Opinion Mining for Software Development: A Systematic Literature Review

BIN LIN, Software Institute - Università della Svizzera italiana, Switzerland
NATHAN CASSEE, Eindhoven University of Technology, The Netherlands
ALEXANDER SEREBRENIK, Eindhoven University of Technology, The Netherlands
GABRIELE BAVOTA, Software Institute - Università della Svizzera italiana, Switzerland
NICOLE NOVIELLI, University of Bari, Italy
MICHELE LANZA, Software Institute - Università della Svizzera italiana, Switzerland

Opinion mining, sometimes referred to as sentiment analysis, has gained increasing attention in software engineering (SE) studies. SE researchers have applied opinion mining techniques in various contexts, such as identifying developers’ emotions expressed in code comments and extracting users’ critics toward mobile apps. Given the large amount of relevant studies available, it can take considerable time for researchers and developers to figure out which approaches they can adopt in their own studies and what perils these approaches entail.

We conducted a systematic literature review involving 185 papers. More specifically, we present 1) well-defined categories of opinion mining-related software development activities, 2) available opinion mining approaches, whether they are evaluated when adopted in other studies, and how their performance is compared, 3) available datasets for performance evaluation and tool customization, and 4) concerns or limitations SE researchers might need to take into account when applying/customizing these opinion mining techniques. The results of our study serve as references to choose suitable opinion mining tools for software development activities, and provide critical insights for the further development of opinion mining techniques in the SE domain.

CCS Concepts: • Software and its engineering → Software development process management; Software libraries and repositories; • Information systems → Sentiment analysis.

Additional Key Words and Phrases: opinion mining, sentiment analysis, software engineering

ACM Reference Format:

1 INTRODUCTION

Opinion mining, the term coined by Dave et al. [27] in 2003, was introduced to refer to “processing a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)”. They proposed a tool to classify product review sentences according to the polarity of the sentiment expressed, i.e., whether these sentences have a positive or negative connotation. Tasks that capture sentiment...
polarity are also called “sentiment analysis” in some other studies [47, 56]. Indeed, the terms “opinion mining” and “sentiment analysis” are often used interchangeably [47, 64].

Since its emergence, opinion mining has evolved and is no longer limited to classifying texts into different polarities. For example, Conrad and Schilder [25] analyzed subjectivity (i.e., whether the text is subjective or objective) of online posts when mining opinions from blogs in the legal domain. Hu et al. [33] adopted a text summarization approach, which identifies the most informative sentences, to mine opinions from online hotel reviews. These new perspectives call for a broader definition of opinion mining. According to Liu [46], “opinion mining analyzes people’s opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes”.

In recent years, opinion mining has attracted considerable attention from software engineering researchers. Studies have seen the usage of opinion mining in collecting informative app reviews, to understand how developers can improve their products and revise their release plans [24, 35, 50, 65, 70, 77]. Researchers have also applied opinion mining techniques to monitor developers’ emotions expressed during development activities [22, 30, 41, 54, 63, 71, 79]. Opinion mining has also been used to assess the quality of software products [17, 28].

Given all these studies, it is important to have an overview of existing opinion mining techniques and their applications in software engineering. In this way, researchers can have a base to advance the field, and tool users can better understand how they can apply the existing techniques and what their limitations are.

We provide a systematic literature review on opinion mining for software development activities. Our contributions are: 1) We provide an overview of opinion mining techniques researchers and developers can use for specific tasks; 2) We present datasets developers can use to train or validate techniques; 3) We report on the results of the tool performance validation, which can serve as a guidance for researchers to conduct performance evaluation and sheds light on the threats when using these tools; 4) We identify the common issues software engineering researchers face when applying opinion mining and indicate the potential solutions; 5) We identify directions for future research in the field.

1.1 Scope of Our Study

Opinion mining is evolving, and covers a wide range of topics. Adopting the categories by Pang and Lee [64], we consider under the umbrella of works related to opinion mining in software development activities those dealing with:

- **Sentiment polarity identification**, to classify the opinions expressed in the text into one of the distinguishable sentiment polarities (e.g., positive, neutral, or negative). Examples include identifying whether developers are expressing positive sentiment in daily communications and discovering whether a code review is expressing negative aspects, which can be associated with specific shortcomings of the source code.

- **Subjectivity detection and opinion identification**, to decide whether a given text contains subjective opinions or objective information. An example is distinguishing whether developers/users are discussing a fact about software or presenting their own point of view.

- **Joint topic-sentiment analysis**, which considers topics and opinions simultaneously and search for their interactions. For example, researchers might analyze which aspects are mentioned in user reviews (e.g., performance, usability) and whether these discussions are positive or negative.

- **Viewpoints and perspectives identification**, to detect the general attitudes expressed in the texts (e.g., political orientations) instead of detailed opinions toward a specific issue or narrow subject. An example of perspectives include general preferences of some platforms/technologies over others.
- Other non-factual information identification, to detect all other types of non-factual information, including e.g., emotion detection, humor recognition, text genre classification. Tasks like identifying what requests users are asking for and extracting knowledge embedded in software documents fall into this category.

1.2 Structure of the Paper
Section 2 presents the relevant surveys, literature reviews and mapping studies. Section 3 presents our research questions and our methodology to conduct the systematic literature review. Section 4 reports the results we obtained. Section 5 discusses the replicability of selected primary studies and the impact of how snowballing is conducted. Section 6 discusses the threats that could affect the validity of our results. Section 7 concludes the paper.

2 RELATED WORK
Given the development of opinion mining techniques, many secondary studies have been conducted to present an overview of this field. In the following we discuss relevant systematic literature reviews (SLR), surveys, and mapping studies on opinion mining.

2.1 Secondary Studies of Opinion Mining in General Domains
As one of the earliest secondary study of opinion mining, Liu [48] defined the problem of opinion mining and presented the key tasks and their corresponding techniques in the literature. This study also specifically discussed the issue of spam detection and quality assessment of online reviews. The survey by Ravi and Ravi [68] presented opinion mining tasks and relevant techniques at a more fine-grained level. That is, all the tasks and subtasks were discussed in the following aspects: the addressed problem, used dataset, selected features, techniques and their performance.

Hemmatian and Sohrabi [32] mainly focused on the categorization of opinion mining techniques. In their survey, opinion mining was classified into four levels: document, sentence, aspect, and concept. They also summarized different types of techniques used in two major opinion mining tasks: aspect extraction and opinion classification. Li et al. [42] classified opinion mining techniques for social multimedia into three categories based the source of opinions: textual sentiment analysis (mining opinions from social media messages), visual sentiment analysis (mining opinions from visual content such as images and videos), and multimodal sentiment analysis (mining opinions from both textual and visual content).

Instead of presenting opinion mining tasks and techniques, Kumar and Nandkumar [38] identified several challenges in opinion mining from literature, such as non-expertise opinions, spam opinions, and opinion trust worthiness. They also summarized the advantages and disadvantages of current opinion mining techniques.

Mäntylä et al. [51] analyzed the evolution of opinion mining studies. More specifically, they observed the number of relevant publications, the number of citations, and popular publication venues over the years. They also run topic modeling techniques on the papers to obtain a thematic overview of the research topics. Moreover, they investigated the research topics of the most cited work in this field.

These studies give a good overview of opinion mining tasks and techniques in general domains. However, our previous studies [36, 44, 59] have demonstrated that the performance of sentiment analysis tools trained on the data from other domains (e.g., SentiStrength trained on social media texts) is not satisfactory when they are used in software engineering related tasks (e.g., identifying sentiment polarity embedded in API discussions). Therefore, a literature review dedicated to the software engineering domain is highly desired.
2.2 Secondary Studies of Opinion Mining in Software Development Activities

We identified eight secondary studies related to opinion mining in software development activities.

The SLR conducted by Sánchez-Gordón and Colomo-Palacios [69] focused on works dealing with emotions of software developers. The authors investigated 66 papers covering 40 discrete emotions expressed by developers and found that while the unreliability of sentiment analysis tools is well recognized, not many works in the literature have leveraged other measures such as self-reported emotions and biometric sensors.

Obaidi and Klünder [62] inspected 80 studies related to sentient polarity and emotion analysis. Their study mainly looked into the application scenarios (i.e., open-source projects, industry, academic) and the motivations (e.g., find best tool, value measurement). They also counted how many times different data types and techniques are used, and listed some frequently mentioned difficulties when analyzing sentiment polarity and emotion in software engineering.

Many SLRs have focused on works related to the analysis of app reviews. Martin et al. [52] conducted a survey on papers related to app store analysis, and identified the aspects which have been explored as well as research trends. Noei and Lyons [58] surveyed 21 papers providing guidelines on how to process, analyze, and use user reviews from app stores. Tavakoli et al. [74] investigated the tools developed for analyzing app reviews and presented the types of information these tools can collect and the challenges these tools are facing. Similarly, the SLR by Genc-Nayebi and Abran [29] involved 24 studies and identified techniques for mining online reviews and challenges in the domain. Moreover, they inspected studies concerning the quality assessment of reviews and spam identification.

The remaining two studies fall into the domain of requirements engineering. Meth et al. [53] analyzed 36 publications regarding automated requirements elicitation, and classified them based on tool category, degree of automation, knowledge reuse, evaluation approach, and evaluation concepts. Wang et al. [78] provided a systematic mapping study which focuses on leveraging crowdsourced user feedback in requirements engineering. The feedback can be either explicit (e.g., directly given in crowd-generated comments) or implicit (e.g., mined from application logs or usage-generated data). The primary studies surveyed in these two studies often mine opinions and emotions toward software products from user comments and overlap with those related to app review analysis.

While all these studies cover various topics of opinion mining in software development, they have a focus on very specific areas, such as emotions [62, 69], sentiment polarity [62], app review analysis [29, 52, 74], and requirements engineering [53, 78]. Our study aims at giving a complete picture of the usage of opinion mining techniques in software development, focusing on the research questions presented in Section 3.1. While other secondary studies mainly aim to give an overview of current development of the techniques and their applications, we have a different goal, i.e., to help researchers and developers better adopt/customize opinion mining tools in their own work. Therefore, in this literature review, we have specifically looked into the datasets available, the performance comparison of the available tools, and the issues specific to tool adoption and customization.

