Time-aware online reputation analysis
Peetz, M.H.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Analyzing Annotation Procedures and Indicators for Reputation Polarity of Online Media

Observational studies, whether in the form of user studies, user panels, or a large-scale log analysis, can help us understand users’ information behavior, how people interact with information. Outcomes of such studies may also inform contrastive experiments related to users’ search experience. Such studies are especially important and informative in the setting of newly emerging search paradigms or newly emerging information sources. Because of this, a large number of observational studies relating to accessing social media have appeared over the past decade. Early studies focused on blogs, bloggers, and blog search [166]. Later studies look at the full spectrum of social media available. Often, one of the key questions asked is who the people using social media are and how they use it. Several observational studies have looked at who uses social media [69, 140]. Other work has looked at how people in general use social media, in particular Twitter [117, 119, 170] or how subgroups of users use social media for support when facing disease [79], communication with loved ones in the army [203], teenagers [63], or coordination of revolutions [152].

By and large these observational studies of users interacting with social media focus on non-professional users, whose information behavior is mostly triggered by private goals, concerns, and interests. In contrast, the information behavior of professionals is especially interesting as there often is a clear, and sometimes even formalized, understanding of the tasks that give rise to their behavior. As a consequence, the information behavior of professionals has been studied extensively, though not for social media professionals. For instance, Huurnink et al. [112] study the search behavior of media professionals to help inform the development of archival video search facilities, while Hofmann et al. [107] analyse the search processes of communication advisors to infer contextual factors that help improve the effectiveness of algorithms for a specific expertise retrieval task. Bron et al. [40] study the interleaved search, research and authoring processes of media studies scholars to inform the design of multiple search interfaces. However, our understanding of professionals’ information behavior related to social media is very limited, as few studies on the topic have been published to date. In this chapter we address this gap. We study the information behavior of professional social media analysts, in
particular reputation analysts.

Reputation analysts distinguish themselves from other social media users in several respects. First, they are consumers of social media data, not contributors. They analyse and annotate data: their goal is to find the reputation of a company based on a pool of data. For that, they filter the single media expressions for relevance and annotate them for reputation polarity. At the end of the process, reputation analysts abstract from this analysis or write reports on the reputation polarity and alert, both are often split by topics. Secondly, they are professionals with a vast background knowledge in the domains in which they are active. This background knowledge is the result of consistent online and offline monitoring and of focused study of the particular domain.

We first look at the procedures of social media analysts which results in our first research question:

RQ1.1 What are the procedures of (social media) analysts in the analysis and annotation of reputation polarity?

Answers to this question can lead to better (semi-manual) algorithms for estimating the reputation polarity of single media expressions. Better algorithms can facilitate the process of estimating reputation polarity, as well as the overall reputation. We use two approaches to answer this question. First, we are analyzing log data of an annotation tool for online media to give quantitative results for four media: Youtube, Google, Facebook and Twitter.

Second, focusing on tweets, we then ask media analysts responsible for annotation about their annotation process. We only focus on tweets for two reasons. First, the benchmarks for automatic classification are for Twitter data and secondly, the metadata is rich but also rather uniform. We find that the annotation process varies over different media types. For tweets in particular, analyzing and understanding the author, topic, and reach of a tweet is vital for estimating the reputation polarity. The reach of a tweet is who and how many people are exposed by the tweet.

Early attempts at automatically estimating the reputation polarity use the sentiment of online conversations [108]. While such approaches are automatic and therefore low-cost approaches, sentiment analysis proves to be prone to delivering incorrect results for reputation polarity [6]. Therefore, single media items are still mainly annotated manually for reputation polarity and (semi)-automatic approaches to the annotation process are needed. A range of initiatives are underway to address this need [6, 7, 280]. For instance, the RepLab 2012–2014 challenges [6, 7, 9] at CLEF provide annotated training and testing data for teams to develop and test systems to estimate reputation polarity.

With rather poor effectiveness of the systems, we cannot declare the automatic estimation of reputation polarity as understood.

This observation naturally leads to our second research question:

RQ1.2 On a per tweet level, what are the indicators that reputation analysts use to annotate the tweet’s impact on the reputation of a company?

We again use two different approaches to answer this question. First, we ask media analysts as to what their indicators are. Secondly, instead of retrospective self-analysis, we record media analysts in their usual annotation environment while thinking aloud. This
allows us to analyse their behavior. Among others, we find several key indicators that help estimating reputation polarity. In particular, the topic of a tweet and the authority—the topical authority—of its author, are vital for the estimation of its reputation polarity. This topical authority does not need to be based on online data, but can in fact be based on offline media and social networks as well. Additionally, the number and type of followers is important; again the reach of a tweet.

Our contribution lies in the analysis of a unique subgroup of professional social media users: reputation analysts. They consume a vast number of social media and need to deduct from individual findings without directly contributing to conversations. In order to understand the procedures and thought processes leading to decisions, we use three datasets (two observational studies and one survey) instead of one, which is common. This allows us to approach the research questions from different angles. We also contribute indicators and feature types that can be used to improve (semi-)automatic estimators for reputation polarity.

We introduce related work on our methodology in Section 2.2. We introduce our three datasets in Section 4.2. Section 4.3 and 4.4 analyse the answers to RQ1.1 and RQ1.2, respectively. We discuss the impact and consequences of our findings and conclude in Section 4.5.

4.1 Methodological Background

We introduce different approaches that have been used to understand subgroups of social media users in Section 4.1.1. We then introduce related work to our log analysis methodology in Section 4.1.2 and to our think aloud protocol in Section 4.1.3. Finally, we introduce annotation interfaces and how they have been analysed in Section 4.1.4.

4.1.1 Understanding Social Media Users

Approaches to understanding subgroups and cultures using social media range from online to offline approaches [175]. danah boyd [63] uses both, online and offline participant-observation for ethnographic community descriptions of teenagers. Examples of offline approaches are phone interviews, either for quantitative analysis [69, 140] or qualitative analysis [79, 203]. Alternative offline approaches are surveys [79] or in person interviews [63]. Online approaches to analyse Twitter often rely on crawled posts, be it user timelines [119, 170], network structures [119], filtered posts based on brands [117] or hashtags [152]. Posts or timelines are then often manually annotated [117, 170]. The network flow of topics can be monitored looking at the propagation of hashes or shingles [152].

Since our user group is a passive consumer, we need to rely on a different set of approaches. We monitor the consuming behaviour using log analysis and the think aloud protocol. Similar to ethnographic studies [63, 79] we do a retrospective survey.
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

4.1.2 Log Analysis

Since the mid-1990’s log data analysis of web sites is considered a typical approach to understand user patterns [20], while the analysis of information retrieval log data goes further back [162]. Early work focusses on time series analysis of access patterns [115]. Commonly used techniques [114] for log data analysis are immediate task cancellations, shifts in direction during task completion, discrepancies during task model and task completion [20], mining association rules of browsing behavior [35], mining user browsing patterns and building hypertext probabilistic grammars [36] or trees [192]. Maximum repeated patterns of user sessions have been analysed in user interfaces [226] in general as well as web interfaces [39]. As an addition to explicit user feedback, such as clicks and requests, Velayathan and Yamada [256] also use implicit feedback, such as dwell time.

In general, we refer to an annotation as a comment and metadata attached to media, such as text, images, or video. An annotation can be as simple as a binary classification (spam/ham) to more advanced annotations like trees (syntax or dependency trees).

4.1.3 Think Aloud

Verbal reports are reports by users before, while, and after they perform a task. Nisbett and Wilson [173] review a number of studies in which verbal reports can give false results, in particular retrospective and interruptive reports. Lewis [141] introduce the think aloud protocol as we know it now, using the concurrent and introspective think aloud protocol, that was later introduced to usability studies [142]. We are using the introspective think aloud protocol, as excellently described in [255]. To the best of our knowledge, there is no prior research on using this method to find indicators for annotations of reputation polarity.

