Recent advancements toward computational extraction and representation of the "meaning" of words and phrases, especially word embeddings [110], may serve as another way to (semi-)automatically compare the contents of multiple news articles.

To conclude, studies in the social sciences researching bias by commission and omission have always compared the analyzed articles with other news articles and/or non-media sources, such as police reports. No approaches from computer science research specifically aim to identify this bias form. However, automated methods, specifically PD, CRA, graph analysis, and more recent also word embeddings, are promising candidates to address this form of bias opening new avenues for unique contributions of well-

established computer science methodology in this area. CRA, for instance, has already been employed by researchers from the social sciences to compare the information contained in two articles.

3.4 Labeling and word choice

When referring to a semantic concept, such as an entity, a geographic position, or activity, authors can *label* the concept and *choose from various words* to refer to it. Instances of bias by labeling and word choice frame the referred concept differently, e.g., simply positively or negatively, or they highlight a specific perspective, e.g., economical or cultural (see Sect. 2.1 for a background on framing). Examples include "coalition forces" or "invading forces," "freedom fighters" or "terrorists" [7], or "immigrant" or "economic migrant." The effects of this form of bias range from concept-level, e.g., a specific politician is shown to be incompetent, to article-level, e.g., the overall tone of the article features emotional or factual words [77, 142].

Content analyses and framing analyses are used in the social sciences to identify bias by labeling and word choice within news articles. Similar to the approaches discussed in previous sections, the manual coding task is once again time-consuming, since annotating news articles requires careful human interpretation. The analyses are typically either *topic-oriented* or *person-oriented.* For instance, Papacharissi and Oliveira used CRA to extract influential words (see Sect. 3.3). They investigated labeling and word choice in the coverage of different news outlets on topics related to terrorism [77]. They found that the New York Times used a more dramatic tone, e.g., news articles dehumanized terrorists by not ascribing any motive to terrorist attacks or usage of metaphors, such as "David and Goliath" [77]. The Washington Post used a less dramatic tone, and both the Financial Times and the Guardian focused their news articles on factual reporting. Another study analyzed whether articles portrayed Bill Clinton, the US president at that time, positively, neutrally, or negatively [142].

The automated analysis of labeling and word choice in news texts is challenging due to limitations of current NLP methods [117], which cannot reliably interpret the frame induced by labeling and word choice, due to the frame's dependency on the context of the words in the text [7]. Few automated methods from computer science have been proposed to identify bias induced by labeling and word choice. Grefenstette et al. [5] devised a system that investigates the frequency of affective words close to words defined by the user, for example, names of politicians. They find that the automatically derived polarity scores of named entities are in line with the publicly assumed slant of analyzed news outlets, e.g., George Bush, the Republican US president at that time, was mentioned more positively in the conservative Washington Times compared to other news outlets.

The most closely related field is *sentiment analysis*, which aims to extract an author's attitude toward a semantic concept mentioned in the text [143]. Current sentiment analysis methods reliably extract the unambiguously stated sentiment [143]. For example, those methods reliably identify whether customers used "positive," such as "good" and "durable," or "negative" words, such as "poor quality," to review a product [143]. However, the highly context-dependent, hence more ambiguous sentiment in news coverage described previously in this section remains challenging to detect reliably [7]. Recently, researchers proposed approaches using *affect analysis*, e.g., using more dimensions than polarity in sentiment analysis to extract and represent emotions induced by a text, and *crowdsourcing*, e.g., systems that ask users to rate and annotate phrases that induce bias by labeling and word choice [57]. We describe these fields in the following paragraphs.

While sentiment analysis presents one of the promising methods to be used for automating the identification of bias by labeling and word choice, the performance of state-of-the-art sentiment extraction on news texts is rather poor (cf. [7, 144]). Two reasons why sentiment analysis performs poorly on news texts [7] are (1) the *lack of large-scale gold standard datasets* and (2) the *high context-dependency* of sentiment-inducing phrases. A gold standard is required to train state-of-the-art sentiment extractors using machine learning [145]. The other category of sentiment extractors use manually [146] or semi-automatically [147, 148] created dictionaries of positive and negative words to score a sentence's sentiment. However, to our knowledge no sentiment dictionary exists that is specifically designed for news texts, and generic dictionaries tend to perform poorly on such texts (cf. [7, 69, 144]). Godbole et al. [147] used WordNet to automatically expand a small, manually created seed dictionary to a larger dictionary. They used the semantic relations of WordNet to expand upon the manually added words to closely related words. An evaluation showed that the resulting dictionary had similar quality in sentiment analysis as solely manually created dictionaries. However, the performance of entity-related sentiment extraction using the dictionary tested on news websites and blogs is missing a comparison against a ground truth, such as an annotated news dataset. Additionally, dictionary-based approaches are not sufficient for news texts, since the sentiment of a phrase strongly depends on its context, for example, in economics a "low market price" may be good for consumers but bad for producers.

To avoid the difficulties of interpreting news texts, researchers have proposed approaches to perform sentiment analysis specifically on quotes [69] or on the comments of readers [149]. The motivation for analyzing only the sentiment contained in quotes or comments is that phrases stated by someone are far more likely to contain an explicit state-

ment of sentiment or opinion-conveying words. While the analysis of quotes achieved poor results [69], readers' comments appeared to contain more explicitly stated opinions and regular sentiment analysis methods perform better: A classifier that used the extracted sentiments from the readers' comments achieved a precision of 0.8 [149].

Overall, the performance of sentiment analysis on news texts is still rather poor. This is attributable to the fact that, thus far, not much research has focused on improving sentiment analysis when compared to the large number of publications targeting the prime use case of sentiment analysis: product reviews. Currently, no public annotated news dataset for sentiment analysis exists, which is a crucial requirement for driving forward successful, collaborative research on this topic.

A final challenge when applying sentiment analysis to news articles is that the one-dimensional positive–negative scale used by all mature sentiment analysis methods likely falls short of representing the complexity of news articles. Some researchers suggested to investigate *emotions* or *affects,* e.g., induced by headlines [150] or in entire news articles [151], whereas investigating the full text seems to yield better results. *Affect analysis* aims to find the emotions that a text induces on the contained concepts, e.g., entities or activities, by comparing relevant words from the text, e.g., nearby the investigated concept, with affect dictionaries [152]. Bhowmick et al. devised an approach that automatically estimates which emotions a news text induces on its readers using features such as tokens, polarity, and semantic representation of tokens [78]. An ML-based approach by Mishne classifies blog posts into emotion classes using features such as n-grams and semantic orientation to determine the mood of the author when writing the text [153].

Semantics derived using word embeddings may be used to determine whether words in an article contain a slant, since the most common word embeddings models contain biases, particularly gender bias and racial discrimination [154, 155]. Bolukbasi describe a method to debias word embeddings [155]; the dimensions that were removed or changed by this process contain potentially biased words; hence, they may also be used to find biased words in news texts.

Besides fully automated approaches to identify bias by labeling and word choice, semi-automated approaches incorporate users' feedback. For instance, NewsCube 2.0 employs *crowdsourcing* to estimate the bias of articles reporting on a topic. The system allows users to collaboratively annotate bias by labeling and word choice in news articles [57]. Afterward, NewsCube 2.0 presents contrastive perspectives on the topic to users. In their user study, Park et al. find that the NewsCube 2.0 supports participants to collectively organize news articles according to their slant of bias [57]. Section 3.8 describes Allsides, a news aggregator that employs crowdsourcing, though not to identify bias by labeling and word choice but to identify spin bias, i.e., the overall slant of an article.

The forms of bias by labeling and word choice have been studied extensively in the social sciences using frame analyses and content analyses. However, to date not much research on both forms has been conducted in computer science. Yet, the previously presented techniques from computer science, such as sentiment analysis and affect analysis, are already capable of achieving reliable results in other similar domains and could therefore make a unique contribution here (and partially have [5]). Crowdsourcing has already successfully been used to identify instances of such bias. Future research should focus on the creation of news-specific methods for the analysis of sentiment and affect, which may not only differentiate between positive and negative connotation but also other properties affecting readers' perception of entities, e.g., emotions.