3 RESEARCH METHOD

Following the guidelines by Kitchenham and Charters [37] to perform our systematic literature review, we present our research questions, search strategy, study selection process, as well as the methodology for data extraction and analysis.
3.1 Research Questions

To help software engineering researchers better conduct opinion mining related studies and assist practitioners in adopting suitable opinion mining approaches in their projects, this literature review aims to understand the following high-level research question (RQ):

RQ: How can opinion mining techniques support software development activities?

To answer this RQ, it is essential to understand what has been accomplished so far with opinion mining techniques in software development activities. Moreover, knowing the limitations of state-of-the-art approaches is needed to improve the existing techniques or propose new approaches. Therefore, to answer this RQ in a more structured manner, we decompose it into the following RQs:

- **RQ1**: In which software engineering activities has opinion mining been applied? We aim to understand the application domains of opinion mining techniques in software engineering, to present an overview on how these techniques are used, thus revealing the potential of opinion mining in software-related tasks.

- **RQ2**: What publicly available opinion mining tools have been adopted/developed to support these activities? We present the opinion mining techniques proposed in the literature categorized by their functionalities, to obtain an overview about which tools can be used for which specific tasks.

- **RQ3**: How often do researchers evaluate the reliability of opinion mining tools when they adopt the tools out-of-the-box? As researchers have already disclosed [36, 44, 59], opinion mining techniques might not achieve satisfactory results when applied in a different context than the one they have been designed for. Thus, it is important to assess the reliability of these tools when used out-of-the-box. We investigate how often opinion mining techniques are evaluated before being applied without any customization in software-related studies.

- **RQ4**: Which opinion mining techniques have been compared in terms of performance and in what contexts? Since opinion mining tools perform differently in different contexts, we summarize the studies in the literature aimed at comparing the performance of different opinion mining tools in specific contexts. This will quickly point researchers and practitioners to studies aimed at identifying the most appropriate tools to use in a given context.

- **RQ5**: Which datasets are available for performance evaluation of opinion mining techniques in software-related contexts and how are they curated? Given that the context might significantly impact the performance of opinion mining tools, we aim to present the available datasets which can be used to either train supervised techniques or validate the tool performance by serving as oracle. To ensure the reliability of the datasets, we only consider the datasets whose correctness has been manually validated by the authors. We exclude datasets which only contain data scraped from online resources without any further sanity check.

- **RQ6**: What are the concerns raised or the limitations encountered by researchers when using opinion mining techniques? Our goal is to summarize the issues encountered during the application of opinion mining techniques in software engineering tasks. We also discuss the potential directions for addressing these issues.

3.2 Relevant Study Identification

The process of identifying relevant studies to be included in our literature review can be seen as Fig. 1.

3.2.1 Search Strategy. We used the following digital libraries to search for primary studies: ACM Digital Library [1], IEEE Xplore Digital Library [4], Springer Link Online Library [11], Wiley Online Library [14], Elsevier ScienceDirect [3], and Scopus [9]. We did not include Google Scholar due to several shortcomings identified by Halevi et al. [31], namely the lack of quality control and clear indexing guidelines, as well as the missing support for data downloads.

Manuscript submitted to ACM
The following search query was used to locate primary studies in these online databases:

("opinion mining" OR "sentiment analysis" OR "emotion") AND ("software") AND ("developer" OR "development")

This query has been defined through a trial-and-error procedure performed by the first author and a discussion among all authors. While Landman et al. [40] pointed out that adding an “OR” operator to the query may reduce the number of results in some databases such as IEEE Xplore, we tested such a feature by comparing the results of queries using the "OR" with the aggregated results of several queries each one using one of the search terms in OR. We did not spot any difference, showing that the "OR" operator is working correctly. We conjecture that the difference with the observations of Landman et al. [40] can be attributed to an update of the search engines after their study.

The goal of the query is to retrieve all relevant studies (i.e., high recall) while keeping reasonable the effort needed to remove false positives in the subsequent manual analysis. The search terms "opinion mining" and "sentiment analysis" have been included since they are often used as synonyms [45]. Emotion analysis is also attracting attention in studies dealing with human aspects of software engineering [61] and, thus, the term “emotion” was included as well. While opinion mining also includes other aspects such as humor detection [16], these topics are not commonly studied in software engineering. Therefore, we do not include the corresponding terms such as “humor” in our query. Concerning the second part of the query, using the term “software engineering” to identify relevant studies resulted to be a too strict searching criterion, while only using “software” resulted in the introduction of too much noise. The (“developer” OR “development”) search condition allowed to reach a fair balance between the number of papers we need to manually inspect and the coverage of relevant studies. While we are aware that some studies might not explicitly include these two terms, this issue has then been mitigated through a snowballing process explained later.

On ACM Digital Library and IEEE Xplore, we conduct the search within its default search box, while in the rest of the databases, due to the large number of irrelevant results returned, we enforced more restrictions when searching. We set the search field of Elsevier ScienceDirect and Scopus to “title, abstract, keywords”, meanwhile the “Subject Area” of Scopus was limited to “Computer Science” to exclude studies out of our interest. We only searched “abstract” of Wiley as multiple-field search is not supported. Also in this case “Computer Science” was used to constrain the subject. For SpringerLink Online Library, we set the Discipline” and “Subdiscipline” to “Computer Science” and “Software Engineering”, obtaining 1,967 papers. Since SpringerLink does not allow search on specific fields, we crawled meta-information of these 1,967 papers and filtered them by using our search query in “title, abstract, keywords”. We acknowledge that enforcing stricter constraints on some databases might lead to the exclusion of relevant studies. However, the backward and forward snowballing performed later on described significantly mitigates this threat.
3.2.2 Study Selection. Based on the search strategy, we identified relevant studies following a process involving study filtering and snowballing, as indicated in Fig. 1. N1 indicates the batch of papers coming from the search query at the end of each step, while N2 shows the number of papers resulting from the snowballing procedure. The lists of papers after each step of study selection can be found in our replication package [43]. Table 1 summarizes the search results. After removing duplicates in the documents returned found in the different databases, we obtained a total of 795 papers.

Table 1. Documents returned by searching databases

<table>
<thead>
<tr>
<th>Source</th>
<th>Returned Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACM Digital Library</td>
<td>340</td>
</tr>
<tr>
<td>IEEE Xplore Digital Library</td>
<td>243</td>
</tr>
<tr>
<td>Springer Link Online Library</td>
<td>19</td>
</tr>
<tr>
<td>Wiley Online Library</td>
<td>29</td>
</tr>
<tr>
<td>Elsevier ScienceDirect</td>
<td>46</td>
</tr>
<tr>
<td>Scopus</td>
<td>580</td>
</tr>
<tr>
<td>Total (excluding duplicates)</td>
<td>795</td>
</tr>
</tbody>
</table>

Study Filtering. The 795 papers went through a two-step filtering process. In the first round, we manually inspected the title and the abstract, and removed unrelated documents. A web app was developed to support this process (source code available in our replication package [43]). The web app assigned a batch of papers to filter to each author, who indicated whether it should be (i) "included in the study", (ii) "discarded", or (iii) "used as a secondary study". The last option was used to indicate that the paper was not a primary study, but rather a literature review, survey or an article introducing the topic. The selected secondary studies have been used in the snowballing process to identify additional primary studies. Each paper was assigned to two of the authors, to reduce the chance that a paper was discarded by mistake. We observed disagreement on 82 papers, which were discussed by all of the authors until a consensus was reached. Then, in the second-round filtering, we downloaded the papers selected as primary studies and each paper was manually inspected by one author to examine if they met our inclusion and exclusion criteria (Table 2).

At the end of first-round filtering, we obtained 127 papers to include, 662 papers to discard, and 6 papers classified as secondary studies. After the second-round filtering on the 127 papers to include, 71 papers remained as primary studies.

Snowballing. Since keyword-based search might result in omitting relevant studies, we also performed a snowballing-based search on the 127 papers selected as primary studies and on the 6 papers tagged as secondary studies. While 56 out of 127 studies were excluded in the second-round filtering, we still included them in the snowballing process as they might contain references to papers we are interested in.

We performed both backward and forward snowballing. Backward snowballing was performed during the second-round filtering, each paper was analyzed by one author and the papers in the references which might be relevant are recorded based on their titles. For forward snowballing, we collected all the papers citing these 133 (127+6) papers from Google Scholar. In the end, we obtained 1,056 new papers after duplication removal. All these papers were fed into our paper filtering process. After the first-round filtering, we marked 268 papers as selected primary studies; and after the second-round filtering, 114 papers were left. Due to the limited human resources, we only applied snowballing once instead of iteratively. Therefore, we discarded those papers labeled as "secondary study" identified during our snowballing process, and no further snowballing was performed on them. In total, 185 studies are included in our study.
Table 2. Inclusion and exclusion criteria.

**Inclusion Criteria**

IC1 The paper must be peer-reviewed and published at conferences, workshops, or journals; to only include papers which have undergone scrutiny by the scientific community.

IC2 The paper must be accessible online (i.e., PDF files available in the selected databases or through Google Search results); to ensure the accessibility of the studies.

IC3 The paper must be included in one of our databases; to prevent including papers from predatory publishing venues. This criterion only applies to the papers collected from the snowballing process described later.

IC4 The study presented in the paper must be related to software development activities (e.g., requirements, design, implementation, testing, documentation, maintenance, team management, etc.); to enforce our research scope listed in Section 1.1.

IC5 The study must adopt at least one opinion mining technique which automatically extracts opinions from texts; to enforce as main research subject “opinion mining”.

**Exclusion Criteria**

EC1 The paper is not in English. Rationale: English is the primary language for published academic studies.

EC2 The technique presented only works for a language other than English. Rationale: we aim to ensure the techniques in the studies are comparable.

