# CLAIMS

What is claimed is:

1. A client computer system comprising at least one processor configured to determine a text fingerprint of a target electronic document so that a length of the text fingerprint is constrained between a lower bound and an upper bound, wherein the lower and upper bounds are predetermined, and wherein determining the text fingerprint comprises:
selecting a plurality of text tokens of the target electronic document;
in response to selecting the plurality of text tokens, determining a fingerprint fragment size according to the upper and lower bounds, and according to a count of the selected plurality of text tokens;
determining a plurality of fingerprint fragments, each fingerprint fragment of the plurality of fingerprint fragments determined according to a hash of a distinct text token of the selected plurality of text tokens, each fingerprint fragment consisting of a sequence of characters, a length of the sequence chosen to equal the fingerprint fragment size; and
concatenating the plurality of fingerprint fragments to form the text fingerprint.

2. The client computer system of claim 1, wherein the at least one processor is further configured to:
send the text fingerprint to a server computer system; and
receive from the server computer system a target label determined for the target electronic document, the target label indicative of a category of documents that the target electronic document belongs to, wherein determining the target label comprises:
retrieving a reference fingerprint from a database of reference fingerprints, the reference fingerprint determined for a reference electronic document belonging to the category, the reference fingerprint selected according

to a length of the reference fingerprint so that the length of the reference fingerprint is between the upper and lower bounds; and

determining whether the target electronic document belongs to the category according to a result of comparing the text fingerprint to the reference fingerprint.

3. The client computer system of claim 2, wherein the category of documents is a spam category.

4. The client computer system of claim 2, wherein the category of documents is a fraudulent document category.

5. The client computer system of claim 1, wherein determining the text fingerprint further comprises determining each character of the sequence of characters according to a distinct group of bits of the hash of the distinct text token.

6. The client computer system of claim 1, wherein selecting the plurality of text tokens comprises:

selecting a preliminary plurality of text tokens of the target electronic document;

determining a count of the preliminary plurality of text tokens; and

in response, when the count of the preliminary plurality of text tokens exceeds a predetermined threshold, prune the preliminary plurality of text tokens to form the selected plurality of text tokens so that the count of the selected plurality of tokens does not exceed the predetermined threshold.

7. The client computer system of claim 6, wherein pruning the preliminary plurality of text tokens comprises selecting a target text token of the preliminary plurality of text tokens into the selected plurality of text tokens according to a hash of the target text token.

8. The client computer system of claim 6, wherein pruning the preliminary plurality of text tokens further comprises:

determining whether the hash of the target text token is divisible by a zoom-out factor; and

in response, when the target text token is divisible by the zoom-out factor, selecting the target text token into the selected plurality of text tokens.

9. The client computer system of claim 1, wherein selecting the plurality of text tokens comprises:

selecting a preliminary plurality of text tokens of the target electronic document;

determining a count of the preliminary plurality of text tokens; and

in response, when the count of the preliminary plurality of text tokens exceeds a predetermined threshold,

determine a plurality of aggregate text tokens, each aggregate text token of the plurality of aggregate text tokens comprising a concatenation of a set of text tokens of the preliminary plurality of text tokens; and

selecting an aggregate token of the plurality of aggregate tokens into the selected plurality of text tokens according to a hash of the aggregate text token.

10. The client computer system of claim 1, wherein the target electronic document is selected from a group consisting of an email message and a hypertext markup language (HTML) document.

11. The client computer system of claim 1, wherein the distinct text token comprises an item selected from a group consisting of a word, an email address, and a uniform resource locator (URL) of the target electronic communication.

12. A server computer system comprising at least one processor configured to perform transactions with a plurality of client systems, wherein a transaction comprises:

receiving a text fingerprint from a client system of the plurality of client systems, the text fingerprint determined for a target electronic document so that a length of the text fingerprint is constrained between a lower bound and an upper bound, wherein the lower and upper bounds are predetermined; and

sending to the client system a target label indicative of a category of documents that the target electronic document belongs to,

wherein determining the text fingerprint comprises:

selecting a plurality of text tokens of the target electronic document;

in response to selecting the plurality of text tokens, determining a fingerprint fragment size according to the upper and lower bounds, and according to a count of the selected plurality of text tokens;

determining a plurality of fingerprint fragments, each fingerprint fragment of the plurality of fingerprint fragments determined according to a hash of a distinct text token of the selected plurality of text tokens, each fingerprint fragment consisting of a sequence of characters, a length of the sequence chosen to equal the fingerprint fragment size; and

concatenating the plurality of fingerprint fragments to form the text fingerprint,

and wherein determining the target label comprises:

retrieving a reference fingerprint from a database of reference fingerprints, the reference fingerprint determined for a reference electronic document belonging to the category, the reference fingerprint selected according to a length of the reference fingerprint so that the length of the reference fingerprint is between the upper and lower bounds; and

determining whether the target electronic document belongs to the category according to a result of comparing the text fingerprint to the reference fingerprint.

