Beware of the Fakes – Overview of Fake Detection Methods for Online Product Reviews

Simon André Scherr(✉), Svenja Polst, and Frank Elberzhager

Fraunhofer Institut für Experimentelles Software Engineering,
Fraunhofer Platz 1, 67663 Kaiserslautern, Germany
simon.scherr@iese.fraunofer.de,
{svenja.polst,frank.elberzhager}@iese.fraunhofer.de

Abstract. Online reviews about products and services, such as reviews in stores, are a valuable source of information for customers. Unfortunately, reviews are contaminated by fake reviews, which may lead to wrong conclusions when including them in the analyses of user feedback. As these fake reviews are not marked as advertisement, they might lead to wrong conclusions for customers. If customers are trusting fake reviews their user experience is significantly lowered as soon as they find out that they were betrayed. Therefore, online stores and social media platforms have to take countermeasures against fake reviews. Thus, we performed a systematic literature review to create an overview of the available methods to detect fake reviews and relate the methods to their necessarily required data. This will enable us to identify fake reviews within different data sources easier in order to improve the reliability of the used customer feedback. We have analyzed 141 methods for fake detection. As the reporting quality of a substantial part lacked understandability in terms of method description and evaluation details, we have provided recommendations for method and evaluation descriptions for future method proposals. In addition, we have performed an assessment in terms of detection effectiveness and quality of those methods.

Keywords: User feedback · Online review · Fake review · Spam · Spammer · Literature study

1 Introduction

Online reviews exist for a tremendous number of products and services, and this number has been growing steadily for years. These reviews are a valuable source of information for customers. The reviews express the current product reputation and they contain requests for improvement. Platforms such as Amazon or app stores provide different ways to let people express their feedback, for example, by star ratings or written text. Such feedback can be valuable for a company to improve their products or to convince other users of using their products or services. However, in case the products or services are of bad quality, users provide critical feedback. In other words, the power of users providing feedback via a review has increased in recent years and can have a big influence on the business success of a company.

G. Meiselwitz (Ed.): HCII 2019, LNCS 11578, pp. 453–467, 2019.
https://doi.org/10.1007/978-3-030-21902-4_32

In the scope of our research activities, we developed Opti4Apps as a quality assurance approach that allows developer to include feedback into their quality assurance and development activities [1]. In previous work, we concentrated on different kinds of text analyses [2]. However, we observed that a certain amount of feedback were fake reviews. In general, more than 10% of reviews are assumed to be fake, for some products this is up to 30% [3, 4]. Reviews for apps are also not free of fakes. Including fake reviews (in the following just called fakes) in a feedback analysis of customers has the risk to lead to wrong conclusions. If customers are trusting fake reviews their user experience is significantly lowered as soon as they find out that they were betrayed. This poor user experience will negatively influence their future visits to a website. To prevent such poor user experiences, online stores and social media platforms have to take countermeasures against fake reviews. Detecting and eliminating the fake reviews will lead to an improved reliability of the user feedback and prevent bad user experience. Therefore, our aim was to get an overview of methods that can find such fake reviews, but also fake reviewers. For this, we performed a systematic literature review.

The paper is structured as follows: Sect. 2 presents the foundations in terms of necessary concepts and definitions for our work. Section 3 continues with describing the systematic literature review process, which we have followed. Our results are described in Sect. 4 and discussed in Sect. 5 followed by our threats to validity. We provide conclusions and possible future work in Sect. 7.

2 Foundations

We consider a review to be fake if the review was written for the purpose of promoting or downgrading a product, service or company [5]. It is possible to distinguish between three types of fake reviews [6]: (1) false opinions, (2) reviews on brands only, and (3) non-reviews. We consider the person or bot that is writing fake reviews to be a faker. The place where the fake review was published is called data source. There are two perspectives to identify fakes. It is possible to identify fake reviews, or to identify fake reviewers. We call these perspectives “review view” or “reviewer view”. The reviewer view assumes that reviews posted by a fake reviewer are likely to be fake reviews.

As methods for detecting fakes work on different data, multiple levels can be identified: (1) review level (2) product level and (3) source level. The first level is the review level. Part of this level are methods that work by checking one review. The product level (2) contains the information from all reviews written about a product on a feedback source and the information about the product. The source level (3) contains the information on the data source. This includes information about all products and the profile information, such as the name of the reviewer being available there.