EC3 The paper is a duplicate or a conference paper extended into a journal article. Rationale: we aim to prevent redundancy.

EC4 The paper is not a full research publication (e.g., abstract only submissions, doctoral symposium articles, presentations, tutorials, posters, forewords, etc.). Rationale: these artifacts are not subjected to the same peer-reviewing process as full research papers.

EC5 The paper does not describe what approach was used to extract opinions/information. Rationale: studies lacking such information are often of low quality and do not provide useful information for answering our RQs.

3.3 Data Extraction and Analysis

To answer the RQ1-RQ5 defined in Section 3.1 and facilitate the data extraction process, we used the data extraction form in Table 3 to collect necessary information from the selected studies. This step was conducted together with the “filtering based on full text”. A Web app was developed to support this activity (source code available in our replication package [43]) and each paper was manually reviewed by one of the authors.

Table 3. Data extraction form.

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>What is the main goal of the whole study?</td>
<td>RQ1</td>
</tr>
<tr>
<td>2</td>
<td>Does the paper propose a new opinion mining approach?</td>
<td>RQ2</td>
</tr>
<tr>
<td>3</td>
<td>Which opinion mining techniques are used (list all of them, clearly stating their name/reference)?</td>
<td>RQ2</td>
</tr>
<tr>
<td>4</td>
<td>Which opinion mining approaches in the paper are publicly available? Write down their name and links. If no approach is publicly available, leave it blank or None.</td>
<td>RQ2</td>
</tr>
<tr>
<td>5</td>
<td>What the researchers want to achieve by applying the technique(s) (e.g., calculate the sentiment polarity of app reviews)?</td>
<td>RQ2</td>
</tr>
<tr>
<td>6</td>
<td>Is the application context (dataset or application domain) different from that for which the technique was originally designed?</td>
<td>RQ3</td>
</tr>
<tr>
<td>7</td>
<td>Is the performance (precision, recall, run-time, etc.) of the technique verified? If yes, how did they verify it and what are the results?</td>
<td>RQ1, RQ4</td>
</tr>
<tr>
<td>8</td>
<td>What success metrics are used?</td>
<td>RQ4</td>
</tr>
<tr>
<td>9</td>
<td>Does the paper replicate the results of previous work? If yes, leave a summary of the findings (confirm/partially confirms/contradicts).</td>
<td>RQ4</td>
</tr>
<tr>
<td>10</td>
<td>Which dataset(s) the technique is applied on?</td>
<td>RQ3</td>
</tr>
<tr>
<td>11</td>
<td>Is/Are the dataset(s) publicly available online? If yes, please indicate their name and links.</td>
<td>RQ3</td>
</tr>
<tr>
<td>12</td>
<td>Write down any other comments/notes here.</td>
<td>-</td>
</tr>
</tbody>
</table>
Given that all extracted information is in free text, we conducted a manual coding process for our data analysis after the data extraction process. This step is important for two reasons: 1) the coding of our extracted information can produce indexes for easing our effort in locating relevant studies, especially considering the large amount of studies we have; 2) the different terminologies used by the authors can be unified, which is essential for answering our RQs.

With the data resulting from the data extraction process, we first identified whether to include the paper by inspecting the answer to No. 12 in Table 3 as we asked the inspectors to take notes here if the paper does not pass the full-text filtering. Then, we identified: 1) the purpose of study (e.g., detecting developers' emotion/sentiment/politeness expressed in software artifacts), 2) whether the approach has been customized, 3) the tools used, 4) whether the approach is available, 5) the type of opinion mining technique (e.g., sentiment polarity analysis), 6) whether the tool is applied in a context different from its origin, 7) & 8) whether the performance of the approach has been verified, 10) the type of dataset (e.g., GitHub issue comments), and 11) whether the dataset is available. As we found that very rarely a study was replicated, therefore we did not collect useful information from No. 9. Not all the information is available for all papers. We used the processes defined in ISO/IEC/IEEE 12207:2017 International Standard [15] for the application domain. More specifically, to identify the relevant process, we compared the purpose of the study to the outcomes, activities, and tasks of each process defined in the ISO/IEC/IEEE 12207:2017 Standard document and then selected the process which matches the best. Additionally, we added the option “team management” to the application domain along with the existing processes in the standard as it is one of the most popular topics in opinion mining software engineering studies, which focuses on developers instead of specific development processes.

Papers excluded by second-round filtering were also included in the coding process, this is to confirm the decision of exclusion based on ICs and ECs as each paper was inspected by one author. In total, 395 papers were included in our coding. At first, we selected first 23 papers in our database for the trial coding (20 out 23 are determined to be included in our study), which was performed by the first two authors. The rest of the authors participated in the discussion until agreement was reached. Then, we equally distributed other 156 papers to all of the authors, namely on average each author was assigned to 26 different papers for reviewing. As there are several open-ended questions to answer during information retrieval (e.g., what the researchers want to achieve by applying the technique(s)?), to reduce the duplicate codes written in different ways, we discussed the codes emerged from the output of this round and unified the phrases expressing same meanings. Finally, we equally distributed the remaining 216 papers to all the authors and finished our coding process. The first author double checked all the coded information and performed the final data organization. While our coded data already provides extensive useful information, we checked the original papers for more detailed information if needed when answering RQ1-RQ5.

To answer RQ6, we inspect the papers proposing or evaluating opinion mining techniques, as we are more interested in the concerns/limitations supported by evidence instead of those based on assumptions. Each paper was manually inspected independently by two of the authors, who extracted insights when the following criteria were satisfied:

- They should be explicitly indicated in the results, discussions, or conclusions.
- They should be relevant to customizing/adopting opinion mining approaches in software engineering.
- They should be supported by data (namely, those proposed without evidence should be discarded).
- They should not describe tool-specific optimizations such as parameter tuning.

We merged the concerns/limitations extracted by the authors and discarded duplicated ones. That is, we removed the same insights from the same paper, but those similar/identical insights were kept if they were extracted from different papers. The extracted insights were then manually categorized based on topic similarity.
4 RESULTS

4.1 RQ1: In which software engineering activities has opinion mining been applied?

In Table 4, we categorize and summarize the papers that apply opinion mining in software engineering activities.

Table 4. Software engineering activities in which opinion mining is applied.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Relevant Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Design Definition Process</strong></td>
<td></td>
</tr>
<tr>
<td>Assessing techniques/services for system implementation</td>
<td>[P99], [P166], [P66], [P62], [P14]</td>
</tr>
<tr>
<td><strong>Knowledge Management Process</strong></td>
<td></td>
</tr>
<tr>
<td>Identifying developers’ assumptions/rationale from communication</td>
<td>[P95], [P9]</td>
</tr>
<tr>
<td>Mining usage knowledge regarding techniques</td>
<td>[P151], [P182], [P3], [P169], [P167], [P43], [P94]</td>
</tr>
<tr>
<td><strong>Quality Assurance Process</strong></td>
<td></td>
</tr>
<tr>
<td>Evaluating software quality from crowd source</td>
<td>[P145], [P61], [P177], [P112], [P60], [P37], [P10], [P176], [P35], [P174]</td>
</tr>
<tr>
<td>Evaluating general user satisfaction</td>
<td>[P152], [P119], [P48], [P29], [P185], [P168], [P51], [P107], [P75], [P92], [P180], [P88], [P56], [P105], [P121], [P98], [P103], [P144], [P53], [P124], [P58], [P11], [P143], [P102], [P181], [P183], [P63], [P85]</td>
</tr>
<tr>
<td>Evaluating user satisfaction toward specific product aspects</td>
<td>[P160], [P134], [P118], [P83], [P38]</td>
</tr>
<tr>
<td>Identifying issues/requests from developer discussions/issue reports</td>
<td>[P114], [P136], [P153], [P108], [P135], [P55], [P25], [P33], [P87], [P140], [P162], [P46], [P56], [P171], [P45], [P150], [P65], [P76], [P90], [P117], [P115], [P77], [P13], [P22], [P116], [P59], [P6], [P86], [P125], [P165], [P42], [P184]</td>
</tr>
<tr>
<td>Identifying issues/requests/other information from user feedback</td>
<td>[P114], [P136], [P153], [P108], [P135], [P55], [P25], [P33], [P87], [P140], [P162], [P46], [P56], [P171], [P45], [P150], [P65], [P76], [P90], [P117], [P115], [P77], [P13], [P22], [P116], [P59], [P6], [P86], [P125], [P165], [P42], [P184]</td>
</tr>
<tr>
<td><strong>Stakeholder Needs and Requirements Definition Process</strong></td>
<td></td>
</tr>
<tr>
<td>Identifying and evolving requirements from other products</td>
<td>[P104], [P101], [P79], [P28]</td>
</tr>
<tr>
<td>Identifying and evolving requirements from user feedback</td>
<td>[P8], [P80], [P21], [P57], [P139], [P106], [P30], [P122], [P78], [P178], [P18], [P7], [P159], [P93]</td>
</tr>
<tr>
<td>Identifying requirements from requirement artifacts</td>
<td>[P1], [P89], [P2]</td>
</tr>
<tr>
<td>Acquiring deeper understanding of requirements</td>
<td>[P142], [P155]</td>
</tr>
<tr>
<td><strong>Team Management</strong></td>
<td></td>
</tr>
<tr>
<td>Relating emotion/sentiment/politeness to performance/behavior</td>
<td>[P24], [P161], [P110], [P156], [P147], [P179], [P17], [P129], [P148], [P47], [P97], [P96], [P32], [P163], [P64]</td>
</tr>
<tr>
<td>Detecting emotion/sentiment/politeness expressed in software artifacts</td>
<td>[P54], [P26], [P52], [P34], [P141], [P157], [P44], [P172], [P131], [P5], [P39], [P132], [P173], [P12], [P82], [P137], [P49], [P138], [P70], [P91], [P120], [P130], [P158], [P36], [P123], [P69], [P113], [P164], [P31], [P41], [P175], [P40]</td>
</tr>
<tr>
<td>Evaluating the trust among team members</td>
<td>[P149], [P27]</td>
</tr>
</tbody>
</table>

4.1.1 Design Definition Process. These activities aim to provide detailed information about the elements which can be used to enable the implementation.

