Detecting Review Spam on Amazon with ReviewAlarm

Author: Anna-Theres Pieper
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT
When making purchasing decisions, customers increasingly rely on opinions posted on the Internet. Businesses therefore have an incentive to promote their own products or demote competitors' products by creating positive or negative spam reviews on platforms like *Amazon.com*. Several researchers propose methods and tools to detect review spam automatically. Reviewskeptic is an automated tool developed by Ott et al (2012) which attempts to identify spam reviews on hotels by employing text-related criteria. This research proposes a framework for detecting also non-hotel spam reviews with reviewskeptic by enhancing the tool with reviewer behavior related and time related criteria derived from the literature. The new framework will be called ReviewAlarm. A ground of truth dataset has been created by the means of a manual assessment and has been used to compare the performance of reviewskeptic and ReviewAlarm on this dataset. With ReviewAlarm we are able to improve the performance of the tool on our dataset. However, this research also reveals several weaknesses about the criteria that are being used in review spam detection. Therefore, we argue that additional qualitative research related to reviewer behavior and the criteria is needed in order to come up with better tools and more generalizable criteria.

KEYWORDS
Review Spam, Spam Detection Tools, Reviewskeptic, Review Spam Detection Criteria, Reviewskeptic

Supervisors:
Dr. Fons Wijnhoven
Dr. Chintan Amrit

7th IBA Bachelor Thesis Conference, July 1st, 2016, Enschede, The Netherlands.

1. AN INTRODUCTION TO REVIEW SPAM AND THE NEED TO DETECT IT

When making purchasing decisions, customers increasingly rely on opinions posted on the Internet (Hu, Bose, Koh & Liu, 2012). Internet users can easily and openly express their opinion about a product or brand by using social media or online product reviews and reach up to millions of potential buyers. With the assistance of opinion mining tools, businesses can retrieve valuable information with regard to product, service and marketing improvements from this kind of user-generated content (Heydari, Tavakoli, Salim & Heydari, 2015).

Online opinions thus can have great impact on brand and product reputation as well as related sales and management decisions. This gives an incentive to businesses to create, for example, positive fake reviews on their own products and negative fake reviews on their competitors' products (Akoglu & Rayana, 2015). There is a variety of ways to spam the internet with fake content. For instance, by hiring professional firms which are specialized in writing spam reviews, by using crowdsourcing platforms to employ review spammers or by using robots to create synthetic reviews. Reviews produced by someone who has not personally experienced the subjects of the reviews are called spam reviews (Heydari et al, 2015). The person who creates the spam review is called an individual review spammer. Individual review spammers working together with other review spammers are group spammers (Mukherjee, Liu & Glance, 2012).

Due to the amount of reviews posted online and the proficiency of review spammers, it is often very hard to detect spam reviews and separate them from trustworthy ones. However, it is essential not only for businesses but also for customers that review spam can be identified and removed in a reliable way. Researchers have suggested a variety of methods and tools to identify spam reviews, review spammers and spammer groups (e.g. Jindal & Liu, 2008; Mukherjee et al, 2012; Xie et al, 2012). One of these tools is *reviewskeptic.com*[1] developed by Ott, Choi, Cardie, & Hancock (2011). The authors claim that reviewskeptic is able to detect spam reviews on hotels based on psychological and linguistic criteria with 90% accuracy. However, hotel reviews are only a fraction of the opinions posted on the Internet. Many reviews are related to individual products, services, brands or stores. Reviewskeptic claims to be a well-working yet very specialized tool for spam review detection. The aim of this research is to assess reviewskeptic's performance on non-hotel reviews and based on the existing literature to give recommendations on how the tool could be enhanced to detect also non-hotel review spam effectively.

The identification of spam reviews will be a relevant research topic as long as opinions will be expressed on the internet. Not only the tools for detection are improving but also the ways of producing review spam are getting more advanced. For example, sellers on Amazon[2] now have the opportunity to provide their products for free or at a discount in exchange for a review. Thereby, the review is still marked as a verified purchase and thus seems more trustworthy to potential buyers and to conventional review spam detection methods. However, the honesty of the reviews obtained in this way is highly questionable (Bishop, 2015). This example shows the importance of developing, testing and improving new methods for spam review detection which can keep up with the novel ways of producing spam reviews constantly.

[1] www.reviewskeptic.com

[2] www.amazon.com

Next to contributing to the existing literature, this research delivers valuable contributions to e-commerce sites, businesses and customers. We argue that e-commerce sites are responsible for preventing spam reviews from appearing on their sites. Yelp[3] – a website for reviewing local businesses – made substantial efforts to ban spam reviews by employing a trade secret algorithm which is likely to be based on reviewer behavior (Mukherjee, Venkataraman, Liu, & Glance, 2013). In contrast to Yelp, Amazon has so far put only limited effort in increasing the reliability of the reviews posted on their site. In 2015, Amazon has taken legal action against individual review spammers as well as companies offering spam reviews online (Gani, 2015). However, this will not solve the problem to its full extent since there are numerous ways to generate spam reviews and various platforms to buy them. This research gives insights for improving the mechanisms of detecting and preventing spam reviews, also on Amazon.

A literature review has been carried out to derive a list of criteria that can be used to identify review spam. Based on this list and recommendations from the literature, a method to manually detect spam reviews has been developed and used to come up with a labeled dataset of 110 Amazon reviews. The labeled dataset has been used as a ground of truth against which reviewskeptic's performance has been compared. For several criteria of the list mentioned before it has been tested in how far these criteria complement and enhance the judgment of reviewskeptic. Recommendations will be given for improving reviewskeptic's performance in detecting non-hotel spam reviews.

To summarize, the contributions of this research are:

1. An extensive list of criteria that can be used to identify spam reviews.
2. An assessment of reviewskeptic's usefulness in identifying non-hotel spam reviews with regard to the ground of truth dataset.
3. An assessment of how criteria from the literature can enhance reviewskeptic's performance.
4. Recommendations on how to increase the performance of reviewskeptic in detecting non-hotel reviews.

The guiding research question:

Which criteria for spam review detection can be used to enable the tool reviewskeptic to detect non-hotel spam reviews more reliably?

In the remainder of this paper first an overview of the related scientific work to the topic of spam review detection will be given. In chapter three we will describe the dataset, the method for manual assessment, the concept of reviewskeptic as well as the additional criteria that have been tested. In chapter four the results of each method will be introduced. Chapter five will discuss these results, name the limitations of this research and give recommendations for future research. In chapter six a conclusion will be drawn.

2. RELATED WORK

A literature search has been conducted on SCOPUS with the search query "*(Fake OR Manipulate* OR Spam) AND (Review) AND (Detect* OR Identify*) AND (Method OR Tool OR Requirements OR Criteria)".* This query includes automated as well as manual detection methods. The initial search revealed 501 articles. After examining titles, abstracts and knowledge domains, the search results have been narrowed to 64 articles. 20 of these articles have been considered for this research based

[3] www.yelp.com

on their quality and relevance. Additionally, the references of these articles have been looked at and some of the sources have been used for the literature review as well. **Appendix A** gives an overview on the literature coverage of methods and tools for spam review detection as well as those papers in which a manual assessment has been employed. The manual assessment is further explained in the methods section. **Table 1** shows the criteria that are related to the methods and tools and their popularity in the literature.

2.1 Web Spam and E-mail Spam

Web spam describes web pages which have been created deliberately for triggering favorable relevance or importance on the Internet. The goal of web spam is mainly to distract search engines to obtain a higher page rank for the target web page (Lai et al, 2010). In web spam it can be differentiated between content spam and link spam (Savage, Zhang, Yu, Chou & Wang, 2015). Several researchers have developed mechanisms to detect content as well as link spam, one of them is TrustRank (Gyöngyi, Garcia-Molina & Pedersen, 2004). TrustRank differentiates between trustworthy and spam intense sites and assigns low trust scores to spam websites (Gyöngyi et al, 2004). However, not only researchers but also search engine providers like Google[4] have taken efforts to increase the quality of search results and tackle spam. In 2011 and 2012 Google has, for example, made major changes to its algorithms which are intended to rank search results based on content quality and natural links between high quality pages ("Panda? Penguin? Hummingbird? A Guide To Feared Google's Zoo", 2014). The mechanisms used for detecting web spam in general have also built the basis for tools that can be used to detect spam reviews in particular.

