CLAIMS

1. A method for detecting harmful content on a network, the method comprising:

receiving, via the network, a URL;

obtaining, from the URL, an HTML document associated therewith;

converting, the HTML document into a text, the converting comprising extracting a respective text portion from an HTML body of the HTML document;

normalizing the text of the HTML document, the normalizing comprising transforming each word of the text to a respective canonic form thereof, thereby generating a plurality of tokens associated with the text;

aggregating each one of the plurality of tokens into a token vector associated with the HTML document;

applying one or more classifiers to the token vector associated with the HTML document to determine a likelihood parameter indicative of the HTML document containing the harmful content,

the one or more classifiers having been trained to determine the harmful content based on a respective training set of data, the respective training set of data comprising a training token matrix having been generated by:

receiving a plurality of training HTML documents;

converting each one of the plurality of training HTML documents, into a respective training text of a plurality of training texts, the converting comprising extracting the respective text portion from an HTML body of each one of the plurality of training HTML document;

normalizing a given one of the plurality of training texts, the normalizing comprising transforming each word of the given one of the plurality of training texts to a respective canonic form thereof, thereby generating a respective plurality of training tokens associated with the given one of the plurality of training texts,

determining, for each one of the respective plurality of training tokens, a respective significance parameter,

the respective significance parameter associated with a given one of the respective plurality of training tokens being determined based on a product of

multiplication of a local frequency of occurrence and an inverse value of a global frequency of occurrence associated therewith,

the local frequency of occurrence being a frequency of occurrence of the given one of the respective plurality of training tokens within the given one of the plurality training texts associated with a respective one of the plurality of training HTML documents; and

the global frequency of occurrence being a frequency of occurrence of the given one of the respective plurality of training tokens within the plurality of training texts of the plurality of training HTML documents;

aggregating respective pluralities of training tokens of the plurality of training HTML documents associated with respective significance parameters into the training token matrix;

in response to the likelihood parameter being equal to or greater than a predetermined likelihood parameter threshold:

determining that the HTML document located under the URL contains the harmful content; and

storing the URL in a database of harmful URLs.

2. The method of claim 1, further comprising restricting access to the URL.

3. The method of claim 1, wherein the normalizing comprises applying one or more natural language processing techniques.

4. The method of claim 3, wherein the normalizing one of the text and the given one of the plurality of training texts further comprises at least one of:

excluding punctuation from each one of the text and the given one of the plurality of training texts;

transforming all letters to lower case; and

morphological analysis.

5. The method of claim 1, wherein the training token matrix is structured such that:

a given row of the training token matrix is representative of a respective one of the plurality of training HTML documents;

a given column of the training token matrix is representative of a respective training token; and

a given element of the training token matrix is representative of a respective significance parameter of the respective token in the respective one of the plurality of training HTML documents.

6. The method of claim 1, wherein:

converting one of the HTML document and a given one of the plurality of training HTML documents into a respective text further comprises extracting respective text portions of the respective text from one or more predetermined fields of the one of the HTML document and a given one of the plurality of training HTML documents; and

normalizing the respective text includes normalizing each one of the respective text portions in parallel.

7. The method of claim 6, wherein the one or more predetermined fields of the one of the HTML document and a given one of the plurality of training HTML documents include at least one of: 'text', 'title', 'meta', and 'url'.

8. The method of claim 1, wherein the generating the plurality of tokens associated with the HTML document further comprises:

identifying, within the plurality of tokens, a sub-plurality of tokens, wherein each one of the sub-plurality of tokens is associated with at least one of a plurality of predetermined classes of the harmful content; and

aggregating the sub-plurality of tokens into the token vector associated with the HTML document.

9. The method of claim 8, wherein the plurality of predetermined classes of the harmful content includes at last one of: business and industry, career and education, arts and entertainment, science, computers and electronics, shopping, reference, people and society, internet and telecom, sports, finance, adult, autos and vehicles, books and literature, health, games, law and government, travel, news and media, home and garden, food and drink, gambling, recreation and hobbies, beauty and fitness, and pets and animals.