13. The server computer system of claim 12, wherein the category of documents is a spam category.

14. The server computer system of claim 12, wherein the category of documents is a fraudulent document category.

15. The server computer system of claim 12, wherein determining the text fingerprint further comprises determining each character of the sequence of characters according to a distinct group of bits of the hash of the distinct text token.

16. The server computer system of claim 12, wherein selecting the plurality of text tokens comprises:

selecting a preliminary plurality of text tokens of the target electronic document;

determining a count of the preliminary plurality of text tokens; and

in response, when the count of the preliminary plurality of text tokens exceeds a predetermined threshold, prune the preliminary plurality of text tokens to form the selected plurality of text tokens so that the count of the selected plurality of tokens does not exceed the predetermined threshold.

17. The server computer system of claim 16, wherein pruning the preliminary plurality of text tokens comprises selecting a target text token of the preliminary plurality of text tokens into the selected plurality of text tokens according to a hash of the target text token.

18. The server computer system of claim 17, wherein pruning the preliminary plurality of text tokens further comprises:

determining whether the hash of the target text token is divisible by a zoom-out factor; and

in response, when the target text token is divisible by the zoom-out factor, selecting the target text token into the selected plurality of text tokens.

19. The server computer system of claim 12, wherein selecting the plurality of text tokens comprises:

selecting a preliminary plurality of text tokens of the target electronic document;

determining a count of the preliminary plurality of text tokens; and

in response, when the count of the preliminary plurality of text tokens exceeds a predetermined threshold,

determine a plurality of aggregate text tokens, each aggregate text token of the plurality of aggregate text tokens comprising a concatenation of a set of text tokens of the preliminary plurality of text tokens; and

selecting an aggregate token of the plurality of aggregate tokens into the selected plurality of text tokens according to a hash of the aggregate text token.

20. The server computer system of claim 12, wherein the target electronic document is selected from a group consisting of an email message and a hypertext markup language (HTML) document.

21. The server computer system of claim 12, wherein the distinct text token comprises an item selected from a group consisting of a word, an email address, and a uniform resource locator (URL) of the target electronic communication.

22. A method comprising employing at least one processor of a client computer system to determine a text fingerprint of a target electronic document so that a length of the text fingerprint is constrained between a lower bound and an upper bound, wherein the lower and upper bounds are predetermined, and wherein determining the text fingerprint comprises:

selecting a plurality of text tokens of the target electronic document;

in response to selecting the plurality of text tokens, determining a fingerprint fragment size according to the upper and lower bounds, and according to a count of the selected plurality of text tokens;

determining a plurality of fingerprint fragments, each fingerprint fragment of the plurality of fingerprint fragments determined according to a hash of a distinct text token of

the selected plurality of text tokens, each fingerprint fragment consisting of a sequence of characters, a length of the sequence chosen to equal the fingerprint fragment size; and

concatenating the plurality of fingerprint fragments to form the text fingerprint.

23. The method of claim 22, further comprising employing the at least one processor to determine a category of documents that the target electronic document belongs to according to the text fingerprint.

24. A method comprising employing at least one processor of a server computer system configured to perform transactions with a plurality of client systems, to:

receive a text fingerprint from a client system of the plurality of client systems, the text fingerprint determined for a target electronic document so that a length of the text fingerprint is constrained between a lower bound and an upper bound, wherein the lower and upper bounds are predetermined; and

send to the client system a target label determined for the target electronic document, the target label indicating a category of documents that the target electronic document belongs to,

wherein determining the text fingerprint comprises:

selecting a plurality of text tokens of the target electronic document;

in response to selecting the plurality of text tokens, determining a fingerprint fragment size according to the upper and lower bounds, and according to a count of the selected plurality of text tokens;

determining a plurality of fingerprint fragments, each fingerprint fragment of the plurality of fingerprint fragments determined according to a hash of a distinct text token of the selected plurality of text tokens, each fingerprint fragment consisting of a sequence of characters, a length of the sequence chosen to equal the fingerprint fragment size; and

concatenating the plurality of fingerprint fragments to form the text fingerprint,

and wherein determining the target label comprises:

retrieving a reference fingerprint from a database of reference fingerprints, the reference fingerprint determined for a reference electronic document belonging to the category, the reference fingerprint selected according to a length of the reference fingerprint so that the length of the reference fingerprint is between the upper and lower bounds; and

determining whether the target electronic document belongs to the category according to a result of comparing the text fingerprint to the reference fingerprint.

