Using information retrieval to evaluate trustworthiness assessment of eshops

Claudio Carpineto, Giovanni Romano, and Davide Lo Re

Fondazione Ugo Bordoni, Rome, Italy
{carpinet, romano, dlore}@fub.it

Abstract. To protect consumers from online counterfeiting, several systems have been recently made available that check whether an ecommerce website is trustworthy or not. In this paper we study how to evaluate and compare trust checkers using an information retrieval methodology to gather suitable data and build a ground truth test collection. The main findings of our experimental evaluation are that the inter-checker agreement was moderate and that the single trust checkers presented relative advantages and disadvantages. Essentially, review-based systems were very precise but largely incomplete, whereas feature-based systems provided assessments for any ecommerce website being submitted but were more prone to errors.

1 Introduction

Online counterfeiting continues to be a thorn in the side of both consumers and enterprises, and there are signs that the problem is worsening despite growing efforts to combat it [7] [6]. One particular but important aspect of this phenomenon is represented by ecommerce websites selling counterfeit goods. These fake websites can attract purchasers and enhance their attempts at deception by using a number of channels of promotion and advertising such as email, social networking, web ads, and search engine optimization techniques. On the other side, better technical countermeasures have begun to appear.

As an anti-counterfeiting aid, some desktop web browsers now remove the full URL from view in the URL bar and display only the domain name, which is usually short and clear for legitimate websites (as opposed to long and messy strings for fake ones). To better protect consumers from being led into a swindle, several research and commercial systems have been recently developed that explicitly assess whether a given website is trustworthy or not. Although these systems employ different paradigms, algorithms, and information sources, they can be roughly grouped in two main categories, namely those based on user reviews and white/black lists (e.g., WOT mywot.com, Trustpilot trustpilot.com, Webutation webutation.net, Scamvoid scamvoid.com), and those making use of website features; e.g., Scamadviser scamadviser.com, [8], [1].

The relative availability of trust checkers raises the question of their evaluation and comparison. To the best of our knowledge, this issue has not been

addressed so far. In this paper we focus on two main research questions: *do current trust checkers agree on each other?*, *which is the best trust checker?*

To answer these questions, we present an information retrieval methodology consisting of three main steps. We first collect data retrieved by major search engines in response to search queries with brand names; then we identify ecommerce websites (whether legitimate or fake) in search results with a suitable classifier and use them for measuring inter-checker agreement; and finally build a ground truth dataset containing manually-labeled legitimate and fake ecommerce websites used for measuring the accuracy of trust checkers.

2 Why assessing trustworthiness of ecommerce websites is a difficult task

Assessing whether an ecommerce website is trustworthy or not may be difficult even for humans. In [1], it is shown that non-experts were inconsistent in discriminating between legitimate and fake ecommerce websites containing discounted offers. To illustrate the difficulty of this task, in Figure 1 we show two websites selling products of two well known luxury brands, namely 'Hugo Boss' (a) and 'Iceberg' (b). At first glance, website (a) has a nice look and feel, offers products with reasonably discounted prices, and shows a well designed navigation system for finding products. It also presents other features that might reassure shoppers that the store is trustworthy. Security seals are in place, a live chat is provided, a Facebook label is displayed. Further, on a more technical side, the brand name (i.e., Hugo Boss) is in the URL's path and not in the domain name, as customary in legitimate ecommerce retailers (except for the official website of the brand). However, on closer inspection, we realized that the security label was appropriated without signing up with its vendor, that the store did not have its own Facebook page, and the live chat did not work. Also, more important, the website did not provide any contact information except for a contact form to be filled in by shoppers. This is a clear sign that the website may be fake. Turning to website (b), we see that it has several nice characteristics to recommend itself. In particular, very detailed contact information are displayed in the footer to increase trust of shoppers, including telephone number and address of physical stores. However, we found that the provided contact information were contained in an image and they turned out to be dummy contacts, on more inquiry; the website did not even offer an email address. This was probably another scam website.

