RevEx: An Online Consumer Reviews Extraction Tool

Julián Alarte
Universitat Politècnica de València
València, Spain
jualal@upv.edu.es

Carlos Galindo
Universitat Politècnica de València
València, Spain
cargaji@vrain.upv.es

Carlos Martín
Universitat Politècnica de València
València, Spain
cmarabe1@upv.edu.es

Josep Silva
Universitat Politècnica de València
València, Spain
jsilva@dsic.upv.es

Abstract

This paper presents RevEx, an online consumer reviews extraction tool. RevEx extracts the comments section for products in webshops. In contrast to other web scraping tools, it can work with heterogeneous web pages automatically, that is, it does not need any additional information apart from the web page itself. In addition, RevEx is a page-level tool, since it only needs to load the web page whose comments are to be extracted. The technique includes a mechanism to group similar DOM nodes together, and then, an algorithm selects the group of DOM nodes that corresponds to the comments of the web page. The results of the empirical evaluation show an average F1 higher than 88%, and perfect results for around 75% of web pages.

CCS Concepts

• **Information systems → Information extraction**; **Document filtering**; **Presentation of retrieval results**.

Keywords

Information Retrieval, Web Mining, Block Detection, Consumer Reviews Extraction

ACM Reference Format:
Julián Alarte, Carlos Galindo, Carlos Martín, and Josep Silva. 2024. RevEx: An Online Consumer Reviews Extraction Tool. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679214

1 Introduction

Block detection is a Web mining discipline that aims to isolate functional blocks from a web page [11], such as the main content, the template, the menu, comments, etc. This article describes a tool called RevEx that is able to automatically extract the consumer reviews of a product in a webshop.

The cost of evaluating products is high [19]. Hence, product reviews provided from other customers can facilitate this process. Heinonen [8] defines the online product review as a way for business managers and customers to connect with other customers. Indeed, there is a large number of factors that can influence a customer's buying decision, e.g., other customers' reviews [12], the quality of the web [4], a product rating [9], etc.

In the literature, there are many sentiment analysis and opinion mining approaches that analyse online consumer reviews through a mining process. However, most of these approaches assume that the consumer reviews are given as an input in plain text and, thus, they do not include a consumer reviews extraction process. Some of them obtain the reviews using scraping techniques (see e.g., [5, 17]), while others use annotated datasets such as [10], which has been later extended [6, 14]. In contrast, RevEx automatically extracts the consumer comments section from heterogeneous web pages without the need to know a priori the web structure, or to preprocess it, which implies a significant improvement in *information acquisition and preprocessing* tasks. To the best of our knowledge, RevEx is the only effective tool capable of extracting consumer reviews from heterogeneous web pages without any previous knowledge about them. It is implemented as a WebExtension and is distributed through the Firefox's Browser Add-ons repository.

To take advantage of RevEx's features, many systems, such as sentiment analysis and opinion mining techniques, can either integrate the tool as a component or use it as a web service. Researchers working in these areas can use RevEx to automatically extract opinions without needing to analyze the web pages beforehand. Moreover, RevEx is not only useful combined with opinion mining techniques, but also for users browsing the Web. For instance, it can facilitate users with functional diversity problems (such as blindness) to isolate the reviews before using accessibility tools to read (or listen to) them.

2 Related work

Most consumer reviews extraction tools are not valid for heterogeneous web pages since they are prepared for particular web pages (e.g., Outscraper[1], or Apify[2]), or the user has to manually indicate the HTML tags or blocks that contain the desired information (e.g.,

This work has been partially supported by the Spanish MCIN/AEI under grant PID2019-104735RB-C41 and by *Generalitat Valenciana* under grant CIPROM/2022/6 (Fasslow). Carlos Galindo was partially supported by the Spanish *Ministerio de Universidades* under grant FPU20/03861.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679214

[1] https://outscraper.com/
[2] https://apify.com/

BrowseAI[3], or the iSocialWeb Product Review Extractor[4]). Two exceptions are ScrapeStorm[5] and Zyte[6]. We compared RevEx with these two commercial tools using 30 product web pages from our benchmark suite (see Section 5.1). RevEx correctly extracted the comments section from 27 web pages, ScrapeStorm from 10 web pages, and Zyte from 3 web pages. Therefore, the efficacy of RevEx was significantly higher than both commercial tools.