3.5 Placement and size allocation

The placement and size allocation of a story indicates the value a news outlet assigns to that story [6, 34]. Long-term analyses reveal patterns of bias, e.g., the favoring of specific topics or avoidance of others. Furthermore, the findings of such an analysis should be combined with frame analysis to give comprehensive insights on the bias of a news outlet, e.g., a news outlet might report disproportionately much on one topic, but otherwise its articles are well-balanced and objective [35].

The first manual studies on the placement and size of news articles in the social sciences were already conducted in the 1960s. Researchers measured the size and the number of columns of articles present in newspapers, or the broadcast length in minutes dedicated to a specific topic, to investigate if there were any differences in US presidential election coverage [79, 80, 156, 157]. These early studies, and also a more recent study conducted in 2000 [34], found no significant differences in article size between the news outlets analyzed. Fewer studies have focused on the placement of an article, but found that article placement does not reveal patterns of bias for specific news outlets [79, 157]. Related factors that have also been considered are the size of headlines and pictures (see also Sect. 3.6 for more information on the analysis of pictures), which also showed no significant patterns of bias [79, 157].

Bias by article placement and size has more recently not been re-analyzed extensively, even though the rise of online news and social media may have introduced significant changes. Traditional printed news articles are a permanent medium, in the sense that once they were printed, their content could not (easily) be altered, especially not for all issues ever printed. However, online news websites are often updated. For example, if a news story is still developing, the

news article may be updated every few minutes (cf. [158]). Such updates of news articles also include the placement and allotted size of previews of articles on the main page and on other navigational pages. To our knowledge, no study has yet systematically analyzed the changes in the size and position of online news articles over time.

Fully automated methods are able to measure placement and size allocation of news articles because both forms can be determined by volumetric measurements (see Sect. 2.3). Printed newspapers must be digitalized first, e.g., using optical character recognition (OCR) and document segmentation techniques [159, 160]. Measuring a digitalized or online article's placement and size is a trivial task. Due to the Internet's inherent structure of linked websites, online news even allows for a more advanced and fully automated measurements of news article importance, such as PageRank [161], which could also be applied within pages of the publishing news outlet. Most popular news datasets, such as RCV1 [162], are text-based and do not contain information on the size and placement of a news article. Future research, however, should especially take into consideration the fast pace in online news production as described previously.

While measuring size and placement automatically is a straightforward task in computer science, only few specialized systems currently exists that can measure these forms of news bias. Saez-Trumper et al. [73] devised a system that measures the importance ascribed to a news story by an outlet by counting the total number of words of news articles reporting on the story. To measure the importance ascribed to the story by the outlet's readers, the system counts the number of Tweets linking to these news articles. One finding is that both factors are slightly correlated. NewsCube's visualization is designed to provide equal size and avoid unfair placement of news articles to "not skew users' visual attention" [27]. Even though the authors ascribe this issue high importance in their visualization, they do not analyze placement and size in the underlying articles.

Research in the social sciences and in computer science benefits from the increasing accessibility of online news, which allows effective automated analysis of bias by taking into consideration article placement and size. Measuring placement and size of articles is a trivial and scalable task that can be performed on any number of articles without requiring cumbersome manual effort. However, most recent studies in the social sciences have not considered including bias by placement and size into their analysis. The same is true for systems in computer science that should similarly include the placement and size of articles as an additional dimension of media bias. With the conclusions that have been drawn based on the analysis of traditional, printed articles, still in need of verification for online media, computer science approaches can here make a truly unique contribution.

3.6 Picture selection

Pictures contained in news articles strongly influence how readers perceive a reported topic [163]. In particular, readers who wish to get an overview of current events are likely to browse many articles and thus view only each article's headline and image. The effects of picture selection even go so far as to influence readers' voting preferences in elections [163]. Reporters or news agencies sometimes (purposefully) show pictures out of context [164, 165], e.g., a popular picture in 2015 showed an aggressive refugee with an alleged ISIS flag fighting against police officers. It later turned out that the picture was taken in 2012, before the rise of ISIS, and that the flag was not related to ISIS [166]; hence, the media had falsely linked the refugee with the terrorist organization.

Researchers from the social sciences have analyzed pictures used in news articles for over 50 years [167], approximately as long as media bias itself has been studied. Basic studies count the number of pictures and their size to measure the degree of importance ascribed by the news outlet to a particular topic (see also Sect. 3.5 for information on bias by size). In this section, we describe the techniques studies use to analyze the semantics of selected images. To our knowledge, all bias-related studies in the social sciences are concerned with political topics. Analyses of picture selection are either *person-oriented* or *topic-oriented*.

Person-oriented analyses ask analysts to rate the articles' pictures showing specific politicians. Typical rating dimensions are [168, 169]:

- *Expression*, e.g., smiling versus frowning
- *Activity*, e.g., shaking hands versus sitting
- *Interaction*, e.g., cheering crowd versus alone
- *Background*, e.g., the country's flags versus not identifiable
- *Camera angle*, e.g., eye-level shots versus shots from above
- *Body posture*, e.g., upright versus bowed torso

Findings are mixed, e.g., a study from 1998 found no significant differences in the selected pictures between the news outlets analyzed, e.g., whether selected pictures of a specific news outlets were in favor of a specific politician [168]. Another study from 1988 found that the Washington Post did not contain significant picture selection bias but that the Washington Times selected images that were more likely favorable toward Republicans [169]. A study of German TV broadcasts in 1976 found that one candidate for German chancellorship, Helmut Schmidt, was significantly more often shown in favorable shots including better camera angles and reactions of citizens than the other main candidate, Helmut Kohl [170].

Topic-oriented analyses do not investigate bias toward persons but toward certain topics. For instance, a recent

study on Belgian news coverage analyzed the presence of two frames [171]: Asylum seekers in Belgium are (1) victims that need protection or (2) intruders that disturb Belgian culture and society. Articles supporting the first frame typically chose pictures depicting refugee families with young children in distress or expressing fear. Articles supporting the second frame chose pictures depicting large groups of mostly male, asylum seekers. The study found that the victim frame was predominantly adopted in Belgian news coverage, and particularly in the French-speaking part of Belgium. The study also revealed a temporal pattern: During Christmas time, the victim frame was even more predominant.

To our knowledge, there are currently no systems or approaches from computer science that analyze media bias through image selection. However, methods in *computer vision* can measure some of the previously described dimensions. Automated methods can identify faces in images [172], recognize emotions [83, 173], categorize objects shown in pictures [174], and even generate captions for a picture [175]. Research has advanced so far in these applications that several companies, such as Facebook, Microsoft, and Google, are using such automated methods in production or are offering them as a paid service. Segalin et al. [82] trained a convolutional neural network (CNN) on the PsychoFlickr dataset to estimate the personality traits of the pictures' authors [176]. To evaluate the classification performance of the system, they compared the CNN's classifications with self-assessments by picture authors and also with attributed assessments by participants of a study. The results of their evaluation suggest that CNNs are suitable to derive such characteristics that are not even visible in the analyzed pictures.

Picture selection is an important factor in the perception of news. Basic research from psychology has shown that image selection can slant coverage toward one direction, although studies in the social sciences on bias by selection in the past concluded that there were no significant differences in picture selection. Advances in image processing research and the increasing accessibility of online news provide completely new avenues to study potential effects of picture selection. Computer science approaches can here primarily contribute by enabling the automated analysis of images on much bigger scale allowing us to reopen important questions on the effect of picture selection in news coverage and beyond.

3.7 Picture explanation

Captions below images and referrals to the images in the main text provide images with the needed textual context. Images and their captions should be analyzed jointly because text can change a picture's meaning, and vice versa [167, 177]. For instance, during hurricane "Katrina" in 2005, two similar pictures published in US media showed survivors wading away with food from a grocery store. The only difference was that one picture showed a black man, who "looted" the store, while the other picture depicted a white couple, who "found" food in the store [84].

Researchers from the social sciences typically perform two types of analyses that are concerned with bias from image captions: jointly analyzing image and caption, or only analyzing the caption, ignoring the image. Only few studies analyze captions and images jointly. For instance, a comparison of images and captions from the Washington Post and the Washington Times found that the captions were not significantly biased [169]. A frame analysis on the refugee topic in Belgian news coverage also took into consideration image captions. However, the authors focused on the *overall* impression of the analyzed articles rather than examining any potential bias specifically present in the picture captions [171].