We have defined a model for classifying fake detection. This data model makes use of the different levels being used for fake detection and the two perspectives on fake data. In addition, we have added the information aspects to the model. The model is described in Fig. 1. The data categories mentioned in the model are building blocks of online platforms who offer the review of products. The building blocks are supposed to

ease the identification of available data in a data source and to select appropriate methods for these data.

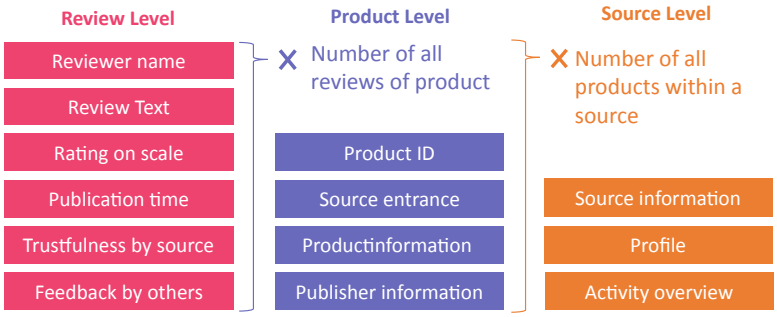

Fig. 1. Data model for method classification

3 Methodology

To capture the state of the art in fake detection in the context of online product and service reviews we have performed a systematic literature review (SLR). The procedure is based on the guidelines provided by Kitchenham [7].

3.1 Planning the SLR

We did not find a review that was conducted systematically, considers all kinds of methods, and covers the most recent ones. A complete and comprehensive description of the methods is necessary to enable us to apply them. Several literature reviews have been published, so far. The study by Sheibani [8] focusses on general terms and definitions rather than methods. Ma and Li put their focus on the challenges and opportunities in the field [9]. Crawford et al. [10] are just covering machine learning methods. The work of Xu is focused on the behavior of fakers that have been identified before [11]. This topic was also investigated in a study of Mukherjee et al. [12]. Heydari et al. provide a comprehensive analysis of fake detection methods [13]. Unfortunately, they have only investigated methods until 2014. Even though the survey of Rajamohana et al. is from 2017 [14] they just analyzed seven methods. They concluded that the field still required future research.

Our goal is to get an understanding of fake detection methods and how these approaches can be applied to different data sources. Therefore, we decided to identify methods, how they work, and how they were evaluated. We considered the data that was used for evaluation as especially important since the characteristics of the data determine whether a method can be applied to another data set. Such characteristics are the language, source of the data set, the domain and the data attributes being used. In addition, we want to analyze how good those fake detection methods are reported. We came up with the following research questions for our SLR:

RQ1: On which aspects do fake detection methods focus on?
RQ2: Which data is used for fake detection?
RQ3: How is the reporting quality of the methods?

3.2 Performing the SLR

Identification of Research. First, we identified the search engines that include publications about fake reviews. These engines were identified and used within the prototyping search phase: ACM Digital Library, Google Scholar, IEEE Xplore Digital Library, Science Direct, Scopus, and Springer Link. Within the prototyping phase we constructed a search query for the SLR. The query was prototyped and revised a couple of times. In the end we agreed on the following query term:

("online review" OR "product review" OR "product recension" OR "online reviews" OR "product reviews" OR "product recensions" OR write-up) AND (fake OR spam OR spammer OR fraud OR deceptive OR manipulation OR "opinion spam")

Our term contains two different major subjects that are connected by a logical 'and'. Each of the major subjects contained various synonyms connected by logical 'or' operators. The first block was used to restrict the results to elements mentioning online reviews and different synonyms. The second one was used to restrict the results to elements mentioning fake or faker and their synonyms.

We decided to apply the query to the title, keyword, and abstract in our prototyping phase. As Google Scholar is not able to offer the abstract field, we executed the search based on the title. Springer Link does not offer a restriction to fields therefore we have not made any restriction. In our prototyping phase it turned out that ACM Digital Library was not able to handle the complexity of our query. Due to this issue, we used this library only for a cross check with our systematically derived sources. We checked after our formal search phase the first 100 results of the result set but did not find new studies to include. An explanation could be that we have used Scopus and Google Scholar, which include the search for ACM content.

We performed our searches in January 2018. In total, the engines found 667 results. An overview of the results per search engine can be seen in Table 1. Google Scholar provided 295 and Scopus 192 results.