3 Using RevEx

This section shows how to install and use RevEx. On the one hand, it is distributed as a ZIP file that can be installed in any Chromium-based or Firefox-based web browser (such as Google Chrome, Microsoft Edge, Opera, Brave, and Mozilla Firefox). On the other hand, it is published by Mozilla as a Firefox add-on. It should be noted that it is exceptionally easy to download, install, and use.

Download. Users can download RevEx from its website[7]. It comes as a single ZIP file that packages the whole add-on. Note that Firefox users can skip this step since RevEx is distributed through the Firefox Browser Add-ons repository.

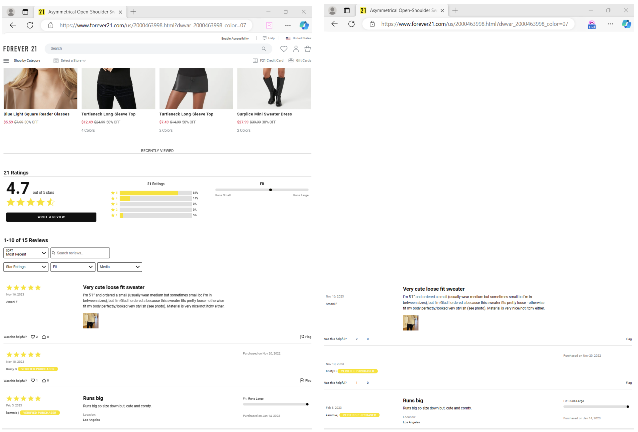
Installation. Mozilla Firefox users only need to search for "review extractor" on the Firefox Add-ons' repository[8] and click the "Add to Firefox" button to install it. Users of Chromium-based web browsers just have to unzip the ZIP file, press the "Load unpacked" button, and finally select the unpacked folder in the Extensions section. Once installed, it automatically adds a new button (the RevEx button: R) to the extensions toolbar.

Usage. Whenever a user wants to extract the comments section of a product's web page, they have to navigate normally to that web page and press the RevEx button. Elements that do not belong to the comments section are hidden, and the only remaining visible element in the web page is the comments section.

Example 3.1. Consider a web page[9] from *forever21*'s website shown in Figure 1(a). This web page contains 15 consumer reviews of the product. By clicking the RevEx button, the tool automatically isolates the comments that are visible on the web page. Figure 1(b) shows the comments extracted by RevEx.

It should be noted that the comments section inferred by RevEx includes images, videos, styles, HTML containers, and all the web components that belong to the original comments section. In consequence, the extracted information contains not only text, but also any kind of information that belongs to user comments.

Once the comments section is extracted, the user can swap between the comments section view and the original web page by pressing the RevEx button again. The main features of RevEx and its functionality on various web pages are illustrated in the following video: https://mist.dsic.upv.es/revex/video-demo/



https://www.forever21.com/us/2000463998.html?dwvar_2000463998_color=07

Blue Light Square Reader Glasses $5.59 $7.99 30% OFF

Turtleneck Long-Sleeve Top $12.49 $24.99 50% OFF 4 Colors

Turtleneck Long-Sleeve Top $7.49 $14.99 50% OFF 2 Colors

Surplice Mini Sweater Dress $27.99 $39.99 30% OFF 2 Colors

21 Ratings

4.7 out of 5 stars

21 Ratings	
★ 5	81%
★ 4	14%
★ 3	0%
★ 2	0%
★ 1	5%

1-10 of 15 Reviews

★★★★★ Nov 16, 2023 Amani F

Very cute loose fit sweater

I'm 5'1" and ordered a small (usually wear medium but sometimes small bc I'm in between sizes), but I'm Glad I ordered a because this sweater fits pretty loose - otherwise fit my body perfectly/looked very stylish (see photo). Material is very nice/not itchy either.