**Assessing techniques/services for system implementation.** Studies have been conducted to mine opinions from online resources to evaluate the strengths and weaknesses of techniques/services. Uddin & Khomh [P166] and Lin et al. [P99] mined Stack Overflow discussions to extract opinions regarding the pros and cons of adopting certain APIs based on different aspects (e.g., usability, compatibility). Not limited to only APIs, Huang et al. [P62] also leveraged Stack Overflow discussions to compare different techniques (e.g., Ant vs Maven). The aspects they used were automatically...
generated by topic modeling techniques, thus being less structured. Ikram et al. [P66] mined tweets to assist in open source software adoption by analyzing developers’ sentiment regarding various aspects. A similar approach has been applied by Ben-Abdallah et al. [P14] to online reviews to help developers select proper cloud service.

4.1.2 Knowledge Management Process. These activities aim to provide opportunities to reuse the existing knowledge about development process, skills and system elements.

Identifying developers’ assumptions/rationale from communication. Li et al. [P95] analyzed discussions from mailing lists to identify assumptions (e.g., a developer guessing what requirements users might have).

This knowledge can be used to infer the rationales behind certain design choices. With similar goals, Alkadhi et al. [P9] identified issues, potential solutions and relevant arguments from development chat messages.

Mining usage knowledge regarding techniques. Several studies have focused on retrieving knowledge about the usage of APIs from online discussions. For example, by analyzing Stack Overflow posts, Uddin et al. [P167] documented how APIs are used and Wang et al. [P169] extracted tips for using APIs. Being wary of potential bad programming practice, Serva et al. [P151] identified those examples associated with discussions having negative sentiment. Other studies have investigated the negative aspects of APIs. For instance, Zhang and Hou [P182] identified discussions on API features from forums which contain negative sentiment. Meanwhile, Ahasanuzzaman et al. [P3] and Li et al. [P94] identified sentences on Stack Overflow mentioning API issues and negative caveats, respectively. From a coarse-grained level, Fucci et al. [P43] classified Stack Overflow posts into 12 types of knowledge, such as functionality, quality, and example.

4.1.3 Quality Assurance Process. These activities aim to identify the issues which might harm software quality and ensure quality requirements are fulfilled. It is worth noting that sometimes the identified issues during these activities can be further processed to refine the requirements, which is highly relevant to the activities in the category “Stakeholder Needs and Requirements Definition Process”. However, in the studies below, such concrete requirement extraction is not conducted.

Evaluating software quality from crowd source. Rahman et al. [P145] extracted opinions about quality or issues of the code from Stack Overflow posts to recommend insightful comments for source code. Hu et al. [P61] analyzed user comments of the same apps from different platforms to evaluate whether the hybrid development tools, which use a single codebase across platforms, manage to deliver products with similar user-perceived quality.

Evaluating general user satisfaction. Studies have been conducted to understand users’ sentiment toward software products by mining their feedback from app reviews [P60, P112], tweets [P177] or free text reviews from other sources [P10, P37]. Durelli et al. [P35] took a further step to investigate whether automated tests in mobile apps impact the overall user satisfaction. Some researchers have investigated the sentiment in support tickets [P16, P174, P176] to reduce ticket escalations and ensure customer satisfaction. These studies do not look into customer feedback from a more fine-grained perspective (e.g., quality aspects, features).

Evaluating user satisfaction toward specific product aspects. Many studies [P48, P51, P53, P58, P63, P75, P88, P102, P105, P107, P121, P124, P143, P152, P168, P180, P181, P183] have classified mobile app reviews into different categories based on the features (e.g., tracking calories), topics (e.g., app theme) or quality aspects (e.g., usability), and then analyzed the sentiment users expressed in these reviews to understand whether the customers are satisfied with the products. A similar technique was also applied to tweets [P11, P29, P50, P98, P103, P121], Google research results [P185], SourceForge user reviews [P144], and online technical review articles [P92]. Keertipati et al. [P85] converted sentiment toward product features into priorities of mobile app feature development, instead of directly presenting it.
Identifying issues/requests from developer discussions/issue reports. Developer discussions in emails [P160] and issue reports [P38, P83, P134] have been analyzed to identify bugs and feature requests. Munaiah et al. [P118] inspected code reviews to identify possibly missed vulnerabilities.

Identifying issues/requests/other information from user feedback. Researchers have proposed classifiers to cluster mobile app reviews into different categories (e.g., feature request, problem discovery, information seeking, user experiences) [P6, P22, P33, P55, P56, P65, P76, P77, P87, P90, P108, P116, P119, P135, P136, P140, P150, P153]. While in different studies the proposed categories can be slightly different, the classified feedback can be further analyzed to identify potential issues, improvement, and new features. A similar approach was applied to tweets [P162], user forums [P86, P117] and OSS mailing-lists [P117]. Some studies have specifically focused on identifying types of issues in app reviews [P13, P25, P42, P45, P46, P114, P165, P171, P184], while others categorize those reviews into different categories concerning quality (e.g., privacy, usability) or topics without explicitly pointing out the issues [P59, P115, P125].

4.1.4 Stakeholder Needs and Requirements Definition Process. These activities aim to help define or refine requirements.

Identifying and evolving requirements from other products. Liu et al. [P104] and Jiang et al. [P79] mined app descriptions to extract requirements related information and recommend new features, while Liu et al. [P101] supported a similar task but only focused on permission-related requirements. Instead of app descriptions, Dalpiaz and Parente [P28] analyzed app reviews of competitors to suggest new features.

Identifying and evolving requirements from user feedback. Mobile app reviews are an important source for requirements elicitation. Several studies have mined app reviews to extract either functional or non-functional requirements [P21, P30, P78, P106, P139, P159]. Similar techniques were also applied to reviews in the format of tweets [P7, P8, P57, P80, P122, P178], Facebook posts [P7], peer-to-peer online review site [P18], and feature requests on SourceForge [P93].

These activities differ from those to “identify issues/requests/other information from user feedback”, as the latter do not aim at eliciting requirements, but rather at assessing the quality of the currently implemented ones.

Identifying requirements from requirement artifacts. Kurtanovic and Maalej [P89] trained a classifier to categorize requirements into functional and non-functional (usability, security, operational, performance). Abad et al. [P2] proposed an approach to extract text from requirements and identify the non-functional requirements related to usability, operability, and performance. They also implemented a prototype ELICA as a mobile app and conducted a case study to illustrate how it might work in real-life scenarios [P1].

Acquiring deeper understanding of requirements. Shi et al. [P155] created an approach to classify sentences in feature requests into six different categories (i.e., intent, explanation, benefit, drawback, example, and trivia). Portugal and do Prado Leite [P142] used sentiment analysis to extract interdependencies among non-functional requirements, focusing on the relationship between the usability-related requirements as well as the requirements of other quality attributes.

4.1.5 Team Management. These activities aims to understand developers’ behavior and performance.

Relating emotion/sentiment/politeness to performance/behavior. The relationship between developers’ feelings and their performance or behavior has been widely studied, including the impacts of developers’ sentiment, emotions, and attitudes on bug/issue fixing efficiency [P32, P64, P129, P179], build success of continuous integration [P161], issue reopening [P24], routine change [P147], activeness of participation [P47, P96], likelihood of introducing bugs [P163], leadership [P17], and productivity [P97, P110]. Reversely, the impact of refactoring activities [P156] and user feedback [P148] on developers’ sentiment were also studied.
Detecting emotion/sentiment/politeness expressed in software artifacts. Several researchers have looked into the feelings of developers expressed in various software artifacts. For example, the sentiment polarity detection (i.e., identifying whether a developer is expressing positive or negative feelings) was applied to code review comments [P12, P137, P138], emails [P5, P39–P41, P49, P54, P91, P137, P164], issue reports [P31, P34, P54, P82, P130, P137], commit messages [P52, P69, P70, P157], commit or pull request comments [P141, P158], requirements documents [P175], and project reports [P113]. Emotions, such as anger, joy, and fear, were detected in issue reports [P31, P44, P120, P131, P132] and GitHub comments [P123, P172, P173]. Specifically, Elbert et al. [P36] detected confusion in code reviews. The politeness of developers was also evaluated in issue reports [P31, P130, P131].

Evaluating the trust among team members. Sapkota et al. [P149] and Maldonado da Cruz et al. [P27] proposed new approaches for estimating trust between developers, leveraging developer interactions and sentiment embedded in pull request or commit comments.

4.2 RQ2: What publicly available opinion mining tools have been adopted/developed to support these activities?

To answer this RQ, we list all the publicly available opinion mining tools we found in the subject software engineering studies. We classify these tools into two major categories: 1) tools for sentiment polarity/emotion/politeness/trust analysis, and 2) tools for artifact content analysis. It is worth noting that while some of these tools are not specifically designed for processing software-related tasks, they are widely used by software engineering researchers. We consider a tool as designed for SE data if it was proposed and evaluated on artifacts generated during software development (e.g., developers’ discussions, documentation) by the original authors.