4.1.4 Annotation Interfaces

Together with the different annotation tasks addressed by our reputation analysts, such as filtering and reputation polarity assignment, come different kinds of interfaces. In the following we split the interfaces into two groups: interfaces for tagging and interfaces for (often linguistic) coding. The main difference between the interfaces is not the how, but the why of annotation. Tagging interfaces are often inherent parts of a system: be it assigning tags to photos, linking photos with other users, or assigning hashtags to tweets. The motivation to tag is implicit: users are looking for a better experience of the system, be it to be connected to more friends or to reach more users with their tweets or photos. Additionally, the annotations themselves are spread over a large user base. Interfaces for coding media exist only for the coding itself: the goal is the coding and not a pleasurable experience of the system. The tasks range from assessing relevance to documents [16, 68, 196] to more advanced linguistic features [60, 77, 169]. Coding interfaces can also be used for media analysis, where the media is being annotated for

---

1http://www.flickr.com
2http://www.facebook.com
3http://www.twitter.com
different stakeholders, reputation polarity, or audience. For coding interfaces the annotators (or assessors) are fewer, but they do more annotations. Most importantly, they often get paid and are therefore explicitly motivated. Additionally, a vital feature of tagging interfaces is that they need to be easy and intuitive to use, while assessors using coding interfaces may be trained.

Now we turn to (publicly available) coding interfaces. For many simple tasks, for example the coding of sentiment or relevance, tailor-made (non-public) interfaces or even spreadsheets are being used [43]. For relevance assessments, Downey and Tice [68] show the interface and its requirements for assessing relevance for topic modeling. They use traditional usability tests to evaluate the interface. Piwowarski et al. [196] have consistently been improving the interface to assess relevance of structured documents, basing their analyses of the assessments on inter-annotator agreement and document coverage. For linguistic features, such as the annotation tree structures or roles, interfaces such as GATE [60], MMAX2 [169], and the ITU treebank annotation tool [77] have been used quite extensively. As to guidelines on how to design such an interface, one of the most important requirements found in early interfaces is that there is a minimum of motion required to make an annotation [274]. Burghardt [43] provides guidelines to assess the usability of those interfaces and shows that many of them do not follow simple usability principles and are often created based on rough intuition of the engineer. As far as we know, only requirement analysis and usability studies of the interfaces themselves have been performed. Analyses with respect to processes of how people annotate a specific problem are missing.

4.1.5 Related Annotation Tasks

While credibility of media is not reputation polarity, it can be an important part. Hilligoss and Rieh [103] introduce a framework for online credibility, finding that social, relational and dynamic frames surrounding the information seeker provide boundaries for their judgments. St. Jean et al. [233] conduct phone interviews with online content contributors to analyse how they establish and assess the credibility of online content. They find that they use one or more of three types of credibility-related judgments intuitive, heuristic, and strategy-based. The intuitive assessment is not random, rather based on an instinct after consuming the information. The heuristic assessment was to stay within familiarity and authority of the author. Strategy-based assessments entail cross-referencing of information and accessing primary sources. There are several differences between our work and previous work on the assessment of online credibility. First, we look at reputation polarity instead of credibility and while credibility (or authority) is an important indicator for reputation polarity, it is not the same. Secondly, we are looking at the indicators used by social media analysts who studied and learnt the assessment of (online) information. One of the indicators we find is in fact user authority, which is similar to credibility, while the other indicators are not only about credibility.
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

4.2 Datasets

The general reputation of a company is determined by media analysts. Media analysts follow offline and online news and digest information about the company and conversations about it. They write daily monitoring or quarterly profiling reports. Social media analysts focus on monitoring and profiling the conversation about a company in social media. Below, we will refer to media analysts and social media analysts together as analysts.

In the following we introduce the datasets, in particular the method of their creation, and the participants. We begin with a questionnaire dataset in Section 4.2.1. Section 4.2.2 introduces a log dataset from an annotation interface, while Section 4.2.3 explains the methodology we used to collect and code videos of analysts in their annotation environment. Section 4.2.4 links the research questions to the datasets.

4.2.1 Questionnaires

We asked a pool of analysts working for a leading Spanish speaking communication and reputation consultancy to fill out a questionnaire about their approach to annotating tweets as part of their daily routine. In total we have 15 responders, of which 12 are between 25 and 34, 2 between 18 and 24, and one between 35 and 44 years old. 7 of the participants are male, the rest female. They worked between 1 and 93 months in the field of ORM (mean: 20.47±17.44). Figure 4.1a shows the analysts’ university background. Interestingly, only 4 analysts have a background in marketing.

The analysts work with companies throughout the entire market spectrum. In detail, they analyse 19 different sectors,4 every analyst analyses companies in 3.8±2.07 unique sectors (min: 1, max: 9), and analyses 5.07±2.98 companies in general (min: 1, max: 13). Figure 4.1b shows the number of analysts per sector. We can see that 80% of the analysts work with a client in the energy sector. Otherwise, the types of sectors, and therefore clients, are very diverse, ranging from government to retail. The actual questionnaire consists of 12 questions (see Table 4.1) and we finish with an open question for comments (kept blank by every participants). The full set-up of the questionnaire can be inspected online.5 Questions with an answer format in the form of a Likert scale (5 point or binary) have a high number of answers (10–15), while open questions, requiring a higher cognitive load, tend to have very few answers (6–8). The two questions where the user was asked to click an image turned out to be difficult: only 4–5 people clicked the images. The reasons could be that the image loading time was too long or the question format was new and participants did not know the procedure.

4The sectors are: Pharmaceutical, Telecom, Law, Finance, Tobacco, Food, Media, Energy, Housing, Mining, Banking, Health, Cosmetics, Car, Government, Tech, Retail, Tourism, and Insurance.

5https://uvafeb.az1.qualtrics.com/SE/?SID=SV_eqR2gDsJKitaDGD
4.2. Datasets

Figure 4.1: *(Questionnaire dataset)* (a) Analysts who participated in the questionnaire with their university background. (b) Analysts and the sectors they are working on. A single analyst may work in multiple sectors and may have multiple university backgrounds.

Table 4.1: Overview over the questions of the questionnaire with additional information. LK stands for 5 point Likert scale, B for binary, OQ for open question, and CI for a clickable image.

<table>
<thead>
<tr>
<th>Topic</th>
<th>ID</th>
<th>Question</th>
<th>Answer format</th>
<th># Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation process</td>
<td>Q1</td>
<td>Below (see Table 4.2) you can find several steps in the process of annotating reputation polarity for a tweet. Please indicate how important this step is in your annotation process.</td>
<td>LK</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Q2</td>
<td>Imagine you had an automatic tool (not 100% reliable) for the steps in the process. Which of the steps below⁶ must be done manually and which steps would you like to have automated?</td>
<td>LK</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Q3</td>
<td>In which steps⁷ of the annotation process do you use background information? Background information can be but is not limited to lists of important users, links to articles or blogs, meaning of hashtags.</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Q4</td>
<td>What kind of background information do you acquire in the annotation process of tweets?</td>
<td>OQ</td>
<td>6</td>
</tr>
</tbody>
</table>
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

Table 4.1: Overview over the questions of the questionnaire with additional information. LK stands for 5 point Likert scale, B for binary, OQ for open question, and CI for a clickable image.