Was this helpful? 2 0

★★★★★ Nov 10, 2023 Kristy G VERIFIED PURCHASER Purchased on Nov 20, 2022

Was this helpful? 1 0

★★★★★ Feb 5, 2023 kammie j VERIFIED PURCHASER

Runs big

Runs big so size down but, cute and comfy.

Location: Los Angeles

Fit: Runs Large

Purchased on Jan 14, 2023

(a) Original product web page

(b) Product web page with the comments section extracted

Figure 1: A product web page from www.forever21.com

4 Internal architecture

4.1 The consumer reviews extraction technique

RevEx implements several novel algorithms in order to extract the comments section. The technique is divided into five stages.

(1) **Assigning weights to some DOM nodes:** RevEx inputs a web page that contains consumer reviews and then an algorithm explores some of the web page's DOM nodes to represent them as points in a four-dimensional Euclidean space ($\mathbb{R}^4$) (in a similar way to [1]). Only those nodes that are located at a depth greater than an empirically computed threshold (explained in Section 5.2) are represented in $\mathbb{R}^4$. To represent a node in $\mathbb{R}^4$, we compute four properties for the node:
   - Its number of children.
   - Its number of descendants.
   - The number of its descendants that are text nodes.
   - Its depth in the DOM tree.

   The combination of these four values forms a point in $\mathbb{R}^4$.

(2) **Grouping the DOM nodes that are equal in $\mathbb{R}^4$:** An algorithm groups the DOM nodes located at the same point in $\mathbb{R}^4$, i.e., the DOM nodes with exactly the same value for the four properties are grouped. The output of this algorithm is a set of groups that contain several DOM nodes. Finally, groups smaller than a given threshold (computed empirically, see Section 5.2) are removed.

(3) **Computing the root of each group:** An algorithm explores the remaining groups of DOM nodes and, for each one, it computes the deepest common ancestor of its DOM nodes.

(4) **Obtaining the root node of the comments section:** For each ancestor computed in the previous stage, an algorithm computes the value of some properties:
   - The number of text words in its descendants that do not belong to a hyperlink.
   - The number of different groups in its descendants.
   - Its depth in the DOM tree.
   - The number of images in its descendants.

[3] https://www.browse.ai/
[4] https://www.isocialweb.agency/en/ai-ecommerce-product-review-extractor/
[5] https://www.scrapestorm.com/
[6] https://www.zyte.com/
[7] RevEx's website: https://mist.dsic.upv.es/revEx/
[8] Direct link: https://addons.mozilla.org/en-US/firefox/addon/review-extractor/
[9] https://www.forever21.com/us/2000463998.html?dwvar_2000463998_color=07

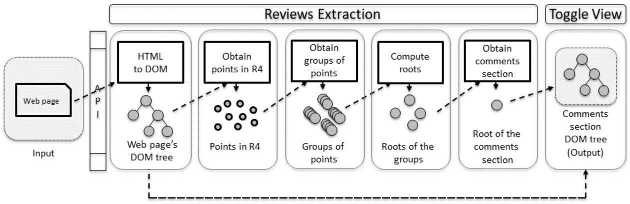


Figure 2: RevEx's process flow diagram

The above properties are combined to assign a weight for each ancestor n of each group, which is computed as:

$$weight(n) = \frac{words(n) \times groups(n) \times depth(n)}{images(n)} \quad (1)$$

The ancestor with a higher weight is considered the root node of the comments section.

(5) **Filtering the comments section:** An algorithm processes the DOM tree of the web page and removes those parts that do not belong to the DOM node computed in the previous stage. This can be done without breaking the visual structure of the original webpage by changing two HTML properties: `node.display = "none";` and `node.style.visibility = "hidden";`. The output of the technique is a well-formed web page that corresponds to the comments section of the original page.

