Fake Review Detection via Exploitation of Spam Indicators and Reviewer Behavior Characteristics

Ioannis Dematis[1](✉), Eirini Karapistoli[2], and Athena Vakali[1]

[1] Informatics Department, Aristotle University of Thessaloniki,
Thessaloniki, Greece
{icdematis,avakali}@csd.auth.gr
[2] CapriTech Limited, 10-12 Mulberry Green,
Old Harlow, Essex CM17 0ET, UK
irene@capritech.co.uk

Abstract. The rapid spread of Internet technologies has redefined E-commerce, since opinion sharing by product reviews is an inseparable part of online purchasing. However, e-commerce openness has attracted malicious behaviors often expressed by fake reviews targeting public opinion manipulation. To address this phenomenon, several approaches have been introduced to detect spam reviews and spammer activity. In this paper, we propose an approach which integrates content and usage information to detect fake product reviews. The proposed model exploits both product reviews and reviewers' behavioral traits interlinked by specific spam indicators. In our proposed method, a fine-grained burst pattern detection is employed to better examine reviews generated over "suspicious" time intervals. Reviewer's past reviewing history is also exploited to determine the reviewer's overall "authorship" reputation as an indicator of their recent reviews' authenticity level. The proposed approach is validated with a real-world Amazon review dataset. Experimentation results show that our method successfully detects spam reviews thanks to the complementary nature of the employed techniques and indicators.

Keywords: Fake review · Reviewer behavior · Spam indicators

1 Introduction

E-commerce has been radically affected by the rapid spread of Web and Internet technologies which enabled tremendous user-generated content (UGC) production and sharing. Consumers publicly and continuously declare and share opinions for purchased products or services and assess quality and value-for-money. A recent study [1] demonstrated that online reviews are quite important to prospective buyers as around 90% of consumers read and incorporate online reviews in their decision-making. Moreover, it has been reported that 88% of consumers trust online reviews as much as personal recommendations.

Such online reviewing impact has opened the floor to "non-honest" activities which aim to either capitalize on or manipulate user reviews for particular products or services.

© Springer International Publishing AG 2018
A M. Tjoa et al. (Eds.): SOFSEM 2018, LNCS 10706, pp. 581–595, 2018.
https://doi.org/10.1007/978-3-319-73117-9_41

It is now evident that professional “spammers” are repeatedly hired to populate the online reviewing space with fake reviews [2, 3] due to competition and/or profit reasons. This large scale of deceptive reviews has emerged as a significant problem attracting the scientific community’s interest. Research efforts mostly aim to improve fake review detection towards re-establishing online opinions validity and credibility.

Fake review detection is a mostly recent research field [4], which initially focused on duplicated review content and review context. Such text analysis was mostly based on machine learning (classifiers) at word or sentence level which targeted the detection of spam reviews by performing supervised learning classification of review content [5–7]. However, the absence of a globally reliable training set of annotated review instances necessary to empower supervised learning approaches led to a shift in research focus [8]. Reviewer behavior was found to hold an abundance of spam indicators including excessive reviewing [9], rating manipulation [10–13], bursty behavior [14–16], etc.

While recent approaches have displayed promising and highly accurate results by featuring a variety of spam indicators, there is a lack of lightweight methods that successfully combine review features and extensive reviewer activity analysis, and work on a fine-grained (product) level, i.e., processing a product’s string of reviews to detect fakes. Additionally, in-depth models are usually hard to adapt and be integrated into functional reviewing sites, while most streamlined and focused approaches result in loss of information.

In this paper, we examine the most important spam indicators relative to review spam and leverage on a reviewer’s behavior characteristics, which are exploited for review labeling in two classes of “honest” or “spam”. Our main goal is to maintain a core part of information associated with online reviewing to feed a generalized methodology, which will be adaptive and computationally effective. The dataset of the proposed work includes commonly available review metadata, which are used to identify bursty review arrival patterns and track reviewer activity. Moreover, past reviewing history is exploited to gain additional indications for a reviewer’s overall reputation, which aids in determining the genuine or deceptive nature of the reviewer’s more recent reviews.

Thus, our main contributions are as follows:

- Proposition of an adaptive fake review detection model which integrates a wide and heterogeneous number of review and reviewer traits.
- Determination of a reviewer’s reputation profile based on reviewer history analysis.
- Inclusion of burst pattern detection not as sole focus but as an additional technique.
- Computational efficiency by implementing a lightweight and non-complex review scoring approach.