4.2.1 Sentiment polarity/Emotion/Politeness/Trust Analysis. Tools in this category (Table 5) are mainly used to analyze the feelings expressed by developers. More specifically, sentiment polarity detection tools predict whether a text contains positive, neutral or negative sentiment. As these tools have been comprehensively compared and evaluated, we kindly invite the readers to refer to Section 4.4 for more information regarding their strengths and weaknesses. Emotion detection tools can extract developers’ emotions from the texts, with different tools being able to detect different types of emotions. For example, DEVA [P73] and MarValous [P68] can detect four emotional states (i.e., excitement, stress, depression, and relaxation), while TensiStrength\(^1\) [75] is used to estimate the strength of stress and relaxation. EmoTxT [P20], instead, can detect whether a text contains the following emotions: joy, love, surprise, anger, sadness, and fear. In addition to what EmoTxT [P20] is capable to detect, NTUA-SLP [18] can also detect if optimism, or pessimism is expressed in the texts, as well as other emotions, i.e. disgust, anticipation, and trust. While these tools in general have good performance, several limitations have been reported. For example, EmoTxT has a relatively low precision and recall for identifying surprise, while NTUA-SLP demonstrated mixed results when predicting the intensity of the emotions. Other issues include the difficulty of handling negations, irony, and sarcasm (DEVA), processing texts with mixed emotions (TensiStrength), and training on a balanced dataset (MarValous). LIWC [6] calculates the percentage of words falling into 90 different dimensions and summarize them into four different perspectives: analytical thinking, clout, authenticity, and emotional tone. However, while LIWC is easy to use and provides a broader range of social and psychological insights, the fact of being a commercial software hinders the adaption and further extention of the tool. Currently, there are not many tools available for measuring the politeness of the text (politeness tool [26]). Unlike other

\(^1\)TensiStrength can be used with its online demo. Standalone tools is also available for free upon request for academic purposes.

Manuscript submitted to ACM
tools which take as input texts, Trust-Framework [27] takes a GitHub repository and calculates the trust score among developers. However, the estimated trust scores have not been verified in real team projects.

4.2.2 **Artifact Content Analysis.** Tools in this category (Table 6) are mainly used to identify the topics or categories of texts from software artifacts. The topics/categories can be either automatically generated (LDA [20] and TwitterLDA [83]), or pre-defined (all the rest). Those tools without pre-defined categories are borrowed from other domains, while the rest are specifically designed for software engineering tasks.

LDA [20] and TwitterLDA [83] are based on Bayesian model. Both of these two tools take a collection of texts as input and output the potential topics of the texts. However, a drawback would be the necessity of knowing the dimension of topics in advance. ARdoc [P136], SURF [P159], MARC 3.0 [P76–P78], and RE-SWOT [P28] are the tools for user review analysis. More specifically, given the user reviews, these tools can be used to classify the reviews into different categories such as feature request and problem discovery (ARdoc, SURF, MARC 3.0), associate reviews to different topics (SURF), identify different types of non-functional requirements such as performance and usability (MARC 3.0), and
Table 6. Publicly available opinion mining tools for artifact content analysis.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Functionality</th>
<th>Based on</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA [20]</td>
<td>automatically extracts topics from texts</td>
<td>English documents</td>
</tr>
<tr>
<td>TwitterLDA [83]</td>
<td>automatically extracts topics from texts</td>
<td>social media texts</td>
</tr>
<tr>
<td>ARdoc [P136]</td>
<td>identifies app reviews related to information giving, information seeking, feature request, and problem discovery</td>
<td>mobile app reviews</td>
</tr>
<tr>
<td>Ticket-Tagger [P83]</td>
<td>classifies issue reports into bug, enhancement, and question</td>
<td>issue reports</td>
</tr>
<tr>
<td>SURF [P159]</td>
<td>identifies app reviews related to information giving, information seeking, feature request, and problem discovery; summarize app reviews based on topics</td>
<td>mobile app reviews</td>
</tr>
<tr>
<td>MARC 3.0 [P76–P78]</td>
<td>identifies app reviews related to bug reports and feature requests; identify topics for functional requests; classify non-functional requests into dependability, reliability, performance, and supportability</td>
<td>mobile app reviews</td>
</tr>
<tr>
<td>RE-SWOT [P28]</td>
<td>analyzes app reviews to suggest new features</td>
<td>mobile app reviews</td>
</tr>
<tr>
<td>DeepTip [P169]</td>
<td>extracts API usage tips</td>
<td>Stack Overflow posts</td>
</tr>
<tr>
<td>POME [P99]</td>
<td>classifies sentences referring to APIs into seven quality aspects (e.g., performance, usability) and determine their sentiment polarity</td>
<td>Stack Overflow posts</td>
</tr>
</tbody>
</table>

extract features classified in strengths, weaknesses, threats, and opportunities (RE-SWOT). While these tools achieve good performance in classifying app reviews based on their categories, several limitations exist. For example, the topic categorization can be coarse-grained (ARdoc). Meanwhile, MARC 3.0 uses only textual information, ignoring other potentially useful meta information such as star ratings and submission time of review. Tools like RE-SWOT do not consider the trend over time, hence the users might not know if some issues have already been addressed. Ticket-Tagger [P83] takes GitHub issues and label them into different categories including bug report, enhancement, and question. However, the recall for enhancement is relatively lower than that of other classes, and there are relatively higher number of false positives for detecting questions. DeepTip [P169] and POME [P99] both analyzed Stack Overflow posts. The former extracted tips on API usage while the latter categories API-related sentences into different quality attributes (e.g., performance, compatibility) and sentiment polarities. While both tools achieve high precision, POME reported a relatively low recall for identifying quality attributes.

4.2.3 Extra Information Related to the Opinion Mining Tools. The results presented in this section provide researchers and developers a reference to the tool they might be able to use in their work. However, we acknowledge that readers might want to have a better understanding of these tools. Therefore, we collected the following information from the original papers proposing these tools: (1) the link to the paper; (2) the link to the tool; (3) the input of the tool; (4) the output of the tool; (5) the core technique used in the tool; (6) the advantages of the tool; and (7) the limitations of the tool. This information can be found in the “supplementary results” page of our online replication package [43]. These supplementary results do not include tools for sentiment polarity analysis, as these tools have the same input and output, and they are extensively compared in the literature (as shown in Section 4.4).

4.3 RQ3: How often do researchers evaluate the reliability of opinion mining tools when they adopt the tools out-of-the-box?

As using opinion mining tools from other domains without performance validation might yield unreliable conclusions [P81], we are interested to see whether software engineering researchers consider addressing this concern when adopting opinion mining tools developed by others and use them to analyze their data. Table 7 lists the tools adopted by other researchers, how often they are used in a domain different from the one they have been designed for, and how
Table 7. Number of tools being adopted, used in different domains, and how often their performance is verified.

<table>
<thead>
<tr>
<th>Tool</th>
<th># Adopted</th>
<th># Used Differently (# Verified / # Unverified)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SentiStrength [76]</td>
<td>15</td>
<td>15 (2/13)</td>
</tr>
<tr>
<td>politeness tool [26]</td>
<td>5</td>
<td>5 (0/5)</td>
</tr>
<tr>
<td>LDA [20]</td>
<td>3</td>
<td>3 (0/3)</td>
</tr>
<tr>
<td>LIMC [34]</td>
<td>3</td>
<td>3 (0/3)</td>
</tr>
<tr>
<td>Senti4SD [P19]</td>
<td>3</td>
<td>3 (1/2)</td>
</tr>
<tr>
<td>Stanford CoreNLP [73]</td>
<td>2</td>
<td>2 (0/2)</td>
</tr>
<tr>
<td>SentiStrength-SE [P74]</td>
<td>2</td>
<td>1 (0/1)</td>
</tr>
<tr>
<td>Watson Natural Language Understanding [13]</td>
<td>1</td>
<td>1 (0/1)</td>
</tr>
<tr>
<td>Rosette [8]</td>
<td>1</td>
<td>1 (0/1)</td>
</tr>
<tr>
<td>TwitterLDA [83]</td>
<td>1</td>
<td>1 (0/1)</td>
</tr>
<tr>
<td>SentiSE [10]</td>
<td>1</td>
<td>0 (0/0)</td>
</tr>
<tr>
<td>Pattern [72]</td>
<td>1</td>
<td>1 (0/1)</td>
</tr>
<tr>
<td>Aylien [2]</td>
<td>1</td>
<td>1 (0/1)</td>
</tr>
<tr>
<td>Syuzhet [5]</td>
<td>1</td>
<td>1 (0/1)</td>
</tr>
<tr>
<td>EmoTstT [P20]</td>
<td>1</td>
<td>0 (0/0)</td>
</tr>
</tbody>
</table>

often the performance is verified when it is used in a different domain. Here, a “different domain” refers to the fact that the type of data in the study is different from that used to customize the tool. For example, Stack Overflow posts and mobile app reviews are considered as a different type of data, despite the fact that they both belong to the “software engineering” domain. In this table, we do not count the cases when the tools are only used to compare the performance with other tools and not chosen as the final tool to support software development activities defined in Section 4.1. To obtain the raw data for this RQ (i.e., the list of papers involved in this RQ and their corresponding performance verification information), please visit the “supplementary results” page of our online replication package [43].

As it can be seen from Table 7, in most of the cases these tools are used in a domain different than the one they have been designed for. What is concerning is that very few researchers try to validate whether these tools can actually produce reliable results in the context of their study. SentiStrength [76] is the most popular opinion mining tool in our subject papers, and of the 15 studies using it in a different context only in 2 cases its performance has been assessed before using it. This is even more problematic since general-purpose sentiment analysis tools such as SentiStrength have been shown to be unreliable in the software engineering context [P81].

4.4 RQ4: Which opinion mining techniques have been compared in terms of performance and in what contexts?

Table 8 and Table 9 present the performance comparisons of sentiment polarity analysis tools and emotion detection tools, respectively. It is worth noting that while EmoTstT [P20] is a tool for emotion detection, the comparison in [P72] was made by mapping emotion states to sentiment polarity (e.g., joy is considered positive). The studies in the table are categorized based on the data type used in the performance evaluation. The tool with the best performance is underlined and the metric used is also indicated. As artifact content analysis tools often deal with different tasks, their performance cannot be directly compared in most of the cases, therefore, no such comparisons can be found for those publicly available tools.

From Table 8 we can see that overall, the tools customized on software related data usually perform better than tools created for general domains. While the performance of different sentiment polarity analysis tools is widely compared...
### Table 8. Performance comparison of sentiment polarity detection tools. Underlined tools has the best performance based on the adopted metric, and tools in bold face are proposed in the literature.