4.2 WebExtension

RevEx is a JavaScript WebExtension. This standard is compatible with most modern web browsers and with the W3C draft community group report[10]. RevEx 1.0 contains 1786 LOC.

A WebExtension consists of a set of files, compressed and packaged in a proper way for its installation and distribution. It contains a file called "manifest.json", which is mandatory and contains basic metadata related to the WebExtension. Moreover, it also contains pointers to the rest of the files included in the WebExtension:

- Background scripts: defines long-running logic.
- Content scripts: contains references to several files of the WebExtension that are loaded into web pages whose URL matches a given pattern.
- Browser action: Defines the interface buttons.

Figure 2 shows the architecture of RevEx. We can observe six modules that implement two functionalities: "Reviews Extraction" (which is executed the first time that the RevEx button is pressed) and "Toggle View" (which is executed on successive button presses). "Reviews Extraction" performs the five-stage process detailed in Section 4.1, and then replaces the current DOM with the computed DOM at stage (5). "Toggle View" swaps the web page displayed (original web page ↔ comments section) by loading the corresponding DOM tree.

4.3 Using RevEx as a library

RevEx can either be used by human users or by automatic systems that need to extract consumer reviews. Consequently, its interface and the output it produces differ in each case:

- **Human users:** RevEx implements a GUI as a WebExtension. Consumer reviews are displayed as a web page and the rest of the elements (those that do not belong to the comments section) are hidden. Therefore, the styles and structure are kept (see the result in Figure 1(b)).
- **Non-human users:** When implemented by other systems, RevEx may produce its output in different ways. For instance, it can output the HTML, marking the root of the comments section with an additional HTML class (i.e., `node.className += "comments_section";`). It can also output the DOM node corresponding to the root of the comments section (the common ancestor computed in stage (4)) without style, or just output the information corresponding to the consumer reviews (text, images, video, etc.).

5 Empirical evaluation

5.1 The benchmark suite

One of the contributions of this work is a new publicly available benchmark suite for consumer review extraction. To the best of our knowledge, there is no publicly available suitable benchmark for this technique because some of them were not prepared for DOM-based techniques or included very few products [10], while others only included reviews from a single website [16]. We also could not adapt a suite of heterogeneous benchmarks prepared for template and content extraction such as TeCo [2] because it includes very few product web pages with consumer reviews.

We have built a dataset of 50 real, heterogeneous product web pages with consumer reviews from 50 different webshops. As a result, all benchmarks have different layouts and page structures. The web pages include different languages (English, Spanish...) to check whether a technique is language-independent. Each benchmark in the suite consists of a single web page and all its resources (CSS, JavaScript, media, images, etc.). Thus, the independence of each benchmark with respect to the evolution (changes in its structure and/or content) of its corresponding web page is guaranteed. For each benchmark in the suite, we have manually marked the consumer comments section by labelling the HTML tag of its root DOM node. We have added to its *class* attribute the className `comments_section`. The labelling inserted in the benchmarks is especially useful for other researchers that want to evaluate or compare their consumer reviews extraction techniques. The suite of benchmarks is publicly available and free[11].

5.2 Experiments

In a first stage, we trained our algorithms to determine the two threshold values (see Section 4.1): (i) the maximum depth of the DOM tree at which DOM nodes can be assigned a weight, and (ii) the minimum size of the groups. We performed experiments to compute the average F_1 for all combinations of both parameters ranging from 0 to 5. The best combination was a value of 2 for the first parameter and a value of 3 for the second. For the training phase, we used 20 randomly selected benchmarks. The remaining 30 benchmarks were used for evaluation. Table 1 shows the results of the empirical evaluation of our technique. The first column (Benchmark) contains