<table>
<thead>
<tr>
<th>Topic</th>
<th>ID</th>
<th>Question</th>
<th>Answer format</th>
<th># Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicators for reputation polarity</td>
<td>Q5</td>
<td>Can you give me a list of typical indicators or features for the reputation polarity of a tweet?</td>
<td>OQ</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Q6</td>
<td>Consider the following indicators for reputation polarity of a tweet. In your annotation process, how important are the indicators for the annotation of reputation polarity?</td>
<td>LK</td>
<td>12</td>
</tr>
<tr>
<td>Indicators for user authority</td>
<td>Q7</td>
<td>Please click on the position of the user profile below (see Figure 4.10a) that is most indicative for the users authority.</td>
<td>CI</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Q8</td>
<td>Is this user (see Figure 4.10a) an authority?</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Q9</td>
<td>Please click on the position of the user profile below (see Figure 4.10c) that is most indicative for the users authority.</td>
<td>CI</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Q10</td>
<td>Is this user (see Figure 4.10c) an authority?</td>
<td>B</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Q11</td>
<td>What is an authoritative Twitter user with respect to a company? Please explain what an authoritative user means to you.</td>
<td>OQ</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Q12</td>
<td>How do you find important and authoritative Twitter users? Please provide a list of steps in the steps in the search process.</td>
<td>OQ</td>
<td>8</td>
</tr>
<tr>
<td>General</td>
<td>Q13</td>
<td>Which of the following statements apply for the analysis of reputation polarity of tweets? (Reputation analysis [is costly, is important for ORM, is fault-tolerant, is tedious, must not have mistakes, is time-consuming])</td>
<td>LK</td>
<td>10/11</td>
</tr>
</tbody>
</table>

4.2.2 Annotation System Logs

In the following we introduce a dataset based on a web interface used for annotating documents for reputation. We first introduce the annotation interface in Section 4.2.2 and then proceed with a description of the log data collected in Section 4.2.2.
4.2. Datasets

Table 4.2: (Questionnaire dataset) Steps taken that lead to an annotation as indicated in the questionnaire.

<table>
<thead>
<tr>
<th>ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Understand the content of linked images</td>
</tr>
<tr>
<td>1</td>
<td>Read the web page linked in the tweet</td>
</tr>
<tr>
<td>2</td>
<td>Read the tweet</td>
</tr>
<tr>
<td>3</td>
<td>Read the replies to a tweet</td>
</tr>
<tr>
<td>4</td>
<td>Read the profile of the author of the tweet</td>
</tr>
<tr>
<td>5</td>
<td>Read the comments in a web page the tweet links to</td>
</tr>
<tr>
<td>6</td>
<td>Look at linked images</td>
</tr>
<tr>
<td>7</td>
<td>Finding tweets</td>
</tr>
<tr>
<td>8</td>
<td>Estimating the topic of the tweet</td>
</tr>
<tr>
<td>9</td>
<td>Determine type of the author of the tweet</td>
</tr>
<tr>
<td>10</td>
<td>Determine the opinion of retweets of the tweet</td>
</tr>
<tr>
<td>11</td>
<td>Determine the opinion of replies to the tweet</td>
</tr>
<tr>
<td>12</td>
<td>Determine the opinion of comments to linked web page</td>
</tr>
<tr>
<td>13</td>
<td>Determine the opinion of a linked web page</td>
</tr>
<tr>
<td>14</td>
<td>Determine the meaning of a hashtag</td>
</tr>
<tr>
<td>15</td>
<td>Determine the audience</td>
</tr>
<tr>
<td>16</td>
<td>Determine if the tweet is important and note</td>
</tr>
<tr>
<td>17</td>
<td>Determine authoritativeness of the author of the tweet</td>
</tr>
<tr>
<td>18</td>
<td>Determine an opinion in the tweet</td>
</tr>
</tbody>
</table>

Annotation Interface

In the following we describe the different parts of the web interface. After logging in, the user starts with a panel to select the project to work on. The project screen then displays the different brands whose reputation is being compared and allows for entering new brands. After clicking on the brand, the user is in the main annotation screen. On the right side of the screen the user can see different tabs, leading to *General*, *Google*, *Youtube*, *Facebook*, *Twitter*, and *Graphics* panels. The *General* panel allows for entering, summarizing, and downloading scores. The *Google*, *Youtube*, *Facebook*, and *Twitter* panels are the actual annotation panels and provide lists of documents to be annotated from the respective media source. The *Graphics* panel combines the scores visually. Our analysis focusses on the annotation panels.

Annotation Panel  We proceed with the main part of the interface, the annotation panel. While there are different panels for different media types, their functionality is principally the same. We explain the functionality of the interface for web results and detail the difference for the media sources afterwards.

Figure 4.2 shows a mock up of the Google panel for the mock company *kumquat*. The panel consists of *global* and *local* areas. The local area is reserved for annotations of a single document, while the global areas allow for annotating multiple documents. For a single document, analysts can see only the title of the document. A click on the
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

Figure 4.2: (Log dataset) Annotation panel for the top Google results for the fictional entity Kumquat.

title expands the document within a frame. Next to the title, one can download and then see the authority score for the document. The authority score for web results is the PageRank, externally estimated. While the previous fields are meant to provide the analysts with information, the next fields we discuss are for annotations: a drop-down menu for the dimensions and their attributes (see Table 2.1), the audience (ranging from general to expert journalists), the reputation polarity (ranging from -2 to 2), and the relevance (ranging from 0 to 5: the default relevance is set to 5). The next two fields are either to save the annotations or delete the document. Deletion means that the document is simply not relevant to the company.

The global actions are either filtering or deleting multiple documents. Deletion can be done by ticking the checkbox before the title of the document and then deleting all of them at once. The filtering of the currently visible documents can be done based on the dimension and the audience. While deletion actually deletes the documents, filtering just offers a new view on the already annotated documents.

Differences between the Annotation Panels  In principle, the panels are the same for different data sources. However, for the expansion of the documents, the panels with Twitter and Facebook data open a new browser window with the original Facebook or Twitter source. Additionally, the authority score depends on the media type: it is the

---

8This is due to restrictions of the Twitter and Facebook APIs, which do not allow framing the original source.
number of views for Youtube videos, the number of fans or members for a Facebook page or group, and the number of followers of the author of a tweet.

Log Data from the Annotation Interface

We collect click data from the annotation interface. Based on this click data we can define sessions and actions, and report some basic statistics.

**Sessions and Actions** A session $s$ is the period of activity between a user logging in and logging out, by herself or automatically after a time out. Each session is unique, but a user may have different sessions. A user action $s_a$ within a session $s$ is a click activated request. An action can be a navigational action, an informational action, or an annotational action. Navigational actions are actions that navigate from one part of the program to another. For example, the google action switches to the Google annotation panel. Informational actions are actions that help the decisions in the annotation, like expanding the media item, requesting further information, or filtering a list of items according to a criterion. Annotational actions are actions where an annotation is assigned to a media item. An annotation session $S$ is a subsession of a session, which is devoid of navigational actions and contains at least one annotation action. An annotation session can be in different modi $m$, depending on where the user navigated to. In fact, as we are not interested in the metadata annotations per se, we define annotation sessions as sessions where $m \in \{\text{Google, Twitter, Youtube, Facebook}\}$.

**Table 4.3: (Log dataset)** The actions that can be logged by the annotation tool.

<table>
<thead>
<tr>
<th>Navigational</th>
<th>Informational</th>
<th>Annotational</th>
</tr>
</thead>
<tbody>
<tr>
<td>login</td>
<td>expanding</td>
<td>delete-single</td>
</tr>
<tr>
<td>login-fail</td>
<td>filter</td>
<td>delete-multi</td>
</tr>
<tr>
<td>google</td>
<td>hide</td>
<td>insert-metadata</td>
</tr>
<tr>
<td>twitter</td>
<td>request-authority</td>
<td>modify-metadata</td>
</tr>
<tr>
<td>youtube</td>
<td>request-metadata</td>
<td>save-annotation</td>
</tr>
<tr>
<td>facebook</td>
<td>graphics</td>
<td></td>
</tr>
<tr>
<td>overview</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Basic statistics** In total we have 127 sessions, 42 contain at least one annotational action. The average number of actions per session is $84.50 \pm 218.74$ and $231.60 \pm 332.52$ for the sessions and sessions with at least one annotation, respectively. The sessions are between 1 and 1308 actions long, and take on average $137943.33 \pm 451822.11$ seconds (removing the outliers: $58347.07 \pm 175098.87$). Sessions with at least one annotation are between 8 and 1308 actions long and take on average $179869.67 \pm 458874.88$ seconds (removing the outliers: $66060.00 \pm 184169.82$ seconds). We could identify at least 11 distinct users, but there can be more because different parts of the company share login data. In total we have 331 annotation sessions.
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

4.2.3 Think Aloud

The think aloud dataset is a set of coded videos of analysts thinking aloud while annotating tweets for reputation polarity. We asked 4 social media analysts to annotate tweets for reputation polarity. The analysts were between 24 and 29 years old (mean: 27, median: 27.5). They have been annotating tweets between 3 weeks and 14 months (mean: 35.5 weeks, median: 39 weeks). We used the introspective think aloud protocol from [255].

