CLAIMS

What is claimed is:

1. A deep learning engine comprising:

a feature extraction module; and

a classification and labelling module,

wherein the feature extraction module is configured to extract both context features and document features from documents simultaneously in one or more data repositories of a computer network or cloud repository, and

wherein the document features include metadata, and wherein the feature extraction module is configured to extract the metadata as at least a portion of the document features, and

wherein the classification and labelling module is configured for content and context aware data classification of the documents in the one or more data repositories of the computer network or cloud repository by business category using neural networks and at least in response to the metadata of the documents in the one or more data repositories of the computer network or cloud repository.

2. The deep learning engine in accordance with Claim 1 wherein the content and context aware data classification of the documents is built from the document features in an iterative process.

3. The deep learning engine in accordance with Claim 2 wherein the document features further comprise user rights, language, document date and document owner, and wherein the classification and labelling module is further configured for content

and context aware data classification of the documents by confidentiality level using the neural networks.

4. The deep learning engine in accordance with Claim 1 wherein the document features further comprise user rights, language, document date and document owner, and wherein the classification and labelling module is further configured for content and context aware data classification of the documents by confidentiality level using the neural networks.

5. The deep learning engine in accordance with Claim 1 wherein the neural networks include convolutional neural networks or recurrent neural networks.

6. The deep learning engine in accordance with Claim 1 wherein the feature extraction module uses term frequency-inverse document frequency (TF-IDF) and latent semantic indexing (LSI) for feature extraction.

7. The deep learning engine in accordance with Claim 1 wherein the feature extraction module uses a word feature embedding approach for feature extraction, and wherein the word feature embedding approach uses word embedding vectors of context and content.

8. The deep learning engine in accordance with Claim 1 wherein the classification and labelling module comprises a supervised classification module.

9. The deep learning engine in accordance with Claim 8 wherein the supervised classification module uses one or more of Random Forest, Naïve Bayes, OnevsRest and XGBoost for supervised classification.

10. The deep learning engine in accordance with Claim 8 wherein the supervised classification module comprises Bidirectional Encoder Representations from Transformers (BERT) fine-tuning module for supervised classification.

11. The deep learning engine in accordance with Claim 10 wherein the BERT fine-tuning module comprises a transformer architecture having a feed-forward neural network with layer norm and multi-head attention.

12. A system for context and content aware data classification, the system comprising:

a deep learning engine comprising a feature extraction module and a classification and labelling module; and

a smart sampling module configured to simultaneously sample a pool of documents in one or more data repositories of a computer network or cloud repository to identify documents or records for content and context aware data classification, wherein the feature extraction module is configured to extract both context features and document features from the documents or records in the one or more data repositories of the computer network or cloud repository, wherein the document features include metadata, and wherein the feature extraction module is configured to extract the metadata as at least a portion of the document features, and

wherein the classification and labelling module is configured for the content and context aware data classification of the documents or records by both business category and confidentiality level using neural networks and at least in response to the metadata of the documents or records in the one or more data repositories of the computer network or cloud repository.

13. The system in accordance with Claim 12 further comprising:

a clustering module for clustering the documents or records in accordance with the context features and document features extracted by the feature extraction module.

14. The system in accordance with Claim 12 wherein the classification and labelling module comprises an autolabelling module for autolabelling of the documents or records.

15. A method for content and context aware data classification by business category and confidentiality level, the method comprising:

simultaneously scanning documents or records in one or more data repositories of a computer network or cloud repository; and

extracting content features and context features of the documents or records utilizing deep learning technologies as convolutional neural networks to associate the documents or records with one or more business categories and one or more confidentiality levels, wherein the content features and context features include metadata of each of the documents or records, and wherein the metadata is extracted as at least a portion of the document features to be used for classification of the documents or records.

16. The method in accordance with Claim 15 wherein the extracting content features and context features of the documents or records comprises extracting content features and context features of the documents or records for further online and offline classification into the one or more confidentiality levels, and wherein the extracting content features and context features of the documents or records comprises extracting content features and context features of the documents or records for further online and offline classification into the one or more business categories.

17. The method in accordance with Claim 15 wherein the extracting content features and context features of the one or more documents or records comprises generating word embedding vectors for model training.

18. The method in accordance with Claim 17 wherein the generating word embedding vectors comprises generating word embedding vectors for each language separately for the model training.

19. The method in accordance with Claim 17 wherein the extracting content features and context features of the one or more documents or records further comprises generating metadata type vectors and data type vectors for model training, the metadata type vectors generated in response to the metadata.

20. The method in accordance with Claim 15 wherein the extracting content features and context features of the one or more documents or records comprises generating

metadata type vectors and data type vectors for model training, the metadata type vectors generated in response to the metadata.