Trust checkers may use people's evaluations and reviews in their assessments, or detect specific mechanisms employed by fake websites such as cloaking [9] and search redirection [4]. A more comprehensive approach consists of training a classifier on a large set of learning features, possibly including the earlier mechanisms. Two recent trust classifiers are [8] and [1], the latter of which makes use of 33 learning features spanning product offer, merchant information, payment methods, website registration data, ecommerce-specific SEO, and relative behavior of website. Unlike users, automatic checkers can take into account a

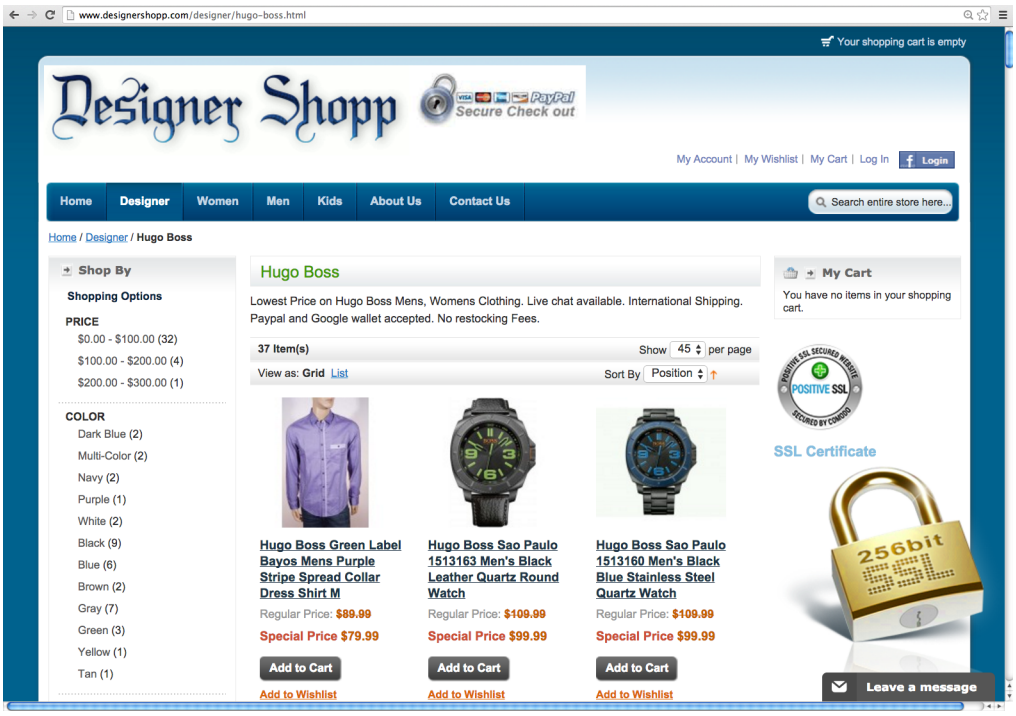
large number of a website's features simultaneously, and they can easily access useful external information that may be not readily available to consumers; e.g., white/black lists, consumer reviews, WHOIS data, Alexa metrics, other checker's assessments, etc. On the other hand, fake websites continue to take steps to increase their similarity to genuine ones, so that sometimes it may be required a deeper understanding of what is behind a feature.