Table 1. Data sources, fields and number of results that we have considered

Data source	Field	Number of results
Google Scholar	Title	295
IEEE Xplore Digital Library	Abstract	54
Science Direct	Abstract	77
Scopus	Abstract	192
Springer Link	All	49
Total results		667

Study Selection. In the beginning, we eliminated duplicates from the results and defined inclusion and exclusion criteria. We considered a result as duplicate if it was included multiple times. We also considered results as a duplicate if we have found a more recent version of the result. The inclusion criteria are described in Table 2. Our exclusion criteria were the opposite the inclusion criteria. A publication had to fulfill all the inclusion criteria to be considered for our SLR. We required a publication to be published from 2007 (T) onwards as the ground-breaking paper for fake detection in online reviews was published in this year [6]. The publication language was required to be fully English (L). The publication type (PT) had to be an article, conference proceeding, journal paper, book chapter, or thesis documents. We required the studies to be focused on online product reviews (OnR). In addition, they had to focus on fake content or spammer (FaSp). Moreover, the papers must present and explain one or multiple methods for detecting fake reviews, fake reviewers, spam or spammers (Me). In the initial version of our criteria, the criteria FaSp and Me were combined to one criterion. In the prototyping phase, we had quality assured our criteria and found out that a distinction into two criteria is necessary to assure a high quality.

Table 2. Inclusion criteria for the selection study phase

Code	Criterion	Definition
T	Time Period of publication	Publication since 2007
L	Language	English
PT	Publication type	Article, Conference Proceeding, Journal Paper, Book Chapter, Thesis
OnR	Online	The paper is focused on product reviews published online
FaSp	Fake or Spammer	The paper is focused on fake reviews, Review Spam or Spammers publishing reviews
Me	Method	The paper has to mention and explain methods or approaches how to identify fake reviews or fake reviewer

We defined the following strategy for the study selection: each paper had to be reviewed by two selectors independently from each other. If one selector decided to include the paper and the other one decided for exclusion, they had to discuss this conflict and come to an agreement. In our selection protocol, we wrote down the selectors' decision, based on which inclusion respectively exclusion criteria the decision was made and to what extent the selector has read the paper.

We defined the following reading strategy for the selection phase (see Fig. 2). First the researchers had to check the meta data, if the result was fitting at all. Then, they should read the title. If the title did not include enough information about the inclusion criteria, they had to read the abstract, then the introduction and conclusion and finally they had to (cross-) read the full paper. We decided to report exclusions by the furthest reading step on which one of the readers excluded it, i.e. if reader one excluded the paper by the abstract and reader two excluded it by reading introduction and conclusion, we reported the paper as excluded by reading introduction and conclusion. In the

end, we had a total number of 139 included results for our extraction phase[1]. Our readers had a very good agreement when selecting the papers. In total, we had only 23 conflicts when selecting the 549 different duplicate-free papers.

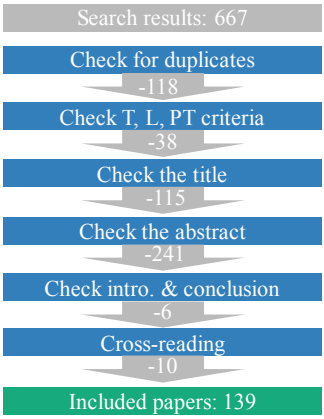

Fig. 2. Reading strategy for selection and number of excluded studies per step

Quality Assessment. We performed a quality assessment for the studies we have included. To be able to extract the methods in a suitable way we defined the following three criteria. (1) We required a result to cover primary research in the area of detection methods. This means that the authors had to present a new or improved method and not just applied a method existing (quality criterion Pri). (2) In addition, we only extracted the methods if they were described in a comprehensible way (quality criterion MethodDec). (3) Furthermore, we required a method evaluation (quality criterion MethodEval) to be present and written in an understandable way providing information of how the method was evaluated, which data was used and what the result was. If the result only fulfilled the first two criteria and not the third one, we did extract it, but marked the impact as low. If it was not fulfilling the first two criteria, we have not performed an extraction. Two persons took decisions for whether a paper not fulfilled the quality criteria. The first proposed a mismatch between our quality criteria and a second researcher had to check if the result really does not fulfill them. We used the second two criteria as a minimum standard for a paper to fulfill. Achieving these criteria does just assure a quality baseline and cannot act as seal of quality. 33 results did not completely fulfill our quality criteria for the extraction. 21 of those results have not provided a new method for fake detection (quality criterion Pri). Five additional results have not described a fake detection method (quality criterion MethodDec). Therefore, these 26 results were not extracted at all. Additional seven results have not fulfilled our quality criterion for method evaluations (MethodEval).