The remainder of this paper is organized as follows. Section 2 offers an overview of existing work in the field of review spam detection. In Sect. 3 the problem definition is laid down. Sections 4 and 5 describe the proposed methodology and experimentation results of our study, respectively. Finally, Sect. 6 concludes this paper summarizing our findings.

2 Related Work

Over the last decade, considerable research has been conducted in the field of opinion spam detection of online reviews. The most relevant literature is summarized with emphasis on detecting spam reviewing activity.

Review text analysis. Identification of fake reviews was initially studied as a task of detecting duplicated review text, since content duplication has been recognized as a common spammer practice, in which the same review is reproduced numerous times (semantically or textually). Indeed, the cosine similarity between review contents is often proposed as an effective detection feature [5, 17]. These duplicate and near-duplicate reviews originally served as the positive class for review content classification approaches [4, 5, 17]. The release of the gold-standard dataset [6, 7] of annotated review instances though, procured by employing the Amazon Mechanical Turk (AMT) service, sparked a new interest in supervised learning. Classifiers, built on the aforementioned gold-standard dataset and based on word n-gram [6, 7, 18] or character n-gram features [19], displayed high detection accuracy across both positive and negative sentiment. However, the reliability of the training set of review instances remains a debatable factor in regards to its applicability on real world review cases, as the knowledge and psychology of AMT workers is allegedly not accurately representative of real professional spammers [20].

Graph-based approaches. Certain studies [21, 22] proposed an heterogenous graph representation to model the interconnections between reviewers, reviews and online stores in order to detect irregularities. Using these interconnections, it is possible to iteratively determine the trustworthiness of reviewers, the reliability of stores and the honesty of reviews. FraudEagle [23], an unsupervised network-based framework, consists of a bipartite network of reviewers and products, with edges representing a positive or negative review rating. It initializes the vertices and then iteratively propagates the respective values across the network via the edges until convergence is achieved, which implies consistent scores between neighboring nodes.

Burst pattern discovery. There has been increasing focus on the aspect of time in regards to studying spam reviewing activity. Given that most reviewers create only a single review for a given product, i.e., singleton review, the authors of [14] observed the bursty arrival pattern of singletons, as well as their temporal correlation to rating, and built a multidimensional time series for each product based on average rating, total number of reviews and the ratio of singleton reviews. A joint anomaly detection on these temporally correlated abnormal sections revealed suspicious singleton review activity. Another study [15] asserted that reviews and reviewers, appearing in the same burst of a product's reviewing activity, are often related and thus, using a graph representation to model author interconnections, successfully identified review spammers. The computational costs of analyzing the entire string of a product's reviews led [9] to only analyze and consider those reviews fallen in bursty time intervals on the grounds that they are most likely to contain suspicious activity.

Rating manipulation analysis. Spammers attempt to promote or demote a product by manipulating its overall ranking. As a result, the identification of the proportion of ratings disagreeing with the majority opinion has already been studied as a standalone detection technique or as part of a wider combination of spam indicators and features [11, 12]. A considerable number of early ratings, as well as extreme ratings, have also been linked to suspicious behavior [8, 11]. Furthermore, spammers have been found to distort their distribution of review scores leaving behind a trail of distributional footprints, which can be used to assist in the discovery of spam reviewers [13].

Group spammers detection. Deceptive reviewers often work in collaboration with each other in order to promote or demote a particular product or service. Using frequent pattern mining, [24] found candidate spammer groups and ranked them with SVM RANK based on a number of group related features. The authors of [25] applied a frequent itemset mining method on Amazon review data to extract candidate groups and rank them according to the probability of spamming. A more recent approach [26] used the co-bursting spammer relations to model a co-bursting network, which successfully detected spammer groups.

In short, most methods utilizing a graph-based model [15, 21, 23] and examining various behavior footprints [8, 11, 12] perform an in-depth analysis of reviewing activity, however their (computational) complexity and/or focus on spammer detection does not enable a dedicated product-level approach akin to already established spam review filtering systems. A few approaches [9, 14] did focus on a product-based analysis by taking as input a product's reviews and identifying burst patterns and suspicious reviews, though they suffer from loss of information by ignoring reviews created outside of bursty time intervals. Moreover, they lack an in-depth analysis of reviewer activity and behavior.