3 Gathering legitimate and fake ecommerce websites

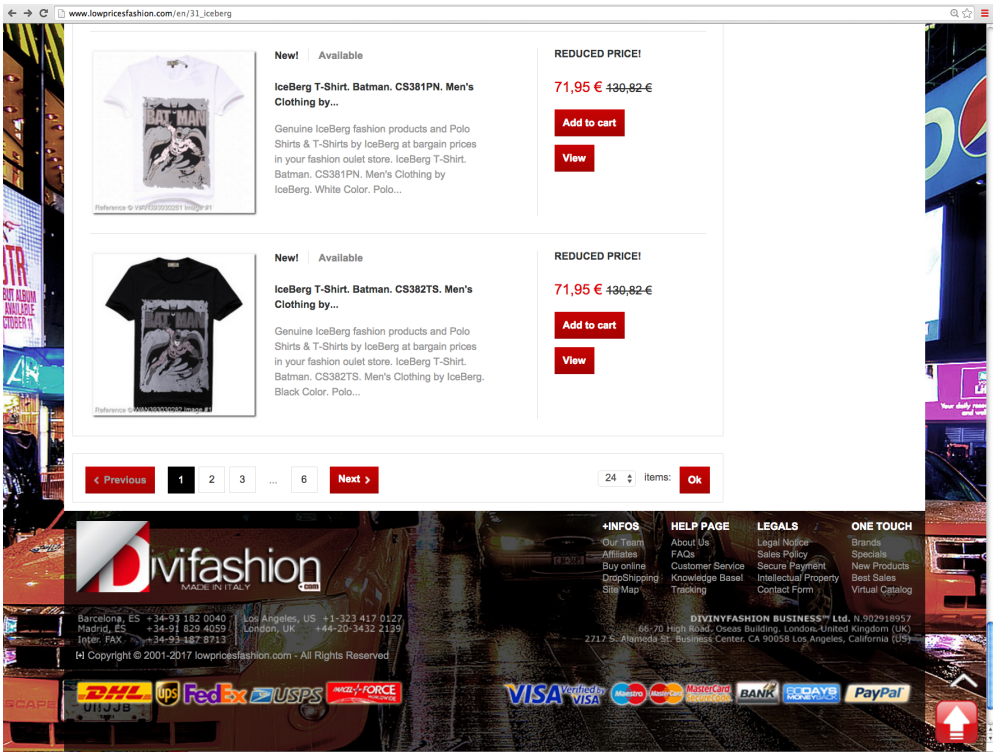
First of all we needed to collect data containing (legitimate and) fake ecommerce websites. One of the major promotion channels used by counterfeiters is manipulation of search engine results. As brand names are extensively searched on the web, usually with a shopping intent [2], online sellers of counterfeits can use search engine optimization techniques to achieve high listings in search results. This way, they may increase traffic on their websites by fraudulently attracting purchasers seeking to buy genuine products or 'complicit' shoppers who are willing to buy replicas. Here we talk about entirely bogus websites, not legitimate sellers that may occasionally sell counterfeit products. Their massive presence in search-engine results has been observed in some recent studies [8] [1].

We collected search results for 39 major shoe brands (e.g., 'Armani', 'Christian Louboutin', 'Gucci', 'Prada', 'Valentino') through the following procedure. We used three types of queries. One neutral query consisting of the brand name followed by 'shoes', to help with ambiguity in brand names and focus on the same product for multi-product brands; one biased query were we added 'cheap' to the brand name and 'shoes', to emphasize that shoppers were seeking discounted offers; one complicit query formed by adding 'replica' to the brand name and 'shoes', to clearly indicate that users were happy with counterfeit products. The queries were automatically submitted to three search engines (i.e., Bing, Google, and Yahoo!) set to the English language, and the first 100 results were saved. In all we collected about 35000 search results. The collected URLs were then post-processed. We performed URL normalization and grouped together the URLs with a same domain name because they usually refer to distinct offers from the same eshop. We kept as a group representative one randomly selected group member. After this operation, the number of URLs reduced to about 24000.

Search results for brand queries usually contain many ecommerce webistes but also many non-ecommerce websites such as shopbots, product catalogues, shop locators, ecommerce blogs, etc. The next step was to identify proper ecommerce websites in search results. This is a research task in its own. We used a specific classifier described in [1] that makes use of 24 classification features about various aspects of the website including product navigation and search, product display, purchase management, and customer service information. In an experimental evaluation, this classifier was able to discriminate between ecommerce and non-ecommerce websites with a classification accuracy of about 90%. By running this classifier on the search results we identified about 10000 ecommerce websites (whether legitimate or fake), with some approximation error.



(a)



(b)

Fig. 1. Screenshots of two presumably fake ecommerce websites, http://www.lowpricesfashion.com/en/31_iceberg (a), and http://www.designershopp.com/designer/hugo-boss.html (b), as of 14 March, 2017.

