CLAIMS

1. A method comprising:

determining one or more tokens from each of a plurality of financial documents from a database;

determining a set of preliminary attribute labels for each of the one or more tokens, wherein the set of attribute labels are associated with one or more attribute types;

training an attribute prediction model comprising an expectation maximization model, wherein the training comprises:

determining the sets of preliminary attribute labels using multiple attribute labeling modules, each module configured to determine preliminary labels for a different attribute type; and

training the expectation maximisation model to determine a confidence value for each of a set of attribute labels for a candidate token based on the set of preliminary labels for the candidate token, wherein the expectation maximisation model is trained using the sets of preliminary attribute labels for the respective one or more tokens, and wherein determining the confidence value for each token comprises:

analyzing correlations between preliminary labels from different attribute labeling modules for the candidate token;

detecting conflicts between the preliminary labels, wherein a conflict occurs when different preliminary labels indicate different attribute types for the same token;

determining a reliability score for each preliminary label based on the analyzed correlations and detected conflicts; and

generating the confidence value by combining the preliminary labels weighted by their respective reliability scores;

providing the set of preliminary attribute labels for each token to the trained attribute prediction model;

generating, by the trained attribute prediction model, for each token, a confidence value for each of the one or more preliminary attribute labels;

determining a plurality of subsets of tokens, each subset being associated with a respective document of the plurality of documents;

determining a set of refined labels for each document based on the confidence values associated with the tokens of a respective subset of tokens of the plurality of subsets of tokens, wherein the set of refined labels comprises a value for one or more attribute types,

wherein determining the set of refined labels for each financial document based on the confidence values comprises:

determining, for each of the one or more attribute types, a respective token of the subset of tokens having a highest confidence value; and

assigning each of the one or more attribute types, the respective token having the highest confidence value; and

training an entity attribute identification model using the refined sets of labels and the respective documents to configure entity attribute identification model to determine entity attributes associated with candidate document.

2. The method of claim 1, wherein each of the one or more attribute types has one or more inherent characteristics determinable from its respective value.

3. The method of claim 2, wherein determining the one or more tokens from each of a plurality of financial documents comprises determining the one or more tokens based on at least one of the one or more inherent characteristics of the one or more attribute types.

4. The method of any one of the preceding claims, wherein determining the sets of preliminary attribute labels comprises providing the one or more tokens to one or more attribute labelling modules to determine, for each token, the sets of preliminary attribute labels and wherein each of the one or more attribute labelling modules is configured to determine preliminary attribute labels of a different attribute type.

5. The method of claim 4, wherein the one or more attribute labelling modules comprises a first digit detector configured to detect a fixed number span of digits within a candidate token, and responsive to determining that the candidate token comprises the fixed number span of digits, assigning the candidate token a first preliminary label of a first attribute type, and responsive to determining that the candidate token does not comprise the fixed number span of digits, assigning the candidate token a first preliminary label indicative of the token not being representative of the first attribute type.

6. The method of claim 5, wherein the one or more attribute labelling modules comprises a first label detector configured to detect a character stream in a proximal token to the candidate token that corresponds to a first label of the first attribute type, and responsive to determining that the proximal token comprises the character stream corresponding to the first label, assigning the candidate token a second preliminary label of the first attribute type, and responsive to determining that the proximal token does not comprise the character stream corresponding to the first label, assigning the candidate token a second preliminary label indicative of the candidate token not being representative of the first attribute type.

7. The method of claim 5 or claim 6, wherein the one or more attribute labelling modules comprises a check sum calculator configured to verify that a span of digits detected by the first digit detector is a number that complies requirements of the first attribute type, and responsive to determining that the number complies with the requirements, assigning the candidate token a third preliminary label of the first attribute type, and responsive to determining that the candidate token does not comply with the requirements, assigning the candidate token a third preliminary label indicative of the token not being representative of the first attribute type.

8. The method of any one of claims 4 to 7, wherein the one or more attribute labelling modules comprises a second digit detector configured to detect a fixed number span of digits within a candidate token, and responsive to determining that the candidate token comprises the fixed number span of digits, assigning the candidate token a first preliminary label of a second attribute type, and responsive to determining that the candidate token does not

comprise the fixed number span of digits, assigning the candidate token a first preliminary label indicative of the candidate token not being representative of the second attribute type.

9. The method of claim 8, wherein the one or more attribute labelling modules comprises a second label detector configured to detect a character stream in a proximal token to the candidate token that corresponds to a second label of the second attribute type, and responsive to determining that the proximal token comprises the character stream corresponding to the second label, assigning the candidate token a second preliminary label of the second attribute type, and responsive to determining that the proximal token does not comprise the character stream corresponding to the second label, assigning the candidate token a second preliminary label indicative of the candidate token not being representative of the second attribute type.

10. The method of claim 8 or claim9, wherein the one or more attribute labelling modules comprises a verification checker configured to verify that a span of digits detected by the second digit detector is a number that corresponds with a registered number for the second attribute type, and responsive to determining that the number corresponds with the registered number, assigning the candidate token a third preliminary label of the second attribute type, and responsive to determining that the candidate token does not comply with the requirements, assigning the candidate token a third preliminary label indicative of the token not being representative of the second attribute type.

11. The method of any one of claims 4 to 10, wherein the one or more attribute labelling modules comprises a third digit detector configured to detect a fixed number span of digits within a candidate token, and responsive to determining that the candidate token comprises the fixed number span of digits, assigning the candidate token a first preliminary label of a third attribute type, and responsive to determining that the candidate token does not comprise the fixed number span of digits, assigning the candidate token a first preliminary label indicative of the token not being representative of the third attribute type.

12. The method of claim 11, wherein the one or more attribute labelling modules comprises a third label detector configured to detect a character stream in a first proximal

token to the candidate token that corresponds to a third label of the third attribute type, and responsive to determining that the first proximal token comprises the character stream corresponding to the third label, assigning the candidate token a second preliminary label of the third attribute type, and responsive to determining that the first proximal token does not comprise the character stream corresponding to the third label, assigning the candidate token a second preliminary label indicative of the candidate token not being representative of the third attribute type.

13. The method of claim 11 or claim 12, wherein the one or more attribute labelling modules comprises a fourth label detector configured to detect a character stream in a second proximal token, proximal to the candidate token and/or the first proximal token, that corresponds to a fourth label of the third attribute type, and responsive to determining that the second proximal token comprises the character stream corresponding to the fourth label, assigning the candidate token a third preliminary label of the third attribute type, and responsive to determining that the second proximal token does not comprise the character stream corresponding to the fourth label, assigning the candidate token a third preliminary label indicative of the candidate token not being representative of the third attribute type.

14. The method of any one of the preceding claims, wherein responsive to the confidence value for each attribute label type of the set of attribute label types for a given token falling short of a threshold confidence level, removing the token from the subset of tokens before determining the set of refined labels for the respective document.

15. The method of any one of the preceding claims, wherein determining the set of refined labels for each document based on the confidence values comprises determining, for each of the one or more attribute types, a token of the subset of tokens having a highest confidence value, and assigning the attribute type the value of the respective token.

16. The method of any one of the preceding claims, wherein the one or more attribute types comprise one or more of: (i) an entity name, (ii) a business name; (iii) a company name; (iv) financial details; (v) a financial institution branch code; and (vi) a financial institution account number.

17. The method of any one of the preceding claims, wherein the documents of the first dataset are accounting documents and a class of the accounting 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.

20. The computer-readable storage medium of claim 21, wherein the computer-readable storage medium is a non-transient computer-readable storage medium.