The vast majority of studies analyze captions without placing them in context with their pictures. Studies and techniques concerned with the text of a caption (but not the picture) are described in the previous sections, especially in the sections for bias by commission and omission (see Sect. 3.3), and labeling and word choice (see Sect. 3.4). We found that most studies in the social sciences either analyze image captions as a component of the main text, or analyze images but disregard their captions entirely [79, 157, 168]. Likewise, relevant methods from computer science are effectively the same as those concerned with bias by commission and omission (see Sect. 3.3), and labeling and word choice (see Sect. 3.4). For the other type of studies, i.e., jointly analyzing images and captions, relevant methods are discussed in Sect. 3.6, i.e., computer vision to analyze the contents of pictures, and additionally in Sects. 3.3 and 3.4, e.g., sentiment analysis to find biased words in captions.

To our knowledge, no study has examined picture referrals contained in the article's main text. This is most likely due to the infrequency of picture referrals.

The few analyses on captions suggest that bias by picture explanation is not very common. However, more fundamental studies show strong impact of captions on the perception of images and note rather subtle differences in word choice. While many studies analyzed captions as part of the regular text, e.g., analyzing bias by labeling and word choice, research currently lacks specialized analyses that examine captions in conjunction with their images.

3.8 Spin: the vagueness of media bias

Bias by spin is closely related to all other forms of media bias and is also the vaguest form. Spin is concerned with the context of presented information. Journalists create the spin of an article on all textual levels, e.g., by supporting a quote with an explanation (phrase level), highlighting certain parts of the event (paragraph level), or even by concluding the

article with a statement that frames all previously presented information differently (article level). The order in which facts are presented to the reader influences what is perceived (e.g., some readers might only read the headline and lead paragraph) and how readers rate the importance of reported information [85]. Not only the text of an article but all other elements, including pictures, captions, and the presentation of the information contribute to an article's overall spin.

In the social sciences, the two primarily used methods to analyze the spin of articles are frame analysis, and more generally content analysis. For instance, one finding in the terrorism analysis conducted by Papacharissi and de Fatima Oliveira [77] (see Sect. 3.2) was that the New York Times often personified the events in their articles, e.g., by focusing on persons involved in the event and the use of dramatic language.

Some practices in *journalism* can be seen as countermeasures to mitigate media bias. *Press reviews* summarize an event by referring to the main statements found in articles by other news outlets. This does not necessarily reveal media bias, because any perspective can be supported by source selection, e.g., only "reputable" outlets are used. However, typically press reviews broaden a reader's understanding of an event and might be a starting point for further research. Another practice that supports mitigation of media bias is opposing commentaries in newspapers, where two authors subjectively elaborate their perspective on the same topic. Readers will see both perspectives and can make their own decisions regarding the topic.

Social media has given rise to new collaborative approaches to media bias detection. Reddit[3] is a social news aggregator, where users post links or texts regarding current events or other topics, and rate or comment on posts by other users. Through the comments on a post, a discussion can emerge that is often controversial and contains the various perspectives of commenters on the topic. Reddit also has a "media bias" thread[4] where contributors share examples of biased articles. Wikinews is a collaborative news producer, where volunteers author and edit articles. Wikinews aims to provide "reliable, unbiased and relevant news […] from a neutral point of view" [178]. However, two main issues are the mixed quality of news (because many authors may have participated) and the low number of articles, i.e., only major events are covered in the English version, and other languages have even fewer articles. Thus, Wikinews currently cannot be used as a primary, fully reliable news source. Some approaches employ crowdsourcing to visualize different opinions or statements on politicians or news topics, for example, the German news outlet Spiegel Online frequently asks readers to define their position regarding two pairs of contrary statements that span a two-dimensional map [179]. Below the map, the news outlet lists excerpts from other outlets that support or contradict the map's statements.

The automated analysis of spin bias using methods from computer science is maybe the most challenging of all forms because its manifestation is the vaguest among the forms of bias discussed. Spin refers to the overall perception of an article. Bias by spin is not, however, just the sum of all other forms but includes other factors, such as the order of information presented in a news article, the article's tone, and emphasis on certain facts. Methods we describe in the following are partially also relevant for other forms of bias. For instance, the measurement of an article's degree of personification in the terrorism in news coverage study [77] is supported by the computation of CRA [85]. What is not automated is the annotation of entities and their association with an issue. Named entity extraction [180] could be used to partially address these previously manually performed tasks.

Other approaches analyze news readers' input, such as readers' comments, to identify differences in news coverage. The rationale of these approaches is that readers' input contains explicitly stated opinions and sentiment on certain topic, which are usually missing from the news article itself. Explicitly stated opinion can reliably be extracted with the help of state-of-the-art NLP methods, such as sentiment analysis. For instance, one method analyzes readers' comments to categorize related articles [86]. The method measures the similarity of two articles by comparing their reader comments, focusing on the entities mentioned and the sentiment expressed in the comments, and country of the comment's author. Another method counts and analyzes Twitter followers of news outlets to estimate the political orientation of the audience of the news outlet [20]. A seed group of Twitter accounts is manually rated according to their political orientation, e.g., conservative or liberal. This group is automatically expanded using those accounts' followers. The method then estimates the political orientation of a news outlet's audience by averaging the political orientation of the outlet's followers in the expanded, group of categorized accounts (cf. [42, 59, 61]).

The news aggregator *Allsides*[5] shows users the most contrastive articles on a topic, e.g., left and right leaning on a political spectrum. The system asks users to rate the spin of news outlets, e.g., after reading articles published by these outlets. To estimate the spin of an outlet, Allsides uses the feedback of users and expert knowledge provided by their staff. *NewsCube 2.0* lets (expert) users collaboratively define and rate frames in related articles [57]. The frames are in turn presented to other users, e.g., a contrast view shows the most contrasting frames of one event. Users can then incrementally improve the quality of coding by refining existing frames.

[3] https://www.reddit.com/.

[4] https://www.reddit.com/r/MediaBias/.

[5] http://www.allsides.com/.

Another method for news spin identification categorizes news articles on contentious news topics into two (opposing) groups by analyzing quotes and nearby entities [181]. The rationale of the approach is that articles portraying a similar perspective on a topic have more common quotes, which may support the given perspective, than articles that have different perspectives. The method extracts weighted triples representing who criticizes whom, where the weight depends on the importance of the triple, e.g., estimated by the position within the article (the earlier, the more important). The method measures the similarity of two articles by comparing their triples.

Other methods analyze frequencies and co-occurrences of terms to find frames in related articles and assign each article to one of the frames. For instance, one method clusters articles by measuring the similarity of two documents using the co-occurrences of the two documents' most frequent terms [182]. The results of this rather simple method are then used for a manually conducted frame analysis. *Hiérarchie* uses recursive topic modeling to find topics and subtopics in Tweets posted by users on a specific issue [183]. A radial treemap visualizes the extracted topics and subtopics. In the presented case study, users find and explore different theories on the disappearance of flight MH-370 discussed in Tweets.

Recently, a Dagstuhl workshop on fake news proposed a number of axes of quantitative computer analysis, such as factuality, readability, and virality, could help users to make more informed judgments about the news items they read [184].

Lastly, manually added information related to media bias, e.g., the overall spin of articles rated by users of Allsides, or articles annotated by social scientists during frame analysis, is in our view, a promising learning dataset for *machine learning*. Other data that exploit the *wisdom of the crowd* might be incorporated as well, e.g., analyzing the Reddit media bias thread.

In our view, the existence of the very concept of spin bias allows drawing two conclusions. First, media bias is a complex model of skewed news coverage with overlapping and partially contradicting definitions. While many instances of media bias fit into one of the other more precisely defined forms of media defined in the news production and consumption process (see Sect. 2.2), some instances of bias do not. Likewise, such instances of bias may fit into other models from the social sciences that are concerned with differences in news coverage, such as the bias forms of coverage, gatekeeping, and statement (Sect. 2.2 briefly discusses other models of media bias), while other instances would not fit into such models. Second, we found that most of the approaches from computer science for identifying, or suitable for identifying, spin bias, omit the extensive research that has been conducted in the social sciences. Computer science approaches currently still address media bias as vaguely defined differences in news coverage and therefore stand to profit from prior research in the social sciences. In turn, there are few scalable approaches to the analysis of media biases in the social sciences significantly hampering progress in the field. We therefore see a strong prospect for collaborative research on automated approaches to the analysis of media bias across both disciplines.