2.2 Review Spam

Jindal and Liu (2007) were the first authors to study the trustworthiness of reviews. They argue that spam reviews are much harder to detect than regular web spam. Jindal and Liu (2008, pp 219) define three types of review spam:

1. *Untruthful opinions* deliberately mislead readers or opinion mining systems by giving undeserving positive or malicious negative reviews to some target objects.
2. *Reviews on brands only* do not comment on the products in reviews specifically but only on the brands, the manufacturers or the sellers of the products.
3. *Non-Reviews* contain advertisements or no opinions.

The main task in review spam detection is to identify the untruthful opinions. The other two types of spam can be identified more easily by a human reader and thus he/she can choose to ignore those (Jindal & Liu, 2008).

2.2.1 Spam Review Detection Methods and Tools

The methods and tools for spam review detection can be grouped into four categories: Review Centric, Reviewer Centric, Time Centric or Relationship Centric.

Review Centric. Jindal and Liu (2008) propose a supervised learning method to identify spam reviews. Their approach is based on finding duplicate and near-duplicate reviews by using a 2-gram based review content comparison method. Additionally, the authors proposed a variety of criteria which can be used to identify spam reviews. Lai et al (2010) employed a similar approach in which a probabilistic language model is used to compute the similarity between pairs of reviews. The authors model the likelihood of one review being generated by the contents of another with the Kullback Leibler divergence measure that estimates the distance between two probability distributions. The research by Ott et al (2011) builds on linguistic as well as psychological criteria of spam reviews and truthful reviews. Since their research is at the heart of this paper it will be described in more detail in chapter three. Ong, Mannino, & Gregg (2014) examined the linguistic differences between truthful and spam reviews. They found that spam reviews concentrate on the information that is provided on the product page and that they are more difficult to read than truthful reviews. The authors suggest that a software should be used to classify the criteria of their findings and separate spam reviews from truthful ones. A similar language based method is proposed by Kim, Chang, Lee, Yu & Kang, (2015) who focus on semantic analysis with FrameNet. FrameNet helped to understand characteristics of spam reviews compared to truthful reviews. The authors use two statistical analysis methods (Normalized Frame Rate and Normalized Bi-frame Rate) to study the semantic frames of hotel reviews and they are able to detect semantic differences in nonspam and spam reviews. Lu, Zhang, Xiao & Li (2013) take on a different approach which aims at detecting the spam review as well as the review spammer at the same time. To achieve this, the authors developed a Review Factor Graph model which incorporates several review related criteria, for example, length of the review and several reviewer related criteria, for example, total helpful feedback rate. Akoglu, Chandy & Faloutsos (2013) use signed bipartite networks to classify reviews. Their tool is called FRAUDEAGLE and it captures the network effects between reviews, users and products. This is used to label reviews as either spam or nonspam, users as either honest or fraud and products as either good or bad quality. FRAUDEAGLE only takes into account the rating and is therefore applicable to a variety of rating platforms. However, Rayana & Akoglu (2015) extended the tool FRAUDEAGLE by expanding its graph representation as well as incorporating meta-information such as the content of the reviews. Their new tool is called SpEagle and achieves more accurate results than the previous one.

Reviewer Centric. Within the reviewer centric methods and tools, Lim et al (2010) identified characteristic behaviors of review spammers with regard to the rating they give and modeled these to detect the review spammer rather than the spam review. Their method is based on the assumption that spam reviewers target specific products and product groups and that their opinion deviates from the average opinion about a product. Based on these characteristics and behaviors, the authors assign a spamming score for each reviewer in a dataset. The method proposed by Savage et al (2015) is also related to anomalies in the rating behavior of a reviewer. The authors focus on differences between a reviewer's rating and the majority opinion. A lightweight statistical method is proposed which uses binomial regression to identify reviewers with anomalous behavior. Wang et al (2015) use a product-review graph model to capture the relationship between reviews, reviewers and products. Additionally, the nodes in their model are assessed according to criteria such as the rating deviation and content similarity. Their algorithm is designed to tackle the problem of computational complexity and inefficiency when identifying spam reviews. A certain amount of reviewers is eliminated during each iteration which significantly speeds up the process.

Time Centric. Several researchers stress the importance of including time-related criteria into the methods and tools to detect review spam. Xie, Wang, Lin & Yu (2012) propose to identify review spam based on temporal patterns of singleton

[4] www.google.com

reviews. Singleton reviews are made by reviewers who only review a single product and they are an indicator of review spam. The authors assume that genuine reviewers arrive in a stable pattern on the reviewing sites whereas spam attacks occur in bursts and are either negatively or positively correlated with the rating. Statistics are used to find such correlations and identify review spammers. This approach is similar to the one by Fei et al (2013). They propose an algorithm which detects bursts in reviews using Kernel Density Estimation. In addition, several criteria are used as indicators for detecting review spammers in review bursts and separating them from nonspam reviewers. Lin et al (2014) use six time sensitive criteria to identify spam reviews, for example, the review frequency of a reviewer. They define a threshold based on the average scores of these criteria and use a supervised method to detect the spam reviews.

Relationship Centric. Mukherjee et al (2012) were the first researchers to examine online reviews based on relationships between the reviewers. With the help of Amazon Mechanical Turk Crowdsourcing[5] they produced a ground of truth labeled dataset for group spam reviews. Mukherjee et al (2012) developed a relation-based approach to detect spammer groups and their tool is called Group Spammer Rank (GSRank). GSRank uses a frequent item set mining method to sets of clusters whose members are likely to work together. The authors used individual spam indicators and group spam indicators to examine the groups that have been found and rank them according to their degree of being a spam group (spamicity). Xu (2013) evaluated anomalies in reviewer behavior with the same set of behavioral criteria as Mukherjee et al (2012) and proposed a hybrid classification/clustering method to detect collusive spammers (group spammers). The members of the groups that are identified show collective behavior such as writing reviews for common target products. Liang et al (2014) work towards detecting review spammers who always work together. Their approach is based on the assumption of TrustRank that trustworthy pages seldom point to spam intense pages. In their multi-edge graph model the authors consider reviewer criteria as well as the relationships between the reviewers. Each node in the model represents a reviewer and each edge represents an inter-relationship between reviewers of one product. These relationships can be conflicting (different opinions) or supportive (same opinion) (Liang et al, 2014). Reviewers who always share supportive opinions with other reviewers are suspicious. For each reviewer an unreliability score is calculated to indicate whether he/she is likely to be a spammer or not. Another approach has been developed by Fayazi, Lee, Caverlee & Squicciarini, (2015) who focus on uncovering crowdsourced manipulation of online reviews. They created a root dataset by identifying products that have been targeted by crowdsourcing platforms with spam reviews. Probabilistic clustering is used to find linkages among reviewers of a product. Hereby, they found groups who often posted reviews for the same products. The authors suggest to integrate the clustering criterion into existing spam review identifiers to increase the performance of these tools.

2.2.2 Criteria for Identifying Review Spam

The overview in **Table 1** includes criteria that have been used to identify spam reviews in the articles introduced before. However, some of the criteria that were mentioned in the literature have been excluded from the list. The reasons for excluding criteria were mainly that they were not generalizable for different product categories, there were conflicting findings over the usefulness of the criteria or they were closely related to one of the other criteria in the list. For example, Ott et al (2011) and Harris (2012) argue that the usage of first person singular pronouns ('I', 'me') is higher in spam reviews than in non-spam reviews. In contrast to that, Akoglu & Rayana (2015) argue that a low percentage of first person pronouns is an indicator of spam reviews. This suggests that the criterion is not generalizable and should not be used in spam review detection. The same counts for the criterion lexical complexity. Whereas Yoo & Gretzel (2009) and Ong et al (2014) found lexical complexity to be higher in spam reviews. Harris' (2012) dataset showed higher lexical complexity in non-spam reviews. We will later present the complexity scores of our dataset and show that there is no significant difference between the complexity of spam reviews and nonspam reviews as measured with the ARI[6]. Furthermore, Kim et al (2015) analyzed semantic frames of hotel reviews, most of them cannot be generalized to all product categories and have therefore been excluded (e.g. non-spam reviews are more likely to include the frame 'building sub-parts').

3. METHOD

In the empirical part, reviewskeptic has first been tested with our dataset from Amazon and then compared against the ground of truth manual assessment. Afterwards, it has been tested in how far adding certain criteria for spam review detection to the analysis would improve the results of reviewskeptic. Amazon has been chosen as a platform because it is relatively transparent regarding the reviewer profiles and provides a large variety of reviews on different products.