The analysts were at their workspace, about to follow their daily routine of writing a daily monitoring report which includes tweets with varying reputation polarity. They were first asked to read and sign an agreement to be filmed in this process and that this data can be used for research purposes. The task is to explain what features they look at to decide the reputation polarity of a tweets, while they are actually annotating the tweets. The experimenter explained the think aloud protocol: as they were non-native English speakers they were allowed to utter their thoughts in their native language (Spanish). After finishing the report or when no tweets needed further annotation, the think aloud process was stopped and the experimenter had an informal interview with the analyst, confirming metadata like age and work experience. Every analysts was thanked with some sweets.

Coding

In total, we have 31 annotated tweets. The codebook for the single tweets was created based on grounded theory [236] and two annotators created a hierarchical codebook in Table 4.3. The codebook has five categories: general, webpage, user, metadata, and text. Every category has subcategories. For example, when an analyst looks at a website, she can look at more specifically the keywords, age, header, etc., of the website.

The tweets were annotated by the same two coders. We measure the inter-coder agreement in two ways: per tweet and for all codes. The inter-coder agreement per tweet analyses for how many tweets there is a complete agreement. The inter-coder agreement for all codes is the kappa inter-coder reliability. We report both agreements because there are many codes decreasing the chances of agreeing anyway. The initial inter-coder agreement was 0.55 at the tweet level and 0.872 at a global level for all codes. We found that most of the disagreement was based on a misunderstanding between the codes keywords and topic and well as the dependency of the codes authority and relevancy of a user on other user codes (the codes are not present in the adjusted codebook in Table 4.11). Recoding with an adjusted codebook, we now have an agreement of 0.65 on a tweet level and 0.924 on a global level. This agreement is acceptable, in particular considering the size of the codebook. In the following we only report on codes both coders agreed on.

4.2.4 Summary

We introduced two research questions in the introduction. For RQ1.1, where we are identifying the procedures of brand analysts in the annotation of reputation polarity, we use log datasets and parts of the questionnaire (Q1–Q4). Section 4.3.3 provides the analysis. For RQ1.2, were we are identifying specific indicators for reputation polarity
4.2. Datasets

- **general**
  - background knowledge

- **webpage**
  - keywords
  - age
  - header (title)
  - relevant parts
  - full webpage
  - relevancy
  - opinions
  - reputation
  - author
  - known

- **user**
  - is company
  - is newspaper/journalist
  - known
  - #followers
  - #tweets
  - tweets in tweetstream
    - * sentiment
    - * reputation

- **metadata**
  - #retweets
  - favorited
  - known
  - related tweets
    - * retweets
      - sentiment
      - reputation
    - * earlier, similar tweets
      - sentiment
      - reputation
    - * replies
      - sentiment
      - reputation
    - * sentiment
    - * reputation

- **text**
  - keywords
  - topic
    - * webcare
    - * other
  - opinion
  - sentiment
  - age

Figure 4.3: *(Think aloud dataset)* Hierarchical codebook used for analyzing the indicators considered by social media analysts before they come to a decision on reputation polarity.

on tweets, we use the codes from the think aloud dataset as well as the second part of the questionnaire (Q5–Q12). Section 4.4 analyses the answers to **RQ1.2**.
4.3  Annotation Procedures for Reputation Polarity

One approach to the definition of reputation polarity focusses on how reputation polarity of a media item can be determined. In this section we analyse the process of annotating for reputation polarity. In particular, we discuss the answer to the question:

**RQ1.1** What are the procedures of (social media) analysts in the analysis and annotation of reputation polarity?

We provide a two-part answer to this question. In Section 4.3.1 we analyse how people come to annotation conclusions based on log data of an annotation interface. This data is based on different kinds of media. We then focus on social media, in particular on Twitter data, and analyse what analysts state to be important procedures in the questionnaire dataset in Section 4.3.3. We summarize the findings in Section 4.3.4.

### 4.3.1 Analysis of Log Data

The overall research question is to understand the process of annotation of reputation polarity. We use log data introduced in Section 4.2.2 to answer the two research questions:

**RQ1.1.a** What are the process actions that lead to annotations?

**RQ1.1.b** Of the different annotation actions, which ones are the most time intensive?

We begin with an analysis of typical annotation sessions to answer **RQ1.1.a** and proceed with **RQ1.1.b** in Section 4.3.2. What actions lead to an annotation? In the following we analyse typical actions in annotation sessions and proceed with the more general analysis of typical sessions.

**Typical Actions** We first analyse typical actions that lead to an annotation decision. An annotation decision is an annotational action, as defined in Section 4.2.2, in particular Table 4.3. The annotational actions are the actual act of annotating a media item (save-annotation), and two kinds of filtering, batch-deletion (delete-multi) and single media item deletion (delete). Table 4.4 shows the three most probable actions that have been logged before one of the three annotation actions with the probability that this action was logged right before the respective annotational action.

Table 4.4: *(Log dataset)* The probability of the most frequent actions to leading to an annotational action. Actions marked with A, I, and N are annotational, informational, and navigational, respectively. Higher probabilities mean that the action occurred more frequently before the annotational action.

<table>
<thead>
<tr>
<th></th>
<th>save-annotation</th>
<th>delete-multi</th>
<th>delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>save-annotation (A)</td>
<td>0.52</td>
<td>0.35</td>
<td>0.46</td>
</tr>
<tr>
<td>request-authority (I)</td>
<td>0.22</td>
<td>0.25</td>
<td>0.36</td>
</tr>
<tr>
<td>expanding (I)</td>
<td>0.18</td>
<td>0.14</td>
<td>0.11</td>
</tr>
</tbody>
</table>

44
4.3. Annotation Procedures for Reputation Polarity

Figure 4.4: (Log dataset) A visualization of the actions of analysts in annotation sub-sessions. The different colors indicate the different actions the analysts performed. To account for a lack of space, we show a sample of 20% of all annotation sessions. C1–C9 indicate the different clusters based on K-Means.

We can see several things in Table 4.4. Both delete and save-annotation are repeated actions in about half of all cases because the actions leading to them are delete and save-annotation in 46% and 53% of all cases. Secondly, for all three annotation types the expanding an information processing action such as request-authority or expanding has been logged in between 35% and 40% of all cases before. Analysts were requesting more information before they proceeded to cast an annotation decision. While save-annotation and delete are embedded in most (87.01% and 25.07%, respectively) annotation sessions, deleting multiple objects happens directly after switching to a different media type. One peculiar thing we can see in Table 4.4 is that request-authority, i.e., requesting the authority for an information object, is not very probable to happen directly before deletions. In fact, looking at the long tail of preceding actions, request-authority has only been logged in 2% of all cases before a delete and never before a delete-multi.

Table 4.5 shows the two preceding actions performed before annotational actions, ranked by probability. It supports the earlier observations.

Typical Annotational Sessions We continue with an analysis of typical annotation sessions. Figure 4.4 displays a clustered subsessions with at least one annotation, for the

---

9Due to legal reasons we cannot publish the entire set of log data.
Table 4.5: (Log dataset) The probability of the most frequent two preceeding actions for transitioning to an annotational action. Actions marked with A, I, and N are annotational, informational, and navigational, respectively. Higher probabilities mean that the two actions occurred more frequently before the annotational action.