In contrast, the proposed method bridges the existing gaps by introducing an effective fake review detection model that operates on a fine-grained (product) level, utilizes burst pattern discovery (detecting suspicious time intervals) as an additional analysis technique and integrates reviewer past and present activity.

3 Problem Definition

Before detailing our approach towards detecting fake reviews, we describe the main concepts of this study and present the issues our method addresses.

To start with, for a given product p we consider a set of n reviews $R = \{r_1, \ldots, r_n\}$ and a set of m reviewers or authors $A = \{a_1, \ldots, a_m\}$ where $m \leq n$, and n, m vary depending on the product. It is apparent that review and reviewer constitute the core entities in our study:

Definition 1 (User Review). *A user review r_i refers to a review written by a user or consumer for a product or a service p based on their experience as a user of the reviewed product. A review usually includes the following information:*

$r_{i,c}$: *A relatively short passage of text or comment expressing the user's experience and judgement of the reviewed product.*

$r_{i,rt}$: *The rating given to the reviewed product with its range typically at the [1, 10] or [1, 5] scales.*

$r_{i,t}$: *The creation date and time of the review.*

$r_{i,a}$: *The author ID of the review.*

Definition 2 (Reviewer). *A reviewer $a(r_i)$ is a person who formally assesses a used product or service p by authoring a review r_i. A reviewer is associated with a set $R_{a,j}$, where $|R_{a,j}| > 0$, defined as the set of all reviews that $a(r_i)$ has written for p.*

Most fake review detection methods focus only on a product's reviews, lacking the deep level (across multiple products) analysis of spammer detection methods. Our goal is to propose a spam review detection approach satisfying the following criterion:

Problem Definition 1 (Fake Review Detection). *Detect spam in online reviews with a model that (1) operates on a product level, (2) exploits all available data relative to reviews and (3) analyzes past and present reviewer activity.*

As we will show in the subsequent section, a hybrid approach combining indicators of spam for review and reviewer, can successfully determine whether the former is fake or honest.

4 Proposed Model

Our approach attempts to create a robust fake review detection system by considering a variety of well-established and accepted by the scientific community spam features linked to both review and reviewer behavior. With regard to the product-level processing, our model receives as input a set of n reviews $R = \{r_1, \ldots, r_n\}$ associated with a product. Then, for each review r_i we extract the necessary information and metadata including review text, review rating, timestamp and reviewer ID, which we first study across some basic spam indicators. We also use burst pattern discovery as a complementary analysis tool to identify bursty time intervals and pinpoint "suspicious" reviews, which we then examine across two additional spam indicators. Thus, our method considers all reviews of a product (no loss of information), while probing further into the most high-risk ones. Lastly, the history of an author's past reviewing activity is taken into account as it can affect their overall reputation as a user and subsequently, as a spam or honest reviewer. During analysis of a review, its author's associated set of past reviews $Hist_{a,j}$ is investigated and studied across a number of features and behavior characteristics as an additional measure of reviewer trustworthiness and ultimately, review spam level. We determine the review spam level by applying a linear weighted scoring function [11] to the review and define a spam score