[1] Due to the large amount of results being found our selection phase results are available to download from http://opti4apps.iese.de/fakes/downloads.html and not listed in the paper.

Data Extraction. We performed the data extraction with the help of a data extraction guideline and an extraction form that was made available to all extractors[2]. The guideline contained the extraction procedure, information for extraction, and our quality assessment criteria as well as a detailed description of each field from our extraction form. To assure the quality of the guideline and the extraction form each extractor should perform one test extraction and give feedback on the guideline and the form. Based on this we improved the guideline in terms of clarity and added additional information that was missed by extractors in the trial phase. The improved guideline was given to the extractors to review the guideline again to maximize the quality. After that, the data extraction took place. Our extraction form and the fields being used are described in the online material. In our extraction phase it turned out that we were not able to access, even with the help of a document retrieval service, five full papers that seemed promising from their descriptions.

Data Synthesis. We investigated how many selected papers were published per year. Then, we calculated the total number of identified methods and the number of papers that describe more than one method. We counted the number of methods that report an assumption for the method and investigated the degree of automation, whether the method detects fake reviewers or fake reviews and which level the method addresses. Moreover, we analyzed how many methods were applied for a certain domain and in which natural language the reviews were written. When extracting the domain, we identified several domains that were overlapping, such as hotel and service or product and book. There are data sets that refer only to books but data sets that included reviews to several different products on Amazon cannot be assigned to a specific domain. As we realized that the domains were not as distinct as we expected, we categorized the domains into the two main domains ‘product’ and ‘service’. We also analyzed the language of the data set and whether the language was mentioned at all.

We investigated the quality of the method descriptions, evaluation descriptions and results regarding completeness and comprehensibility. We created a checklist on how aspects should be reported by authors proposing fake detection methods. The checklist can be seen in Table 3, our results per method are available online. Aim of the checks is to distinguish methods being presented and evaluated in a detailed and clear manner from those being insufficiently presented and evaluated from the perspective of a person that is seeking for fake detection methods in order to apply them. The checklist is not suitable to rank methods in terms of detection quality or performance and should more serve as requirements of elements that should be checked when analyzing methods. Core aspects of the checklist are the description details and the applicability of the method, as well as the reported evaluation steps and results. If a method did not fulfil the criteria, we reported a 0, if the method fulfilled the criteria, we reported a 1. If a criterion like the reporting of assumptions was not applicable to the method, we have not reported a value to keep it neutral. We calculated a score in percent of the fulfilled and applicable criterions to provide a ranking of the methods.

[2] We have made the extraction guideline, results and form downloadable from http://opti4apps.iese.de/fakes/downloads.html.

Table 3. Checklist for method description and evaluation assessment

Name	Definition
View	Is it clear if the method tries to detect spam or spammers?
Level	Can we map the method to our data levels?
Assumptions	If assumptions are made, are they explained? Do they preserve a realistic scenario?
Language	Is the natural language named for which the method was proposed
Data categories	Are the used data attributes described in a way that we can classify the methods within our classification scheme?
Degree of automation	Is the degree of automation clear?
Data source	Is the data source for the data to be analyzed named?
Domain	Is the domain of reviews named?
Replicable	Could a third party create an evaluation for a different method, which can be compared if they have access to the data set?
Results	Are the results clearly reported and backed by evaluation metrics?
Own fake data	If manually inserted fake data is used, is it explained how this data has been created?
Data set	Is the data set described clearly i.e. which elements were used and how they got there?

4 Results

In total, we identified 141 methods from 108 papers. 22 of these papers described more than one method. The maximum number of methods described in a single paper was seven. Since the number of identified methods is big, we do not describe the identified methods in this publication. The method descriptions are available under the download link mentioned above. Figure 3 shows the number of selected papers per year. In the year 2007, only the initial paper by Jinadal about fake detection methods was published [6]. The figure shows that the importance of fake detection methods has greatly increased since then.