[10] https://browserext.github.io/browserext/

[11] See https://mist.dsic.upv.es/revEx/

Benchmark	DOM nodes							Text (words)						Runtime
	Total	Com.	Ret.	OK	Recall	Precision	F_1	Com.	Ret.	OK	Recall	Precision	F_1	
averyaustin.com	2458	329	329	329	100.00 %	100.00 %	100.00 %	545	545	545	100.00 %	100.00 %	100.00 %	249 ms.
be.green	3153	1703	1678	1678	98.53 %	100.00 %	99.26 %	1398	1393	1393	99.64 %	100.00 %	99.82 %	1310 ms.
bike-discount.de	6674	101	101	101	100.00 %	100.00 %	100.00 %	184	184	184	100.00 %	100.00 %	100.00 %	990 ms.
bonprix.es	3622	162	162	162	100.00 %	100.00 %	100.00 %	154	154	154	100.00 %	100.00 %	100.00 %	238 ms.
comefruta.es	2437	186	186	186	100.00 %	100.00 %	100.00 %	94	94	94	100.00 %	100.00 %	100.00 %	246 ms.
de.myprotein.com	7342	1107	1107	1107	100.00 %	100.00 %	100.00 %	696	696	696	100.00 %	100.00 %	100.00 %	1449 ms.
gourmetfoodstore.com	1683	184	56	0	0.00 %	0.00 %	0.00 %	113	11	0	0.00 %	0.00 %	0.00 %	143 ms.
hsnstore.eu	6229	567	1013	0	0.00 %	0.00 %	0.00 %	245	1270	0	0.00 %	0.00 %	0.00 %	1369 ms.
huntoffice.ie	2351	338	338	338	100.00 %	100.00 %	100.00 %	290	290	290	100.00 %	100.00 %	100.00 %	387 ms.
jeffbanksstores.co.uk	2778	941	919	919	97.66 %	100.00 %	98.82 %	399	398	398	99.75 %	100.00 %	99.87 %	449 ms.
lulus.com	1045	212	565	212	100.00 %	37.52 %	54.57 %	315	630	315	100.00 %	50.00 %	66.67 %	103 ms.
madridhifi.com	5053	244	244	244	100.00 %	100.00 %	100.00 %	99	99	99	100.00 %	100.00 %	100.00 %	683 ms.
majestic.co.uk	3536	1006	1006	1006	100.00 %	100.00 %	100.00 %	291	291	291	100.00 %	100.00 %	100.00 %	646 ms.
matalan.co.uk	1237	374	374	374	100.00 %	100.00 %	100.00 %	122	122	122	100.00 %	100.00 %	100.00 %	112 ms.
mylittlewardrobe.com.au	4559	351	1152	0	0.00 %	0.00 %	0.00 %	172	520	0	0.00 %	0.00 %	0.00 %	833 ms.
next.es	1271	212	212	212	100.00 %	100.00 %	100.00 %	72	72	72	100.00 %	100.00 %	100.00 %	142 ms.
overclockers.co.uk	2617	152	152	152	100.00 %	100.00 %	100.00 %	127	127	127	100.00 %	100.00 %	100.00 %	133 ms.
pharmabuy.es	1163	172	172	172	100.00 %	100.00 %	100.00 %	39	39	39	100.00 %	100.00 %	100.00 %	106 ms.
powerplanetonline.com	9709	240	240	240	100.00 %	100.00 %	100.00 %	608	608	608	100.00 %	100.00 %	100.00 %	936 ms.
puzzlemaster.ca	2076	609	609	609	100.00 %	100.00 %	100.00 %	969	969	969	100.00 %	100.00 %	100.00 %	297 ms.
scorer.es	2037	275	229	229	83.27 %	100.00 %	90.87 %	60	41	41	68.33 %	100.00 %	81.19 %	302 ms.
shopcoffee.co.uk	3820	555	555	555	100.00 %	100.00 %	100.00 %	402	402	402	100.00 %	100.00 %	100.00 %	1582 ms.
sundae-muse.com	3345	1177	1177	1177	100.00 %	100.00 %	100.00 %	528	528	528	100.00 %	100.00 %	100.00 %	614 ms.
tcompanyshop.com	2141	537	537	537	100.00 %	100.00 %	100.00 %	716	716	716	100.00 %	100.00 %	100.00 %	129 ms.
tea-and-coffee.com	1526	468	468	468	100.00 %	100.00 %	100.00 %	117	117	117	100.00 %	100.00 %	100.00 %	175 ms.
thegourmetbox.in	2461	260	260	260	100.00 %	100.00 %	100.00 %	128	128	128	100.00 %	100.00 %	100.00 %	143 ms.
thejewellershop.com	2346	223	220	220	98.65 %	100.00 %	99.32 %	172	172	172	100.00 %	100.00 %	100.00 %	137 ms.
voromotors.com	3367	337	337	337	100.00 %	100.00 %	100.00 %	414	414	414	100.00 %	100.00 %	100.00 %	449 ms.
winechateau.com	3422	441	441	441	100.00 %	100.00 %	100.00 %	312	312	312	100.00 %	100.00 %	100.00 %	416 ms.
woodcraft.com	3699	218	218	218	100.00 %	100.00 %	100.00 %	91	91	91	100.00 %	100.00 %	100.00 %	1040 ms.
Averages	3305	456	502	416	89.27 %	87.92 %	88.09 %	329	381	311	88.92 %	88.33 %	88.25 %	527 ms.