The dataset consists of the complete set of reviews for three products: a smartphone case, a smartphone charger and a smartphone screen protector. All data has been collected manually from Amazon. In total, data about 125 reviewers has been collected. For each reviewer the following information has been recorded:

<Reviewer Name> <Review Date> <Average Rating for Product> <Reviewer Rating for Product> <Amazon Verified Purchase (Yes/No)> <Review Text> <ARI> <Reviewskeptic Label (Nonspam/Spam)>

Whereas reviewskeptic uses only the review text to label a review as either spam or nonspam, the other information has been important when the additional criteria were added to the analysis.

3.1 Manual Assessment

The key challenge in assessing methods and tools for spam review detection is the absence of a ground of truth that could be used to compare against. Several researchers solved this problem by employing a manual assessment approach to create a ground of truth labeled dataset. Ott et al (2011) discovered that for a human judge it is very hard to determine whether a review is a spam review or not. However, their research is based on a review centric approach only. Inspired by other researchers (Harris, 2012; Liang et al, 2014; Lin et al , 2014; Mukherjee et al 2012; Lu et al, 2013; Ott et al 2012; Xie et al, 2012) we carried out a manual assessment approach which integrates all of the dimensions of review spam detection and therefore is more reliable.

[5] www.mturk.com

[6] The automated readability index (ARI) is a measure to judge the complexity of a written text which correlates to the US grade level that is needed to understand the text (Olewinski, 2016).

Table 1 Overview of the Criteria for Spam Review Detection

	Criterion	Sources
	1. Review Centric Criteria	
C1.1	**Review content similarity:** If a reviewer's reviews are duplicates or near duplicates to each other or to other reviews, this may be an indicator of review spam.	*Akoglu & Rayana, 2015; Fei et al, 2013; Jindal & Liu, 2008; Lai et al, 2010; Lim et al, 2010; Lin et al, 2014; Lu et al, 2013; Mukherjee et al, 2012; Wang et al, 2015*
C1.2	**Length of the reviews:** Depending on the review spammer's strategy, very short or very long review texts may be indicators of review spam.	*Akoglu & Rayana, 2015, Fayazi et al, 2015; Jindal & Liu, 2008; Lai et al; 2010; Lu et al, 2013*
C1.3	**Brand mention:** If a reviewer mentions the brand name of the reviewed product often this may be an indicator of review spam.	*Harris, 2012; Jindal & Liu, 2008*
C1.4	Excessive **use of numerals, capitals and all capital words** are indicators of non-reviews and synthetic reviews.	*Akoglu & Rayana, 2015; Jindal & Liu, 2008*
C1.5	Excessive **use of positive and negative sentiment words** may be an indicator of review spam.	*Jindal & Liu, 2008; Lu et al, 2013*
C1.6	Excessive **use of superlatives in review text** may be an indicator of review spam.	*Ott et al, 2011*
C1.7	**Focus on external aspects of the products** may be an indicator of review spam.	*Ott et al, 2011*
C1.8	If **personal relationship information** is emphasized in a review this may be an indicator of spam.	*Kim et al, 2015; Ott et al, 2011*
C1.9	If there is high **similarity between the review and the product description** this may be an indicator of review spam.	*Jindal & Liu, 2008*
C1.10	**Order of good and bad reviews:** When a bad review was written just after a good review and vice versa this may be an indicator of review spam.	*Jindal & Liu, 2008*
	2. Reviewer Centric Criteria/Criteria	
C2.1	**Rating deviation:** If the majority of a reviewer's reviews deviates from the average opinion this may be an indicator of review spam.	*Akoglu & Rayana, 2015; Fei et al, 2013; Jindal & Liu, 2008; Liang et al, 2014; Lim et al, 2010; Lu et al, 2013; Mukherjee et al, 2012; Savage et al, 2015*
C2.2	**Helpfulness rating of reviews**: If the majority of a reviewer's reviews are voted to be helpful he/she is less likely to be a spammer.	*Fayazi et al, 2015; Jindal & Liu, 2008; Lai et al, 2010; Liang et al, 2014; Lu et al, 2013*
C2.3	**Number of written reviews:** Depending on the review spammer's strategy, a larger number of reviews or a very low number of reviews may be an indicator of review spam.	*Akoglu & Rayana, 2015; Lai et al, 2010; Lu et al, 2013; Xie et al, 2012; Xu, 2013;*
C2.4	**Verified purchase:** If the majority of a reviewer's reviews is based on verified purchases he/she is less likely to be a spammer.	*Fayazi et al, 2015; Fei et al, 2013; Xie et al, 2012*
C2.5	**Reviewer rating variety:** If a reviewer gives only favorable ratings or only unfavorable ratings this may be an indicator of review spam.	*Jindal & Liu, 2008; Liang et al, 2014; Lim et al, 2010; Xu, 2013;*
C2.6	**Reviewer active duration:** If a reviewer has not long been a member of a site this may be an indicator of review spam.	*Akoglu & Rayana, 2015; Wang et al, 2015*
C2.7	**Real name:** If the name of a reviewer is verified by Amazon he/she is less likely to be a spammer.	*Fayazi et al, 2015*
C2.8	If a reviewer writes **multiple reviews on one product** this may be an indicator of review spam.	*Lim et al, 2010*
	3. Relationship Centric Criteria/Criteria	
C3.1	**Group content similarity**: See C1.1.	*Mukherjee et al, 2012; Xu, 2013*
C3.2	**Group deviation**: See C2.1	*Mukherjee et al, 2012; Xu, 2013*
C3.3	**Group member content similarity**: See C1.1.	*Mukherjee et al, 2012; Xu, 2013*
C3.4	**Group early time frame:** See C4.1.	*Mukherjee et al, 2012; Xu, 2013*
C3.5	**Group time window:** See C4.2.	*Mukherjee et al, 2012; Xu, 2013*
C3.6	**Group size ratio**: The ratio of group size to the total number of reviewers for a product can indicate spamming.	*Mukherjee et al, 2012; Xu, 2013*
C3.7	**Group size**: Large groups are less likely to be together by chance. Large group size may be an indicator of review spam.	*Mukherjee et al, 2012; Xu, 2013*

C3.8	**Group support count** measures the number of products towards which a group has worked together and may be an indicator of review spam.	*Mukherjee et al, 2012; Xu, 2013*
C3.9	**Individual member coupling in a group** measures how closely a member works with the other members of the group and may be an indicator of review spam.	*Mukherjee et al, 2012*
	4. Time Centric Criteria/Criteria	
C4.1	**Early time frame:** If a reviewer writes reviews shortly after the reviewed products have been launched this may be an indicator of review spam.	*Akoglu & Rayana, 2015; Jindal & Liu, 2008; Liang et al, 2014; Lim et al, 2010; Lu et al, 2013; Mukherjee et al, 2012; Wang et al, 2015;*
C4.2	**Reviewer time window:** If the reviewer writes multiple reviews in a short time period this may be an indicator of review spam.	*Akoglu & Rayana, 2015; Lim et al, 2010; Lin et al, 2014; Lu et al, 2013; Wang et al, 2015;*
C4.3	**Time between arrival and review:** If a reviewer writes a review shortly after arriving on the site this may be an indicator of review spam.	*Xie et al, 2012*
C4.4	**Burstiness of singleton reviews:** A singleton review which has been written in a time interval where the ratio of singleton reviews, the number of total reviews as well as the average rating showed abnormal patterns, is more likely to be a spam review.	*Xie et al, 2012*

First of all, it had to be decided who would perform the task of the manual assessment. Lim et al (2010), Liang at (2014) and Lu et al (2013) employed student evaluators who are familiar with online shopping as their human judges. Mukherjee et al (2012) chose for industry experts from Rediff and ebay. In this research, two business students from the University of Twente with experience in online shopping will act as human judges. One common method is to provide the human judges with signals or decision rules to identify a spam review or a review spammer (Liang et al, 2014; Lu et al, 2013; Mukherjee et al, 2012; Fei et al, 2013). The signals suggested in the literature have been complemented with information from *wikihow.com*[7], some of the criteria in **Table 1**, as well as observations that have been made during the collection of the dataset. **Appendix B** displays the list of signals that the human judges have been provided with during this research. To promote the internal consistency of our approach, it was very important that the signals that will be given to the human judges are aligned with the criteria mentioned in **Table 1**. However, we could only include criteria that are visible for the human judges. For example, the criterion time between arrival and review *(C4.3)* is not visible on the reviewer profile; therefore it had to be excluded from the signal list. Furthermore, the criteria have been reformulated to be easier to understand for the human judges. Even though the list of criteria and signals is very useful when labeling reviews, the human judges should only treat these signals as a help for making the labeling decision. They should get a complete picture of the reviewer and base their decision on a combination of the different signals and their personal, intuitive judgment. For example, a person who always writes first reviews *(S3)* could be a lead user. However, if his/her reviews are additionally very short, display a one-sided opinion and are non-verified purchases, it is very likely that this reviewer is a spammer.