<table>
<thead>
<tr>
<th>Compared Tools</th>
<th>Adopted Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>issue reports</strong></td>
<td></td>
</tr>
<tr>
<td>SentiStrength, SentiStrength-SE, SentiCR, Senti4SD, SentiMoji [P23]</td>
<td>overall accuracy</td>
</tr>
<tr>
<td>SentiStrength, SentiStrength-SE, SentiCR, Senti4SD [P128]</td>
<td>micro-average F1</td>
</tr>
<tr>
<td>SentiStrength, Alchemy (Watson NLU), NLTK, Stanford CoreNLP [P81]</td>
<td>weighted kappa</td>
</tr>
<tr>
<td>SentiStrength, NLTK, Watson NLU, Microsoft Text Analytics API [P84]</td>
<td>weighted kappa</td>
</tr>
<tr>
<td>SentiStrength, SentiStrength-SE, SentiSW [P34]</td>
<td>overall accuracy</td>
</tr>
<tr>
<td>SentiStrength, SentiStrength-SE, NLTK, Stanford CoreNLP [P74]</td>
<td>overall F1</td>
</tr>
<tr>
<td>SentiStrength, SentiStrength-SE, NLTK, Stanford CoreNLP [P100, P109, P146]</td>
<td>overall accuracy</td>
</tr>
<tr>
<td>SentiStrength-SE, Senti4SD, EmoTxT [P72]</td>
<td>overall accuracy</td>
</tr>
</tbody>
</table>

| **Stack Overflow posts** | |
| SentiStrength, SentiStrength-SE, SentiCR, Senti4SD, SentiMoji [P23] | overall accuracy |
| SentiStrength, SentiStrength-SE, SentiCR, Senti4SD [P128] | micro-average F1 |
| SentiStrength, SentiStrength-SE, SentiCR, Senti4SD [P166] | micro-average F1 |
| SentiStrength, SentiStrength-SE, Senti4SD [P19] | overall F1 |
| SentiStrength, SentiStrength-SE, NLTK, Stanford CoreNLP [P100, P109, P146] | overall accuracy |
| SentiStrength-SE, Senti4SD, EmoTxT [P72] | overall accuracy |

| **code reviews** | |
| SentiStrength, SentiStrength-SE, SentiCR, Senti4SD, SentiMoji [P23] | overall accuracy |
| SentiStrength, SentiStrength-SE, SentiCR, Senti4SD [P128] | micro-average F1 |
| SentiStrength, SentiStrength-SE, SentiCR, Senti4SD [P12] | micro-average F1 |
| SentiStrength-SE, Senti4SD, EmoTxT [P72] | overall F1 |

| **GitHub comments** | |
| SentiStrength, SentiCR, Senti4SD, Alchemy (Watson NLU), NLTK, Stanford CoreNLP [P67] | weighted kappa |

| **mobile app reviews** | |
| SentiStrength, SentiStrength-SE, NLTK, Stanford CoreNLP [P100, P109, P146] | overall accuracy |

### Table 9. Performance comparison of emotion detection tools. Underlined tools has the best performance based on the adopted metric, and tools in bold face are proposed in the literature.

<table>
<thead>
<tr>
<th>Compared Tools</th>
<th>Adopted Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>issue reports</strong></td>
<td></td>
</tr>
<tr>
<td>TensiStrength, DEVA [P73]</td>
<td>average F1</td>
</tr>
</tbody>
</table>

| **issue reports + Stack Overflow posts (mixed)** | |
| DEVA, MarValous [P68] | average F1 |

on issue reports, Stack Overflow posts, and code reviews, more attention should be given to GitHub comments and mobile app reviews. The performance on these two types of data has not been verified for the latest development of sentiment polarity analysis tool (e.g., SentiMoji).
While an overall performance comparison result is given in the table, in practice tool users might have specific focus and preference. For example, when the amount of data is huge, precision might be more important than recall to avoid noise. Another scenario is when analyzing users’ complaints from app reviews, a tool which can better identify negative sentiment is preferred. Therefore, to allow readers to check specific metrics, we have aggregated the comparison results for different metrics, which can be found on the “supplementary results” page of our replication package [43].

4.5 RQ5: Which datasets are available for performance evaluation of opinion mining techniques in software-related contexts? How are they curated?

We present the datasets that can be used by researchers to evaluate or customize opinion mining techniques for software engineering tasks. More specifically, we list in which paper the dataset was presented, which type of dataset were used, the number of data points in the dataset, the categories used in the dataset, and the number of data points falling into each category. We separate these datasets based on what purpose they can be used for: 1) datasets for sentiment polarity/emotion/politeness detection (Table 10), and 2) datasets for sentiment artifact content analysis (Table 11). If several datasets are presented in one paper, their dataset ID would be formatted as “DSN-X”, where X denotes the index of the dataset in the paper (e.g., DS9-1 refers to the first dataset in paper #9). It is worth noting that we do not include datasets whose download link is no longer valid or whose access needs to be requested to the authors. The original paper of DS1 presents three groups of data, all with manually annotated emotions. However, in the first two groups of the data, only raw annotations were given (i.e., emotions assigned by different annotators) and the conflicts were not resolved. Therefore, here we only include the data in group 3 in which the conflicts were addressed by the authors.

Most of the datasets included have been manually labeled by at least two evaluators, however, how conflicts were resolved varies. DS1, DS14, DS18-1, DS18-2, DS20, and DS21 were labeled by three evaluators, a label was assigned only when at least two of them agreed, thus no extra process was needed to resolve the disagreements. Similarly, DS5, DS6, DS13, DS15, and DS17 were also labeled by three people, but when conflicts emerged after labeling, a majority voting criterion was applied. It is worth noting for DS5, if opposite labels were provided, the corresponding data point was discarded. DS2, DS4, DS7, DS8, DS11, and DS16 were labeled by 4, 3, 3, 2, 2, and 2 evaluators, respectively. Discussion sessions were held afterwards to determine the final labels for those data points with conflicted labels. DS2 also discarded those data points on which no agreement could be reached. For DS9-1, DS9-2, and D23, each data point was labeled by two people. When there was a disagreement, the final label was decided by a third person. Similarly, DS19 were labeled by two Ph.D. students, and the conflicts were resolved by two of the authors other than the two students. Four evaluators labeled each data point of DS10, and the dataset used the label “contradictory” to annotate the conflicts. For DS12, the first 100 data points were labeled by two people, as the agreement was reached, the remaining data points were labeled by only one person. We are not able to identify how conflicts were resolved for DS3 and DS22, while we know that these datasets were labeled by two evaluators. For D22, the authors mentioned that a guideline (publicly available online) was given to minimize disagreements.

4.6 RQ6: What are the concerns raised or the limitations encountered by researchers when using/customizing opinion mining techniques?

In this RQ, we discuss the concerns and the limitations of using and customizing opinion mining tools for software engineering tasks. Meanwhile, we discuss the potential directions to address these issues.
Table 10. Datasets available for sentiment polarity/emotion/politeness detection.

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Presented by</th>
<th>Data Type</th>
<th>Data Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS1</td>
<td>Ortu et al., 2016 [P133]</td>
<td>JIRA issue comments</td>
<td>4,000 sentences</td>
</tr>
<tr>
<td>DS2</td>
<td>Ebert et al., 2017 [P56]</td>
<td>Gerrit code reviews</td>
<td>792 comments</td>
</tr>
<tr>
<td>DS3</td>
<td>Williams &amp; Mahmoud, 2017 [P177]</td>
<td>tweets</td>
<td>1000 tweets</td>
</tr>
<tr>
<td>DS4</td>
<td>Ahmed et al., 2017 [P4]</td>
<td>Gerrit code reviews</td>
<td>1,600 comments</td>
</tr>
<tr>
<td>DS5</td>
<td>Calefato et al., 2018 [P19]</td>
<td>Stack Overflow posts</td>
<td>4,423 posts</td>
</tr>
<tr>
<td>DS6</td>
<td>Novielli et al., 2018 [P127]</td>
<td>Stack Overflow posts</td>
<td>4,800 posts</td>
</tr>
<tr>
<td>DS7</td>
<td>Islam &amp; Zibran, 2018 [P73]</td>
<td>JIRA issue comments</td>
<td>1,795 comments</td>
</tr>
<tr>
<td>DS8</td>
<td>Ding et al., 2018 [P34]</td>
<td>GitHub comments</td>
<td>3,000 comments</td>
</tr>
<tr>
<td>DS9-1</td>
<td>Lin et al., 2018 [P100]</td>
<td>mobile app reviews</td>
<td>341 sentences</td>
</tr>
<tr>
<td>DS9-2</td>
<td>Lin et al., 2018 [P100]</td>
<td>Stack Overflow posts</td>
<td>1,500 sentences</td>
</tr>
<tr>
<td>DS10</td>
<td>Kaur et al., 2018 [P84]</td>
<td>JIRA issue comments</td>
<td>500 comments</td>
</tr>
<tr>
<td>DS11</td>
<td>Imtiaz et al., 2018 [67]</td>
<td>GitHub comments</td>
<td>589 comments</td>
</tr>
<tr>
<td>DS12</td>
<td>Sapkota et al., 2020 [P149]</td>
<td>GitHub comments</td>
<td>616 comments</td>
</tr>
</tbody>
</table>

4.6.1 Using/Customizing tools for sentiment polarity/emotion/politeness/trust analysis. We identify the following concerns/limitations and potential solutions:

**Tool performance is often unsatisfactory.** Researchers have found that when applying sentiment polarity and emotion analysis tools on software-related data, the accuracy of their output is often unsatisfactory when the domain of application is not the one the tools have been designed for [P67, P81, P84, P100, P148, P182]. What is more concerning is that these tools even do not agree each other, meaning the results or conclusions might change by applying different tools on the same data [P81]. Similar issues also hold for emotion analysis tools [P170] and politeness detection tools [P67] developed in other domains. Therefore, we recommend that when adopting opinion mining tools designed for non-software engineering contexts, researchers carefully evaluate the reliability and suitability of these tools, as suggested by [P100, P128].