4 Reliability, generalizability, and evaluation

This section discusses how automated approaches for analyzing media bias should be evaluated. Therefore, we first describe how social scientists measure the reliability and generalizability of studies on media bias.

The reliability and generalizability of manual annotation in the social sciences provides the benchmark for any automated approach. Best practices in social science research can involve both the careful development and iterative refinement of underlying codebooks, as well as the formal validation of inter-coder reliability. For example, as discussed in Sect. 2.3, a smaller, careful inductive manual annotation aids in constructing the codebook. The main deductive analysis is then performed by a larger pool of coders where the results of individual coders, their agreement on the assignment of codes can be systematically compared. Standard measures for inter-coder reliability, e.g., the widely used Krippendorff's Alpha [185], provide estimates for the reliability and robustness of the coding. Whether coding rules, and with these the quality of annotations, can be generalized beyond a specific case is usually not routinely analyzed because, by virtue of the significant effort required for manual annotation, the scope of such studies is usually limited to a specific question or context. Note, however, that the usual setup of a small deductive analysis, conducted on a subset of the data, implies that a codebook generated in this way can generalize to a larger corpus.

Computer science approaches for the automated analysis of media bias stand to profit a lot from a broad adoption of their methods by researchers across a wider set of disciplines. The impact and usefulness of automatized approaches for substantive cross-disciplinary analyses, however, hinges critically on two central questions. First, compared to manual methodologies, how reliable are automated approaches? Specifically, broad adoption of automated approaches in social sciences applications is only likely if the automated approaches identify at least close to the same instances of bias as manual annotations would.

Depending on which kind of more or less subtle form of bias is analyzed, the results gained through manual annotation might represent a more or less difficult benchmark

to beat. Especially in complex cases, manual annotation of individual items may systematically perform better in capturing subtle instances relevant to the analysis question than automated approaches. Note that, for example, currently no public annotated news dataset for sentiment analysis exists (see Sect. 3.4). The situation is similar for most of the applications reviewed in this article, i.e., there is currently a dearth of standard benchmark datasets. Meaningful validation would thus require as a first step the careful (and time-intensive) development of such datasets across a range of relevant contexts.

One way to counter the present lack of evaluation datasets is to not solely rely on manual content analysis for annotation. For simple annotation tasks, such as rating the subjective slant of a news picture, crowdsourcing can be a suitable alternative to content analysis. This procedure requires less effort than conducting a full content analysis, including creation of a codebook and refining it until the ICR is sufficiently high (cf. [186]). Another way is to reuse existing datasets, something which has already been done successfully to create learning datasets in the context of biased language. For instance, Recasens et al. [187] use bias-related revisions from the Wikipedia edit history to retrieve presumably biased single-word phrases. The political slant classification of news articles and outlets crowdsourced by users on web services such as Allsides (see Sect. 3.8) may serve as another comparison baseline.

Another way to evaluate the performance of bias identification methods, is to manually analyze the automatically extracted instances of media bias, e.g., through crowdsourcing or (typically fewer) specialized coders. However, evaluating the results of an automated approach this way decreases the comparability between approaches, since these have to be evaluated in the same way manually again. Generating annotated benchmark datasets, on the other hand, requires greater initial effort, but the results can then be used multiple times to evaluate and compare multiple approaches.[6]

The second central question is how well automated approaches generalize to the study of similar forms of bias in different contexts than those contexts for which they were initially developed. This question pertains to the external validity of developed approaches, i.e., is their performance dependent on a specific empirical or topical context? Out-of-sample performance could be tested against benchmark datasets not used for initial evaluation; however, as emphasized before, such datasets must still be developed. Hence, systematically testing the performance of approaches across many contexts is likely also infeasible for the near future simply because the costs of generating benchmark datasets is too high. Ultimately, it would be best practice for benchmark studies to establish more generally whether or not specific characteristics of news are related to the performance of the automated approaches developed.

[6] The SemEval series [76] are a representative example from computer science where with high initial effort comprehensive evaluation (and training and test) datasets are created, allowing a quantitative comparison of the performance of multiple approaches afterward.

5 Conclusion

News coverage strongly influences public opinion. However, at times, the news coverage of media outlets is far from objective, a phenomenon called media bias. Media bias can potentially negatively impact the public, since biased news coverage may influence elections or public opinion on societal issues. Recent trends, such as social bots that automatically write news posts, or the centralization of media outlet ownership, have the potential to further amplify the negative effects of biased news coverage. News consumers should be able to view different perspectives of the same news topic [26]. Especially in democratic societies, unrestricted access to unbiased information is crucial for citizens to form their own views and make informed decisions [46, 123], e.g., during elections. Since media bias has been, and continues to be, structurally inherent in news coverage [27, 56, 58], the detection and analysis of media bias is a topic of high societal and policy relevance—especially, if these analyses and associated tools and platforms help news consumers to become more aware of instances of media bias.

Researchers from the social sciences have studied media bias extensively over the past decades, resulting in a comprehensive set of methodologies, such as content analysis and frame analysis, as well as models to describe media bias. One of these models, the "news production and consumption process," describes how journalists turn events into news articles. The process defines nine forms of media bias that can occur during the three phases of news production: in the first phase, "gathering of information," the bias forms are (1) event selection, (2) source selection, and (3) commission and omission of information. In the second phase, "writing," the bias forms are (4) labeling and word choice. In the third phase, "editing," the bias forms are (5) story placement, (6) size allocation, (7) picture selection, and (8) picture explanation. Lastly, bias by (9) spin is a form of media bias that represents the overall bias of a news article and essentially combines the other forms of bias, including minor forms not defined specifically by the news production and consumption process.

For each of the forms of media bias, we discussed exemplary approaches being applied in the social sciences, and described the automated methods from computer science that have been used, or could best be used, to address the particular form of bias. We summarize the findings of our comprehensive review of the current status-quo as follows:

F1. Only few approaches in computer science address the analysis of media bias. The majority of these approaches analyze media bias from the perspective of regular news consumers and neglect both the established approaches and the comprehensive models that have already been developed in the social sciences. In many cases, the underlying models of media bias are too simplistic, and their results when compared to models and results of research in the social sciences do not provide additional insights.

F2. The majority of content analyses in the social sciences do not employ state-of-the-art text analysis methods from computer science. As a result, the manual content analysis approaches conducted by social scientists require exacting and very time-consuming effort, as well as significant expertise and experience. This severely limits the scope of what social scientists can study and has significantly hampered progress in the field.

F3. In our view, there is therefore a lot of potential for interdisciplinary research on media bias among computer scientists and social scientists. Useful state-of-the-art approaches from computer science are available for each of the nine forms of media bias that we discussed. On the one hand, methodologies and models of media bias in the social sciences can help computer scientists make the automated approaches more effective. Likewise, the development of automated methods to identify instances of specific forms of media bias can help make content analysis in the social sciences more efficient by automating more tasks.

Media bias analysis is a rather young research topic within computer science, particularly when compared with the social sciences, where the first studies on media bias were published more than 60 years ago [29, 177]. Our first finding (F1) is that most of the reviewed computer science approaches treat media bias vaguely, and view it only as "differences of [news] coverage" [149], "diverse opinions" [188], or "topic diversity" [26]. The majority of the existing approaches neglect the state of the art developed in the social science. They do not make use of models describing different forms of media bias or how biased news coverage emerges in the news production and consumption process [6, 27] (Sect. 2.2). Also, approaches in computer science do not employ methods to analyze the specific forms of bias, such as content analysis [34] and frame analysis [88] (Sect. 2.3). Consequently, many state-of-the-art approaches in computer science are limited in their capability for identifying instances of media bias. For instance, matrix-based news aggregation (MNA) organizes articles and topics in a matrix to facilitate showing differences in international news topics, but the approach can neither determine whether there are actual differences, nor can MNA enforce finding differences [117]. Likewise, *Hiérarchie* finds subtopics in news posts that may or may not refer to differences caused by media bias [183]. To overcome the limitations in identifying bias, some approaches, such as *NewsCube 2.0* [57] and *Allsides* (Sect. 3.8), outsource the task of identifying media bias to users, e.g., by asking users to manually rate the slant of news articles.