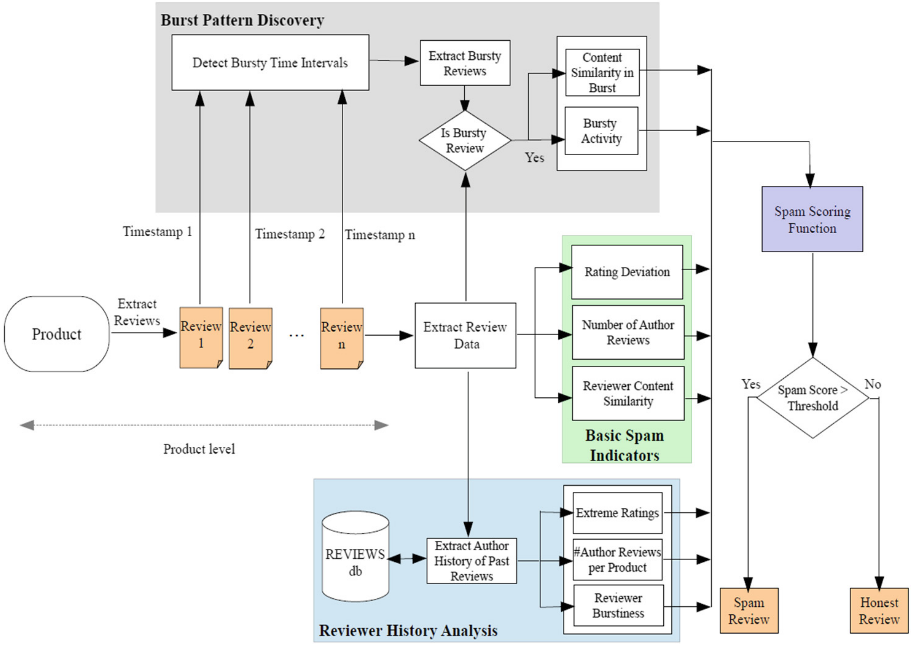

Fig. 1. Overview of the proposed method's workflow.

threshold to which we compare each review's accumulated score. Thus, our method outputs as fake those reviews whose score exceeds the threshold and as honest those reviews whose score does not exceed the threshold. On overview of the proposed method can be seen in Fig. 1.

4.1 Basic Spam Indicators

This section presents and describes the 3 basic spam indicators used in our model to detect spam in online reviews.

- **Rating Deviation (RD)**

A spam review will typically aim at increasing or decreasing a product's overall rank by manipulating its mean score towards a particular direction and, consequently, deviates from the mean.

Considering $S_{mean}(p)$ as the mean rating of a product p and normalizing according to a 5-star rating scale, the rating deviation score [0, 1] of a review r_i is found to be:

$$RD(r_i) = \frac{|r_{i,rt} - S_{mean}(p)|}{4} \tag{1}$$

- **Number of Reviews (NR)**

It is a common spammer practice to create multiple reviews for the same product in order to exert greater influence on public opinion and manipulate the mean rating.

Naturally, the spam score of a review r_i created by reviewer $a(r_i)$ should also be affected by the number of reviews $|R_{a,j}|$ the author has contributed for the same product:

$$NR(a(r_i)) = \left|R_{a,j}\right| \tag{2}$$

- **Content Similarity (CS)**

Spammers often reproduce the same review text as authoring original content would prove time consuming. Therefore, we can detect spammers by considering the overall content similarity of their reviews. In accordance with the existing literature [5, 17], we use the cosine similarity for this purpose.

The content similarity score [0, 1] of a reviewer $a(r_i)$, attributed to review r_i, is the average of the similarities of each review $r_j \in R_{a,j}$:

$$CS(a(r_i)) = Avg\left(\frac{\sum_{z=1}^{|R_{a,j}|} cosine\left(r_j, r_z\right)}{\left|R_{a,j}\right|}\right), j \neq z \tag{3}$$

4.2 Burst Pattern Detection

Spammers typically create a large quantity of reviews in a reasonably short time period in order to quickly negate the effects of and dominate honest opinions. Such excessive posting can lead to the appearance of sudden increases in a product's reviewing activity, creating "bursts" or peaks in certain time intervals. Our model incorporates a burst pattern detection technique, which has already been used successfully in the past [9], as a means of narrowing down the most suspicious time intervals and, subsequently, the most potentially harmful reviews. While the authors of [9] only considered these reviews, missing the rest of a product's reviews, we believe that they should not be the sole focus of a detection model as spam could also exist outside of bursts as well. So, we merely subject these reviews, as well as their respective reviewers, to further analysis with 2 additional spam indicators. Thus, our method investigates all reviews of a product for the existence of spam, analyzing more thoroughly those created in bursty time intervals.

The algorithm for burst pattern discovery is presented below.