<table>
<thead>
<tr>
<th>save-annotation</th>
<th>delete-multi</th>
<th>delete</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.46 save-annotation (A), save-annotation (A)</td>
<td>0.21 expanding (I), expanding (I)</td>
<td>0.38 delete (A), delete (A)</td>
</tr>
<tr>
<td>0.20 expanding (I), request-authority (I)</td>
<td>0.10 delete-multi twitter (N)</td>
<td>0.17 delete (A), expanding (I)</td>
</tr>
<tr>
<td>0.14 save-annotation (A), expanding (I)</td>
<td>0.10 save-annotation (A), expanding (I)</td>
<td>0.13 save-annotation (A), expanding (I)</td>
</tr>
</tbody>
</table>

4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

lack of space we show a sample (20%). Every line is a subsession, and the colors code the different actions analysts took in the sequence going from left to right. The clustering is based on K-Means and frequency vectors of the actions, meaning that subsessions with similar actions are grouped together. We can see very distinct clusters for deletion and save-annotation sessions in Figure 4.4. We can see annotations without further exploring the data (C3 and C7), informational annotations in clusters where the item is looked at (expanded) and the annotated (C4) and even more explorative annotations, including requests for authority.

Figure 4.5 displays all annotation subsessions, grouped by different media modi. Figure 4.5a and Figure 4.5d display the annotation subsessions for Facebook and Twitter. They prominently feature sessions without informational actions and few deletions that happen within a streak of saving annotations. Figure 4.5b shows the annotation sessions for Youtube data. It features prominently deletions and annotations with a preceding informational action, we call this explorative deletions and explorative annotations. Figure 4.5c shows the sessions for Google results. We can see that the again, deletions are few and often preceded by an expansion of the webpage, while the annotations are preceded by both, requests for authority and expansions.

Figure 4.6 shows a transition graph from one action to another, where the thickness of the arrow indicates the probability that the action that is being pointed to follows the previous one. Transitions with very low probability are left out, unless the transitions would be the only incoming transition for the following action node. Those graphs in Figure 4.6 supports our earlier findings. Figure 4.6a shows that people do not explore in the Facebook annotation mode. We can also see that people rarely do not at all delete after annotations, in fact they either annotate or they delete. We can find a similar pattern in in Twitter subsessions, see Figure 4.6d: here, however, people do occasional change between save-annotation and delete, showing that there are more incidental deletes. Additionally, we can see that the expanding and request-authority are disconnected from the annotational actions. For Youtube (Figure 4.6b) and Google
4.3. Annotation Procedures for Reputation Polarity

Figure 4.5: *(Log dataset)* A visualization of the actions of analysts in the subsessions. The different colors indicate the different actions the analysts performed. We use the same color coding as for Figure 4.4.

(Figure 4.6c) subsessions, this is different. Exploration of the data and annotating the data goes hand in hand as people alternate between annotational and explorational actions. Again, analysts do not seem to save-annotations after delete and vice versa, analysts are expanding before and after deletion.

In general, we can identify specific recurring action patterns surrounding annotations for all annotation modes. For the Google subsessions, it is expanding, request-authority, save annotation. This means that to assess the reputation polarity, dimensions, and audience for websites, the content as well as an authority score (Page-Rank, externally computed) is important. For the Youtube subsession it is expanding, save-annotation, while for Facebook and Twitter it is only save-annotation, without further explorative actions. For Youtube media, the analysts first look at the video before annotating, while for Facebook and Twitter data, the annotations are saved immediately, without further explorative actions: social media texts are short.
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

![Figure 4.6: (Log dataset) Transition graphs for the actions in the different modi. Arrow thickness indicates the probability that one action follows another. Transitions with low probability are left out for clarity, unless no incoming transition would happen. We use the same color coding as for Figure 4.4.](image)

4.3.2 Annotation Difficulty

We want to learn more about reputation polarity. To this end we look at annotation difficulty. As a proxy for this we use the time it takes to perform an annotation. As a caveat, we do not have the actual time for annotation actions. Under the assumption that the annotational subsessions are for annotations only, we can, however, measure the time it takes for annotations to be done within the length of a subsession. This gives rise to two definitions of approximate annotation time, session approximation and transition approximation. **Session approximation** is the average time needed to perform an annotational action: \[ \frac{\text{session length}}{\# \text{annotational actions}} \]. **Transition approximation** is the time passed between two annotational actions and before a certain action.

On the one hand, the transition approximation allows us to identify the time it takes to perform a single action and not just the average. On the other hand, this is very prone to outliers: sessions with breaks from annotations may yield different results. We therefore remove outliers: if the annotation time for a single action is more than two standard deviations away from the mean for that session, we remove this action. Table 4.6 shows the average approximation times it takes analysts to annotate. Again, we filter...
4.3. Annotation Procedures for Reputation Polarity

Table 4.6: Average time in seconds needed to complete an annotation task.

<table>
<thead>
<tr>
<th></th>
<th>Facebook</th>
<th>Youtube</th>
<th>Google</th>
<th>Twitter</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>time session</td>
<td>1658.81</td>
<td>388.90</td>
<td>1221.17</td>
<td>1315.62</td>
<td>1134.08</td>
</tr>
<tr>
<td>±</td>
<td>3406.23</td>
<td>±424.33</td>
<td>±2519.67</td>
<td>±2476.28</td>
<td>±2433.51</td>
</tr>
<tr>
<td># actions in session</td>
<td>17.00 ±20.82</td>
<td>20.41 ±21.70</td>
<td>30.83 ±34.05</td>
<td>12.33 ±14.64</td>
<td>22.02 ±26.79</td>
</tr>
<tr>
<td>avg. session approx.</td>
<td>47.85 ±47.39</td>
<td>49.41 ±42.57</td>
<td>86.44 ±86.31</td>
<td>64.07 ±62.47</td>
<td>64.49 ±63.26</td>
</tr>
<tr>
<td>avg. trans. approx.</td>
<td>37.94 ±29.07</td>
<td>35.71 ±26.08</td>
<td>60.80 ±55.93</td>
<td>63.64 ±51.75</td>
<td>49.13 ±41.51</td>
</tr>
<tr>
<td>avg. trans. approx., delete</td>
<td>15.65 ±10.85</td>
<td>42.51 ±54.99</td>
<td>39.88 ±48.32</td>
<td>23.57 ±43.51</td>
<td>33.95 ±41.85</td>
</tr>
<tr>
<td>avg. trans. approx., save</td>
<td>44.66 ±29.86</td>
<td>36.85 ±26.35</td>
<td>67.96 ±59.95</td>
<td>68.10 ±51.25</td>
<td>58.97 ±54.30</td>
</tr>
</tbody>
</table>

Figure 4.7: (Log dataset) Distributions of annotation times in seconds. Figure 4.7a shows the raw times, while Figure 4.7b has all times outside of two standard deviations removed.

out “extreme” sessions, that are sessions with average values greater than two standard deviations from the mean of all average approximations. We can see that deletion takes significantly (based on a t-test) less time than saving ($p < 0.001$), unless in the Youtube mode, where it takes more time, though not significantly. Also, saving the annotations in the Youtube and Facebook mode takes significantly (based on a t-test) less time than in the Twitter and Google mode ($p < 0.01$). Figure 4.7 shows the distributions of the annotation times, in general, motivating our decision to remove the outliers.

It comes as a surprise that the save-annotation annotations for Twitter takes so long: after all a tweet is only 140 characters long. However, to analyse the details of a tweet in depth, an analyst needs to browse the profiles on Twitter, which may take time. Another surprise is that annotating Youtube videos is rather fast. In fact, for many professional brands there are not that many videos outside of the brand’s own advertisement videos. An informed analyst should know those videos, which leads to short annotation times.
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

Summary

The process actions that lead to annotation decisions depend on the media mode: for Google and Youtube subsessions the actions are informational, in particular looking at the media item in particular. For Facebook and Twitter media not much explorative action has been taken before the annotation. We find that saving the annotation (i.e. finishing the annotation) is more time consuming for Twitter and Google media. For search results this makes sense because articles are long. For (short) tweets, this suggests that the information seeking and exploration happens outside the annotation tool and can not be explained using log data. Below we dive into the process of annotating Twitter media in particular.