The next step was to build a ground-truth dataset containing eshops labeled as legitimate or fake. As this is a time-consuming process that requires trained personnel, the size of the ground truth dataset was small. Starting from the large dataset containing eshops described above, we manually labeled randomly selected items until a balanced dataset containing 255 legitimate and fake ecommerce websites was found (by downsampling the class containing legitimate websites).

4 Trust checkers

We used four trust checkers: WOT, Trustpilot, Scamadviser, and RI.SI.CO. [1].[1] WOT and Trustpilot are mainly based on lists of websites and on user reviews. Scamadviser essentially uses WHOIS data, but the details of both the features and the algorithm are not disclosed. RI.SI.CO. is an SVM classifier with a large set of features. The interfaces to the four systems are shown in Figure 2. The three commercial system take an URL as an input and return a weighted assessment of thrustworthiness, whereas the input to RI.SI.CO. is a brand name and its output consists of a list of fake eshops. In Figure 2, we show (top left) the output of RI.SI.CO. for the query 'Christian Louboutin' (i.e., a luxury shoe brand). We also show the outputs of Scamadviser (right top) and WOT (left bottom) for one (highlighted) result of RI.SI.CO. Scamadviser rated it as moderately unsafe, with a score of 52% in a range from 0% to 100% safeness), while for WOT it was clearly untrustworthy; i.e., 9 out of 100. Trustpilot was not able to assess the website retrieved by RI.SI.CO.; we show the output of Trustpilot (right bottom) for Zalando, a legitimate website receiving four stars out of five.

Strictly speaking, the assessment weights returned by WOT, Trustpilot, and Scamadviser are scores, not probabilities. In order to compare the four systems, we had to convert assessment weights into binary classification values. For WOT and Scamadviser we used a simple splitting criterion based on equal width intervals; i.e, threshold = 0.5. While more powerful, supervised methods are conceivable [3], we will see in Section 1 that it yielded results close to the theoretical upper bound of performance. For Trustpilot, we converted any score into the class "legitimate" (the rationale is explained below).

5 Inter-checker agreement

The first experiment was aimed to measure the consistency of rating across different checkers. One practical constraint concerned the time necessary to gather the rates. Because APIs were available only for WOT, it was not possible to run all the checkers through the whole dataset containing about 10000 ecommerce websites. To keep the times manageable, we used a subset containing 632

[1] RI.SI.CO. is available at http://uibm-ici.fub.it/risico with a password-protected access. It was developed for and in cooperation with the Directorate-General for the Fight against Counterfeiting - Italian Patent and Trademark Office.