Algorithm 1 Algorithm to detect bursty time intervals for a product associated with n reviews $R = \{r_1, \ldots, r_n\}$. Inputs are the corresponding review creation dates $T = \{d_1, \ldots, d_n\}$ and the time window dt, which divides the product's timeline into intervals $\{I_1, \ldots, I_k\}$ of duration dt, where I_j is the number of reviews posted during the j-th interval. dt is set to 7 days [9]. Output is whether I_j is bursty.

```
Input: $T = \{d_1, \ldots, d_n\}, dt$
Output: whether interval $I_j$ is bursty
$len = d_n - d_1$                                   // Measured in days
$k = \#Intervals = \frac{len}{dt}$
$I = \{I_1, \ldots, I_k\}$
$Avg(I_j) = \frac{n}{k}, 1 \le j \le k$             // Average number of reviews per interval
for $j = 1 : k$ do
    if $I_j > Avg(I_j)$ then
        if $j = 1$ & $I_j > I_{j+1}$ then $I_j \leftarrow$ Bursty
        else if $1 < j < k$ & $I_{j-1} < I_j > I_{j+1}$ then $I_j \leftarrow$ Bursty
        else if $j = k$ & $I_j > I_{j-1}$ then $I_j \leftarrow$ Bursty
end for
```

We then extract the reviews fallen in bursty intervals and apply the 2 following spam indicators to them.

- **Content Similarity in Burst (CSBu)**

A high enough similarity score between a review and other reviews of the same "burst" could indicate that a review is suspiciously resembling other reviews.

We thus calculate the cosine similarity between r_i and all other $I_j - 1$ reviews of the same burst:

$$CSBu(r_i) = \begin{cases} \frac{\sum_{z=1}^{I_j} cosine(r_i, r_z)}{I_j - 1} - 0.5, & \frac{\sum_{z=1}^{I_j} cosine(r_i, r_z)}{I_j - 1} > 0.5 \\ 0, \; otherwise \end{cases} \tag{4}$$

Assuming that a similarity score of 0.5 is considered normal, we have modified the CSBu metric so as to only affect those reviews that display higher similarity than normal to not penalize reviews simply for being posted in a bursty time interval.

- **Bursty Activity (BuA)**

A spammer is expected to create large numbers of reviews in small bursts of activity to quickly manipulate the general opinion. We assume that an honest reviewer would create at most 2 bursty reviews, so the bursty activity score for a reviewer $a(r_i)$, and subsequently for his/her reviews, is measured as:

$$BuA(a(r_i)) = \begin{cases} 1, & \textit{bursty reviews} > 2 \\ 0, & \textit{otherwise} \end{cases} \tag{5}$$

4.3 Reviewer Reputation

There is ample available information in regards to author past reviewing activity, which could empower our model to better evaluate a reviewer's overall reputation and, ultimately, the trustworthiness of his/her review(s), via a reviewer-level analysis. This leads us to the following definition:

Definition 3 (Author Reputation). *Author reputation refers to a reviewer's general trustworthiness based on their behavior and activity across their past reviews.*

A reviewer $a(r_i)$ is associated with a set of reviews $Hist_{a,j}$, his/her past reviewing history, across a number of distinct products, which our model exploits by considering 3 addition reviewer history-based spam indicators.

- **Extreme Rating (EXR)**

Most spammers resort to extreme ratings (e.g. 1 or 5 in a 5-star scale) in order to rapidly increase or decrease the mean score of a product.

To this end, the amount of extreme ratings on a 5-star scale among all past ratings $RS_{a,j}$ of an author $a(r_i)$ is collected, and divided by the total number of given ratings $|RS_{a,j}|$ leading to the reviewer's ratio [0, 1] of extreme ratings, which ultimately adds to his/her overall reputation score:

$$EXR(a(r_i)) = \frac{|RS_{a,j} \in \{1,5\}|}{|RS_{a,j}|} \tag{6}$$

- **Number of Reviews per Product (NRP)**

Due to the impact of excessive reviewing, we also consider a reviewer's relevant behavior on past reviewed products. To this end, we measure the average number of reviews a reviewer $a(r_i)$ writes per product by dividing the size of his reviewing history $Hist_{a,j}$ with the number of reviewed products $n_{a,p}$:

$$NRP(a(r_i)) = \frac{|Hist_{a,j}|}{n_{a,p}} \tag{7}$$

- **Reviewer Burstiness (RBu)**

Spammers tend to create all their reviews in great volume and in a short time window (burst) in order to quickly dominate honest reviews. Examining a time window of δ = 30 days [8], the burstiness score of a reviewer $a(r_i)$ is measured like so:

$$RBu(a(r_i)) = \begin{cases} 0, LR(a(r_i)) - FR(a(r_i)) > \delta \\ 1 - \frac{LR(a(r_i)) - FR(a(r_i))}{\delta}, \; otherwise \end{cases} \tag{8}$$

where $LR(a(r_i))$ indicates creation date of the reviewer's last and more recent review, while $FR(a(r_i))$ represents the creation date of the first written review by this reviewer account.