Content analysis and frame analysis both require significant manual effort and expertise (F2). Especially time-intensive are the tasks of systematic screening of texts and their subsequent annotation, tasks that can only be performed by human coders [34, 88]. Currently, in our view, the execution of these tasks cannot be improved significantly by employing automated text analysis methods due to the lack of mature methods capable of identifying specific instances of media bias, which follows from F1. This limitation, however, may be revised once interdisciplinary research has resulted in more advanced automated methods. Other tasks, such as data gathering, or searching for relevant documents and phrases, are already supported by basic (semi-)automated methods and tools, such as content analysis software [99]. However, clearly the full potential of the state of the art in computer science is not yet being exploited. The employed techniques, e.g., keyword-based text matching to find relevant documents [100], or frequency-based extraction of representative terms to find patterns [99], are rather simple compared to the state of the art in automated text analysis. Few of the reviewed tools used by researchers in the social sciences employ methods proven effective in natural language processing, such as resolution of co-references or synonyms, or finding related article using an event-based search approach.

In our view, combining the expertise of both the social sciences and computer science results in valuable opportunities for interdisciplinary research (F3). Reliable models of media bias and manual approaches for the detection of media bias can be combined with methods for automated data analysis, in particular, with state-of-the-art text analysis and natural language processing approaches. NewsCube [27], for instance, extracts so-called "aspects" from news articles, which refer to the "frames" defined by social scientists [39]. Users of NewsCube became more aware of the different perspectives contained in news coverage on specific topics, than users of Google News. In this article, we showed that promising automated methods from computer science are available for all forms of media bias as defined by the news production and consumption process (see Sect. 3). For instance, studies concerned with bias by source selection or the commission and omission of information, investigate how information is reused in news coverage [42, 61, 138]. Similarly to these studies, methods from plagiarism detection aim to identify instances of information reuse in a set of documents, and these methods yield reliable results for pla-

giarism with sufficient textual similarity [125, 126]. Finally, recent advancements text analysis, particularly word embeddings [110] and deep learning [189], open a promising area of research on media bias. Thus far, few studies use word embeddings and deep learning to analyze media bias in news coverage. However, the techniques have proven very successful in various related problems (cf. [75, 134, 190, 191]), which lets us anticipate that the majority of the textual bias forms could be addressed effectively with such approaches.

We believe that interdisciplinary research on media bias can result in three main benefits. First, automated approaches for analyzing media bias will become more effective and more broadly applicable, since they build on the substantial, theoretical expertise that already exists in the social sciences. Second, content analyses in the social sciences will become more efficient, since more tasks can be automated, or supported by automated methods from computer science. Finally, we argue that news consumers will benefit from improved automated methods for identifying media bias, since the methods can be used by news aggregators to detect and visualize the occurrence of potential media bias in real time.

Acknowledgements This work was supported by the Carl Zeiss Foundation and the Zukunftskolleg program of the University of Konstanz. We thank the anonymous reviewers for their valuable comments that significantly helped to improve this article.

References

1. Mossberger, K., Tolbert, C.J., McNeal, R.S.: Digital Citizenship: The Internet, Society, and Participation. MIT Press, Cambridge (2007)
2. Urban, C.D.: Examining our credibility: perspectives of the public and the press. Urban1999-URBEOC, Asne Foundation (1999)
3. Crossman, J.: Aussies turn to social media for news despite not trusting it as much, *Crossman Communications*, Nov-2014. http://crossman.com.au/?p=3853
4. University of Michigan.: News bias explored—The art of reading the news (2014). http://umich.edu/~newsbias/. Accessed 01 Aug 2018
5. Grefenstette, G., Qu, Y., Shanahan, J., Evans, D.: Coupling niche browsers and affect analysis for an opinion mining application. In: Proceedings of 12th International Conference on Rech. d'Information Assistee par Ordinateur (2004)
6. Baker, B.H., Graham, T., Kaminsky, S.: How to Identify, Expose and Correct Liberal Media Bias. Media Research Center, Alexandria (1994)
7. Oelke, D., Geißelmann, B., Keim, D.A.: Visual analysis of explicit opinion and news bias in german soccer articles. EuroVis Workshop on Visual Analytics, EuroVA 2012, Vienna, Austria, June 4–5 June 2012. https://doi.org/10.2312/PE/EuroVAST/EuroVA12/049-053
8. Gentzkow, M., Shapiro, J.M: Media bias and reputation. J. Polit. Econ. **114**(2), 280–316 (2006). https://doi.org/10.1086/499414
9. Bucher, H.J., Schumacher, P.: The relevance of attention for selecting news content. An eye-tracking study on attention patterns in the reception of print and online media. Communications **31**, 347 (2006)
10. Bernhardt, D., Krasa, S., Polborn, M.: Political polarization and the electoral effects of media bias. J. Public Econ. **92**(5), 1092–1104 (2008)
11. Napolitan, J.: The Election Game and How to Win it. Doubleday, New York (1972)
12. Meyrowitz, J.: No Sense of Place: The Impact of Electronic Media on Social Behavior. Oxford University Press, Oxford (1986)
13. Tye, L.: The Father of Spin: Edward L. Bernays and the Birth of Public Relations. Macmillan, London (2002)
14. Amos, A., Haglund, M.: From social taboo to 'torch of freedom': the marketing of cigarettes to women. Tob. Control **9**(1), 3–8 (2000)
15. Lutz, A: These 6 corporations control 90% of the media in America. *Bus. Insider* (2014). http://www.businessinsider.com/these-6-corporations-control-90-of-the-media-in-america-2012-6. Accessed 12 Jan 2017
16. Esser, F., Reinemann, C., Fan, D.: Spin doctors in the United States, Great Britain, and Germany Metacommunication about media manipulation. Harv. Int. J. Press. **6**(1), 16–45 (2001)
17. Straubhaar, J.D.: Media Now: Communication Media in Information Age. Thomson Learning, Boston (2000)
18. Bakshy, E., Messing, S., Adamic, L.A.: Exposure to ideologically diverse news and opinion on Facebook. Science **348**(6239), 1130–1132 (2015)
19. An, J., Cha, M., Gummadi, K.P., Crowcroft, J., Quercia, D.: Visualizing media bias through Twitter. In: Proceedings of ICWSM SocMedNews Workshop (2012)
20. Golbeck, J., Hansen, D.: Computing political preference among twitter followers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1105–1108 (2011)
21. Sunstein, C.R.: The law of group polarization. J. Polit. Philos. **10**(2), 175–195 (2002)
22. Barberá, P., Jost, J.T., Nagler, J., Tucker, J.A., Bonneau, R.: Tweeting from left to right: is online political communication more than an echo chamber? Psychol. Sci. **26**(10), 1531–1542 (2015)
23. Facebook, "Company Info." (2016)
24. Bui, C.: How online gatekeepers guard our view: news portals' inclusion and ranking of media and events. Glob. Media J. **9**(16), 1–41 (2010)
25. Wanta, W., Golan, G., Lee, C.: Agenda setting and international news: media influence on public perceptions of foreign nations. Journal. Mass Commun. Q. **81**(2), 364–377 (2004)
26. Munson, S.A., Zhou, D.X., Resnick, P.: Sidelines: an algorithm for increasing diversity in news and opinion aggregators. In: ICWSM (2009)
27. Park, S., Kang, S., Chung, S., Song, J.: NewsCube: delivering multiple aspects of news to mitigate media bias. In: Proceedings of CHI'09, SIGCHI Conference on Human Factors Computer System, pp. 443–453 (2009)
28. Munson, S.A., Lee, S.Y., Resnick, P.: Encouraging reading of diverse political viewpoints with a browser widget. In: ICWSM (2013)
29. White, D.M.: The 'gate keeper': a case study in the selection of news. J. Bull. **27**(4), 383–390 (1950)
30. Williams, A.: Unbiased study of television news bias. J. Commun. **25**(4), 190–199 (1975)
31. Harcup, T., O'neill, D.: What is news? Galtung and Ruge revisited. J. Stud. **2**(2), 261–280 (2001)