The inter-evaluator agreement between the two judges has been calculated with Cohen's Kappa which also acted as a quality indicator for the proposed manual assessment approach. The table provided in Viera and Garret (2005) has been used to interpret the Kappa values in this research. **Table 2** displays the different ranges of kappa values and their interpretation.

[7]http://www.wikihow.com/Spot-a-Fake-Review-on-Amazon

Table 2 Interpretation of Kappa Values (as cited in Viera & Garrett, 2005)

Kappa	Agreement
<0	Less than change
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement

3.2 Reviewskeptic

During the study in which reviewskeptic was developed, Ott et al (2011) focused their efforts on identifying the first type of review spam, untruthful opinions. The authors argue that for a human reader it is very hard to identify these untruthful opinions because they often sound authentic. The authors developed a ground of truth (gold standard) dataset containing 400 nonspam and 400 spam reviews. The dataset was created by gathering spam reviews via Amazon Mechanical Turk, a crowdsourcing platform where tasks such as the writing of a spam review are posted and those who fulfill them receive a small amount of money upon completion. By integrating psychological criteria as well as computational linguistics, Ott et al (2011) tested three approaches for identifying spam reviews in their gold standard dataset. Firstly, a text categorization approach has been applied to the dataset. N-gram based classifiers have been used to label the reviews as either spam or nonspam. Secondly, assuming that spam reviews reflect psychological effects of lying, a Linguistic Inquiry and Word Count (LIWC) Software has been used to analyze the spam reviews. This software counts and groups the occurrence of keywords into psychological dimensions such as functional aspects of the text (number of words, misspelling, etc.), psychological processes, personal concerns, and spoken category (e.g. filler words). Thirdly, the authors explored the relationships between spam reviews and imaginative writing as well as nonspam reviews and informative writing. Part-of-Speech tags have been used to separate the reviews by genre (imaginative or informative). Machine learning classifiers have been trained on the three approaches and their performance has been assessed.

The authors found that there is indeed a relationship between spam reviews and imaginative writing. However, this approach

of identifying spam reviews is outperformed by a combination of the psycholinguistic approach and n-gram based criteria. The main observations were that nonspam reviews tend to include more sensorial and concrete language than spam reviews. Additionally, nonspam reviews are more specific about spatial configurations. This means that in the nonspam reviews the reviewer was more likely to comment, for example, on the size of the room or the bathroom. Furthermore, Ott et al (2011) observed that the focus of the spam reviews was more on aspects external to the object being reviewed. Based on this, the tool reviewskeptic has been developed.

Reviewskeptic is a machine learning classifier trained with a hotel review dataset. Its performance on the identification of hotel reviews is of course higher than with non-hotel reviews. The problem with non-hotel reviews is that many different product categories exist and reviewskeptic would need to be trained on these categories to be able to detect spam reviews reliably. However, some of the lexical criteria used by reviewskeptic (e.g. focus on details external to the product being reviewed *(C1.7)* and excessive use of superlatives and sentiment words *(C1.5; C1.6)*) are not product related and can therefore be generalized over different product categories. This led us to the assumption that even though being trained on hotel reviews, reviewskeptic must be able to detect at least some of the non-hotel spam reviews as well. We therefore researched in how far reviewskeptic is able to detect non-hotel spam reviews and in how far the tool can be complemented and improved with criteria that are not related to the review text. The advantage of doing this is that reviewskeptic as a well-researched tool already exists and therefore, no entirely new tool needs to be developed.

3.3 ReviewAlarm

It has been tested in how far three of the criteria of **Table 1** complemented and enhanced the judgment of reviewskeptic. In order to test the different criteria, a classifier has been built in WEKA and 10 fold cross validation has been used to compare the performances of the different criteria. WEKA provides open source machine learning algorithms for data mining ("Weka 3 - Data Mining with Open Source Machine Learning Software in Java", 2016).

First of all, the criterion *Rating Deviation (C2.1)* has been tested. This criterion is used in many approaches in different ways (Akoglu & Rayana, 2015; Fei et al, 2013; Jindal & Liu, 2008; Liang et al, 2014; Lim et al, 2010; Lu et al, 2013; Mukherjee et al, 2012; Savage et al, 2015). For our dataset we have calculated the rating deviation that each unique review has with the majority opinion on the respective product. The rating deviation, dev_n, for each review has been calculated as follows; where *n* is the review number, $av_p$is the average rating for the respective product and r_n is the review rating.

$$dev_n = av_p - r_n$$

The second criterion that has been assessed is the *Amazon Verified Purchase label (C2.4).* Fayazi et al (2015), Fei et al (2013) and Xie et al (2012) have used the AVP as an indicator to calculate 'spamicity' scores and rank suspicious reviewers. Again, we did not look at the ratio of AVPs for each reviewer but assessed the predictive strength of this criterion in relation to the individual reviews. Each review in the dataset is either labeled with *yes* or *no* with regard to being a verified purchase. Amazon (2016) claims 'when a product review is marked "Amazon Verified Purchase," it means that the customer who wrote the review purchased the item at Amazon. Customers can add this label to their review only if we can verify the item being reviewed was purchased at Amazon.com'. However it has been observed that also spam reviews are marked as Amazon Verified Purchases which lead to doubts about this criterion and lead us to research the criterion in further detail.

Furthermore, we will apply a similar approach as Xie et al (2012) and check whether spam reviews in our dataset can be identified by detecting bursts in singleton reviews *(C4.4).* In order to do so, one week time intervals have been constructed and the number of reviews that appear in each interval, the average rating of the reviews in each interval and the ratio of singleton reviews in each interval has been calculated. The average rating (avr_i) has been calculated by adding up all individual ratings in a time interval (r_n) and dividing them by the number of individual ratings (n) in the interval. The ratio of Singleton Reviews (rsr_i) in each interval has been calculated by dividing the number of singleton reviews (sr_n) by the number of non-singleton-reviews (nsr_n).

$$avr_i = (r_1 + \cdots + r_n)/n$$

$$rsr_i = sr_n/nsr_n$$

Bursty time intervals are those in which abnormal patterns can be observed on all three dimensions (average rating, number of reviews and ratio of singleton reviews). We will apply the decision rule to label all singleton reviews which occur in bursty time intervals as spam reviews. By comparing the labels with our labeled dataset we can assess the effectiveness of this approach.

4. RESULTS

In the following, the results of the experiments will be introduced along with some observations that have been made while the data has been collected. These observations could give valuable insights into reviewer behavior and guide future research directions.

4.1 Observations during Data Collection

Our approach is very different in that the data has been collected manually by copy pasting it from the Amazon website. Until now, researchers have only used data that has been collected automatically. We believe that the manual collection of our data has added a lot of value to our research as we were able to make observations which other researchers would have missed because they did not read every single review which they analyzed. Our observations open up some questions that should be handled in further research to get a better understanding of reviewer behavior and spammers.

First of all, the literature does not give any advice on how to treat reviewers who write reviews in exchange for free product samples or discounts. Lately this has become very popular on Amazon. Our observation showed that these people give positive reviews very often. Research is needed to get a general picture and maybe use this behavior as an indicator of spam in the future (Gani, 2015).

Furthermore, we observed that the reviewer burstiness *(C4.2)* can be a dangerous criterion to use. During the data collection it became apparent that also nonspammers have a tendency to write many of their reviews within a short period of time. It can be seen that a lot of people have several bursts in the review behavior. However, we argue that this criterion is safe to use when all of a reviewer's reviews have been written on the same date.

Another observation that has been made is that some of the review text related criteria may lead to false results as well. Especially for the criterion length of the review text *(C1.2)* it can be observed that nonspammers sometimes also prefer to write short review texts. Especially on products which already

have a lot of reviews, we argue that the reviewers write only a short text to add their rating to the others because they know that a potential buyer will not read all of the reviews that have been submitted.

4.2 Results Manual Assessment

The human judges evaluated the reviews independently from each other and came to the results that are displayed in **Table 3**. The inter-evaluator agreement has been calculated with Cohen's Kappa. There is no universal scale to interpret Kappa values, however a scale that is widely used is provided (Table 2: as cited in Viera & Garrett, 2005). A Kappa value of 1 indicates perfect agreement whereas a Kappa value of 0 or less indicates agreement by chance. The Kappa value between the two human judges in this research was **0.485** which means that the agreement among the two judges is moderate and it indicates that our results can be used with confidence.