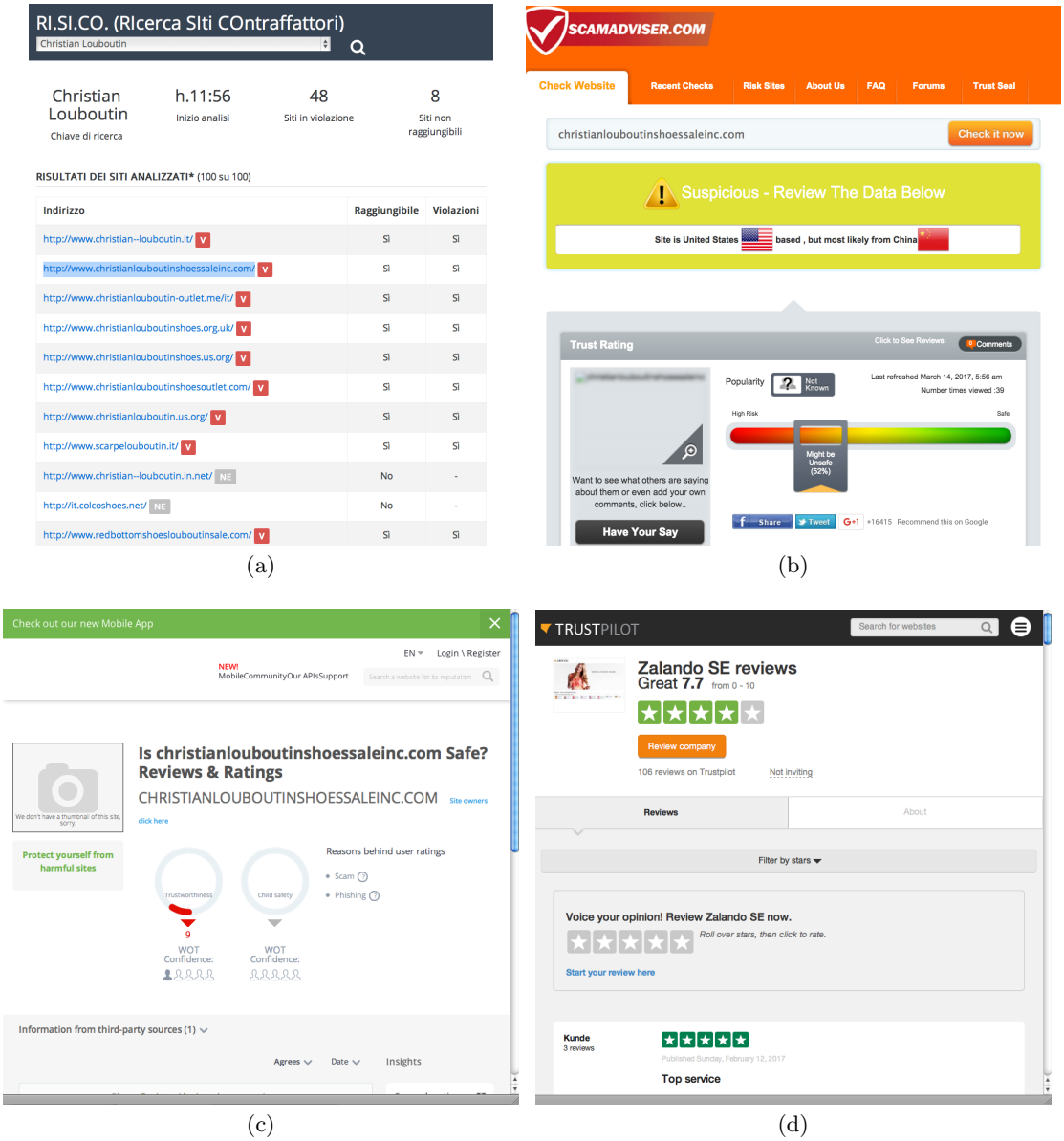


Fig. 2. Screenshots of RI.SI.CO. (a) for 'Christian Louboutin' shoe brand, Scamadviser (b) and WOT (c) for the fake ecommerce website 'www.christianlouboutinshoessaleinc.com', and Trustpilot (d) for the legitimate ecommerce website 'www.zalando.com'.

randomly extracted URLs. We submitted these URLs to the checkers and collected the binarized assessments, used to measure the inter-checker agreement. First of all we noted that Trustpilot and WOT were able to assess only a subset of the URLs, respectively 81 and 351 URLs. We removed Trustpilot from this evaluation and considered the 351 URLs assessed by all three remaining systems. The percent agreement (calculated by averaging the number of agreement scores divided by the total number of scores for each URL) was 88%, which can be seen as a moderate agreement for this simple measure of interrater reliability [5]. More specifically, we found that only for 232 URLs out of 351 the three checkers returned the same class label, which means that in more than one third of cases they were not able to make a unanimous decision. An analysis of pairwise consistency showed that RI.SI.CO. and WOT agreed 323 times, while the agreement between RI.SI.CO. and Scamadviser was equal to that between WOT and Scamadviser and was much lower; i.e., 246. From these findings, it is clear that Scamadviser made more unique decisions. On the whole, this experiment suggests that the predictions made by trust checkers are different, but it does not tell us if they are correct. This question is answered in the next section.