32. McCarthy, J.D., McPhail, C., Smith, J.: Images of protest: dimensions of selection bias in media coverage of Washington demonstrations, 1982 and 1991. Am. Soc. Rev. **61**(3), 478–499 (1996)
33. Mullainathan, S., Shleifer, A.: Media bias. In: National Bureau of Economic Research, vol. 9295, pp. 127 (2002)
34. D'Alessio, D., Allen, M.: Media bias in presidential elections: a meta-analysis. J. Commun. **50**, 133–156 (2000)
35. Druckman, J.N., Parkin, M.: The impact of media bias: how editorial slant affects voters. J. Polit. **67**(4), 1030–1049 (2005)
36. Gerber, A.S., Karlan, D., Bergan, D.: Does the media matter? A field experiment measuring the effect of newspapers on voting behavior and political opinions. Am. Econ. J. Appl. Econ. **1**(2), 35–52 (2009)
37. Gentzkow, M., Glaeser, E.L., Goldin, C.: The rise of the fourth estate How newspapers became informative and why it mattered. In: Glaeser, E.L., Goldin, C. (eds.) Corruption and Reform: Lessons from America's Economic History, pp. 187–230. University of Chicago Press, Chicago (2006)
38. De Vreese, C.H.: News framing: theory and typology. Inf. Des. J. Doc. Des. **13**(1), 51–62 (2005)
39. Iyengar, S.: Is Anyone Responsible? How Television Frames Political Issues. University of Chicago Press, Chicago (1994)
40. Zaller, J.: The Nature and Origins of Mass Opinion. Cambridge university Press, Cambridge (1992)
41. Kahneman, D., Tversky, A.: Choices, values, and frames. Am. Psychol. **39**(4), 341 (1984)
42. Groseclose, T., Milyo, J.: A measure of media bias. Q. J. Econ. **120**, 1191–1237 (2005)
43. Sunstein, C.R.: Echo Chambers: Bush v. Gore, Impeachment, and Beyond. Princeton University Press, Princeton (2001)
44. Frey, D.: Recent research on selective exposure to information. Adv. Exp. Soc. Psychol. **19**, 41–80 (1986)
45. Mutz, D.C.: Facilitating communication across lines of political difference: the role of mass media. Am. Polit. Sci. Assoc. **95**(01), 97–114 (2001)
46. Mullainathan, S., Shleifer, A.: The market for news. Am. Econ. Rev. **95**, 1031–1053 (2005)
47. N. Newman, D. A. L. Levy, and R. K. Nielsen, "Reuters Institute Digital News Report 2015," *Available SSRN 2619576*, 2015
48. De Marzo, P., Vayanos, D., Zwiebel, J.: Persuasion bias. Soc. Influ. Unidimen. Opin. Q. J. Econ. **118**, 909–967 (2003)
49. Lakoff, G.: Women, Fire, and Dangerous Things. What Categories Reveal About Mind. The University of Chicago Press, Chicago (1987)
50. Kull, S., Ramsay, C., Lewis, E.: Misperceptions, the media, and the Iraq war. Polit. Sci. Q. **118**(4), 569–598 (2003)
51. DellaVigna, S.. Kaplan, E.: The fox news effect: media bias and voting. Q. J. Econ. **122**(3), 1187–1234 (2007). https://doi.org/10.3386/w12169
52. Larcinese, V., Puglisi, R., Snyder, J.M.: Partisan bias in economic news: evidence on the agenda-setting behavior of US newspapers. J. Public Econ. **95**(9), 1178–1189 (2011)
53. Scheufele, D.A.: Agenda-setting, priming, and framing revisited: another look at cognitive effects of political communication. Mass Commun. Soc. **3**(2–3), 297–316 (2000)
54. Entman, R.M.: Framing: toward clarification of a fractured paradigm. J. Commun. **43**(4), 51–58 (1993)
55. Entman, R.M.: Framing bias: media in the distribution of power. J. Commun. **57**(1), 163–173 (2007)
56. Herman, E.S., Chomsky, N.: Manufacturing Consent: The Political Economy of the Mass Media. Random House, New York (2010)
57. Park, S., Ko, M., Kim, J., Choi, H., Song, J.: NewsCube 2.0: an exploratory design of a social news website for media bias mitigation. In: Workshop on Social Recommender Systems (2011)
58. Herman, E.S.: The propaganda model: a retrospective. J. Stud. **1**(1), 101–112 (2000)
59. MacGregor, B.: Live, Direct, and Biased?: Making Television News in the Satellite Age. Arnold, London (1997)
60. Baron, D.P.: Persistent media bias. J. Public Econ. **90**(1), 1–36 (2006)
61. Gentzkow, M., Shapiro, J.M.: What drives media slant? Evidence from US daily newspapers. Econometrica **78**(1), 35–71 (2010)
62. Gilens, M., Hertzman, C.: Corporate ownership and news bias: newspaper coverage of the 1996 Telecommunications Act. J. Polit. **62**(02), 369–386 (2000)
63. D'Angelo, P., Kuypers, J.A.: Doing News Framing Analysis: Empirical and Theoretical Perspectives. Routledge, Abingdon (2010)
64. Besley, T., Prat, A.: Handcuffs for the grabbing hand? Media capture and government accountability. Am. Econ. Rev. **96**(3), 720–736 (2006). https://doi.org/10.1257/aer.96.3.720
65. Paul, R., Elder, L.: The Thinker's Guide for Conscientious Citizens on how to Detect Media Bias & Propaganda in National and World News. Foundation Critical Thinking (2004)
66. F. Esser, "Editorial structures and work principles in British and German newsrooms," *Eur. J. Commun.*, 1998
67. Boczkowski, P.J.: The processes of adopting multimedia and interactivity in three online newsrooms. J. Commun. **54**, 197–213 (2004)
68. Sundar, S.S.: Exploring receivers' criteria for perception of print and online news. J. Mass Commun. Q. **76**(2), 373–386 (1999)
69. Balahur, A. et al.: Sentiment analysis in the news. arXiv Prepr. arXiv1309.6202 (2013)
70. Vallone, R.P., Ross, L., Lepper, M.R.: The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the Beirut massacre. J. Pers. Soc. Psychol. **49**(3), 577 (1985)
71. Gruenewald, J., Pizarro, J., Chermak, S.M.: Race, gender, and the newsworthiness of homicide incidents. J. Crim. Justice. **37**, 262 (2009)
72. Bourgeois, D., Rappaz, J., Aberer, K.: Selection bias in news coverage: learning it, fighting it. In: Companion of the The Web Conference 2018 on The Web Conference 2018—WWW'18 (2018)
73. Saez-Trumper, D., Castillo, C., Lalmas, M.: Social media news communities: gatekeeping, coverage, and statement bias. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (2013)
74. Sanderson, M.: Duplicate detection in the Reuters collection. In: Technical Report Department of Computer Science University Glasgow. G12 8QQ, UK (1997)
75. Agirre, E, et al.: SemEval-2016 task 1: semantic textual similarity, monolingual and cross-lingual evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (2016)
76. Gipp, B.: Citation-based plagiarism detection. In: Gipp, B. (ed.) Citation-Based Plagiarism Detection, pp. 57–88. Springer, Berlin (2014)
77. Papacharissi, Z., de Fatima Oliveira, M.: News frames terrorism: a comparative analysis of frames employed in terrorism coverage in U.S. and U.K. newspapers. Int. J. Press. **13**(1), 52–74 (2008)
78. Bhowmick, P.K., Basu, A., Mitra, P.: Reader perspective emotion analysis in text through ensemble based multi-label classification framework. Comput. Inf. Sci. **2**, 64–74 (2009)
79. Stovall, J.G.: The third-party challenge of 1980: news coverage of the presidential candidates. J. Mass Commun. Q. **62**(2), 266 (1985)
80. Stempel, G.H.: The prestige press meets the third-party challenge. J. Mass Commun. Q. **46**(4), 699–706 (1969)