## CLAIMS

1. A client computer system comprising at least one processor configured to determine a text fingerprint of a target electronic document so that a length of the text fingerprint is constrained between a lower bound and an upper bound, wherein the lower and upper bounds are predetermined, and wherein determining the text fingerprint comprises:

selecting a plurality of text tokens of the target electronic document, wherein selecting the plurality of text tokens comprises:

selecting a preliminary plurality of text tokens of the target electronic document,

determining a count of the preliminary plurality of text tokens, and

in response, when the count of the preliminary plurality of text tokens exceeds a predetermined threshold, prune the preliminary plurality of text tokens to form the selected plurality of text tokens so that a count of the selected plurality of tokens does not exceed the predetermined threshold;

in response to selecting the plurality of text tokens, determining a fingerprint fragment size according to the upper and lower bounds, and according to the count of the selected plurality of text tokens;

determining a plurality of fingerprint fragments, each fingerprint fragment of the plurality of fingerprint fragments determined according to a hash of a distinct text token of the selected plurality of text tokens, each fingerprint fragment consisting of a sequence of characters, a length of the sequence chosen to equal the fingerprint fragment size; and

concatenating the plurality of fingerprint fragments to form the text fingerprint.

2. The client computer system of claim 1, wherein the at least one processor is further configured to:

send the text fingerprint to a server computer system; and

receive from the server computer system a target label determined for the target electronic document, the target label indicative of a category of documents that the target electronic document belongs to, wherein determining the target label comprises:

retrieving a reference fingerprint from a database of reference fingerprints, the reference fingerprint determined for a reference electronic document belonging to the category, the reference fingerprint selected according to a length of the reference fingerprint so that the length of the reference fingerprint is between the upper and lower bounds; and

determining whether the target electronic document belongs to the category according to a result of comparing the text fingerprint to the reference fingerprint.

3. The client computer system of claim 2, wherein the category of documents is a spam category.

4. The client computer system of claim 2, wherein the category of documents is a fraudulent document category.

5. The client computer system of claim 1, wherein determining the text fingerprint further comprises determining each character of the sequence of characters according to a distinct group of bits of the hash of the distinct text token.

6. The client computer system of claim 1, wherein pruning the preliminary plurality of text tokens comprises selecting a target text token of the preliminary plurality of text tokens into the selected plurality of text tokens according to a hash of the target text token.

7. The client computer system of claim 6, wherein pruning the preliminary plurality of text tokens further comprises:

determining whether the hash of the target text token is divisible by a zoom-out factor; and

in response, when the target text token is divisible by the zoom-out factor, selecting the target text token into the selected plurality of text tokens.

8. The client computer system of claim 1, wherein selecting the plurality of text tokens further comprises, when the count of the preliminary plurality of text tokens exceeds the predetermined threshold:

determining a plurality of aggregate text tokens, each aggregate text token of the plurality of aggregate text tokens comprising a concatenation of a set of text tokens of the preliminary plurality of text tokens; and

selecting an aggregate token of the plurality of aggregate tokens into the selected plurality of text tokens according to a hash of the aggregate text token.

9. The client computer system of claim 1, wherein the target electronic document is selected from a group consisting of an email message and a hypertext markup language (HTML) document.

10. The client computer system of claim 1, wherein the distinct text token comprises an item selected from a group consisting of a word, an email address, and a uniform resource locator (URL) of the target electronic communication.

11. A server computer system comprising at least one processor configured to perform transactions with a plurality of client systems, wherein a transaction comprises:

receiving a text fingerprint from a client system of the plurality of client systems, the text fingerprint determined for a target electronic document so that a length of the text fingerprint

is constrained between a lower bound and an upper bound, wherein the lower and upper bounds are predetermined; and

sending to the client system a target label indicative of a category of documents that the target electronic document belongs to,

wherein determining the text fingerprint comprises:

selecting a plurality of text tokens of the target electronic document, wherein selecting the plurality of text tokens comprises:

selecting a preliminary plurality of text tokens of the target electronic document,

determining a count of the preliminary plurality of text tokens, and

in response, when the count of the preliminary plurality of text tokens exceeds a predetermined threshold, prune the preliminary plurality of text tokens to form the selected plurality of text tokens so that a count of the selected plurality of tokens does not exceed the predetermined threshold;

in response to selecting the plurality of text tokens, determining a fingerprint fragment size according to the upper and lower bounds, and according to the count of the selected plurality of text tokens;

determining a plurality of fingerprint fragments, each fingerprint fragment of the plurality of fingerprint fragments determined according to a hash of a distinct text token of the selected plurality of text tokens, each fingerprint fragment consisting of a sequence of characters, a length of the sequence chosen to equal the fingerprint fragment size; and

concatenating the plurality of fingerprint fragments to form the text fingerprint,

and wherein determining the target label comprises:

retrieving a reference fingerprint from a database of reference fingerprints, the reference fingerprint determined for a reference electronic document belonging to the category, the reference fingerprint selected according to a length of the reference fingerprint so that the length of the reference fingerprint is between the upper and lower bounds; and

determining whether the target electronic document belongs to the category according to a result of comparing the text fingerprint to the reference fingerprint.