Table 1: Results of the empirical evaluation

the domain name of the web page. For each benchmark, we computed metrics for the DOM nodes and the text retrieved by our technique. Hence, the first block of columns shows the results with respect to the amount of DOM nodes retrieved. Similarly, the second block of columns corresponds to the obtained values with respect to retrieved text words. Total shows the number of nodes of the web page's DOM tree; Com. show the number of real DOM nodes and the number of words of the comments section; Ret. shows the number of nodes/words retrieved by the tool; OK shows the number of DOM nodes/words correctly retrieved by the tool; Recall shows the number of correctly retrieved nodes/words divided by the total number of nodes/words in the comments section; Precision shows the number of correctly retrieved nodes/words divided by the number of retrieved nodes/words; finally, F_1 shows the F_1 metric, computed as $(2PR)/(P+R)$, where P and R stand for the precision and recall (respectively). The last column (Runtime) corresponds to the runtime of the algorithm measured in milliseconds, which was computed as the average of repeating each experiment 10 times. All the experiments were performed in the same hardware with all processes stopped. The first iteration was discarded to avoid the effect of library loads, etc. The standard deviation of the runtime for the 10 repetitions was 5.43 ms.

The experiments reveal an average recall close to 90%, and an average precision and F_1 about 88%. These values are high when compared to other block detection techniques. For instance, RevEx obtains similar values to the best template detection techniques (e.g., the F_1 metric obtained by TemEx is 88.46 % [3]), and to the best content extraction techniques (e.g., ConEx obtains an F_1 value of 84.54 % for retrieved DOM nodes [1]). It should be noted that some of the best block detection techniques are based on text extraction [18, 22, 23]. However, RevEx is a DOM-based technique that extracts the part of the DOM tree that corresponds to the comments section. Hence, it not only extracts the text of the reviews but also the associated multimedia elements, and it is also language-independent.

6 Conclusions

RevEx is a tool that can automatically extract consumer reviews from an heterogeneous web page (whose structure is unknown a priori). This, together with its speed, allows the tool to work online. The technique employed at the core of RevEx is not tied to the type of information it includes (text, images, videos, animations, etc.), but it is instead based on the representation of the DOM nodes as points in a Euclidean space $\mathbb{R}^4$. By comparing the points' positions, the technique is able to infer which of them correspond to consumer reviews. Our experiments reveal an F_1 above 88%, which is an excellent value for a block detection technique.

References

[1] J. Alarte and J. Silva. Page-level main content extraction from heterogeneous webpages. *ACM Trans. Knowl. Discov. Data*, 15(6), jun 2021.

