WHAT IS CLAIMED IS:

1. A method for classifying text blocks of a data collection response, comprising:

(a) receiving, via a proxy server, the data collection response that was scraped from a data collection target according to a data collection request wherein the request was received by a scraping agent from a requesting user device and the request was executed through the proxy by the scraping agent against the data collection target;

(b) extracting a plurality of text blocks from the data collection response; for each text blocks in the plurality of text blocks:

(c) obtaining a path describing a location of the respective text block within the data collection response wherein the received data collection response is in HTML or MHTML format and wherein the path is an XPath of the text block within the data collection response;

(d) collecting, within the data collection response via the path, meta attributes describing the text block's display and functional characteristics within a page specified by the data collection response;

e) deriving classification attributes from the text block, the path and the meta attributes, the classification attributes being derived from at least one of a group selected from HTML tags, classes, ids and variables of the text block, textual attributes of the text block, style attributes of the text block, and the path, wherein prior to deriving the classification attributes, irrelevant text blocks were eliminated from the plurality of text blocks by determining, for each respective text block, that a style attribute among the meta attributes collected in (d) indicates that the respective text block is not visible within the page, or that the respective text block contains no text or a single character;

(f) executing a trained machine learning classification model against the classification attributes to determine a classification for the text block;

(g) constructing a dataset with the plurality of text blocks and corresponding classifications determined in (f) for each text block in the plurality of text blocks; and

(h) communicating the dataset to the requesting user device via the scraping agent.

2. The method according to claim 1, further comprising rendering the data collection response in the MHTML format to extract an HTML file, wherein the extracting (b) comprises the extracting the plurality of text blocks from the HTML file.

3. The method according to any one of the preceding claims, in particular according to claim 2., wherein collecting (d) comprises collecting the meta attributes from HTML and non-HTML parts of the data collection response.

4. The method according to any one of the preceding claims, in particular according to claim 1, wherein a datapoint comprises a text block, the associated meta attributes, the corresponding path and the source HTML element within the data collection response.

5. The method of claim 4., wherein further comprising pre-processing of the datapoint to derive classification attributes from the associated meta attributes, and assigning the associated classification attributes to the corresponding text block.

6. The method according to claim 5, wherein the classification attributes identified are registered and assigned to the text block in a numerical format.

7. The method according to any one of the preceding claims, in particular according to claim 1, wherein the meta attributes include style information.

8. The method according to any one of the preceding claims, in particular according to claim 1, further comprising joining separate text blocks from the plurality of text blocks.

9. The method according to any one of the preceding claims, in particular according to claim 1, wherein the executing (f) comprises applying the classification attributes to a plurality of machine learning classification models, each of the plurality of machine learning classification models trained to identify whether the text block belongs to a category.

10. The method of claim 9, wherein the plurality of machine learning classification models each determine a classification probability indicating a likelihood that the text block belongs to the category that the respective machine learning classification model is trained to detect.

11. The method according to any one of the preceding claims, in particular according to claim 1, wherein the trained machine learning classification model employed is at least one of the following, though not limited to: Bag of words, Naive Bayes algorithm, Support vector machines, Logistic Regression, Random Forest classifier, or Extreme Gradient Boosting Model.

12. The method according to any one of the preceding claims, in particular according to claim 1, wherein a classification decision at a classification platform is submitted for quality assurance wherein the classification assigned is examined and confirmed through crowd-sourcing.

13. The method of claim 12 wherein the classification decision subjected to quality assurance is categorized as correct and becomes a part of future machine learning classification model training and is incorporated into the corresponding training set.

14. The method according to any one of the preceding claims, in particular according to claim 1, wherein the data collection response includes non-textual information.

15. The method of claim 14, wherein content delivered within the non-textual information is processed by the trained machine learning classification model.

16. The method according to any one of the preceding claims, in particular according to claim 1, wherein the communicating (h) is executed via a mediating component such as a scraping agent.

17. The method according to any one of the preceding claims, in particular according to claim 1, wherein a data collection response is not submitted for classification if obtaining (b) returns no identifiable text blocks, and the request is re-submitted as a data collection request.