CLAIMS

What is claimed is:

1. A method for content and context aware data classification, the method comprising:

scanning a plurality of unstructured documents in one or more network data repositories of a computer network in need of confidentiality management, wherein the plurality of unstructured documents comprise documents having at least two different formats;

extracting content features and context features from the plurality of unstructured documents into one or more term frequency-inverse document frequency (TF-IDF) vectors and one or more latent semantic indexing (LSI) vectors; and

classifying the plurality of unstructured documents into a number of category classifications by machine learning the extracted content features and context features of the plurality of unstructured documents at a file management platform of the computer network, each of the category classifications being associated with one or more confidentiality classifications and one or more business category classifications, wherein the one or more business category classifications comprise at least one of the following category classifications: Corporate Entity, Legal/Compliance, Client/Customer, Accounting/Tax, Employee/Personal Document, and Transaction.

2. The method according to claim 1, wherein the step of extracting content features and context features of the plurality of unstructured documents into the one or more TF-IDF vectors and the one or more LSI vectors comprises:

for each document of the plurality of unstructured documents, extracting content features from content data of the document by conducting one or more natural language processing (NLP) steps comprising:

recognizing language of the content data of the document;

splitting content data of the document into tokens of words;

removing stop words and punctuation marks from the tokens of words;

transforming each of the tokens of words into a word stem, word base or word root; and

extracting content features of the document by extracting all the transformed tokens of words of the content data.

3. The method according to claim 1 or 2, wherein the step of extracting content features and context features of the plurality of unstructured documents into the one or more TF-IDF vectors and the one or more LSI vectors comprises:

for each document of the plurality of unstructured documents, extracting context features from metadata of the document by conducting one or more natural language processing (NLP) steps comprising:

splitting the metadata of the document into tokens of word strings, wherein the tokens of word strings are divided by symbols in the metadata;

retaining tokens of word strings that contain alphabets and/or digits; and

extracting context features of the document by extracting all the retained tokens of words of the metadata.

4. The method according to any one of claims 1 to 3, wherein the step of extracting content features and context features of the plurality of unstructured documents into the one or more TF-IDF vectors and the one or more LSI vectors further comprises:

vectorizing the content features and the context features of the plurality of unstructured documents into one or more bag-of-words (BOW) representations;

converting the one or more bag-of-words (BOW) representations respectively into the one or more term frequency-inverse document frequency (TF-IDF) vectors; and

projecting the one or more TF-IDF vectors into a latent semantic indexing (LSI) space based on a desired level of retained variance to transform the one or more TF-IDF vectors into the one or more LSI vectors.

5. The method according to claim 4, further comprising:

storing the content features, the context features, the BOW representations and the LSI vectors of the plurality of unstructured documents in a database of the file management platform.

6. The method according to any one of claims 1 to 5, wherein the step of classifying the plurality of unstructured documents comprises:

clustering the one or more LSI vectors into the number of category classifications.

7. The method according to any one of claims 1 to 6, further comprises:

splitting the plurality of unstructured documents into a labeled set and an unlabeled set;

randomly splitting the labeled set into a raw labeled training set and a raw test set based on a predetermined ratio; and

merging the unlabeled set with the raw labelled training set to form a raw training set.

8. The method according to claim 7, wherein the step of machine learning the extracted content features and context features comprises:

predicting one or more clustering models based on machine learning the one or more LSI vectors of the documents in the raw training set, each clustering model clustering the one or more LSI vectors into the number of category classifications;

selecting a clustering model from the one or more clustering models based on a confirmation of its corresponding category classifications by an administrator of the computer network;

evaluating the selected clustering model by applying the selected clustering model to cluster the documents in the raw test set; and

if the evaluation of the selected clustering model produces same category classifications as the labels of the raw test set, classifying all of the plurality of unstructured documents into the number of category classifications using the selected clustering model, and if the evaluation of the selected clustering model produces different category classifications from the labels of the raw test set, selecting another clustering model from the one or more clustering models.

9. The method according to claim 8, further comprising:

validating the classified category classifications by one or more owners of the plurality of unstructured documents.

10. The method according to claim 5, further comprising:

storing the content features, the context features, the BOW representation and the one or more LSI vectors of the one or more documents into a plurality of interlinked tables in the database of the file management platform.

11. The method according to claim 10, wherein the plurality of interlinked tables are each HBASE tables.

12. The method according to claim 10 or 11, wherein the step of storing the content features, the context features, the BOW representations and the LSI vectors of the plurality of unstructured documents into the plurality of interlinked tables comprises:

storing metadata of the plurality of unstructured documents into a first table, the metadata including file paths, file sizes, file extensions, file owners, predicted category classifications and confidentiality classifications predicted by the machine learning, and/or validated category classifications and confidentiality classifications validated by the file owner, wherein each row of the first table corresponds to a file path of each of the plurality of unstructured documents;

storing the metadata of the plurality of unstructured documents into a second table, wherein the second table supplements the first table, and wherein each row of the second table corresponds to a concatenated data of a file owner with a hashed file path of each of the plurality of unstructured documents owned by the file owner; and/or

storing the content features of the plurality of unstructured documents into a third table.

13. The method according to claim 12, further comprising:

generating a visualization of one or more of the category classifications, confidentiality classification, language, and/or document type of each of the plurality of unstructured documents based on at least the first table.