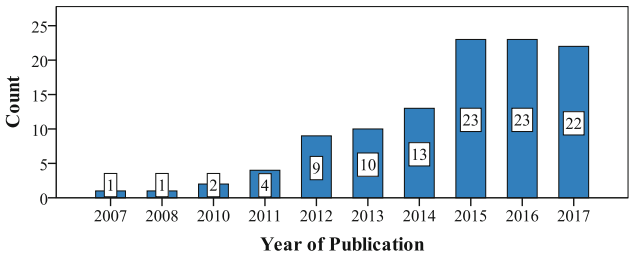

Fig. 3. Number of extracted studies per year

4.1 Aspects of Methods (RQ1)

We have analyzed the distribution of the domains. The reviews were in 32.6% of the methods about products, in 20.6% about a service and in 3.5% about products as well as services. Most methods (43.3%) did not report the domain.

In 2011, 2013 and 2014 data sets with fake and non-fake reviews were developed by Ott [15, 16], Li [17] and their colleagues. They collected reviews from various hotel rating sites (Expedia.com, Hotels.com, Orbitz.com, Priceline.com, TripAdvisor.com and Yelp.com) and extended them with self-created fake reviews. These sets were made available to the public and later used by other authors. Nine studies used the data set of Ott et al. and seven the one by Li et al. Reviews from Amazon were used for the evaluation of 34 methods. Next to Amazon, Yelp.com was the most often used source of reviews (21 methods). TripAdvisor was used seven times. Twenty-three methods used reviews from various sources, three of them used even seven different sources. However, the majority (84 methods) uses only one source.

More than the half of the methods (56.0%) address the detection of fake reviews, about one-third of the methods (53 methods, 37.6%) aim at detecting fake reviewers and only six (4.3%) address reviews as well as reviewers. Methods for fake reviews investigate, for instance, the review length the appearance of duplicates, the sentiment, readability scores and inclusion of hyperlinks in a review test. For the identification of fake reviewers, the source level is required by 45 out of 53 methods (88.2%). The level required by the methods about fake reviews is more divers; the review level is the most often used level (45.6% of fake reviews methods), the product level the second most (35.4%) and the source level is used by 19.0%. None of the methods addressing fake reviews as well as fake reviewers uses the product level. These methods used equally either review or source level.

4.2 Data Used for Fake Detection (RQ2)

One goal of the systematic literature review was to map methods to required data (see Fig. 1. for data model). The idea is to have a list of all data necessary to apply a method. This list could be matched with the data available in a certain data source, such as Amazon or Google Play Store. For 17 methods (12.1%) we were not able to identify data. For one more method, we could identify only one data, however, it was obvious that more, but unidentifiable data were required.

Reviews mostly consist of the review text, a rating on a defined scale, the name of the reviewer, trustfulness information added by other reviewers (e.g. helpfulness rating), trustfulness ratings by the system (e.g. verified purchase) and publication date of the review. Most methods (89 methods, 63.1%) use the review text as input. These kinds of methods perform linguistic analyses such as readability analyses (e.g. [18]) or they use the text to identify duplicates. There are methods that consider exact duplicates, but also partially related reviews. In 53 methods (37.6%), the text is the only data required for the method. The rating on a rating scale and the publication date are also often subject of a method, the rating in 32 methods and the publication date in 15 methods. The ratings given by a reviewer could be compared to the average ratings to the reviewed products. If the ratings often diverge from the average, the reviewer is

considered suspicious. The publication date could be used in several ways, too. It could be compared to the date a product enters a source (i.e. a product is set on an online platform). If the publication date is shortly after the product entrance date, the review is suspicious by several methods (e.g. [19]). The publication date could also be used to analyze the number of reviews a reviewer writes within a certain time.

Three methods used the trustfulness rating provided by the source. In all three cases the rating was the verified purchase information. The feedback added by other reviewers was used in two methods. For nine methods it was clear that an identification of the reviewer was needed, however, it was not clear which type of data was used for the identification.

According to our data model, on the product level, there are four types of data; a product ID (required by 13 methods), information about the publisher respectively the product brand (6 methods), the information added by the publisher (e.g. product description) (2 methods), date of entering the source (1 method).

The methods aiming at identifying fake reviewer often require the activity overview of a user. This mostly contains all reviews written to all products by a specific reviewer. The methods often extract the time of publishing a review and the rating on a rating scale. Often more information of the product is required, such as the brand and the rating of other reviewers. The methods then consider a reviewer suspicious if his ratings often diverge from the ratings of other reviewers or when he mostly reviews product of a certain brand. The activity overview of a reviewer was used in 31 methods. The profile of a reviewer was required for nine methods.