Based on these results, the ground of truth dataset has been created by eliminating all entries where the human judges disagreed. This has been done to ensure the reliability of the ground of truth data. The ground of truth dataset consists of 110 reviewers of which 56 are spammers and 54 are non-spammers.

Table 3 Results from Manual Assessment

Cohen's Kappa		**Judge 2**		
Judge 1	**0.485**	spam	nonspam	Total
	spam	56	6	**63**
	nonspam	8	54	**62**
	Total	**65**	**60**	125

4.2.1 Characteristics of Reviews.

To validate the experimentation with the criteria that have been suggested in 3.3, the characteristics of spam reviews and nonspam reviews have been analyzed with regard to these criteria. **Table 4** shows the ratio of the non-verified purchases per category, the mean deviation per category as well as the mean ARI (complexity measure) per category.

Table 4 Characteristics of Categories

	Spam	**Nonspam**
AVP Ratio	0.22	0.02
Mean Deviation	0.92	1.29
Mean ARI	5.01	4.32

The analysis shows that the ratio of non-verified purchases is significantly higher for spam reviews which suggested that the criterion *Amazon Verified Purchase (C2.4)* label may indeed be helpful. Furthermore, the outcome of this analysis questions one of the major assumptions made in the literature *(C2.1 Rating Deviation)*. Even though not statistically significant (*p=0.106*), the rating deviation of nonspam reviews in our dataset is greater than the rating deviation of spam reviews. This indicates that the average rating of the analyzed products is dominated by spammers. As we will later show, a classifier can be built with this rating deviation which gives moderate results for our dataset. This phenomenon shows how dangerous it can be to generalize criteria over different products and platforms. To add to the findings in the literature, we also looked at the complexity of the review text by calculating the ARI for each review in the ground of truth dataset. However, no significant difference could be found (*p=0.601*). Therefore, we advise not to use ARI as a criterion in review spam detection. The SPSS output for the statistical calculations can be found in **Appendix C** and **D**.

4.3 Results Reviewskeptic

The labels given by reviewskeptic have been compared to the ground of truth dataset. The metrics that have been used to assess the performance are the accuracy, the kappa statistic, precision & recall, as well as Area Under the Curve (AUC).

The accuracy is equal to the percentage of reviews that have been labeled correctly. The data shows that reviewskeptic labeled **53.6%** of the reviews correctly and **46.4%** incorrectly. The Kappa statistic is **0.08** which indicates an agreement close to chance **(Table 2)**. In the confusion matrix **(Table 5)** it can be observed that the main problem of reviewskeptic is that a lot of spam reviews are identified as nonspam reviews. A high recall rate indicates that reviewskeptic identified most of the relevant results. For nonspam reviews that was the case (Recall=**0.759**), however, for the spam reviews the recall was unacceptably low (Recall=**0.321**). The low general precision score (Precision=**0.550**) indicates that reviewskeptic returned a substantial number of irrelevant results (Powers, 2007). AUC annotates the area under the receiver operating curve. This curve plots the true positives against the false positives. Essentially, the AUC says how well reviewskeptic separated the two classes spam and nonspam and it is the most commonly used metric in evaluating classifiers. The AUC for reviewskeptic is **0.481**, which again indicates a judgment close to a random guess (Fawcett, 2006). The detailed accuracy of reviewskeptic is displayed in **Table 6**.

Table 5 Confusion Matrix Reviewskeptic

Classified as	**Spam**	**Nonspam**
Spam	18	38
Nonspam	13	41

As predicted, the performance of reviewskeptic on non-hotel reviews is indeed lower than on hotel reviews. The reviews which have been identified correctly and incorrectly have been analyzed in further detail to get a picture of the way reviewskeptic is working and to be able to give recommendations on how to improve the tool.

Many of the spam reviews that have been identified correctly have in common that they include overly positive language. This indicates that reviewskeptic is well trained on the extensive use of positive adjectives such as 'perfect' and 'great' *(C1.6)*. Furthermore, it can be observed that the correctly identified spammers often talk about things that are external to the product such as delivery, family and friends, or their prior expectations *(C1.7; C1.8)*.

Example 1: 'I baby my phone a lot and will surely buy screen protector that gives 100 percent protection for my phone. Luckily, I found this great product. This is the best screen protector I had ever bought for my Sony Z3. It looks superb and it is easy to install. I also bought one for my friend. And she loves it very much. Recommend it.'

Table 6 **Detailed Accuracy by Class Reviewskeptic**

	True Positive	False Positive	Precision	Recall
Spam	0.321	0.241	0.581	0.321
Nonspam	0.759	0.679	0.519	0.759
Weighted average	0.536	0.456	0.550	0.536

Nevertheless, the majority of spam reviews in the dataset has not been detected by reviewskeptic. A major drawback of the tool is that it is trained on specific words and word combinations. However, the tool misses out on very short spam reviews. It can be observed that spam reviews such as 'Good', 'Excellent', 'Worked beautifully' have not been identified as spam reviews even though they clearly meet the criteria of short review text and overly positive language *(C1.2;C1.6)*.

***Example 2**: 'I love this high quality product. couldn't wait to put the screens on my phone and it looks **really great**. it is solid and gives protection to my iphone 6**. strongly recommend** this product.'*

All in all, it can be concluded that reviewskeptic employs some of the basic criteria of review centric spam detection, however the accuracy of non-hotel review spam detection is not significantly better than detection by chance. Therefore, reviewskeptic should be supported by more criteria to be able to detect non-hotel spam reviews more reliably.

4.4 Results ReviewAlarm

To be able to come up with a useful training dataset, the number of reviews in the two classes had to be balanced. Two random spam reviews have been removed from the dataset so that the number of spam reviews is now equal to the number of nonspam reviews (54).

Afterwards, the criterion burstiness of singleton reviews has been applied and the reviews have been labeled according to the related analysis. **Figure 1** shows an example of how the criterion has been operated. All reviews of the three products have been categorized according to one week time intervals. Intervals in which all three dimensions (number of reviews, average rating and ratio of singleton reviews) showed abnormal patterns have been marked and all singleton reviews in these windows have been labeled as spam. The judgment of this procedure has then been added as another attribute to the dataset.

The final dataset includes the following information:

*<Deviation from Average Rating- **DEV**> <Amazon Verified Purchase- **AVP** (Yes/No)> <Reviewskeptic Label- **RS** (Spam/Nonspam)><Burstiness of Singleton Reviews Label- **TS** (Spam/Nonspam)> <Ground of Truth Label (Spam/Nonspam)*

The dataset has been uploaded into WEKA and the different classifiers provided in WEKA have been tested on it. The *rules.PART* classifier showed reasonable results with regards to the metrics mentioned before and has therefore been used to build ReviewAlarm.

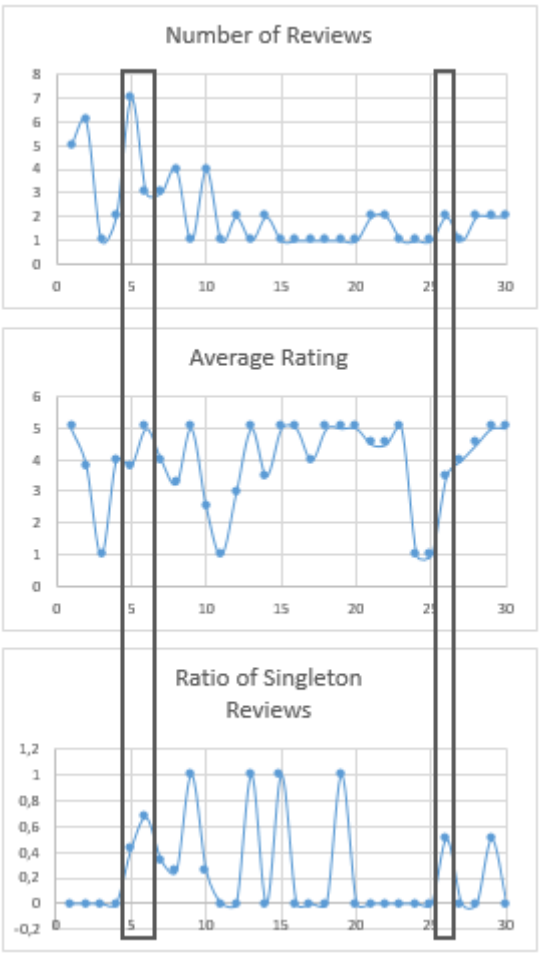

Figure 1 Burstiness of Singleton Reviews on Product 3

PART is a method for creating decision lists based on partial decision trees (Frank & Witten, 2016). The classifier calculates a set of rules that it uses to classify the reviews as spam or nonspam. The rules are based on an automatic analysis of the data which tries to find combinations of attributes which have high predictive strengths in order to label the subjects. PART was especially useful in this case because only the labels assigned by certain methods have been used in the classifier and not the methods themselves.