81. Waldman, P., Devitt, J.: Newspaper photographs and the 1996 presidential election: the question of bias. J. Mass Commun. **75**(2), 302–311 (1998)
82. Segalin, C., Perina, A., Cristani, M., Vinciarelli, A.: The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. IEEE Trans. Affect. Comput. **2**, 268 (2017)
83. Busso, C., et al.: Analysis of emotion recognition using facial expressions, speech and multimodal information. In: Proceedings of The 6th International Conference on Multimodal Interfaces, pp. 205–211 (2004)
84. Sommers, S.R., Apfelbaum, E.P., Dukes, K.N., Toosi, N., Wang, E.J.: Race and media coverage of Hurricane Katrina: analysis, implications, and future research questions. Anal. Soc. Issues Public Policy **6**(1), 39–55 (2006)
85. Christian, D., Froke, P., Jacobsen, S., Minthorn, D.: The Associated Press Stylebook and Briefing on Media Law. The Associated Press, New York (2014)
86. Abbar, S., Amer-Yahia, S., Indyk, P., Mahabadi, S.: Real-time recommendation of diverse related articles. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1–12 (2013)
87. Hamborg, F., Meuschke, N., Gipp, B.: Bias-aware news analysis using matrix-based news aggregation. *Int. J. Digit. Libr*
88. Van Gorp, B.: Strategies to take subjectivity out of framing analysis. Doing news Fram. Anal. Empir. Theor. Perspect. pp 100–125 (2010)
89. Hunter, J.E., Schmidt, F.L., Jackson, G.B.: Meta-Analysis: Cumulating Research Findings Across Studies, vol. 4. Sage Publications, Inc, New York (1982)
90. Oliver, P.E., Maney, G.M.: Political processes and local newspaper coverage of protest events: from selection bias to triadic interactions. Am. J. Sociol. **106**(2), 463–505 (2000)
91. McCarthy, J., Titarenko, L., McPhail, C., Rafail, P., Augustyn, B.: Assessing stability in the patterns of selection bias in newspaper coverage of protest during the transition from communism in Belarus. Mob. An Int. Q. **13**(2), 127–146 (2008)
92. Davis, M.S., Goffman, E.: Frame analysis: an essay on the organization of experience. Contemp. Sociol. **4**, 599–603 (1975)
93. Matthes, J.: What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990–2005. J. Mass Commun. Q. **86**(2), 349–367 (2009)
94. Cappella, J.N., Jamieson, K.H.: Spiral of Cynicism: The Press and the Public Good. Oxford University Press, Oxford (1997)
95. Neuendorf, K.A.: The Content Analysis Guidebook. Sage Publications, New York (2016)
96. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977)
97. Vaismoradi, M., Turunen, H., Bondas, T.: Content analysis and thematic analysis: implications for conducting a qualitative descriptive study. Nurs. Health Sci. **15**(3), 398–405 (2013)
98. Schreier, M.: Qualitative Content Analysis in Practice. Sage Publications, Thousands Oaks (2012)
99. Lowe, W.: Software for content analysis–A review. *Cambridge Weather. Cent. Int. Aff. Harvard Identity Proj.* (2002)
100. Stemler, S.: An overview of content analysis. Pract. Assess. Res. Eval. **7**(17), 137–146 (2001)
101. Leetaru, K., Schrodt, P.A.: GDELT: global data on events, location and tone, 1979–2012. In: Annual Meeting International Studies Association (2013)
102. Tsagkias, M., De Rijke, M., Weerkamp, W.: Linking online news and social media. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 565–574 (2011)
103. LexisNexis, "LexisNexis Police Reports," 2017. [Online]. https://policereports.lexisnexis.com/search/search. Accessed 25 Sep 2018
104. McGregor, M.J., Wiebe, E., Marion, S.A., Livingstone, C.: Why don't more women report sexual assault to the police? Can. Med. Assoc. J. **162**(5), 659–660 (2000)
105. Julinda, S., Boden, C., Akbik, A.: Extracting a repository of events and event references from news clusters. COLING **2014**, 14 (2014)
106. Blei, D.M.: Probabilistic topic models. Commun. ACM **55**(4), 77–84 (2012)
107. Best, C., van der Goot, E., Blackler, K., Garcia, T., Horby, D.: Europe media monitor. Technical Report EUR 22173 EN (2005)
108. Manning, C.D., Raghavan, P., Schutze, H.: An introduction to information retrieval. *Online* (2009)
109. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. **24**(5), 513–523 (1988)
110. Le, Q., Mikolov, T.: Distributed Representations of Sentences and Documents. Int. Conf. Mach. Learn. ICML **32**, 2014 (2014)
111. Maimon, O., Rokach, L.: Introduction to knowledge discovery and data mining. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 1–15. Springer, Berlin (2009)
112. McKeown, K.R., et al.: Tracking and summarizing news on a daily basis with Columbia's Newsblaster. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 280–285 (2002)
113. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2007)
114. Hamborg, F., Lachnit, S., Schubotz, M., Hepp, T., Gipp, B.: Giveme5W: main event retrieval from news articles by extraction of the five journalistic w questions. In: Proceedings of the iConference (2018)
115. Hamborg, F., Breitinger, C., Schubotz, M., Lachnit, S., Gipp, B.: Extraction of main event descriptors from news articles by answering the journalistic five W and one H questions. In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL) (2018)
116. Evans, D.K., Klavans, J.L., McKeown, K.R.: Columbia newsblaster: multilingual news summarization on the Web. In: Demonstration Papers at HLT-NAACL 2004, pp. 1–4 (2004)
117. Hamborg, F., Meuschke, N., Gipp, B.: Matrix-based news aggregation: exploring different news perspectives. In: Proceedings of ACM/IEEE Joint Conference of Digital Libraries, vol. 10, no. 17 (2017)
118. Mitchell, R.: Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media, Inc. (2015)
119. Paliouras, G., Mouzakidis, A., Moustakas, V., Skourlas, C.: PNS: A personalized news aggregator on the web. In: Tsihrintzis, G.A., Virvou, M. (eds.) Intelligent Interactive Systems in Knowledge-Based Environments. Springer, Berlin, pp. 175–197 (2008)
120. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conferencce on Web Search and Data Mining, pp. 441–450 (2010)
121. Hamborg, F., Meuschke, N., Breitinger, C., Gipp, B.: Newsplease: a generic news crawler and extractor. In: Proceedings of the 15th International Symposium of Information Science, pp. 218–223 (2017)
122. Radev, D.R., Jing, H., Budzikowska, M.: Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In: Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization, pp. 21–30 (2000)
123. Hamborg, F., Meuschke, N., Aizawa, A., Gipp, B.: Identification and analysis of media bias in news articles. In: Proceedings of the 15th International Symposium of Information Science (2017)
124. Agran, P.F., Castillo, D.N., Winn, D.G.: Limitations of data compiled from police reports on pediatric pedestrian and bicycle motor vehicle events. Accid. Anal. Prev. **22**, 361 (1990)