One common reason for sentiment polarity misclassification is the domain-specific vocabulary [P16, P50, P126]. For example, the occurrence of the word “issue” from issue trackers might mislead the general-domain sentiment analysis tools and the predictions tend to be more negative than it should be. A possible solution is to tune the dictionary to include more domain-specific vocabularies [P16] or train on software engineering data [P4, P177]. Currently, there is a lexicon for emotional arousal in software engineering [P111], which can be considered when customizing emotion detection tools for software-related tasks. SentiStrength-SE also provides a list of domain-specific terms containing...
no sentiments in software engineering context [P74], which has been proven more effective than general-domain dictionaries when identifying sentiment polarity in software-related contexts [P71]. Another challenge is the detection of irony and sarcasm [P50, P73, P74, P166]. Islam and Zibran [P74] pointed out that a potential solution is “combining the dictionary-based lexical method with machine learning”, as done in other domains. The existence of decreasing comparative terms (e.g., little problems) often poses challenges for natural language processing based techniques [P75].

**Tool performance varies on different data.** Even within the software engineering domain, different datasets can still result in unstable performance of sentiment polarity analysis tools [P23, P100, P166]. Similarly, the agreement of the predictions produced by different tools also vary on different datasets [P72]. Moreover, even when the data are extracted from the same domain, classifiers might still achieve different performance for different types of text. For example, when detecting confusion in code comments, the comment types (i.e., inline and general comments) can impact the precision and recall [P36]. Therefore, especially in the case of supervised techniques, it is recommended to leverage datasets from the same data source on which it will be applied when training an approach [P128]. Another fact worth noting is that when extracting emotions, the results are more reliable on sentences expressing “joy” (compared to on sentences expressing “anger”) and on team-level aggregated texts (compared to individual texts) [P170].

**Retraining a tool with software-related data requires substantial effort.** As opinion mining tools often do not perform well in software engineering contexts, researchers sometimes retrain existing approaches with software-related data. However, manually building a training set for supervised approaches (e.g., those based on deep learning) can
be exhausting and it does not guarantee that the retrained approach will get better performance [P100]. However, researchers have found that training the model with a mixture of software-related and unrelated data can be a solution: pre-training the approach with data from other domains (social media [P23], Google News [P15], Wikipedia English [P146]) can significantly improve the performance.

Neutral sentiment is difficult to identify. Researchers have found that often neutral texts are mistakenly classified as positive or negative, while the opposite occurs much more rarely [P100]. Shen et al. [P154] confirmed that both machine learning approaches and lexicon & rule-based approaches have difficulties in correctly identifying neutral texts. Therefore, when evaluating the performance of a sentiment polarity analysis tool, the dataset containing only positive and negative sentiments are insufficient, since the real challenge comes when neutral items are part of the dataset [P100]. Applying some balancing techniques (e.g., oversampling and undersampling) might to a certain extent improve the low performance caused by dominant neutral texts [P15, P109].

Neutral sentiment is difficult to identify. Researchers have found that often neutral texts are mistakenly classified as positive or negative, while the opposite occurs much more rarely [P100]. Shen et al. [P154] confirmed that both machine learning approaches and lexicon & rule-based approaches have difficulties in correctly identifying neutral texts. Therefore, when evaluating the performance of a sentiment polarity analysis tool, the dataset containing only positive and negative sentiments are insufficient, since the real challenge comes when neutral items are part of the dataset [P100]. Applying some balancing techniques (e.g., oversampling and undersampling) might to a certain extent improve the low performance caused by dominant neutral texts [P15, P109].

Human created gold set for tool customization/evaluation may be unreliable. When creating datasets for tool customization or evaluation, one issue is that sometimes there are no clear guidelines, thus the gold set might contain some noise (e.g., “bug report” is mistakenly labeled as negative) [P67, P128]. Another issue is subjectivity during data labeling. Studies have found that when it comes to GitHub comments, people have low agreement regarding sentiment and politeness [P67]. Moreover, it is easier for evaluators to agree on emotions like love and sadness than others [P120]. Therefore, clear guidelines are needed for the labeling process and it is necessary to distinguish the objective report of facts (e.g., there is a bug) from the affective state expressed in the text [P128].

User ratings are not always in line with the sentiment expressed. In app review analysis, user ratings are not reliable as a proxy for the user sentiment. While the reviews with one or two stars are negative in most of the cases, reviews with high ratings may also contain issues [P114]. The sentiment expressed in the reviews can be more accurately captured by sentiment analysis tools than star ratings [P107].

4.6.2 Using/Customizing tools for artifact content analysis. We identify the following concerns/limitations and potential solutions: Single data source may not be enough for mining user feedback. Researchers have found that tweets provide more objective opinions related to apps compared to reviews on app stores [P121]. Besides, software companies often use social media to collect bug reports and feature requests. Thus, tweets, especially those from company support accounts can be a useful source for mining opinions about software products [P136]. Meanwhile, most of the reviews do not contain valuable or actionable information for researchers to improve their apps [P98]. Therefore, it is suggested to look into different data sources to gather more comprehensive feedback.

The artifact content can belong to multiple categories. During data labeling, researchers have found that a small portion (around 1.1%) of user reviews are related to more than one type of requirement [P106]. Thus, when researchers need to customize an approach, multi-class classification might be necessary. As a workaround, splitting the text into multiple parts (e.g., sentences) has also been adopted [P106].

Data for training is often unbalanced. When training a classifier for identifying various types of user requests, classifiers usually perform badly on the minority types [P87, P93, P106]. Using both project-specific keywords (e.g., those...
mined from project description and unlabeled user requests) and non-project-specific keywords (e.g., those derived from requirements ontologies and taxonomies) as features for training classifiers can improve the performance to a certain extent [P93].

The quality of datasets affects the performance of the automatic approach for classifying user reviews. If the size of the training set is small, traditional machine learning approaches outperform deep learning [P162]. Meanwhile, when only the data with highly confident labeling (i.e., two evaluators agree on the same class) are used, the performance of machine learning approaches also improve in review classification [P87]. This indicates the importance of the balance between quantity and quality of the training set. Another important factor is the annotation guide, when category definitions are misunderstood or apparently have similar meanings, misclassification is more likely to happen [P56].

Same words can be used to identify different topics/attributes. When identifying quality attributes mentioned in app reviews, same keywords might correspond to different attributes (e.g., the “fast” in “fast loading” refers to performance, while the “fast” in “the app is easy and I can do things fast with it” is more related to usability) [P75]. A potential solution could be “analyzing keywords more than one term” [P25, P46, P75]. For example, using both bi-grams and tri-grams as features to train classifiers might help correctly classify “fast loading” as performance and classify “do things fast” as “usability”. However, this does not guarantee same phrase will not convey several different meanings [P46].

The various choices of vocabulary negatively impact the performance of user review classification. Different users might use different keywords and linguistic patterns to explain the same issue, which can lead to review misclassification [P25]. One potential way to address this issue is to include more instances of reviews in the training set [P25]. At the same time, errors of spelling and grammatical structure and non-standard sentences can also affect the performance [P80], which can be addressed by adding spell checker during preprocessing [P56]. Meanwhile, there are vocabulary mismatches between different populations (e.g., the technical vocabulary used by developers vs. informal lexicon in the reviews [P25, P105]).

The information provided by users can become invalid due to software evolution. Researchers have noticed that some reviews become outdated as they describe already removed features or technologies used by the apps, and a potential way to solve this issue is to correlate reviews with the app change logs [P116].

Data provided by the source can be incomplete. App reviews provided by Google Play Store are incomplete, and researchers have found that using incomplete reviews might bias the findings. It is recommended to collect user reviews continuously for a long time period [P125].

Sentences discussing the interested subjects can be hard to locate. When mining opinions for APIs, the precision drops when the API mention is more than one sentence away from the related opinions, or several APIs are mentioned together [P166]. For the former, it is recommended also considering four surrounding neighboring sentences as well [P182].

5 DISCUSSION

In this section, we discuss the replicability issue of these studies we spotted during the analysis of 185 papers. Additionally, we point out the potential directions for future work.
5.1 Replicability of Selected Studies

During our study, we spotted a few issues which might hinder the replicability of opinion mining-related software engineering studies. First of all, if we take a look at techniques in Section 4.2, we can easily find that there are much more tools available for sentiment polarity and emotion detection than artifact content analysis, while the latter is also widely used in software engineering activities (Section 4.1). Indeed, lots of proposed approaches for artifact content analysis are not open-source, which also leads to the fact that researchers are often unable to compare their approach with relevant ones (Section 4.4). Besides, when we extracted available tools and datasets, we found many links in the papers to be invalid, in particular when those artifacts were hosted on personal homepages. Thus, it is recommended to store the artifacts on reliable third-party services such as Zenodo\(^2\), Figshare\(^3\), and GitHub\(^4\). Moreover, the artifacts provided in the paper often lack proper documentation, which makes it hard to comprehend the resources.

5.2 Impact of One-Round Snowballing

As snowballing is a very expensive activity, iterative snowballing is rarely performed. However, only conducting a single snowballing round is a threat to the completeness of the relevant primary studies we identified. Therefore, to have a basic idea how many papers we might miss in our study, we randomly sampled 10% of the selected papers (i.e., \([114 \times 0.1] = 12\)) obtained from the first snowballing round and conducted a second-round backward and forward snowballing. After removing the papers already collected in our previous selection process, we got 387 studies (denoted as Set 1). Also, we randomly took 10% of the secondary studies abandoned during our paper selection after first-round snowballing (i.e., \([11 \times 0.1] = 2\)) and inspected whether the cited papers in these studies can be a potential primary study in our literature review. This leads to 24 new studies (denoted as Set 2). We followed the same process as described in Section 3.2.2 to filter these 411 papers based on title and abstract. As a result, we found that 26 studies (25 from Set 1 and 1 from Set 2) might fit into our study scope. The list of papers before and after filtering can be found in the “supplementary data” page of our replication package [43]. When inspecting these 26 papers, we found that 12 are obtained by snowballing a single paper. This fact indicates that if we miss a study addressing a specific issue when conducting keyword-based searching, even if we can include it in the first-round snowballing, we might still miss many relevant studies. By inspecting the venues of these 26 papers, we found that 11 were not published in software-specific conferences or journals, which made them less likely to be relevant for software engineering researchers.

