CLAIMS

1. A method for detecting harmful content on a network, the method being executable by a computing device comprising a processor, the method comprising:

receiving, by the processor, via the network, a URL;

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

converting, by the processor, the HTML document into a text;

normalizing, by the processor, the text associated with the HTML document, thereby generating a plurality of tokens associated therewith;

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

applying, by the processor, one or more classifiers to the token vector associated with the HTML document to determine a likelihood parameter indicative of the URL being associated with 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, by the processor, a plurality of training HTML documents;

converting, by the processor, each one of the plurality of training HTML documents, into a respective training text of a plurality of training texts;

normalizing, by the processor, a given one of the plurality of training texts, thereby generating a respective plurality of training tokens associated therewith,

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

aggregating, by the processor, 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:

identifying, by the processor, the URL as being associated with the harmful content; and

storing, by the processor, 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, by the processor, 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;

transforming each word of each one of the text and the given one of the plurality of training texts to a respective canonic form thereof; 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 HTML documents into a respective text 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 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 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, by the processor, 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, by the processor, 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. The method of claim 1, wherein a given significance parameter associated with a given one of the plurality of training tokens is indicative of a combination 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.

11. The method of claim 10, wherein the given significance parameter associated with the given one of the plurality of training tokens is determined based on a product of multiplication of the local frequency of occurrence and an inverse value of the global frequency of occurrence associated

therewith.

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

receiving, by the processor, a plurality of training HTML documents;

converting, by the processor, each one of the plurality of training HTML documents, a respective training text of a plurality of training texts,

normalizing, by the processor, a given one of the plurality of training texts, thereby generating a respective plurality of training tokens associated therewith;

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

aggregating, by the processor, 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, by the processor, the one or more classifiers based on the respective training set of data;

using, by the processor, the one or more classifiers to determine a likelihood parameter that a given HTML document is associated with the harmful content.

13. The method of claim 12, wherein the normalizing comprises applying, by the processor, one or more natural language processing techniques.

14. The method of claim 12, 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.

15. The method of claim 12, 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.

16. The method of claim 12, 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.

17. The method of claim 12, wherein the generating the respective plurality of training tokens further comprises:

identifying, by the processor, 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, by the processor, the sub-plurality of training tokens into the training token matrix.

18. The method of claim 17, 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.

19. The method of claim 12, wherein a given significance parameter associated with a given one of the plurality of training tokens is indicative of a combination 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.

20. The method of claim 19, wherein the given significance parameter associated with the given one of the plurality of training tokens is determined based on a product of multiplication of the local frequency of occurrence and an inverse value of the global frequency of occurrence associated therewith.

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

a processor;

a non-transitory computer-readable medium comprising instructions;

the processor, upon executing the instructions, being configured to:

receive, via the network, a URL;

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

convert the HTML document into a text;

normalize the text associated with the HTML document, thereby generating a plurality of tokens associated therewith;

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 URL being associated with 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;

normalizing a given one of the plurality of training texts, thereby generating a respective plurality of training tokens associated therewith,

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 indicative of a combination 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;

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

identify the URL as being associated with the harmful content; and

store the URL in a database of harmful URLs.

22. The system of claim 21, wherein the processor is configured to determine the given significance parameter associated with the given one of the plurality of training tokens based on a product of multiplication of the local frequency of occurrence and an inverse value of the global frequency

~~of occurrence associated therewith.~~