Taking into consideration the above 3 history-based spam indicators, we propose measuring a reviewer's reputation by adding the accumulated indicator scores. Thus, we introduce the following combined method that models trustworthiness or reputation for a reviewer $a(r_i)$. Each generated score is multiplied by a respective weight according to the desired impact of the indicator on the final score:

$$Rep(a(r_i)) = \frac{1}{2}\,EXR(a(r_i)) + \frac{1}{2}\,NRP\,(a(r_i)) + RBu(a(r_i)) \tag{9}$$

A low score is indicative of good reputation, while a high score is implying suspicious behavior.

4.4 Spam Scoring Function

We now introduce our linear weighted scoring function, which combines the individual scores generated by each previously mentioned indicator and outputs an overall spam score for each review. Thus, the spam score of a review r_i, written by a reviewer $a(r_i)$, is measured by the following method:

$$S(r_i) = RD(r_i) + \frac{1}{3}NR(a(r_i)) + 1.5\,CS(a(r_i)) + 2\,CSBu(r_i) + BuA(a(r_i)) + Rep(a(r_i)) \tag{10}$$

The weights of our model's indicator scores are empirically selected based on feature significance as well as value range. Content Similarity in Burst (CSBu) has a value of [0, 0.5] so we give it a weight of 2 to increase its impact, while Extreme Rating (EXR) is considered the weakest indicator, since an honest reviewer could also resort to extreme ratings, and is given a smaller weight. The two spam features (NR, NRP) linked to excessive reviewing are given relatively low weights to counterbalance their potentially high values. Finally, we believe that reviewer Content Similarity (CS) provides strong evidence of spam so we increase its weight accordingly.

Finally, a defined threshold separates the fake reviews from the genuine reviews. After examining the expected score values for honest reviews, as well as for spam reviews, we set the threshold to 3. Thus, reviews with spam scores exceeding the threshold are marked as fake, while reviews with spam scores lower than the threshold are considered genuine.

5 Experimental Analysis

We will now evaluate the effectiveness of the proposed methodology. We conduct experiments on a dataset of real-world reviews and report our findings.

5.1 Dataset

We procured the Amazon review dataset, crawled by [4], to conduct our experiments. The initial dataset is comprised of 5.838.041 reviews of 1.230.915 products created by 2.146.057 reviewers. To facilitate experiments, we sample this dataset to acquire a smaller and easier to evaluate dataset. We exclude from the sampling process those products with less than 5 reviews as lacking attention from users. Our final dataset is comprised of 244.882 reviews, 175.146 reviewers and 13.768 Amazon products.

5.2 Evaluation by Supervised Text Classification

Evaluation has always been a significant barrier in developing highly reliable review spam detection systems. The difficulty stems from the absence of real-world ground truth data of spam reviews necessary for evaluation and model building. A common solution is employing human evaluators and experts to annotate review instances. However, this method includes human subjectivity in the evaluation process.

In this paper, we utilize a different evaluation approach already used successfully in the past [8, 15]. It relies on supervised text classification of the reviews labeled by our method, which are used to represent the positive and negative class, respectively. We iterate over all products in our dataset and score their reviews. Then, all reviews are ranked in descending order, with the top-2000 representing the positive (spam) class and the bottom-2000 representing the negative (honest) class. We choose the top-2000 reviews, as they are heavy spam cases and feature more spam-like text. A Naïve Bayes classifier is then built on these reviews based on UNIGRAM features and the Bag-of-Words model. We perform 10-fold cross validation and report the results. Given the limitation that it is sometimes hard to determine review authenticity by content alone, classification accuracy won't be completely representative of our actual accuracy nor will it allow for a safe comparison to other methods. It will however indicate whether our model is effective and has accurately labeled the evaluation reviews. Accuracy is measured with the established metrics of precision, recall and F-score to ensure consistency with other works in the field.

5.3 Experimentation Results

In order to display the impact of all employed techniques of our model, we first evaluate the effectiveness of the 3 basic review spam indicators. Then, we perform fake review detection with the addition of burst pattern detection. Finally, we include reviewer reputation in the detection process and observe its impact.