10. A method for detecting harmful content on a network using one or more classifiers, the method comprising training the one or more classifiers based on a respective training set of data by executing the steps of:

receiving a plurality of training HTML documents;

converting each one of the plurality of training HTML documents, to a respective training text of a plurality of training texts, the converting comprising extracting a respective text portion from an HTML body of each one of the plurality of training HTML document;

normalizing a given one of the plurality of training texts, the normalizing comprising transforming each word of the given one of the plurality of training texts to a respective canonic form thereof, thereby generating a respective plurality of training tokens associated with the given one of the plurality of training texts;

determining, for each one of the respective plurality of training tokens, a respective significance parameter,

the respective significance parameter associated with a given one of the respective plurality of training tokens being determined based on a product of multiplication of a local frequency of occurrence and an inverse value of a global frequency of occurrence associated therewith,

the local frequency of occurrence being a frequency of occurrence of the given one of the respective plurality of training tokens within the given one of the plurality training texts associated with a respective one of the plurality of training HTML documents; and

the global frequency of occurrence being a frequency of occurrence of the given one of the respective plurality of training tokens within the plurality of training texts of the plurality of training HTML documents;

aggregating respective pluralities of training tokens of the plurality of training HTML documents into a training token matrix of the respective training set of data, wherein:

a given row of the training token matrix is representative of a respective one of the plurality of training HTML documents;

a given column of the training token matrix is representative of a respective training token; and

a given element of the training token matrix is representative of a respective significance parameter of the respective token in the respective one of the plurality of training HTML documents;

training the one or more classifiers based on the respective training set of data;

using the one or more classifiers to determine a likelihood parameter that a given HTML document contains the harmful content.

11. The method of claim 10, wherein the normalizing further comprises applying one or more natural language processing techniques.

12. The method of claim 10, wherein the determining the likelihood parameter is based on respective decisions generated by the one or more classifiers, and each one of the respective decisions being assigned a respective decision weight.

13. The method of claim 10, wherein the one or more classifiers includes at least one of: a Random Forest classifier, an XGB classifier, a Logistic Regression classifier, and an SVM classifier.

14. The method of claim 10, further comprising identifying, in the respective plurality of training tokens, tokens non-indicative of the harmful content and deleting the tokens non-indicative of the harmful content therefrom.

15. The method of claim 10, wherein the generating the respective plurality of training tokens further comprises:

identifying, within the respective plurality of training tokens, a sub-plurality of training tokens associated with at least one of a plurality of predetermined classes of harmful content; and

aggregating the sub-plurality of training tokens into the training token matrix.

16. The method of claim 15, wherein the plurality of predetermined classes of the harmful content includes at last one of: business and industry, career and education, arts and entertainment, science, computers and electronics, shopping, reference, people and society, internet and telecom, sports, finance, adult, autos and vehicles, books and

literature, health, games, law and government, travel, news and media, home and garden, food and drink, gambling, recreation and hobbies, beauty and fitness, pets and animals.

17. A system for detecting harmful content on a network, the system comprising a computing device, the computing device further comprising:

at least one processor;

at least one non-transitory computer-readable medium comprising instructions, which, when executed by the at least one processor, cause the system to:

receive, via the network, a URL;

obtain, from the URL, an HTML document associated therewith;

convert the HTML document into a text by extracting a respective text portion from an HTML body of the HTML document;

normalize the text associated with the HTML document by transforming each word of the text to a respective canonic form thereof, thereby generating a plurality of tokens associated with the text;

aggregate each one of the plurality of tokens into a token vector associated with the HTML document;

apply one or more classifiers to the token vector associated with the HTML document to determine a likelihood parameter indicative of the HTML document containing the harmful content,

the one or more classifiers having been trained to determine the harmful content based on a respective training set of data, the respective training set of data having been generated by:

receiving a plurality of training HTML documents;

converting each one of the plurality of training HTML documents, into a respective training text of a plurality of training texts, the converting comprising extracting the respective text portion from an HTML body of each one of the plurality of training HTML document;

normalizing a given one of the plurality of training texts, the normalizing comprising transforming each word of the given one of the plurality of training texts to a respective canonic form thereof, thereby generating a respective plurality of training tokens associated with the given one of the plurality of training texts,

determining for each one of the respective plurality of training tokens, a respective significance parameter, a given significance parameter associated with a given one of the plurality of training tokens being determined based on a product of multiplication of:

a local frequency of occurrence of the given one of the plurality of training tokens within the given one of the plurality training texts associated with a respective one of the plurality of training HTML documents; and

a global frequency of occurrence of the given one of the respective plurality of training tokens within the plurality of training texts of the plurality of training HTML documents;

aggregating respective pluralities of training tokens of the plurality of training HTML documents associated with respective significance parameters into a training token matrix of the respective training set of data, wherein:

a given row of the training token matrix being representative of a respective one of the plurality of training HTML documents;

a given column of the training token matrix being representative of a respective training token; and

a given element of the training token matrix being representative of a respective significance parameter of the respective token in the respective one of the plurality of training HTML documents; and

in response to the likelihood parameter being equal to or greater than a predetermined likelihood parameter threshold:

determining, by the processor, that the HTML document located under the URL contains the harmful content; and

store the URL in a database of harmful URLs.