125. Kim, J.W., Candan, K.S., Tatemura, J.: Efficient overlap and content reuse detection in blogs and online news articles. In: Proceedings of the 18th international conference on World wide web, pp. 81–90 (2009)
126. Meuschke, N., Gipp, B.: State-of-the-art in detecting academic plagiarism. Int. J. Educ. Integr. **9**(1), 50 (2013)
127. Zu Eissen, S.M., Stein, B.: Intrinsic plagiarism detection. In: European Conference on Information Retrieval, pp. 565–569 (2006)
128. Hoad, T.C., Zobel, J.: Methods for identifying versioned and plagiarized documents. J. Am. Soc. Inf. Sci. Technol. **54**(3), 203–215 (2003)
129. Shivakumar, N., Garcia-Molina, H.: SCAM: a copy detection mechanism for digital documents. In: In Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries (1995)
130. Gipp, B., Taylor, A., Beel, J.: Link proximity analysis-clustering websites by examining link proximity. In: International Conference on Theory and Practice of Digital Libraries, pp. 449–452 (2010)
131. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 497–506 (2009)
132. Hanna, M.: Keywords in news and journalism studies. J. Stud. **15**(1), 118–119 (2014)
133. Šarić, F., Glavaš, G., Karan, M., Šnajder, J., Bašić, B.D.: Takelab: systems for measuring semantic text similarity. Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, Association for Computational Linguistics, pp. 441–448 (2012)
134. Rychalska, B., Pakulska, K., Chodorowska, K., Walczak, W., Andruszkiewicz, P.: Samsung Poland NLP Team at SemEval-2016 Task 1: Necessity for diversity; combining recursive autoencoders, WordNet and ensemble methods to measure semantic similarity. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016) (2016)
135. Brychcin, T., Svoboda, L.: UWB at SemEval-2016 Task 1: semantic textual similarity using lexical, syntactic, and semantic information. In: International Workshop on Semantic Evaluation (2016)
136. Spitz, A., Gertz, M.: Breaking the news: extracting the sparse citation network backbone of online news articles. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 274–279 (2015)
137. The Media Insight Project. The Personal News Cycle: How Americans Get Their News (2014)
138. Haes, J.W.H.: September 11 in Germany and the United States: reporting, reception, and interpretation. Cris. Commun. Lessons Sept. **11**, 125–132 (2003)
139. Smith, J., McCarthy, J.D., McPhail, C., Augustyn, B.: From protest to agenda building: description bias in media coverage of protest events in Washington, DC. Soc. Forces **79**(4), 1397–1423 (2001)
140. Shalaby, W., Zadrozny, W., Jin, H.: Beyond word embeddings: learning entity and concept representations from large scale knowledge bases. Inf. Retr. J. (2018). https://doi.org/10.1007/s10791-018-9340-3#citeas
141. Corman, S.R., Kuhn, T., McPhee, R.D., Dooley, K.J.: Studying complex discursive systems. Hum. Commun. Res. **28**(2), 157–206 (2002)
142. Niven, D.: Tilt?: The Search for Media Bias. Greenwood Publishing Group, Westport (2002)
143. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)
144. Kaya, M., Fidan, G., Toroslu, I.H.: Sentiment analysis of turkish political news. In: Proceedings of the 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology, vol. 01, pp. 174–180 (2012)
145. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. EMNLP **4**, 412–418 (2004)
146. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168–177 (2004)
147. Godbole, N., Srinivasaiah, M., Skiena, S.: Large-scale sentiment analysis for news and blogs. ICWSM **7**(21), 219–222 (2007)
148. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. LREC **10**, 2200–2204 (2010)
149. Park, S., Ko, M., Kim, J., Liu, Y., Song, J.: The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, pp. 113–122 (2011)
150. Strapparava, C., Mihalcea, R.: Semeval-2007 task 14: affective text. In: Proceedings of the 4th International Workshop on Semantic Evaluations, pp. 70–74 (2007)
151. Grefenstette, G., Qu, Y., Shanahan, J., Evans, D.: Coupling niche browsers and affect analysis for an opinion mining application. In: Proc. 12th Int. Conf. Rech. d'Information Assist. par Ordinat., pp. 186–194 (2004)
152. Subasic, P., Huettner, A.: Affect analysis of text using fuzzy semantic typing. IEEE Trans. Fuzzy Syst. **9**, 483–496 (2001)
153. Mishne, G.: Experiments with mood classification in blog posts. In: Proc. ACM SIGIR 2005 Work. Stylist. Anal. Text Inf. Access (2005)
154. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science (80-) **356**, 183 (2017)
155. Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: Advances in Neural Information Processing Systems, pp. 4349–4357 (2016)
156. Stempel, G.H., Windhauser, J.W.: The prestige press revisited: coverage of the 1980 presidential campaign. Journal. Mass Commun. Q. **61**(1), 49 (1984)
157. Stovall, J.G.: Coverage of 1984 presidential campaign. Journal. Mass Commun. Q. **65**(2), 443 (1988)
158. Cohen, N.S.: At work in the digital newsroom. Digit. J. **2**, 94 (2018)
159. Mori, S., Nishida, H., Yamada, H.: Optical Character Recognition. Wiley, Hoboken (1999)
160. Jain, A.K., Bhattacharjee, S.: Text segmentation using Gabor filters for automatic document processing. Mach. Vis. Appl. **5**(3), 169–184 (1992)
161. Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford InfoLab (1999)
162. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: a new benchmark collection for text categorization research. J. Mach. Learn. Res. **5**, 361–397 (2004)
163. Rosenberg, S.W., Bohan, L., McCafferty, P., Harris, K.: The image and the vote: the effect of candidate presentation on voter preference. Am. J. Pol. Sci. **30**, 108–127 (1986)
164. Rusk, D.: How the internet misled you in 2015 (2015). https://www.bbc.com/news/world-35051618. Accessed 18 Nov 2005
165. Estrin, J.: The Real Story About the Wrong Photos in #BringBackOurGirls (2014). https://lens.blogs.nytimes.com/2014/05/08/the-real-story-about-the-wrong-photos-inbringbackourgirls/. Accessed 18 Nov 2005

166. Dearden, L.: The fake refugee images that are being used to distort public opinion on asylum seekers (2015). https://www.independent.co.uk/news/world/europe/the-fake-refugee-images-that-are-being-used-to-distort-public-opinion-on-asylum-seekers-10503703.html. Accessed 18 Nov 2005
167. Kerrick, J.S.: News pictures, captions and the point of resolution. Journal. Mass Commun. Q. **36**(2), 183–188 (1959)
168. Waldman, P., Devitt, J.: Newspaper photographs and the 1996 presidential election: the question of bias. Journal. Mass Commun. Q. **75**(2), 302–311 (1998)
169. Kenney, K., Simpson, C.: Was coverage of the 1988 presidential race by Washington's two major dailies biased? Journal. Mass Commun. Q. **70**(2), 345–355 (1993)
170. Kepplinger, H.M.: Visual biases in television campaign coverage. Commun. Res. **9**(3), 432–446 (1982)
171. Van Gorp, B.: Where is the frame? Victims and intruders in the Belgian press coverage of the asylum issue. Eur. J. Commun. **20**(4), 484–507 (2005)
172. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
173. De Silva, L.C., Miyasato, T., Nakatsu, R.: Facial emotion recognition using multi-modal information. In: Proceedings of 1997 International Conference on Information, Communications and Signal Processing, 1997. ICICS, vol. 1, pp. 397–401 (1997)
174. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
175. Feng, Y., Lapata, M.: Automatic caption generation for news images. IEEE Trans. Pattern Anal. Mach. Intell. **35**(4), 797–812 (2013)
176. Segalin, C., Cheng, D.S., Cristani, M.: Social profiling through image understanding: personality inference using convolutional neural networks. Comput. Vis. Image Underst. **156**, 34 (2017)
177. Kerrick, J.S.: The influence of captions on picture interpretation. Journal. Mass Commun. Q. **32**(2), 177–182 (1955)
178. Wikinews. Main Page—Wikinews, The free news source you can write. (2015)
179. Spiegel Online: Übertreibt Horst Seehofer seine Attacken? Das sagen die Medien. http://www.spiegel.de/politik/deutschland/uebertreibt-horst-seehofer-seine-attacken-das-sagen-die-medien-a-1076867.html. Accessed 05 Sep 2017 (2016)
180. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investig. **30**(1), 3–26 (2007)
181. Park, S., Lee, K., Song, J.: Contrasting opposing views of news articles on contentious issues. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 340–349 (2011)
182. Miller, M.M.: Frame mapping and analysis of news coverage of contentious issues. Soc. Sci. Comput. Rev. **15**(4), 367–378 (1997)
183. Smith, A., Hawes, T, Myers, M.: Hiérarchie: interactive visualization for hierarchical topic models. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 71–78 (2014)
184. Fuhr, N. et al.: An information nutritional label for online documents. In: ACM SIGIR Forum, pp. 46–66 (2018)
185. Hayes, A.F., Krippendorff, K.: Answering the Call for a Standard Reliability Measure for Coding Data. Commun. Methods Meas. **1**(1), 77–89 (2007)
186. Hripcsak, G., Rothschild, A.S.: Agreement, the F-measure, and reliability in information retrieval. J. Am. Med. Inform. Assoc. **12**(3), 296–298 (2005)
187. Recasens, M., Danescu-Niculescu-Mizil, C., Jurafsky, D.: Linguistic models for analyzing and detecting biased language. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 1650–1659 (2013)
188. Munson, S.A., Resnick, P.: Presenting diverse political opinions: how and how much. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1457–1466 (2010)
189. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436 (2015)
190. Kumar, A. et al.: Ask me anything: dynamic memory networks for natural language processing. arXiv (2015)
191. dos Santos, C.N., Gatti, M.: Deep convolutional neural networks for sentiment analysis of short texts. Coling-2014 (2014)