This result suggests that iterative snowballing plays an important role in the completeness of selected primary studies. However, many databases currently do not provide a convenient way for automatically collecting the papers during snowballing. We acknowledge this common issue in literature reviews, and we would recommend that the search engines could provide an easy way for researchers to download the citing and cited papers. Meanwhile, our result is only based on a small set of samples, we are not sure if performing snowballing on the rest of the studies will lead to similar amount of new papers, especially when we had the rather extreme case that one paper alone introduced 12 new relevant studies. We would also recommend that future researchers working on literature reviews could conduct similar sampling to provide more quantitative insights on the impact of multi-round snowballing. While our study might not include all relevant studies, the research questions we investigated are not highly dependent on the completeness of the samples. Instead, we believe that given the large number of papers included and the in-depth analysis of these studies, our literature review can still provide valuable information regarding opinion mining in software development.

\(^2\)https://zenodo.org
\(^3\)https://figshare.com
\(^4\)http://github.com

Manuscript submitted to ACM
6 THREATS TO VALIDITY

Wohlin et al. [80] list the potential threats researchers might face during software engineering research.

**Threats to construct validity** concern the relation between theory and observation. We only select papers indexed in our chosen databases. There might be relevant studies in other databases, however, we have included most popular ones. Besides, including search engines like Google Scholar might introduce a large amount of noise including not peer-reviewed work and low-quality papers. Another threat is that the search string might not cover all the studies which fit in our search scope. This is mitigated by our backward and forward snowballing process. We only conducted one-round snowballing, which might still miss some relevant papers. Nevertheless, snowballing requires huge amount of human effort, and conducting a second round can be impractical. We believe that most relevant studies were included based on the expertise that the authors have in this domain. Moreover, the large number of papers included in this study can already bring rich information to readers and answer the research questions with sufficient details. Another threat is that we did not apply extra quality assessment criteria on the primary studies we selected. While quality assessment criteria is sometimes used in literature review studies, many criteria are rather subjective. As we only selected peer-reviewed papers, many papers with major design flaws should have been filtered out. However, we acknowledge that some peer-reviewed studies might still contain significant flaws. Our in-depth analysis of the primary studies through the lens of various research questions can mitigate this issue.

**Threats to internal validity** concern external factors we did not consider that could affect the variables and the relations being investigated. The databases we used are constantly indexing more papers, and they function like black boxes, meaning we are not able to tell whether their search algorithm would change at some point. However, as we take all the results returned and conducted snowballing, we believe that most relevant papers are included in our study. Another issue is that papers are dynamically indexed in these databases. We might not be able to replicate the search results even if the same search strategy is employed. For example, some papers might be indexed in the database much later than their real publication date. Therefore, it is possible to find more papers in the future even if the publication date range remains unchanged. These factors threat the replicability of our study.

**Threats to external validity** concern the generalizability of our findings. We only focused on opinion mining techniques designed for artifacts written in English. While English is used as a “lingua franca” in global software development [49], we acknowledge that developers might create software artifacts (e.g., user interfaces, user manuals) in a language other than English. In fact, researchers have found that industry projects are more likely to contain comments and identifiers in more than one language compared to open source software projects [67]. Additionally, developers might communicate in other languages as well. As coping with multi-lingual texts remains one of the key challenges in natural language processing, it would be interesting to investigate the relevant studies in the future. Besides, all the selected studies are directly associated with software development processes or developers. This choice was taken as our goal was assisting researchers and developers in adopting/customizing relevant approaches in software development activities. Our paper search was performed until early 2020. We acknowledge that additional opinion mining tools and datasets have been released [19, 23, 39, 82] and more performance comparisons have been conducted [21, 23, 60, 81].

**Threats to conclusion validity** concern the relations between the conclusions and our analyzed data. In our study, each paper was inspected by one author, and the corresponding coding was verified by the first author without further examination due to the large amount of studies in our work. While this did not guarantee the correctness of our coding, we did take extra caution when writing the paper and re-recheck all studies for which something was unclear.

Manuscript submitted to ACM
7 CONCLUSIONS

7.1 Summary

In this study, we conducted a systematic literature review involving 185 papers related to opinion mining for software engineering. We first presented fine-grained categories of software development activities in which opinion mining is applied and described what these activities are. We then summarized publicly available opinion mining tools in the subject papers and explained in which context these tools are created. We later investigated whether the performance of these tools are evaluated when adopted in other studies, and we found that very few researchers evaluate tool performance when these tools are used in a domain different from the one they have been designed for. We also presented the contexts in which these tools are compared, so that researchers and developers can refer to corresponding studies to figure out which tool might work the best for their own data. We next presented 23 publicly available software-related datasets which can be used to evaluate and customize new opinion mining approaches in the software engineering domain. In the end, we highlighted the concerns and limitations researchers and developers face when adopting and customizing opinion mining tools in software engineering and indicated potential solutions.

7.2 Insights for Tool Adoption Practices

Our study is by far the largest literature review regarding opinion mining in software development activities, and the results of RQ6 highlight some good practices for using opinion mining tools in this context:

- Use the tool trained and/or evaluated on the same data type of the task.
- When using tools trained on other domain, careful verification of tool performance is necessary.
- Do not expect 100% accuracy of the opinion mining tools, especially when texts contain irony and sarcasm.
- When modeling users’ attitude, consider using emotions instead of sentiment polarities. However, be aware that some emotions such as joy are easier to capture than others.
- When collecting users’ feedback, aggregate the information from various sources (e.g., twitter, mobile app stores).
- When analyzing users’ reviews, give more weight to the sentiment expressed in the reviews than user ratings, and also pay attention to the validity of the reviews (whether the information is outdated).

7.3 Directions for Future Work

Given the issues we identified for using existing opinion mining tools for software engineering tasks, we list potential directions for future work, with an aim of advancing this domain.

Opinion mining for other software development activities. While opinion mining has been applied to many software-related tasks, there are still some areas which opinion mining has not set foot into. An example is the application in the human resource management process. Human resource managers and project leaders can mine discussions in open source project artifacts to understand developers’ desired tasks and their capabilities, and this information can be taken into account for recruitment, promotion, and task assignment. Moreover, opinions embedded in user feedback can be leveraged for some more specific tasks, such as identifying the need of ending certain system elements (corresponding to the disposal process defined in ISO/IEC/IEEE 12207:2017 International Standard [15]), as well as selecting the optimal software architecture (corresponding to the architecture definition process) and data structure (corresponding to the design definition process).

Productivity enhancement based on monitored developer feelings. Many studies have investigated the sentiment polarity, emotions, and politeness expressed by developers in software artifacts Section 4.1.5. However, few of
them have converted these insights into actionable items. Future researchers could investigate how these measured emotions of developers can be used to enhance productivity. For example, when constant negative emotions are detected from developers, team managers might need to help boost developers’ mood and pay more attention to work-life balance. We would expect controlled experiments to evaluate whether the proposed actions are effective.

**Performance improvement of sentiment polarity analysis.** Inspirations to improve sentiment polarity analysis tools can be distilled from the results in Section 4.6.1. Researchers can focus on constructing vocabularies for specific domains, such as issue reports and app reviews. Also, researchers can integrate several datasets from other domains for pre-training the classifier. As the performance of sentiment analysis tools varies on different datasets, it would also be helpful to design a self-adaptive tool which can adjust the approach based on the type of data it deals with.

**Validation of user feedback.** Section 4.6.2 pointed out a challenge researchers face when identifying opinions from user feedback, namely that many opinions are not valid anymore due to software updates. Therefore, it is necessary to propose an approach to distinguish still valid opinions from outdated ones. This is not trivial as many feedback are not associated with specific versions, therefore, researchers need to rely on other information such as the published date of the feedback and update logs of software for the classification.

**Fine-grained classification of opinion topics.** Researchers have managed to identify whether users are expressing requests (e.g., [P77, P108]) or describing issues and extract tips regarding how to use APIs (e.g., [P169]). However, these classifications are coarse-grained. Given the large amount of feedback available, it is necessary to further categorize the user feedback in order to reduce developers’ manual effort. Topic modeling techniques have been used to address this issue (e.g., [P62]), however, topics automatically generated by these approaches are sometimes not very meaningful. Some researchers have already tried to define taxonomies for types of app reviews [P136]. However, more well-defined taxonomies are needed for other purposes, such as concrete types of API usage tips. Researchers can then classify these opinions in a more fine-grained and meaningful level.

**ACKNOWLEDGMENTS**

We gratefully acknowledge the financial support of the Swiss National Science Foundation for the projects PROBE (SNF Project No. 172799) and SENSOR (SNF-JSPS Project No. 183587).

**LIST OF SELECTED STUDIES**


Manuscript submitted to ACM
Opinion Mining for Software Development: A Systematic Literature Review


Manuscript submitted to ACM


Manuscript submitted to ACM.


Man, submitted to ACM.

Jing Li, Aixin Sun, and Zhenchang Xing. 2018. To Do or Not To Do: Distill crowdsourced negative caveats to augment api documentation.


Sherlock Licorish and Stephen MacDonell. 2014. Relating IS developers’ attitudes to engagement. ACIS.


Manuscript submitted to ACM
Lin, et al.

Nicol Novielli, and Emitza Guzman (Eds.). IEEE, 20–26. https://doi.org/10.1109/AffectRE.2018.00010


Manuscript submitted to ACM
REFERENCES


[10] [n. d.]. SentiSE. https://github.com/amiangshu/SentiSE.


Opinion Mining for Software Development: A Systematic Literature Review


Manuscript submitted to ACM


Opinion Mining for Software Development: A Systematic Literature Review


Manuscript submitted to ACM