CLAIMS

1. A method comprising:

a) determining a plurality of documents associated with an entity;

b) for each document:

i) providing the document to a numerical representation generation model;

ii) generating, by the numerical representation generation model, a numerical representation of the document; and

iii) determining a document score for the document based on the numerical representation, wherein the document score comprises one or more confidence scores for corresponding one or more attributes of the document;

c) providing the document scores for the documents to a clustering module;

d) determining, by the clustering module, one or more clusters, each cluster being associated with a class of the documents;

e) outputting, by the clustering module, a cluster identifier indicative of the class of each document; and

f) associating each document with its respective cluster identifier.

2. The method of claim 1, wherein each document is associated with a corresponding label of a plurality of labels, and the method further comprises:

determining a dataset for each label of the plurality of labels, the dataset comprising the documents associated with the label; and

performing steps b) to e) for each dataset separately.

3. The method of claim 1 or claim 2, wherein the numerical representation is a multi-dimensional vector, and wherein the method further comprises:

providing the numerical representation to a dimensionality reduction model to determine the document score.

4. The method of claim 3, wherein the dimensionality reduction model performs Principal Component Analysis (PCA) to generate a dimensionally reduced numerical representation of the document, and the method further comprises:

multiplying the dimensionally reduced numerical representation by the variance-ratio to determine the document score.

5. The method of any one of the preceding claims, ~~wherein the numerical representation comprises one or more confidence scores for corresponding one of more attributes of the document,~~ wherein the one or more attributes comprise: (i) amount; (ii) entity; (iii) due date; (iv) bill date; (v) invoice number; (vi) tax amount; and/or (vii) currency.

6. The method of any one of the preceding claims, comprising:

applying a filter to the document to blur image data of the document before providing the filtered document to the numerical representation generation model.

7. The method of any one of the preceding claims, wherein one or more of the cluster identifiers are indicative of a low confidence score class of document for which one or more low confidence scores have been allocated and the method further comprises:

selecting the documents of the one or more low confidence score classes for label review.

8. The method of any one of the preceding claims, wherein one or more of the cluster identifiers are indicative of a low confidence score class of document for which one or more low confidence scores have been allocated and the method further comprises:

retraining a model used to generate the confidence scores using documents from the one or more low confidence score classes of document.

9. The method of any one of the preceding claims, wherein determining, by the clustering module, one or more clusters, comprises:

determining a plurality of histogram bins based on the document scores;

determining a bin score for each document; and

grouping histogram bins into clusters.

10. The method of claim 9, wherein grouping histogram bins into clusters comprises determining the clusters as respective local minima of the histogram bins.

11. The method of any one of claims 1 to 8, wherein determining, by the clustering module, one or more clusters, comprises performing k-means clustering, density-based spatial clustering (DBSCAN) or hierarchical clustering.

12. The method of any one of the preceding claims, wherein each document comprises character data, and the method further comprising:

for each document:

providing the character data to an attribute determination model;

determining an attribute associated with the document based on the character data; and

associating the document with the determined attribute as the label.

13. The method of any one of the preceding claims, wherein each document comprises image data, and the method further comprises:

for each document:

extracting the image data from the document;

providing the image data to an attribute determination model;

determining, by the attribute determination module, an attribute associated with the document based on the image data; and

associating the document with the determined attribute as the label.

14. The method of any one of claims 1 to 12, wherein each document comprises image data, and the method further comprises:

for each document:

extracting the image data from the document;

providing the image data to an image-based numerical representation generation module;

determining by the image-based numerical representation generation module, an image-based numerical representation of the document;

providing the character data to a character-based numerical representation generation module;

determining by the character-based numerical representation generation module, a character-based numerical representation of the document;

providing, to a consolidated numerical representation generation module, the image-based numerical representation of the document and the character-based numerical representation of the document; and

generating, by the consolidated numerical representation generation module, a combined numerical representation of the character data and the image data of the document;

providing the combined numerical representation to an attribute prediction module; and

determining, by the attribute prediction module, the attribute associated with the document; and

associating the document with the determined attribute as the label.

15. The method of any one of claims 12 to 14, wherein the attribute is an entity associated with the document.

16. The method of any one of any one of the preceding claims, wherein the documents are derived from previously reconciled accounting documents of an

accounting system, each of which has been associated with a respective entity, and wherein the label of each document is indicative of the respective entity.

17. The method of any one of the preceding claims, wherein the document is an accounting document and the class of documents includes one or more of: (i) an invoice; (ii) a credit note; (iii) a receipt; (iv) a purchase order; and (v) a quote.

18. A system comprising:

one or more processors; and

memory comprising computer executable instructions, which when executed by the one or more processors, cause the system to perform the method of any one of claims 1 to 17.

19. A computer-readable storage medium storing instructions that, when executed by a computer, cause the computer to perform the method of any one of claims 1 to 17.