Table 7 shows how the different criteria that have been analyzed added to the accuracy of reviewskeptic and therefore improved the tool.

Table 7 Detailed Accuracy of ReviewAlarm

Criteria Combination	Identified Correctly	Identified Incorrectly	Kappa	True Positive	False Positive	AUC	Precision	Recall
RS	53.6%	46.4%	0.08	0.536	0.456	0.481	0.550	0.536
RS+DEV	55.56%	44.44%	0.11	0.556	0.444	0.517	0.586	0.513
RS+AVP	54.63%	45.37%	0.09	0.546	0.454	0.497	0.587	0.546
RS+TS	56.48%	43.52%	0.13	0.565	0.435	0.561	0.582	0.565
RS+DEV+AVP	50.93%	49.07%	0.02	0.509	0.491	0.561	0.509	0.509
RS+DEV+TS	**65.74%**	**34.26%**	**0.32**	**0.657**	**0.343**	**0.677**	**0.680**	**0.657**
RS+AVP+TS	62.04%	37.96%	0.24	0.620	0.380	0.581	0.680	0.620
RS+AVP+TS+DEV	**64.82%**	**35.18%**	**0.30**	**0.648**	**0.352**	**0.676**	**0.678**	**0.648**

RS = Reviewskeptic Label, DEV = Deviation from average rating, AVP = Amazon Verified Purchase Label, TS = Time Series Label

Two combinations of criteria were performing better than the others (RS+DEV+TS; RS+AVP+TS+DEV). Both of these classifiers apply logic rules and get decent results. However, we propose the classifier **RS+DEV+TS** which employs the attributes label of reviewskeptic, rating deviation and burstiness of singleton reviews. This is due to the fact that the training dataset should contain the least noise possible. The former analysis showed that a substantial amount of spam reviews is marked as Amazon Verified Purchase. Therefore using the criterion AVP would increase the amount of noise in the dataset. Additionally, the AVP does not increase the performance metrics and therefore it is advisable to choose for the classifier with the highest scores on performance metrics, least noise and the lowest possible number of criteria.

The classifier **RS+DEV+TS** is based on four rules. The rules are visualized as partial decision trees in **Figure 2**.

1. *TS = nonspam AND deviation <= 1.8 AND RS = nonspam:* ***nonspam*** *(58.0/26.0, 55% true positives)*

The **first rule** says to label a review as nonspam if it is labeled as nonspam by the burstiness of singleton reviews criterion, has a rating deviation smaller than or equal to 1.8 and is labeled as nonspam by reviewskeptic. The PART classifier chooses the threshold values based on its analysis of the data and based on the partial decision trees which can be built from the data. At first, rule one does not seem to be aligned with the findings that have been reported beforehand since we initially found the rating deviation to be higher for nonspam reviews. From the raw data you would rather expect a rule which labels reviews as nonspam when they have a rating deviation above a certain threshold value. However, it seems that in 55 % of the cases, the reviews which have been labeled as nonspam by both other criteria and which have a rating deviation smaller than or equal to 1.8 are indeed nonspam reviews. The predictive strengths of this rule is still questionable but it leads to 32 correctly identified instances which is already more than half of the correctly identified instances by reviewskeptic alone.

2. *TS = nonspam AND deviation <= 1.8:* ***spam*** *(24.0/9.0, 62.5% true positives)*

The **second rule** says to label the remaining reviews as spam if they are labeled as nonspam by the burstiness of singleton reviews criterion and the deviation is smaller than or equal to 1.8. The dataset shows that the criterion burstiness of singleton reviews labels a vast amount of the reviews as nonspam. Therefore, a rule was needed to filter out the spam reviews which have been falsely identified as nonspam. It seems that after the first iteration of the data, our first assumption about the deviation holds true again and that the spam reviews in this fraction of the data have a smaller deviation than the nonspam reviews. The second rule has a predictive strength of 62.5% and leads to 15 correctly identified instances.

3. *TS = nonspam:* ***nonspam*** *(14.0/1.0, 93% true positives)*

The **third rule** says to label all remaining reviews which are labeled as nonspam by the burstiness of singleton reviews criterion as nonspam. After the filtering of rule one and two, the predictive strength of the burstiness of singleton reviews criterion gets significantly stronger. The remaining dataset can be labeled based on this criterion only with a false positive rate of 7%.

4. *: spam (12.0, 100% true positives)*

The **fourth rule** suggests to label all remaining reviews as spam. This rule is 100% accurate.

As the results in **Table 7** show, the performance of reviewskeptic has been increased by adding the criteria deviation and burstiness of singleton reviews. The accuracy improved by 12.14 percentage points (accuracy of ReviewAlarm = **65.74%**). Additionally, we were able to increase the Kappa value by 0.24 which shows that the new approach is no longer just a random guess (kappa of ReviewAlarm = **0.32**). The most important measure in assessing classifiers, the AUC, has also increased to **0.677**. The new classifier framework will be called ReviewAlarm.

5. DISCUSSION & LIMITATIONS

In the following, the results of this research will be discussed, the limitations to it will be pointed out and directions for future research will be given.

5.1 Discussion

Our research shows that it is possible to complement a text based tool for review spam detection with reviewer related and time related criteria and enhance the performance of the original tool.

A problem which can be observed is that in the new classifier, reviewskeptic is only part of one rule. The classifier built its rules based on the predictive strengths of the criteria. The fact that reviewskeptic has been used only for one rule as well as its weak stand-alone performance question reviewskeptic's

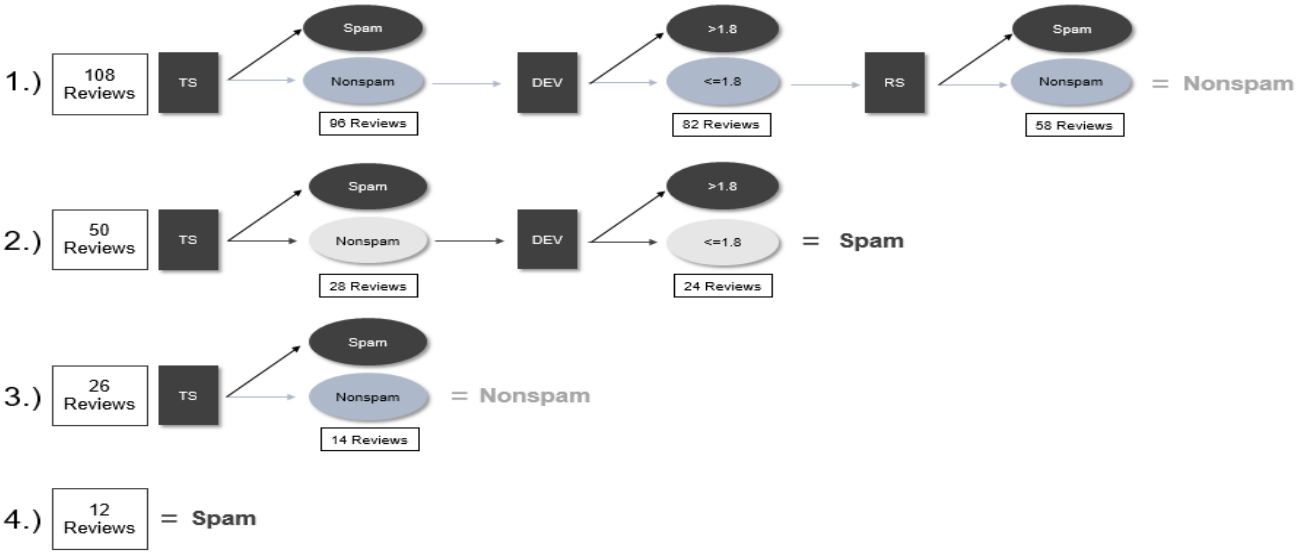

Figure 2 Rules of PART Classifier ReviewAlarm

usefulness for detecting non-hotel review spam in general. In order to check whether reviewskeptic actually disturbs the accuracy of the other criteria, their stand-alone performance without reviewskeptic has been tested. However, all criteria performed worse than in combination with reviewskeptic. Therefore, we can conclude that reviewskeptic adds at least some value to the identification of non-hotel spam reviews.