4.3 Reporting Quality Assessment (RQ3)

We applied our checklist for fake detection method description and evaluation assessment to the methods. Figure 4 shows the results how well the methods have been evaluated according to our checklist. On average the methods achieved a score of 64%. 101 methods achieved a score being higher than 50%, but only 56 achieved more than 70%. The criteria which were fulfilled mostly were the view (97%) and level (96%) description. On the opposite we only got the natural language mentioned in 16% of the methods. In 4.9% of the methods the review language was Chinese and in 9.8% it was English. Methods that reported the use of a data set by Ott et al. [16] or Li et al. [17] were considered to have reported English reviews even if the language was not directly mentioned in the study.

35% of the evaluations contain a description which data was used, which goes beyond mentioning the data source and some general remarks. 37 methods did not mention any data source of the reviews. 24% of the method evaluations use self-created fake data describe in a reproducible way how this data was inserted. The number of these methods does not include the self-created fake reviews by Li et al. [17] and Ott et al. [16]. Adding the methods that rely on these data sets, a total of 22 methods were evaluated on self-created fakes. Most of the methods did not report an assumption (68.8%). Methods that detect fake reviewers had more often an assumption (39.6% of fake reviewer methods) compared to methods that address fake reviews (24.1% of fake review methods). The degree of automation of the methods is high as 62.4% of the

methods are fully automated and 7.8% are semi-automated. However, there are many methods that do not report the automation degree (28.4% of all methods).

Three method proposals achieved a score of 100%. These methods were namely presented by Ahsan et al. [20], Heydari et al. [19] and Sandulescu and Ester [21].

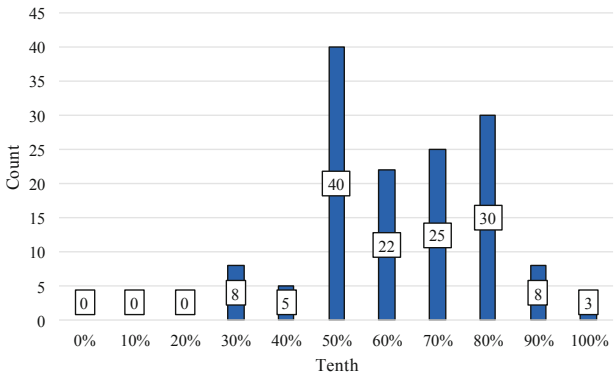

Fig. 4. Quality assessment results grouped by tenths

5 Discussion

We identified required data for 88% of all methods but we assume that the list of identified data is not complete. It seems reasonable that several methods need to somehow distinguish users. However, the identification of users is mostly not described. Methods aiming at identifying fake reviewer by their activities in the source platform often do not describe how they obtain the activities. The activities could be obtained by a publicly available activity overview linked to a profile or by generating a large set of data within the source.

We identified three major problems: (1) documentation of the method, (2) documentation of the evaluation, and (3) description of the evaluation data. This implies that there is still a lot of improvement potential not only for better methods but also for better reporting and evaluation of them.

Methods are documented in very different levels of detail. The publications range from detailed algorithms and theoretical mathematical background for the method like in Ye and Akoglu [22] to just mentioning that the method solves fake detection with machine learning [23]. Many methods do not report information that is necessary for applying the method to an own data set. Authors need to mention the data source for which the method is designed, as it might use unique characteristics of that source. We also noticed that authors frequently do not describe the target language for their detection method. This is a problem especially if the method uses linguistic approaches for the detection. Recently publications in the field of natural language processing methods actively investigate the problem that usually those methods are highly depending on the language being used and cannot be just adapted to new languages [24].

Our results lead to several recommendations about how to report a fake detection method. The method should be described in a way that other people get an understanding of the process of detecting the fakes. This means finding a balance between a high-level description not providing any details and a description being full of complex mathematical formulas which require comprehensive expertise in this subject area. We recommend to present algorithms or processes as flow charts, activity diagrams or sequence diagrams to describe the methods in a nutshell. To identify if a method is applicable for certain use case it is important to mention the natural language, domain, source of the data and characteristics of the data source that could restrict the applicability of the method. We would also like to see a description how automated the method is. In case of semi-automatic methods, it should be clear which parts are automated and which not.

Even though authors started to reuse data sets, we identified six methods that were not evaluated by their authors or were not described in a comprehensible way and 37 evaluation descriptions did not contain information about the data set or even the data source being used. The problem of test data generation was mentioned as critical issue in the SLR of Ma and Li in 2012 [9]. Recent publications in the field posed several standard metrics how to evaluate a method. The metrics precession, recall, F-measure and accuracy have been widely accepted [25]. This in connection with a gold standard for data is a step forward in enabling evaluations to be comparable.