12. The server computer system of claim 11, wherein the category of documents is a spam category.

13. The server computer system of claim 11, wherein the category of documents is a fraudulent document category.

14. The server computer system of claim 11, wherein determining the text fingerprint further

comprises determining each character of the sequence of characters according to a distinct group of bits of the hash of the distinct text token.

15. The server computer system of claim 11, wherein pruning the preliminary plurality of text tokens comprises selecting a target text token of the preliminary plurality of text tokens into the selected plurality of text tokens according to a hash of the target text token.

16. The server computer system of claim 15, wherein pruning the preliminary plurality of text tokens further comprises:

determining whether the hash of the target text token is divisible by a zoom-out factor; and

in response, when the target text token is divisible by the zoom-out factor, selecting the target text token into the selected plurality of text tokens.

17. The server computer system of claim 11, wherein selecting the plurality of text tokens further comprises, when the count of the preliminary plurality of text tokens exceeds the predetermined threshold:

determining a plurality of aggregate text tokens, each aggregate text token of the plurality of aggregate text tokens comprising a concatenation of a set of text tokens of the preliminary plurality of text tokens; and

selecting an aggregate token of the plurality of aggregate tokens into the selected plurality of text tokens according to a hash of the aggregate text token.

18. The server computer system of claim 11, wherein the target electronic document is selected from a group consisting of an email message and a hypertext markup language (HTML) document.

19. The server computer system of claim 11, wherein the distinct text token comprises an item selected from a group consisting of a word, an email address, and a uniform resource locator (URL) of the target electronic communication.

20. A method comprising employing at least one processor of a client computer system to determine a text fingerprint of a target electronic document so that a length of the text fingerprint is constrained between a lower bound and an upper bound, wherein the lower and upper bounds are predetermined, and wherein determining the text fingerprint comprises:

selecting a plurality of text tokens of the target electronic document, wherein selecting the plurality of text tokens comprises:

selecting a preliminary plurality of text tokens of the target electronic document,

determining a count of the preliminary plurality of text tokens, and

in response, when the count of the preliminary plurality of text tokens exceeds a predetermined threshold, prune the preliminary plurality of text tokens to form the selected plurality of text tokens so that a count of the selected plurality of tokens does not exceed the predetermined threshold;

in response to selecting the plurality of text tokens, determining a fingerprint fragment size according to the upper and lower bounds, and according to the count of the selected plurality of text tokens;

determining a plurality of fingerprint fragments, each fingerprint fragment of the plurality of fingerprint fragments determined according to a hash of a distinct text token of the selected plurality of text tokens, each fingerprint fragment consisting of a sequence of characters, a length of the sequence chosen to equal the fingerprint fragment size; and

concatenating the plurality of fingerprint fragments to form the text fingerprint.

21. The method of claim 20, further comprising employing the at least one processor to determine a category of documents that the target electronic document belongs to according to the text fingerprint.

22. A method comprising employing at least one processor of a server computer system configured to perform transactions with a plurality of client systems, to:

receive a text fingerprint from a client system of the plurality of client systems, the text fingerprint determined for a target electronic document so that a length of the text fingerprint is constrained between a lower bound and an upper bound, wherein the lower and upper bounds are predetermined; and

send to the client system a target label determined for the target electronic document, the target label indicating a category of documents that the target electronic document belongs to,

wherein determining the text fingerprint comprises:

selecting a plurality of text tokens of the target electronic document, wherein selecting the plurality of text tokens comprises:

selecting a preliminary plurality of text tokens of the target electronic document,

determining a count of the preliminary plurality of text tokens, and

in response, when the count of the preliminary plurality of text tokens exceeds a predetermined threshold, prune the preliminary plurality of text tokens to form the selected plurality of text tokens so that a count of the selected plurality of tokens does not exceed the predetermined threshold;

in response to selecting the plurality of text tokens, determining a fingerprint fragment size according to the upper and lower bounds, and according to the count of the selected plurality of text tokens;

determining a plurality of fingerprint fragments, each fingerprint fragment of the plurality of fingerprint fragments determined according to a hash of a distinct text token of the selected plurality of text tokens, each fingerprint fragment consisting of a sequence of characters, a length of the sequence chosen to equal the fingerprint fragment size; and

concatenating the plurality of fingerprint fragments to form the text fingerprint,

and wherein determining the target label comprises:

retrieving a reference fingerprint from a database of reference fingerprints, the reference fingerprint determined for a reference electronic document belonging to the category, the reference fingerprint selected according to a length of the reference fingerprint so that the length of the reference fingerprint is between the upper and lower bounds; and

determining whether the target electronic document belongs to the category according to a result of comparing the text fingerprint to the reference fingerprint.