For the reviewer reputation scoring phase, we use the entire non-sampled Amazon dataset, which contains ample information regarding reviewer history across a range of distinct products, as our sample dataset may not feature enough information.

Table 1 reports the results of our model's effectiveness after performing 10-fold cross-validation of the classification of our dataset reviews. Surprisingly, the inclusion of burst pattern discovery seems to be lowering accuracy by 1% compared to the results of the basic spam indicators. The difference, however, is small enough to be attributed to the limitations of review text classification so no real conclusion can be made. The addition of reviewer reputation though displayed a considerable improvement in detection accuracy, reporting nearly 75%. Considering again the limitations of our evaluation method, this is a very positive result, which attests to the importance of reviewer reputation in discovering spam reviews. This makes us confident that complementing basic spam indicators and burst pattern discovery with analysis of reviewer past activity allows our model to successfully detect harmful fake reviews.

Table 1. Results of 10-fold cross validation for different combinations of indicators.

Method	Precision	Recall	F-score
Basic	67.6	66.2	65.4
Basic + burst pattern	66.9	65.2	64.3
Basic + burst pattern + reviewer reputation	75.2	75	**74.9**

On top of supervised text classification as an evaluation method, we present a thorough examination of 5 unique review scoring cases. Table 2 displays the respective scores of a sample of 5 reviews of our dataset for all 8 employed spam indicators. The first review has accumulated a very high spam score due to its author's extensive reviewing (NR = 37) on the same product. We also observe that the CS score is quite low, which means that the reviewer created reviews of distinct content to obfuscate their activity. The second, fourth and fifth reviews on the table feature scores close to the defined threshold and are mostly the result of duplicated or near-duplicated content (NR $>$ 1 and CS $\approx$ 1). Three of the reported reviews are also unreasonably similar to other reviews of the same bursty time interval (CSBu $>$ 0), which we discover thanks to our burst pattern technique. Finally, the inclusion of reviewer past history analysis truly shines with the detection of the second sample review, which is a singleton review. Owing mostly to the extremely high NRP score, our method revealed the reviewer's past spamming activity, which in turn weighs down on their recent review.

Table 2. Review scoring examples for 5 spam reviews of the Amazon dataset.

RD	CS	NR	CSBu	BuA	EXR	NRP	RBU	Spam score
0.03	0.3	37	0.0	1	1.0	1.0	0.26	20.69
0.05	1.0	2	0.0	0	0.68	1.01	0.0	3.06
0.15	0.0	1	0.0	0	1.0	57	0.0	29.48
0.07	0.98	3	0.29	1	1.0	1.0	0.0	5.13
0.06	0.99	2	0.49	0	1.0	1.0	0.0	4.22

Overall, our proposed model has displayed positive detection accuracy on the Amazon review dataset. We detected 6.168 fake reviews (2.5% of reviews), that constitute both serious and minor cases of review spamming. In reality, spam percentage is even higher, due to singleton reviews. While we have detected singletons, there are more that can only be captured by specialized techniques [14], which are not our focus. Moreover, we have found that most spam is owed to reviewers reproducing the same (or marginally altered) review twice or thrice, leading to a spam score close to the defined threshold. The most extreme cases of spamming, featuring high spam scores, are those of a reviewer creating multiple reviews for a single product and putting the effort to author dissimilar content in order to avoid detection.

6 Conclusion

In this paper, we propose a new approach for detecting spam reviews. We exploit a variety of different spam indicators on a product level relative to both review and reviewer behavior in order gather and utilize every bit of available information. Moreover, our model features additional analysis features based on burst pattern discovery, which enables the identification of suspicious time intervals and reviews. Finally, we measure reviewer reputation, by examining their history of past reviews and activity, to better determine the authenticity of their more recent reviews. The evaluation of our proposed method was performed on a dataset of Amazon product reviews and the experimentation results showed that our combined method is effective in detecting harmful fake reviews.

As future work, we plan to modify the introduced methodology to better account for singleton spam reviews. While these reviews as individual pieces of content lack the influence on a product's overall rating and popularity, however, in unison they could pose a real threat to unsuspecting review readers and consumers.

Acknowledgments. The authors acknowledge research funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 691025.