4.3.3 Analysis of Questionnaire

We found that we can not entirely explain the procedures of annotating tweets by looking at the log data. While we do have an idea that for example the authority of a website is very important, due to the limitations of the log data, we do not understand what is important for tweets, in particular, what analysts look at right before the annotation. In this section we therefore focus on the procedures for the annotation of tweets. We seek to answer the question

RQ1.c What are the procedures of brand analysts in the annotation of reputation polarity of tweets?

To this end, we analyse the answers to questions Q1–Q4 (see Table 4.1) of the questionnaire in Section 4.2.1.

Figure 4.8 shows the participants answers to questions Q1–Q3, based on the steps in Table 4.2. Question Q1 asks participants about the importance of steps in the annotation process. We can see in the first part of Figure 4.8 (the graph labelled Importance) that the participants agreed about the importance of the steps by considering all steps to be important. In particular everyone agreed that step 8, 9, and 18 (Estimating the topic of the tweet, Determine the type of author of the tweet, Determine an opinion in the tweet, respectively) are the only steps in the process that are considered important or extremely important by every participant. On average, however, step 9 and 18 were given the highest importance. The participants could also manually input further steps that they deemed important. The additional steps are:

1. Count the number of mentions on the topic
2. Evaluate the polarity
3. Find if the tweet has given rise to contents on other networks like Facebook, YouTube, blogs, etc.
4. Assess the number of followers [of] the user who posted the tweet

The second point adds the overall task to the list of steps. The other three steps are interesting because they can be summarized as the additional step

Determine the reach of the tweet.
4.3. Annotation Procedures for Reputation Polarity

Table 4.7: *(Questionnaire dataset)* The answers to Q4: *What kind of background information do you acquire in the annotation process of tweets?*

<table>
<thead>
<tr>
<th>ID</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>You acquire information about the polarity of the tweet and the topic. In addition, it’s useful to know the importance of the tweet and the author.</td>
</tr>
<tr>
<td>A2</td>
<td>Determine authoritativeness of the author. Determine if the content is important or not. Select the most relevant conversations and identify possible risks.</td>
</tr>
<tr>
<td>A3</td>
<td>Previous informations.</td>
</tr>
<tr>
<td>A4</td>
<td>Context and outstanding.</td>
</tr>
<tr>
<td>A5</td>
<td>Lists of important users and links to articles or blogs.</td>
</tr>
<tr>
<td>A6</td>
<td>The type of the author is very important.</td>
</tr>
</tbody>
</table>

Question Q2 asks the participants in how far they think that the steps should be automated.\(^{10}\) Steps that according to our participants should be (at least mostly) done automatically are step 7 and 15, *Finding tweets*, and *Determine the audience*, respectively. The earlier step is more an overall selection of tweets step, while the step determining the audience may be used as an indicator for reputation polarity. Section 4.4 goes into depth as to why this is indeed an important indicator. Finding and filtering tweets is an ongoing topics of research, as reflected by the previous Weps [5] and RepLab [6, 7] challenges. Determining the intended audience of a tweet is not. This is a problem is two ways: a) It is hard to define the reputation without having an understanding of its components; and b) The actual underlying motivation for this research was to embed the indicators into (semi)-automatic approaches for the estimation of reputation polarity. Without previous work this will prove harder. As a contrast, steps that should be done manually are step 14, 0, and 8: *Determine the meaning of a hashtag*, *Understand the content of linked images*, and *Estimating the topic of a tweet*, respectively. Finally, even though analysts do not necessarily deem the automation of all the processes important, finding automatic approaches for all steps are active fields of research [95, 138, 204], in particular estimating the topic for a tweet in the reputation scenario [232].

The final plot in Figure 4.8 shows if the participants use background information for the steps in the process (Q3). For most of the steps in the process, more than a third (4) participants use background information. In particular, we can see that more than half (7) of the participants use background information for step 18 and 14 (*Determine an opinion in the tweet* and *Determine the meaning of a hashtag*). Standing out are the steps 7 and 15: *Finding tweets* and *Determine the audience*. Here, the participants claim to use no background information. This correlates with the desire for automation: for both steps, the participants would like to have automated approaches.

Table 4.7 shows the six answers to the open question about the kind of background information the analysts acquire. We can see that four out of six answers consider the author as important. Three answers state that the importance or authority of the author is important. A5 is particularly interesting because there are *lists* of important authors.

---

\(^{10}\) We did not include the “trivial” steps that involved reading.
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

To summarize, we find that the process of annotation consists of many different steps that are all considered important by the majority of the participants. In particular, the most important actions are to determine the topic, author, and reach of the tweet. We find that of these two, the participants would like to determine the audience of a tweet automatically. We also find that background information is used a lot for all but the actions that the participants would like to see automated. Background information is used particularly to determine opinions and meaning of a hashtag in a tweet.

4.3.4 Summary

To summarise, the analysis of the log data (see Section 4.2.2) shows that the annotation process varies over different media types. Section 4.3.1 shows that information processing and reading is important for media types that contain a lot of information, such as Youtube and Google media. While for tweets there are few information processing actions happening within the annotation tool, our questionnaire (see Section 4.2.1, Q1–Q4) shows that the annotation process is still rich. Section 4.4.1 shows that in particular, the
4.4 Indicators for Reputation Polarity

In order to automatically assess the reputation polarity of tweets and thereby support analysts, we would like to understand the analysts’ information processing and exploring process in more detail. Additionally, the indicators can provide a general idea towards the definition of reputation polarity. In this section we seek to find and understand the indicators pertaining to a single tweet for its reputation polarity. In particular, we ask

RQ2 On a per tweet level, what are the indicators that reputation analysts use to annotate the tweet’s impact on the reputation of a company?

We answer this question based on the questionnaires detailed in Section 4.4.1. In Section 4.4.2 we analyse videos of analysts who annotate tweets according to their reputation polarity. We summarize and connect the findings using the Questionnaire (see Section 4.2.1) and Think aloud datasets (see Section 4.2.3) in Section 4.4.3.

4.4.1 Analysis of Questionnaire

A first approach towards understanding how analysts measure reputation polarity is to simply ask them. In this section we answer:

RQ1.2.a What are the measures at the individual tweet level that analysts use to annotate the reputation of a company as stated by the analysts themselves?

We ask analysts using a questionnaire (see Section 4.2.1) where we use several approaches to understand the indicators. We ask them to write down indicators in free text as well as multiple choice on a preselected list of indicators that proved successful for automated classification. Additionally, previous informal interviews showed that the authority of the author is vital. We then try to understand what this authority is based on, using two approaches: for one we ask the analysts to click on the most authoritative part of a profile, for the other, we ask two open questions. We detail the results below.