6 Classification accuracy on ground-truth dataset

Using the ground truth dataset, we evaluated the classification accuracy of the four checkers. The results are shown in Table 1. For the systems returning trustworthiness scores, we show the performance values using both a predefined threshold value = 0.5 and the value that maximizes the (a posteriori) global classification accuracy.[2] The main findings of this experiment are that the review-based systems (Trustpilot and WOT) made their assessments on only a subset of the URLs, in proportions similar to those reported above in the inter-checker agreement analysis, but they were more precise than the feature-based systems (Scamadviser and RI.SI.CO.). In particular, Trustpilot (under the extensive interpretation) and WOT achieved an overall accuracy of, respectively, 100% and 93%, versus 77% of Scamadviser and 87% of RI.SI.CO.. Table 1 also suggests that evaluation of trustworthiness was, in general, more difficult for fake than for legitimate eshops. One possible interpretation is that legitimate eshops usually have only legitimacy features in place, whereas fake websites may or may not have illegitimacy features.

Choosing a system-specific threshold value may have a great impact not only on Trustpilot, as already mentioned, but also on Scamadviser, whose global accuracy improved from 77% to 81% with a threshold value equal to 0.75 rather than 0.5. By contrast, it did not affect much the performance of WOT. Its score distribution was such that the classification accuracy with a threshold = 0.5 was nearly the same as that with the optimal threshold value (i.e., 0.52), with as

[2] We noticed that the optimal threshold of Trustpilot is set assuming that whenever the system returns an assessment the website is trustworthy, regardless of the number of stars. In other words, it seems that Trustpilot assesses the convenience of purchase or the quality of service for otherwise legitimate ecommerce websites.

many as 49 fake websites receiving a WOT score = 1 (in a range from 1 to 100). Note also that optimizing the threshold value for global accuracy will lower the recall on legitimate or fake eshops.

Table 1. Performance of WOT, Trustpilot, Scamadviser (with predefined and optimal threshold value) and RI.SI.CO. on ground truth dataset.

	WOT	**Trustpilot**	**Scamadviser**	**RI.SI.CO.**
Number of assessments	55%	14%	100%	100%
ACCURACY OF ACTUAL ASSESSMENTS				
Legitimate and fake eshops	93% (94%)	66% (100%)	77% (81%)	87%
Only legitimate eshops	96% (95%)	100% (100%)	92% (86%)	89%
Only fake eshops	92% (93%)	0% (0%)	64% (77%)	85%

7 Conclusions

Using an evaluation methodology inspired by information retrieval, we found that trustworthiness assessments of eshops made by existing checkers are characterized by moderate inter-consistency and varying classification accuracy. This research suggests that there is much room for performance improvement and that combination of existing methods holds potential for developing better solutions.

References

1. Claudio Carpineto and Giovanni Romano. Learning to detect and measure fake ecommerce websites in search-engine results. Submitted.
2. Jeffrey P. Dotson, Ruixue Rachel Fan, McDonnel Feit Elea, Jeffrey D. Oldham, and Yi-Hsin Yeh. Brand attitudes and search engine queries. *Journal of Interactive Marketing*, 37:105–116, 2017.
3. Elizabeth A. Freeman and Gretchen G. Moison. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, 217:48–58, 2008.
4. N. Leontiadis, T. Moore, and N. Christin. Measuring and analyzing search-redirection attacks in the illicit online prescription drug trade. In *Proceedings of USENIX Security 2011, San Francisco, CA, USA*, 2011.
5. Mary L. McHugh. Interrater reliability. *Biochemia Medica*, 22(3):276–282, 2012.
6. NetNames. The risks of the online counterfeit economy. Technical report, 2016.
7. OECD/EUIPO. *Trade in Counterfeit and Pirated Goods: Mapping the Economic Impact*. OECD Publishing, Paris, 2016.
8. John Wadleigh and Jake Drew andTyler Moore. The e-commerce market for "lemons": Identification and analysis of websites selling counterfeit goods. In *In WWW '15*, pages 1188–1197, 2015.
9. D. Y. Wang, M. Der, M. Karami, L. Saul, D. McCoy, S. Savage, and G. M. Voelker. Search + seizure: The effectiveness of interventions on seo campaigns. In *In IMC'14, Ney York, NY, USA*, pages 359–372. ACM Press, 2014.