As mentioned above it was not always clear which data was used for the evaluation. Some authors used a mixture of real data from websites enriched with self-made fake data, e.g. Banerjee et al. [26]. This data generation is a challenge. As Zhang et al. [25] reported it is complex to generate realistic fake data. Also, the literature survey of Heydari et al. [13] complained about lack of gold standard data sets. Recent years showed an improvement in that area. In addition, the source data itself might contain fake data that might or might not be detected by the method. We consider the artificial creation of fake data to evaluate a method as a threat to validity for the methods. First methods might be optimized to detect the artificial fake data sets, which might be different to real data.

6 Threats to Validity

While preparing and performing our SLR we have identified and mitigated several possible threats to validity. To capture the relevant sources for establishing a state of the art in fake and spam detection in the context of online reviews it is crucial to not miss relevant sources. We tried to achieve this by prototyping our search term in an iterative way, adding a lot of synonyms, to keep the result set relatively broad. Within our selection phase we put emphasis on only deselecting obviously not matching results within the title stage. In addition, every selection decision was performed by two independent people. This should prevent that the opinion of one single researcher is able to influence the selection decision.

Another identified threat is that errors might occur while searching. This might be entering a malformed search term, using the wrong field codes or copying not all the

results. Therefore, we checked the search results with our trial searches to ensure not having a wrong term entered.

To perform SLR the extraction phase is crucial. A wrong extraction guideline or a not matching extraction sheet reduces the quality a lot. Therefore, we quality assured and prototyped our guideline as well as the extraction form. In addition, we focused in our extraction phase on copying citations from our results into our forms. This should prevent the addition of personal interpretation or opinion of the primary research. Our listing of methods faces the problem that it was up to the authors to define what a method exactly is. It might be the case that an author proposes a set of methods that would be proposed by a different author as a single method.

7 Conclusion and Future Work

We performed a systematic literature review on the topic of fake detection methods within online reviews. Our search led to initially 139 papers being included. 141 methods were extracted from a result set of 108 different papers. These methods tried to detect fakes with various approaches ranging from analyzing the content of a review to analyzing the entire user behavior. We have mapped the methods to two different views. Identifying fakes and identifying fakers. Furthermore, we classified the data level and data categories the methods are using. We observed that most methods were using the review level, followed by the product level. This seems naturally as more methods were focused on detecting fakes compared to detecting fakers.

Our analysis of the method descriptions and evaluations revealed that the descriptions lack information, which is necessary to apply the methods to other data sets. The fact that data sources and review language was missing quite frequently is a huge problem in reproducing the evaluation. The current status of available fake detection methods makes it hard to prevent poor user experience as untruthful reviews cannot be detected easily. This lowers significantly the potential of online reviews as source of trustful information. We were clearly able to identify Amazon, Yelp, and TripAdvisor as leading platforms of investigation. The data set provided by Ott et al. [16] and Li et al. [17] became powerful standard data sets to be used.

Despite the found shortcomings the results show many promising opportunities to continue our investigation. In the context of user feedback analysis removing fake reviews is the first step to raw data quality assurance. The second step would be to identify indicators how reliable or trustworthy the different feedback entries are. Measuring reliability and trustworthiness is yet a huge challenge for the research community that has just been addressed in some smaller focus areas.

Acknowledgments. The research described in this paper was performed in the project Opti4Apps (grant no. 02K14A182) of the German Federal Ministry of Education and Research. We would like to thank the students Selina Meyer, Lisa Müller, Sadaf Alvani, Phil Stüpfert and Lukas Zerger, who contributed to the systematic literature review.