14. A method for content and context aware data classification, the method comprising:

accessing a plurality of unstructured documents in one or more network data repositories of a computer network in need of confidentiality management to identify a document from the plurality of unstructured documents in a network data repository of the one or more network data repositories, wherein the plurality of unstructured documents comprise documents having at least two different formats;

scanning content data and metadata of the identified unstructured document;

extracting content features and context features of the identified unstructured document based on the content data and metadata;

converting the content features and context features into term frequency–inverse document frequency (TF-IDF) vectors;

by Latent Semantic Indexing (LSI), transforming the TF-IDF vectors to LSI vectors to reduce dimensionality; and

predicting one or more category classifications for the identified unstructured document by machine learning the extracted content features and context features of the identified unstructured document at a file management platform of the computer network in response to the LSI vectors, wherein each of the category classifications is associated with one or more confidentiality classifications and one or more business category classifications, wherein the one or more business category classifications comprise at least one of the following category classifications: Corporate Entity,

Legal/Compliance, Client/Customer, Accounting/Tax, Employee/Personal Document, and Transaction.

15. The method according to claim 14, wherein the predicted one or more category classifications and the associated one or more confidentiality classifications are confirmed by an administrator of the computer network.

16. The method according to claim 14 or 15, wherein the confirmed one or more category classifications and the associated one or more confidentiality classifications are validated by the owner of the document.

17. A method for content and context aware data security anomaly detection, the method comprising:

machine learning content and context data of a plurality of unstructured documents in one or more network data repositories of a computer network to determine category classification and confidentiality classification for each of the plurality of unstructured documents, wherein the plurality of unstructured documents comprise documents having at least two different formats;

receiving a request from a user to access one of the plurality of unstructured documents stored in one of the one or more network data repositories, the request including content or context data of the document;

identifying a confidentiality classification and a category classification of the document in response to the machine learning;

determining a security access control level of the user based on lightweight directory access protocol (LDAP) extraction; and

detecting a data security anomaly if the security access control level of the user is lower than the confidentiality classification of the document.

18. The method according to claim 17, wherein the content or context data of the document comprises a file path of the document.

19. The method according to claim 17 or 18, further comprising:

granting an access to the user if the security access control level of the user is not lower than the confidentiality classification of the document.

20. The method according to claim 17 or 18, further comprising:

generating a data security anomaly report in response to the detection of the data security anomaly; and

presenting the data security anomaly report to an administrator of the computer network, the owner of the document, and/or the user.

21. A machine learning engine for content and context aware data classification, the machine learning engine comprising:

a model training module configured to:

predict one or more clustering models based on machine learning latent semantic indexing (LSI) vectors of a plurality of unstructured documents in a raw training set, wherein the plurality of unstructured documents comprise documents having at least two different formats, and wherein each clustering model clusters one or more of the LSI vectors into a number of category classifications, each of the category classifications being associated with one or more confidentiality

classifications and one or more business category classifications, wherein the one or more business classifications comprise at least one of the following category classifications: Corporate Entity, Legal/Compliance, Client/Customer, Accounting/Tax, Employee/Personal Document, and Transaction, and

select a clustering model from the one or more clustering models based on a confirmation of its corresponding category classifications by an administrator of a computer network;

a model evaluation module configured to evaluate the selected clustering model by applying the selected clustering model to cluster the documents in the raw test set; and

a data classification engine configured to:

if the evaluation of the selected clustering model produces same category classifications as the labels of the raw test set, classify all of the one or more documents into the number of category classifications using the selected clustering model, and

if the evaluation of the selected clustering model produces different category classifications from the labels of the raw test set, indicate to the feature clustering module to select another clustering model from the one or more clustering models.

22. A file management platform system for content and context aware data classification, the file management platform system comprising:

a feature engineering module configured to:

split a plurality of unstructured documents in one or more network data repositories of a computer network into a labeled set and an unlabeled set, wherein the

plurality of unstructured documents comprise documents having at least two different formats, and wherein the labeled set is randomly split into a raw labeled training set and a raw test set based on a predetermined ratio, and wherein the unlabeled set is merged with the raw labeled training set to form a raw training set,

extract content features and context features of the documents in the raw training set,

vectorize the content features and the context features of the documents in the raw training set into one or more bag-of-words (BOW) representations,

convert the one or more bag-of-words (BOW) representations of the documents in the raw training set respectively into one or more term frequency-inverse document frequency (TF-IDF) vectors, and

project the one or more TF-IDF vectors of the documents in the raw training set into a latent semantic indexing (LSI) space based on a desired level of retained variance to transform the one or more TF-IDF vectors to one or more LSI vectors; and

a machine learning engine comprising:

a model training module configured to:

predict one or more clustering models based on machine learning the one or more LSI vectors of the plurality of unstructured documents in the raw training set, each clustering model clustering the one or more LSI vectors into a number of category classifications, each of the category classifications being associated with one or more confidentiality classifications and one or more category classifications, and

select a clustering model from the one or more clustering models based on a confirmation of its corresponding category classifications by an administrator of the computer network;

a model evaluation module configured to evaluate the selected clustering model by applying the selected clustering model to cluster the plurality of unstructured documents in the raw test set; and

a data classification engine configured to:

if the evaluation of the selected clustering model produces same category classifications as the labels of the raw test set, classify all of the plurality of unstructured documents into the number of category classifications using the selected clustering model, and

if the evaluation of the selected clustering model produces different category classifications from the labels of the raw test set, indicate to the feature clustering module to select another clustering model from the one or more clustering models.

23. A computer readable medium comprising instructions which, when executed by a processor, make the processor perform a method for content and context aware data classification according to any one of claims 1 to 16.

24. The computer readable medium according to claim 23, further comprising instructions which, when executed by the processor, make the processor perform a method for content and context aware data security anomaly detection according to any one of claims 17 to 20.