Although, we were able to increase the performance of reviewskeptic by developing a framework for ReviewAlarm for detecting non-hotel spam reviews, we were not able to come up with an innovative and reliable way to detect review spam. The performance of ReviewAlarm is similar to that of methods developed by other researchers (e.g. Akoglu et al (2013), Xu (2013).

The main challenge in spam review detection is that there is no ultimate truth. We will never know for sure whether the reviews in our ground of truth dataset are really spam or nonspam. Due to this reason, the external validity of the research results is limited. There is a variety of different ways to create spam reviews and as this research shows, generalizing criteria may be dangerous. The ReviewAlarm framework may work well on our dataset, however it may not work on products where other spamming strategies have been applied. One of the major drawbacks of ReviewAlarm is that it is trained with a dataset where the spammers dominate the average rating. Therefore, especially the criterion rating deviation may lead to false results on other datasets. Savage et al (2015) propose an iterative process to correct for this situation. The authors reduce the contribution from each reviewer to the average rating of the product based on the reviewer's proportion of non-majority ratings. It was out of scope for this research to test whether this iterative process can lead to improvements in the labeling and we therefore encourage other researchers to implement the process proposed by Savage et al (2015) when using the ReviewAlarm framework.

The other criterion that has been proposed to be used in combination with reviewskeptic is burstiness of singleton reviews. This criterion is very interesting because it has a perfect precision on detecting spam reviews; it does not produce any false positives, i.e. spam which is actually nonspam. The recall however is rather low. This indicates that it is wise to use the criterion burstiness of singleton reviews in combination with other criteria. The perfect match for the criterion burstiness of singleton reviews would be a criterion which has very high precision on nonspam reviews.

5.2 Limitations & Future Directions

One of the main limitations of this research was the size and nature of the dataset. Because the data was collected manually and the time for this research was restricted to ten weeks, only 108 reviewers have been analyzed in detail. A bigger dataset would have been more beneficial for having more reliable results. However, we think that the manual collection of the data has added a lot of value to our understanding of the issues related to spam review detection. Furthermore, we argue that 108 is still a representative number of reviewers. With regard to the nature of the dataset one should note that the three products that have been analyzed are from the same product category (smartphone accessories). This further limits the external validity as the research results cannot be generalized for different product categories.

Another limitation which may have created noise in the dataset is that only two human judges manually labeled the ground of truth dataset. Even though a Kappa value of 0.485 is moderate, there is still room for improvement. A larger scaled research could employ three or more evaluators to get more reliable results.

As a conclusion from our research, we would like to encourage other researchers to critically look at the criteria that are being used in review spam detection. This research shows that it can lead to increased noise in the dataset to apply certain criteria. It is possible that this is the case for a variety of criteria being used in the literature and therefore, scientific research and validation is needed for each criterion in relation to reviewer behavior. We propose future research to investigate reviewer behavior in a qualitative study. Qualitative research will give valuable insights into the actual behavior of reviewers, spammers and nonspammers. Until now, many assumptions about reviewer behavior have been made based on quantitative studies. However, quantitative studies which do not correspond to an ultimate ground of truth can be biased. To overcome this, qualitative research can be used to validate or invalidate criteria and build better classifiers. We suggest to have interviews with a representative sample of spammers and nonspammers. The pool of interviewees should include both females and males, respondents from different age groups as well as respondents from different locations. Furthermore, it is essential that respondents from the different platforms that are used to obtain spam reviews are interviewed.

Another issue that should be investigated is the relationship between reviews that are written in exchange for product samples or discounts and the sentiment and rating given in the respective review. We argue that these reviews far too often display a positive opinion about a product. The opinions expressed in these reviews seem to be biased and therefore harm other customers by manipulating their purchasing decision.

Furthermore, we propose to repeat our study with a fully automatic classifier, a bigger dataset and more human judges. The fully automatic classifier must be able to apply the criteria of reviewskeptic; Linguistic Inquiry and Word Count (LIWC) as well as N-gram based classifiers. The classifier should be trained with the gold standard dataset from Ott et al (2012) in order to imitate the way reviewskeptic is working. Additionally, the classifier should be able to apply the iterative process to correct for biased rating deviations by Savage et al (2015). Furthermore, the criterion burstiness of singleton reviews should be automated with the suggestions of Xie et al (2012) and be incorporated into the classifier. We suggest to carry out the new research in different ways. Firstly, the training dataset of Ott et al (2012) should be used to train the classifier and then one should train the classifier with a dataset of non-hotel reviews. One can test whether a new training set works better on its own or in combination with the training set from Ott et al (2012). Also our training dataset can be used to train classifiers after correcting for the biased rating deviation.

6. CONCLUSIONS

With regard to the research question, it can be concluded that the performance of reviewskeptic in detecting non-hotel spam reviews has been improved by adding the criteria rating deviation and burstiness of singleton reviews. However, the application of reviewskeptic in its current state to identify non-hotel spam reviews is limited. The classifier ReviewAlarm produces decent yet improvable results. In spam review detection the aim should be to identify the highest number of spam reviews that is possible to be able to provide potential customers with a realistic picture of the product they are informing themselves about. This research provides a framework for a semi supervised method for spam review detection. Until now, the classifier ReviewAlarm only uses the

labels assigned by certain methods as well as the deviation as a numeric attribute. A fully automatic version of the classifier still needs to be built.

Detecting spam reviews is crucial in today's knowledge sharing economy to protect consumers from falsely advertised products and to ensure fair competition among sellers. Further research in the field of review spam detection is needed to improve current tools and develop new ones based on new findings. Furthermore, we argue that it is important to make research on spam review detection not publically available. Otherwise, there is a high risk that spammers will adjust their strategies to the literature available and not be detected by the algorithms.

Based on our findings, we recommend Amazon to investigate their reviewing policies and to improve the Amazon Verified Purchase Label. We highly encourage Amazon to take *Yelp.com* as an example and with the help of researchers from the field to develop an own algorithm to filter out at least the obvious spam reviews. We argues that in the long term, Amazon will profit from adjusting their policies since customers will have more trust in the site and will be more satisfied with the products they purchase.

7. ACKNOWLEDGMENTS

Hereby, I would like to thank Fons Wijnhoven and Chintan Amrit for supervising my bachelor thesis and giving valuable feedback. I would like to thank Olivia Plant and Jarno Bredenoord for the great cooperation within our bachelor circle and for their input. Furthermore, I would like to say thank you to Bouke Willem Bommerson for teaching me how to use WEKA, Cvetanka Koceva for acting as a human judge during the manual assessment and Karolina Vaschenko for proof-reading my thesis.

8. REFERENCE LIST

Akoglu, L., Chandy, R., & Faloutsos, C. (2013). Opinion Fraud Detection in Online Reviews by Network Effects. *ICWSM*, *13*, 2-11.

Amazon Verified Purchase (2016). *Amazon.com*. Retrieved 10 June 2016, from https://www.amazon.com/ gp/ community-help/amazon-verified-purchase

Bishop, T. (2015). *Operator of 'Buy Amazon Reviews' responds to Amazon suit, insists he's doing nothing wrong - GeekWire*. *GeekWire*. Retrieved 9 May 2016, from http://www.geekwire.com/2015/operator-ofbuyamazon-reviews-com-responds-to-amazon-suit-insists-hes-doing-nothing-wrong/

Dixit, S., & Agrawal, A. J. (2013). Survey on review spam detection. *Int J Comput Commun Technol ISSN (PRINT)*, *4*, 0975-7449.

Fayazi, A., Lee, K., Caverlee, J., & Squicciarini, A. (2015, August). Uncovering Crowdsourced Manipulation of Online Reviews. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 233-242). ACM. DOI : 10.1145/2766462.2767742

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874. DOI: 10.1016 /j.patrec.2005.10.010

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting Burstiness in Reviews for Review Spammer Detection. *ICWSM*, *13*, 175-184.

Frank, E., & Witten, I. H. (1998, July). Generating accurate rule sets without global optimization. In *ICML* (Vol. 98, pp. 144-151).

Gani, A. (2015). *Amazon sues 1,000 'fake reviewers'. the Guardian*. Retrieved 9 May 2016, from https://www .theguardian.com/technology/2015/oct/18/amazon-sues-1000-fake-reviewers

Gyöngyi, Z., Garcia-Molina, H., & Pedersen, J. (2004, August). Combating web spam with trustrank. *In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30* (pp. 576-587). VLDB Endowment.