References

1. The Impact of Online Reviews on Customers' Buying Decisions [Infographic]. http://www.business2community.com/infographics/impact-online-reviews-customers-buying-decisions-infographic-01280945#k4Q7iGGLamrml8iA.97
2. Ott, M., Cardie, C., Hancock, J.: Estimating the prevalence of deception in online review communities. In: Proceedings of the 21st International Conference on World Wide Web, pp. 201–210. ACM (2012)
3. Wang, Z.: Anonymity, social image, and the competition for volunteers: a case study of the online market for reviews. B.E. J. Econ. Anal. Policy **10**(1), 1–33 (2010)
4. Jindal, N., Liu, B.: Opinion spam and analysis. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 219–230. ACM (2008)
5. Lin, Y., Zhu, T., Wang, X., Zhang, J., Zhou, A.: Towards online anti-opinion spam: spotting fake reviews from the review sequence. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 261–264. IEEE (2014)

6. Ott, M., Choi, Y., Cardie, C., Hancock, J.T.: 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, Portland, Oregon, USA, pp. 309–319 (2011)
7. Ott, M., Cardie, C., Hancock, J.T.: Negative deceptive opinion spam. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, USA, pp. 309–319 (2013)
8. Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R.: Spotting opinion spammers using behavioral footprints. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 632–640. ACM (2013)
9. Heydari, A., Tavakoli, M., Salim, N.: Detection of fake opinions using time series. Expert Syst. Appl. **58**, 83–92 (2016)
10. Jindal, N., Liu, B., Lim, E.-P.: Finding unusual review patterns using unexpected rules. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1549–1552. ACM (2010)
11. Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., Lauw, H.W.: Detecting product review spammers using rating behaviors. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 939–948. ACM (2010)
12. Savage, D., Zhanga, X., Yua, X., Choua, P., Wang, Q.: Detection of opinion spam based on anomalous rating deviation. Expert Syst. Appl. **42**(22), 8650–8657 (2015)
13. Feng, S., Xing, L., Gogar, A., Choi, Y.: Distributional footprints of deceptive product reviews. ICWSM **12**, 98–105 (2012)
14. Xie, S., Wang, G., Lin, S., Yu, P.S.: Review spam detection via temporal pattern discovery. In: Proceedings of the 18th ACM International Conference on Knowledge Discovery and Data Mining, pp. 823–831. ACM (2012)
15. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. ICWSM **13**, 175–184 (2013)
16. Ye, J., Kumar, S., Akoglu, L.: Temporal opinion spam detection by multivariate indicative signals. In: ICWSM, pp. 743–746 (2016)
17. Lau, R.Y., Liao, S., Kwok, R.C.W., Xu, K., Xia, Y., Li, Y.: Text mining and probabilistic language modeling for online review spam detecting. ACM Trans. Manag. Inf. Syst. **2**(4), 1–30 (2011)
18. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2, pp. 171–175. Association for Computational Linguistics (2012)
19. Fusilier, D.H., Montes-y-Gómez, M., Rosso, P., Cabrera, R.G.: Detection of opinion spam with character n-grams. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9042, pp. 285–294. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18117-2_21
20. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: Fake review detection: classification and analysis of real and pseudo reviews. Technical report UIC-CS-2013–03, University of Illinois at Chicago (2013)
21. Wang, G., Xie, S., Liu, B., Yu, P.S.: Review graph based online store review spammer detection. In: 2011 IEEE 11th International Conference on Data Mining (ICDM), pp. 1242–1247. IEEE (2011)
22. Fayazbakhsh, S., Sinha, J.: Review spam detection: a network-based approach. Final Project Report: CSE 590 (2012)
23. Akoglu, L., Chandy, R., Faloutsos, C.: Opinion fraud detection in online reviews by network effects. ICWSM **13**, 2–11 (2013)

24. Mukherjee, A., Liu, B., Wang, J., Glance, N., Jindal, N.: Detecting group review spam. In: Proceedings of the 20th International Conference Companion on World Wide Web, pp. 93–94. ACM (2011)
25. Mukherjee, A., Liu, B., Glance, N.: Spotting fake reviewer groups in consumer reviews. In: Proceedings of the 21st International Conference on World Wide Web, pp. 191–200. ACM (2012)
26. Li, H., Fei, G., Wang, S., Liu, B., Shao, W., Mukherjee, A., Shao, J.: Bimodal distribution and co-bursting in review spam detection. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1063–1072. International World Wide Web Conferences Steering Committee (2017)