References

1. Elberzhager, F., Holl, K.: Towards automated capturing and processing of user feedback for optimizing mobile apps. Procedia Comput. Sci. **110**, 215–221 (2017)
2. Scherr, S., Elberzshager, F., Holl, K.: An automated feedback-based approach to support mobile app development. In: Proceedings - 43rd Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2017, Vienna (2017)
3. Tuttle, B.: 9 Reasons Why You Shouldn't Trust Online Reviews. http://business.time.com/2012/02/03/9-reasons-why-you-shouldnt-trust-online-reviews/. Accessed 03 Feb 2012
4. Weise, K.: A Lie Detector Test for Online Reviewers. https://www.bloomberg.com/news/articles/2011-09-29/a-lie-detector-test-for-online-reviewers. Accessed 29 Sept 2011
5. Jindal, N., Liu, B.: Opinion spam and analysis. In: WSDM 2008 Proceedings of the 2008 International Conference on Web Search and Data Mining, Palo Alto, California, USA (2008)
6. Jindal, N., Liu, B.: Analyzing and detecting review spam. In: IEEE International Conference on Data Mining, Omaha, NE, United States (2007)
7. Kitchenham, B.: Guidelines for performing systematic literature reviews in software engineering (2007)
8. Sheibani, A.: Opinion mining and opinion spam: a literature review focusing on product reviews. In: 6th International Symposium on Telecommunications (IST 2012), Shiraz (2012)
9. Ma, Y., Li, F.: Detecting review spam: challenges and opportunities. In: 8th International Conference Conference on Collaborative Computing: Networking, Applications and Worksharing, Collaboratecom, Pittsburgh (2012)
10. Crawford, M., Khoshgoftaar, T., Prusa, D., Richter, N., Al Najada, H.: Survey of review spam detection using machine learning techniques. J. Big Data **2**, 24 (2015)
11. Xu, C.: Detecting collusive spammers in online review communities. In: International Conference on Information and Knowledge Management, Proceedings, San Francisco (2013)
12. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What yelp fake review filter might be doing? In: Seventh International AAAI Conference on Weblogs and Social Media, Cambridge (2013)
13. Heydari, A., Tavakoli, M., Salim, N., Heydari, Z.: Detection of review spam: a survey. Expert Syst. Appl. **42**(7), 3634–3642 (2015)
14. Rajamohana, S., Umamaheswari, K., Dharani, M., Vedackshya, R.: A survey on online review SPAM detection techniques. In: 2017 International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT), Coimbatore (2007)
15. Ott, M., Cardie, C., Hancock, J.: Estimating the prevalence of deception in online review communities. In: WWW 2012 Proceedings of the 21st International Conference on World Wide Web, Lyon (2012)
16. Ott, M., Cardie, C., Hancock, J.: Negative deceptive opinion spam. In: NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference, Atlanta (2013)
17. Li, J., Ott, M., Cardie, C., Hovy, E.: Towards a general rule for identifying deceptive opinion spam. In: 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore (2014)
18. Banerjee, S., Chua, A.: A linguistic framework to distinguish between genuine and deceptive online reviews. In: Proceedings of the International Multi Conference of Engineers and Computer Scientists, Hong Kong (2014)

19. Heydari, A., Tavakoli, M., Salim, N.: Detection of fake opinions using time series. Expert Syst. Appl. **58**, 83–92 (2016)
20. Ahsan, M.N.I., Nahian, T., Kafi, A., Hossain, M., Shah, M.: An ensemble approach to detect review spam using hybrid machine learning technique, Dhaka, Bangladesh (2016)
21. Sandulescu, V., Ester, M.: Detecting singleton review spammers using semantic similarity, Florence (2015)
22. Ye, J., Akoglu, L.: Discovering opinion spammer groups by network footprints. In: Appice, A., Rodrigues, P.P., Santos Costa, V., Soares, C., Gama, J., Jorge, A. (eds.) ECML PKDD 2015. LNCS (LNAI), vol. 9284, pp. 267–282. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23528-8_17
23. Xi, Y.: Chinese review spam classification using machine learning method. In: 2012 International Conference on Control Engineering and Communication Technology, Liaoning (2012)
24. Hogenboom, A., Bal, M., Frasincar, F., Bal, D.: Towards cross-language sentiment analysis through universal star ratings. In: Uden, L., Herrera, F., Bajo Pérez, J., Corchado Rodríguez, J. (eds.) Knowledge Management in Organizations: Service and Cloud Computing. AISC, vol. 172, pp. 69–79. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-30867-3_7
25. Zhang, D., Zhou, L., Kehoe, J., Kilic, I.: What online reviewer behaviors really matter? Effects of verbal and nonverbal behaviors on detection of fake online reviews. J. Manag. Inf. Syst. **33**(2), 456–481 (2016)
26. Banerjee, S., Chua, Y., Kim, J.: Let's vote to classify authentic and manipulative online reviews: the role of comprehensibility, informativeness and writing style. In: 2015 Science and Information Conference (SAI), London (2015)