[2] J. Alarte and J. Silva. Hybex: A hybrid tool for template extraction. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 205–209, New York, NY, USA, 2022. Association for Computing Machinery.

[3] J. Alarte, J. Silva, and S. Tamarit. What web template extractor should i use? a benchmarking and comparison for five template extractors. *ACM Trans. Web*, 13(2), mar 2019.

[4] S. Aren, M. Güzel, E. Kabadayı, and L. Alpkan. Factors affecting repurchase intention to shop at the same website. *Procedia - Social and Behavioral Sciences*, 99:536–544, 2013. The Proceedings of 9th International Strategic Management Conference.

[5] L. Chen, L. Qi, and F. Wang. Comparison of feature-level learning methods for mining online consumer reviews. *Expert Systems with Applications*, 39(10):9588–9601, 2012.

[6] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, page 231–240, New York, NY, USA, 2008. Association for Computing Machinery.

[7] T. Gottron. Evaluating content extraction on HTML documents. In *Proceedings of the 2nd International Conference on Internet Technologies and Applications (ITA'07)*, pages 123–132. National Assembly for Wales, sep 2007.

[8] K. Heinonen. Consumer activity in social media: Managerial approaches to consumers' social media behavior. *Journal of Consumer Behaviour*, 10(6):356–364, 2011.

[9] M. Hossin, Y. Mu, J. Fang, and A. N. Kofi Frimpong. Influence of picture presence in reviews on online seller product rating: Moderation role approach. *KSII TIIS*, 13, 12 2019.

[10] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA, 2004. Association for Computing Machinery.

[11] D. Insa, J. Silva, and S. Tamarit. Using the words/leafs ratio in the DOM tree for content extraction. *The Journal of Logic and Algebraic Programming*, 82(8):311–325, 2013.

[12] A. Johan. Product ranking: Measuring product reviews on the purchase decision. *Business & Economic Review*, 4, 06 2021.

[13] J. Leonhardt, A. Anand, and M. Khosla. Boilerplate removal using a neural sequence labeling model. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 226–229, New York, NY, USA, 2020. Association for Computing Machinery.

[14] Q. Liu, Z. Gao, B.-Q. Liu, and Y. Zhang. Automated rule selection for aspect extraction in opinion mining. In *International Joint Conference on Artificial Intelligence*, 2015.

[15] S. S. Modi and S. B. Jagtap. Multimodal web content mining to filter non-learning sites using nlp. In A. Pandian, T. Senjyu, S. M. S. Islam, and H. Wang, editors, *Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCBI - 2018)*, pages 23–30, Cham, 2020. Springer International Publishing.

[16] J. Ni, J. Li, and J. McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197, 2019.

[17] S. Saumya, J. P. Singh, A. M. Baabdullah, N. P. Rana, and Y. K. Dwivedi. Ranking online consumer reviews. *Electronic Commerce Research and Applications*, 29:78–89, 2018.

[18] F. Sun, D. Song, and L. Liao. Dom based content extraction via text density. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 245–254, New York, NY, USA, 2011. ACM.

[19] R. M. Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552, 2018.

[20] N. Utiu and V.-S. Ionescu. Learning web content extraction with dom features. In *2018 IEEE 14th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 5–11, 2018.

[21] T. Vogels, O. Ganea, and C. Eickhoff. Web2text: Deep structured boilerplate removal. *CoRR*, abs/1801.02607, 2018.

[22] T. Vogels, O.-E. Ganea, and C. Eickhoff. Web2text: Deep structured boilerplate removal. In *European Conference on Information Retrieval*, pages 167–179. Springer International Publishing, 2018.

[23] T. Weninger, W. Henry Hsu, and J. Han. CETR: Content Extraction via Tag Ratios. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, editors, *19th International Conference on World Wide Web (WWW'10)*, pages 971–980. ACM, apr 2010.

[24] H. Zhang and J. Wang. Boilerplate detection via semantic classification of textblocks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.