We first asked the participating analysts an in an open question to provide indicators for reputation polarity (Q5). Table 4.8 shows the curated list of indicators provided by our participants. To summarize, we see the author of the tweet, its authority, the sentiment of the tweet, and hashtags are indicators mentioned in at least three answers. Surprising indicators are the frequency of posts and the type of followers. The latter indicator is shared by two answers and opens the question in how far a user is defined by its followers or whether this is an indicator of authority.

author, topic, and reach of the tweet can and are better be analysed on the Twitter page of the tweet itself. We also find that within the annotation process background information is very important for steps such as understanding the opinion in related media or the tweet itself, or the general meaning of hashtags. While finding and filtering is already implemented in the tool, the answers to our questionnaire show that the additional functionality of determining the audience would be appreciated by the annotators.
4. Analyzing Annotation Procedures and Indicators for Reputation Polarity

Table 4.8: *(Questionnaire dataset)* The answers to Q5: *Can you give me a list of typical indicators or features for the reputation polarity of a tweet?*

<table>
<thead>
<tr>
<th>ID</th>
<th>Answer</th>
</tr>
</thead>
</table>
| A1 | sentiment of a tweet  
hashtag of a tweet  
importance of the topic to the company  
authority of the author  
sentiment of the replies |
| A2 | author  
vocabulary  
links  
profile information  
type of followers  
number of retweets  
active conversation |
| A3 | hashtags  
bio of the author  
sentiment of the tweet. |
| A4 | author  
relevance in the country, region  
authority  
hashtag  
topic  
relation with the company (stakeholder type)  
sentiment  
date  
number of retweets  
number of mentions of the author  
author profile validity  
type of followers  
How often the author post comments about the company or topic? |
| A5 | It depends on type of company, the sector, the incident |
4.4. Indicators for Reputation Polarity

For Q6 we then ask the analysts to rate the importance of low-level features used in [190] for automatic annotation of reputation polarity, as well as sentiment and reputation polarity of linked webpages. Figure 4.9 shows the answer distribution per feature/measure. Most features, except for if the author enabled location services and the time zone the user posts from and the number of hashtags in a post were deemed important. The most important features are in two groups: (1) the social graph (number of friends and followers), and (2) the reputation and sentiment of the post, linked webpage and its comments, and the retweets of a tweet. It seems that important indicators for reputation polarity are the reputation polarity and sentiment of related data. Interestingly, while we can see that the reputation polarity of comments of a linked webpage is between an important and extremely important indicator, their sentiment is not considered a vital indicator.

This leads to the second approach to answering the research question: we want to understand what makes a Twitter user an authority on a certain topic. We chose two users, the technology magazine @wired and an (anonymous) venture capitalist, @anonymous. We asked the participating analysts to click on the part of the profile that indicates the authority of the user. We then asked the participants if they consider the users authorities. Figure 4.10 shows the original images of the profile next to an image overlaid with the click distribution. We see that the participants consider the number of followers the most important indicator for the authority of the users. Other indicators are the related media and the author’s biography.

The control question if the users were considered an authority was answered positively: all participating analysts agree that @wired is an authority (66.7% strongly agree), while the result for @anonymous was more ambiguous (58% agree, 33% disagree, 8% don’t know). This coincides with the click statistics for Q9: while the clicks for Q7 where primarily on the follower number, the clicks for Q9 are on more diverse positions as well.

For Q11 we asked the participating analysts what an authoritative author means to them, i.e., what is an authoritative user with respect to a company. Table 4.9 shows the answers to the open question, Q11. Let us begin with the indicators that are hard to measure. Here, we observe that offline authority (A1) and being an opinion leader (A4, A6, A8) defines an authority. Follow-up research needs to understand what constitutes offline authority. As to more easily measurable indicators, whether the account is verified (A4, A5), the number of tweets (A2), retweets and replies (A6), and the reach of the opinions (A8) are mentioned as indicators for an authority. This corresponds with the analysis of Q7–9, where the number of followers, related media, and the biography of the author were found to be important indicators.

For Q12 we ask the open question how participants find authoritative users. Table 4.10 shows the free text answers to Q12. The participants search on different tools and combining the findings (A3, A4, A5). Additionally, the social network structure is being used (A1, A2, A4). Finally, as mentioned before, the user needs to be verified (A2, A4) and the participants use background information (A1, A2, A4), that could in form of the offline reputation (A2).

To summarize our findings from this questionnaire, we find that for our participating analysts two groups of indicators are most important for determining reputation polarity: (1) user authority, and (2) sentiment and reputation polarity of the tweet itself and related
### Analyzing Annotation Procedures and Indicators for Reputation Polarity

#### Figure 4.9: (Questionnaire dataset) The answers to Q6: Consider the following indicators for reputation polarity of a tweet. In your annotation process, how important are the indicators for the annotation of reputation polarity?

We can see the indicators on the $y$-axis and graded importance on a 5 point Likert scale on the $x$-axis. The numbers at position $(x, y)$ are the number of analysts considering the indicator $y$ as $x$ important. The darker a cell, the more agreement by analysts, the red dot is the mean.

---

**4.4.2 Analysis of Think Aloud Videos**

One of the limits of the questionnaire is that the indicators were mainly listed without an example at hand (except for Q7 and Q9). In the following we want to understand which indicators analysts *use* to annotate tweets for reputation polarity. The *used* indicators may be different from the ones *stated* without context. This leads to the following question:

**RQ2.b** What are the measures on a tweet level used to annotate the reputation of a...
4.4. Indicators for Reputation Polarity

Figure 4.10: (Questionnaire dataset) The click distributions of Q7 and Q9: Please click on the position of the user profile below (Figure 4.10a or 4.10c) that is most indicative for the users authority.

Figure 4.10a and 4.10c show the original user profile, while Figure 4.10b and 4.10d show the original profile overlaid with a heatmap for the number of clicks. The redder the heat of the overlay, the more analysts clicked on this position.

We observe that in 87.1% of all tweets (27 out of 31) the annotator looks at the text and in 58% (18 out of 31) of the cases at the user. Additionally, whenever the tweet had a company as used by social media analysts?
### Table 4.9: (Questionnaire dataset) The answers to Q11: What is an authoritative Twitter user with respect to a company? Please explain what an authoritative user means to you.

<table>
<thead>
<tr>
<th>ID</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Is a person who has a lot of followers or authority offline because he is an important activist, a political,</td>
</tr>
<tr>
<td>A2</td>
<td>Number of tweets, relation between following and followers</td>
</tr>
<tr>
<td>A3</td>
<td>[Authorities] are users with different attributes, is very important [for] your online reputation and digital identity. The brands and companies have greater visibility</td>
</tr>
<tr>
<td>A4</td>
<td>An important person in your industry, you need to verify that your account is real.*</td>
</tr>
<tr>
<td>A5</td>
<td>It’s just a user who has verified his account. Not necessarily an authority.</td>
</tr>
<tr>
<td>A6</td>
<td>A person who usually talks about topics of the industry and has a significant number of RTs and replies. In addition, this person is asked by other users when a topic is a trend.</td>
</tr>
<tr>
<td>A7</td>
<td>An authoritative Twitter user is important for a company when his tweets influence the company’s reputation</td>
</tr>
<tr>
<td>A8</td>
<td>Is an opinion leader in certain areas, and its opinions can reach a wide audience</td>
</tr>
</tbody>
</table>

---

link the link was followed. The most important *textual indicator* is the **topic** (64.5% of all tweets) of the tweet followed by the occurrence of specific **keywords** (19.4% of all tweets) within the tweet. One particular topic, *webcare*, was pointed out several times for its effect on the reputation polarity. One comment of an analyst about the reputation polarity of a tweet about *webcare*:

> In this country the service of the company is very bad, so we know it is going to be negative.

In other countries, the service can be better, leading to a different prior towards the reputation polarity. The most important **user indicators** are all based on authority: the number of followers (25.6% of all tweets) as well as the type of user (in total: 12, in particular: *is company* = 3, *is newspaper/journalist* = 6, *is known* = 3). The **metadata indicators** are related to the reach of a tweet. For 29% of all tweets the number of retweets was looked at, and also the number of times the tweet was favorited (2 out of 31). As mentioned before, when the tweet had a link, the link was followed. Here, the full text was not always important (37.5% of all tweets with links), but sometimes only the header (37.5% of all tweets with links). Additionally, if the text was very long, the analyst mainly looked at the relevant parts. Interestingly, automatic approaches for the estimation of reputation polarity have so far often ignored the content of the linked webpage [6]. In general, and independent from the root nodes, the most frequent indicators are the topic of the tweet, the number of retweets and followers and certain keywords.

Next, we look at frequent combinations of codes. The most frequent combination is **text and user**: in more than half of all videos the annotators looked at the user and
4.4. Indicators for Reputation Polarity

Table 4.10: (Questionnaire dataset) The answers to Q12. How do you find important and authoritative Twitter users? Please provide a list of steps in the steps in the search process.