Heydari, A., Tavakoli, M., Salim, N., & Heydari, Z. (2015). Detection of review spam: A survey. *Expert Systems With Applications*, *42*(7), 3634-3642. DOI: 10.1016/j.eswa.2014.12.029

How to Spot a Fake Review on Amazon. (2016). *wikiHow*. Retrieved 29 April 2016, from http://www.Wikihow .com/Spot-a-Fake-Review-on-Amazon

Hu, N., Bose, I., Koh, N. S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, *52*(3), 674-684. DOI: 10.1016/j.dss.2011.11.002

Jindal, N., & Liu, B. (2007, May). Review spam detection. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1189-1190). ACM.

Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (pp. 219-230). ACM.

Kim, S., Chang, H., Lee, S., Yu, M., & Kang, J. (2015, October). Deep Semantic Frame-Based Deceptive Opinion Spam Analysis. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (pp. 1131-1140). ACM. DOI: 10.1145 /2806416.2806551

Lai, C. L., Xu, K. Q., Lau, R. Y., Li, Y., & Jing, L. (2010, November). Toward a language modeling approach for consumer review spam detection. In *e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference on* (pp. 1-8). IEEE. DOI: 10.1109/icebe.2010.47

Liang, D., Liu, X., & Shen, H. (2014, October). Detecting spam reviewers by combing reviewer criterion and relationship. In *Informative and Cybernetics for Computational Social Systems (ICCSS), 2014 International Conference on* (pp. 102-107). IEEE.

Lim, E., Ngyuen, V., Jindal, N., Liu, B., & Lauw, H. (2010). Detecting Product Review Spammers using Rating Behaviors. *19th ACM International Conference on Information and Knowledge Management Toronto - CIKM'10.*

Lin, Y., Zhu, T., Wu, H., Zhang, J., Wang, X., & Zhou, A. (2014, August). Towards online anti-opinion spam: Spotting fake reviews from the review sequence. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on* (pp. 261-264). IEEE.

Lu, Y., Zhang, L., Xiao, Y., & Li, Y. (2013, May). Simultaneously detecting fake reviews and review spammers using factor graph model. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 225-233). ACM.

Mukherjee, A., Liu, B., & Glance, N. (2012, April). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web* (pp. 191-200). ACM. DOI: 10.1145 /2187836.2187863

Olewinski, K. (2016). *Readability Tests and Formulas-Blog | Ideosity*. *Ideosity.com*. Retrieved 15 June 2016, from https://www.ideosity.com/readability-tests-and-formulas/#ARI

Ong, T., Mannino, M., & Gregg, D. (2014). Linguistic characteristics of shill reviews. *Electronic Commerce Research And Applications*, *13*(2), 69-78. DOI: 10.1016/j.elerap.2013.10.002

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, June). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 309-319). Association for Computational Linguistics.

Panda? Penguin? Hummingbird? A Guide To Feared Google's Zoo. (2014). *WooRank Blog*. Retrieved 9 May 2016, from http://blog.woorank.com/2014/08/panda-penguin - a-guide-to-feared-googles-zoo/

Powers, D. (2007). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*. Adelaide: Flinders University of South Australia.

Rayana, S., & Akoglu, L. (2015, August). Collective Opinion Spam Detection: Bridging Review Networks and Metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 985-994). ACM. DOI: 10.1145/ 2783258 .2783370

Savage, D., Zhang, X., Yu, X., Chou, P., & Wang, Q. (2015). Detection of opinion spam based on anomalous rating deviation. *Expert Systems With Applications*, *42*(22), 8650-8657. DOI: 0.1016/j.eswa.2015.07.019

Viera, A. & Garret, J. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Fammed*, *37*(5), 360-363. Retrieved from http://www1.cs.columbia.edu/ ~julia/courses/CS6998/Interrater_agreement.Kappa_statistic.pdf

Wang, Z., Hou, T., Li, Z., & Song, D. (2015). Spotting fake reviewers using product review graph. *Journal Of Computational Information Systems*, *11*(16), 5759-5767.

Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (2016). *Cs.waikato.ac.nz*. Retrieved 10 June 2016, from http:// www.cs .waikato.ac. nz /ml/-weka/

Xie, S., Wang, G., Lin, S., & Yu, P. S. (2012, August). Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 823-831). ACM.

Xu, C. (2013, November). Detecting collusive spammers in online review communities. In *Proceedings of the sixth workshop on Ph. D. students in information and knowledge management* (pp. 33-40). ACM. DOI: 10.1145/2513166.2513176

Yoo, K. & Gretzel, U. (2009). Comparison of Deceptive and Truthful Travel Re-views. In W. Höpken, U. Gretzel & R. Law, *Information and communication technologies in tourism, 2009* (1st ed., pp. 37-47). Springer Vienna. DOI: 10.1007/978-3-211-93971-0_4

9. APPENDICES

A. Literature Overview

	Review Spam Detection Methods				**Manual Assessment**
Publication	Review Centric	Reviewer Centric	Time Centric	Relationship Centric	
Rayana & Akoglu, 2015	**X**				
Akoglu et al, 2013	**X**				
Fayazi et al, 2015				**X**	
Fei et al, 2013			**X**		
Harris, 2012					**X**
Jindal & Liu, 2008	**X**				
Kim et al, 2015	**X**				
Lai et al, 2010	**X**				
Li et al, 2013	**X**				
Liang et al, 2014				**X**	**X**
Lim et al, 2010		**X**			**X**
Lin et al, 2014			**X**		
Lu et al, 2013	**X**				**X**
Mukherjee et al, 2012				**X**	**X**
Ong et al, 2013	**X**				
Ott et al, 2011	**X**				**X**
Savage et al, 2015		**X**			
Wang et al, 2015		**X**			
Xie et al, 2012			**X**		**X**
Xu, 2013				**X**	

B. List of Signals for Human Judges

	Signal	Sources
	When reading the **Reviews** pay close attention to:	
S1 C1.3 C1.9	**Reviews that read like an advertisement:** the brand name is mentioned very often and the reviewer talks mainly about criteria that are given in the product description	*Lu et al, 2013*
S2 C1.6	**Reviews which are full of empty and superlative adjectives:** e.g. the product is amazing, it is the best product.	*Mukherjee et al, 2012; Fei et al, 2013*
S3 C1.5	**Reviews which express a purely one-sided opinion:** e.g. the product is amazing, I like everything about the product.	*Mukherjee et al, 2012; Fei et al, 2013*
S4 C1.2	**Reviews which are either very short or very long:** If the review text is very short it may be that the reviewer just wanted to boost or lower the star rating and thus may be a spammer, if the review text is very long it may be that the reviewer is trying to be more authentic.	("How to Spot a Fake Review on Amazon", 2016)
S5 C1.10	**Positive reviews which are written directly after a negative one, negative reviews which are written directly after a positive one.**	*Jindal & Liu, 2008*
	When visiting **the reviewer's profile** pay close attention to:	
S6 C2.1	**Reviewers who always give an opposite opinion to the majority rating.**	*Liang et al, 2014*
S7 C2.3	**Reviewers who write only one review.**	*Xie et al, 2012*
S9 C2.5	**Reviewers who only give positive ratings and reviewers who only give negative ratings.**	
S10 C1.1 C3.1	**Reviewers who write many duplicate reviews or near duplicate reviews:** Including duplicate reviews from the same reviewer on different products, duplicate reviews from different reviewers on the same product, and duplicate reviews from different reviewers on different products.	*Liang et al, 2014*; *Lu et al, 2013*
S11 C4.1	**Reviewers who write many first reviews.**	*Liang et al, 2014*
S12 C4.2	**Reviewers who made many reviews within a short period of time.**	*Lu et al, 2013; Mukherjee et al, 2012; "How to Spot a Fake Review on Amazon", 2016*
S13 C2.4	**Reviewers who make reviews about non-verified purchases:** Be careful, just because a purchase is verified it does not mean that the reviewer is not a spammer.	*"How to Spot a Fake Review on Amazon", 2016*
S14	**Reviewers who write reviews in exchange for free product samples or discounts:** People who write reviews in exchange for free product samples or discounts are considered as review spammers since their opinion is often biased.	*Own Observation*
S15	**Reviewers who have a wish list which dates back quite long:** Reviewers with a wish list which was updated over the years and includes different products from different categories are less likely to be spammers.	*Own Observation*

C. Independent Samples Mann Whitney U Test for Rating Deviation by Label

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of Deviation is the same across categories of Label.	Independent-Samples Mann-Whitney U Test	,106	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.

D. Independent Samples Mann Whitney U Test for ARI by Label

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of ARI is the same across categories of Label.	Independent-Samples Mann-Whitney U Test	,601	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is ,05.