<table>
<thead>
<tr>
<th>ID</th>
<th>Answer</th>
</tr>
</thead>
</table>
| A1 | The followers  
The information of the profile, background information and replies or retuits.  
Following/Followers |
| A2 | Number of followers and offline reputation.  
The quality of their website. |
| A3 | Manual search,* google |
| A4 | I found any brand mentions in search.twitter.com,  
then I use followerwonk.com and many others to verify the user.  
Define keywords related with the topic.  
Search on Twitter.  
analyse the top tweets about the topic.  
Search on other tools.  
Determine the differences on the results provided by Twitter and the other tool.  
Select the tweets in common  
analyse the conversation around the tweets: replies and RT.  
An important and authoritative Twitter user depends of topic. |
| A5 | Using the search tools available, and looking for the most difunded. |

* busqueda manual

the text. We find two surprising combinations. First, the combination of topic and is newspaper/journalist: here, in most of the cases when the annotator found the newspaper indicator important, the topic was also an important indicator. Second, whenever the topic was customer support, the annotator looked at the user. This again supports the earlier quotation, where the user and the location are very important indicators in the context of webcare.

In summary, we saw that the most important indicators were the topic of the tweet (in terms of general topic and keywords), the reach (in terms of retweets and followers), and the authority of the user (in terms of followers and author type). Additionally, while ignored in current automatic approaches, the content of the linked webpages is important.

4.4.3 Summary

We used two approaches to answering RQ1.2, using a questionnaire (see Section 4.2.1, Q5–Q12) and an analysis of videos taped while analysts are annotating tweets (see Section 4.2.3). According to the outcomes of the questionnaire in Section 4.4.1, analysts consider user authority as well as sentiment and reputation polarity of the tweet itself and related media to be important indicators for the reputation polarity of a tweet. They state that they measure user authority in terms of number of followers, offline authority, and whether the user is an opinion leader for the topic of the tweet. While observing
Figure 4.11: (Think aloud dataset) Hierarchical codebook used for analyzing the indicators considered by social media analysts before they come to a decision on reputation polarity. The number next to the indicator shows the number of tweets for specific codes where both coders agreed. This codebook contains less codes than the codebook displayed in Figure 4.3 because codes with zero tweets were not included.

analysts during the annotation of tweets, in Section 4.4.2 we find that the most important indicators are the topic of the tweet, the reach, and again, the authority of the user. Combining the results from both datasets, we see the need for estimating the (topical) authority of a user. In particular, we need computational methods to determine the type of a user and the offline reputation, which may be mutually dependent. We see that authority is currently often measured by the number of followers. However, this number is not topic dependent (users talking about two topics may combine followers), nor is it a reliable authority measure alone [11]. This study points out that offline authority is one of the major indicators, the prior background knowledge of an expert is therefore very important for the annotation of reputation polarity. Future work should emphasize on a clear definition of what constitutes offline authority.
4.5 Conclusion

This work studied the information behavior of professional reputation analysts annotating social media. We analysed three datasets to understand the annotation process of analysts and the indicators for reputation polarity of media. We collected the three datasets based on the analysts’ annotation behavior. For one, we used log data of an annotation interface as a non-invasive method of logging. We then used retrospective questionnaires for the analysts to self report their annotation approach. Finally, we filmed experts following the think aloud protocol.

Based on the diverse datasets, we reported several findings. Looking at the annotation process, we found that the importance of annotation actions differs among media: information processing and reading are important actions for information-rich media such as Google and Youtube. For Twitter media, determining the authority, topic, and reach of

<table>
<thead>
<tr>
<th>Tags</th>
<th>counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>16</td>
</tr>
<tr>
<td>text, topic</td>
<td>13</td>
</tr>
<tr>
<td>metadata</td>
<td>10</td>
</tr>
<tr>
<td>metadata, # retweets</td>
<td>9</td>
</tr>
<tr>
<td>text</td>
<td>8</td>
</tr>
<tr>
<td>metadata</td>
<td>8</td>
</tr>
<tr>
<td>metadata, # retweets</td>
<td>7</td>
</tr>
<tr>
<td>metadata</td>
<td>7</td>
</tr>
<tr>
<td>text, topic</td>
<td>7</td>
</tr>
<tr>
<td>text</td>
<td>6</td>
</tr>
<tr>
<td>metadata</td>
<td>5</td>
</tr>
<tr>
<td>text, topic</td>
<td>5</td>
</tr>
<tr>
<td>metadata</td>
<td>5</td>
</tr>
<tr>
<td>metadata, # retweets</td>
<td>5</td>
</tr>
<tr>
<td>text</td>
<td>4</td>
</tr>
<tr>
<td>metadata, # retweets</td>
<td>4</td>
</tr>
<tr>
<td>user</td>
<td>3</td>
</tr>
<tr>
<td>user, # followers</td>
<td>3</td>
</tr>
<tr>
<td>metadata</td>
<td>3</td>
</tr>
<tr>
<td>text, age</td>
<td>3</td>
</tr>
<tr>
<td>metadata, # retweets</td>
<td>3</td>
</tr>
<tr>
<td>text, keywords</td>
<td>3</td>
</tr>
<tr>
<td>text</td>
<td>3</td>
</tr>
<tr>
<td>user</td>
<td>3</td>
</tr>
<tr>
<td>text, topic</td>
<td>3</td>
</tr>
<tr>
<td>text, topic, webcare</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4.11: *(Think aloud dataset)* The most common co-occurring codes.
a tweet is important. Focussing on the indicators for reputation polarity, in particular tweets, using the questionnaire as well as the think aloud dataset, we found that the author of a tweet is important, in particular her authority in offline and online life with respect to the topic of the tweet. We also found that the reach of a tweet is an important indicator for reputation polarity.

This work is relevant for three groups of people: researchers trying to find algorithms to automate the annotation of reputation polarity, annotation interface designers, and business and economics researchers. For (semi-)automatic reputation analysis the findings are interesting because determining topical authority has in fact been studied \[232, 269\], but it has not been taken into account for the classification of reputation polarity. Furthermore, the reach of a tweet has only been used by looking at the number of followers \[190\]. Aral and Walker \[12\] and Aral \[11\] emphasizes that the high number of followers does not necessarily entail being an opinion leader and that content spreads through chains of influential people. And, while there are several approaches to identify opinion leaders (or social influencers) \[34, 237, 238, 272\], none have been used to help with the automatic estimation of reputation polarity. For annotation interface designers, identifying key processes in the annotation procedure may guide new semi-automatic annotation software that relies on the analysts’ vast knowledge of the topic and company but helping with the tediousness of the annotation. Finally, the findings are interesting for researchers in the area of online reputation management, where simple reputation indicators may correlate with financial performance. Additionally, with this work we contributed a new approach to understanding a classification task in general. Earlier approaches solved the reputation polarity classification problem in a purely data-driven way—similar to Hofmann et al. \[107\] for information retrieval, we are looking at how expert annotators, humans, assess media for its reputation polarity.

This study has a number of limitations. For one, due to the relative lack of experts, a large part of the study is qualitative, but we believe that merging three different data sources and answering the research questions from different angles helps reduce this limitation. Furthermore, we are looking at analysts from a single company. We believe, however, based on informal interviews with other companies in the same space, that other companies use similar annotation procedures. Finally, this study focuses on tweets and tweets are not the only online medium used as a proxy for reputation polarity.

Future work can go into two directions. For one, future research can address the limitations of this study using data from more data sources (possibly for quantitative analyses) as well as including more companies. We also propose to analyse how indicators differ between media sources. Secondly, future work can make use of the findings of this study. Future research can focus on topical followers, their activity, as well as opinion diversity. As background information is important and fully automatic approaches are not necessarily going to be entirely accurate, future work can entail automating tedious steps and assisting analysts in their daily work. For fully automatic approaches, the prior knowledge of the expert on offline authority of an author is something that might be captured by models based on multiple data sources, online (e.g., Twitter, Facebook, News) and offline (e.g., newspapers, radio interviews, interviews with related stakeholders).