

CLAIMS

What is claimed is:

1. A system comprising one or more computers and one or more storage devices storing instructions that when executed by the one or more computers cause the one or more computers to implement:

a master AI subsystem for training a machine learning processing pipeline, the machine learning processing pipeline comprising a plurality of machine learning components to process an input document, wherein the plurality of machine learning components comprise a data balancing component, and wherein the data balancing component comprises at least two of an informative down-sampling implementation, a down-up sampling implementation, or a minority-class-oriented active sampling implementation,

wherein each of at least two of the plurality of machine learning components is provided with at least two candidate implementations, and

wherein the master AI subsystem is to train the machine learning processing pipeline by selectively deploying the at least two candidate implementations for each of the at least two of the plurality of machine learning components.

2. The system of claim 1, wherein the plurality of machine learning components further comprise a file conversion component, a data grouping component, a domain look-up component, a document parser, a tokenization component, a feature generation component, a hyper-parameter selection component, a reference search component, and a standardization component.

3. The system of claim 2, wherein the file conversion component provides a plurality of file converters that each converts the input document from a source file type to a target file type, and wherein the master AI subsystem is to select one of the plurality of file converters based on the source type.

4. The system of claim 3, wherein the data grouping component is to:
identify, in the input document, one or more data items corresponding to a same meaning but in different formats; and
group the one or more data items into a common group, wherein the master AI subsystem is to process data items according to groups.

5. The system of claim 4, wherein the master AI subsystem is to select one of the at least two of the informative down-sampling implementation, the down-up sampling implementation, or the minority-class-oriented active sampling implementation based on a test run on the data items in the input document using each of the at least two of the informative down-sampling implementation, the down-up sampling implementation, or the minority-class-oriented active sampling implementation.

6. The system of claim 5, wherein the domain look-up component comprises a plurality of domain knowledge bases, and wherein the master AI subsystem is to receive the data items of the input document and to look up the plurality of domain knowledge bases based on the received data items.

7. The system of claim 6, wherein the document parser is to generate a document object model (DOM) tree based on the data items of the input document, and wherein each node of the DOM tree comprises one of a sentence or a paragraph.

8. The system of claim 7, wherein the tokenization component comprises a universal tokenizer and an entropy-based on-demand tokenizer for generating tokens, and wherein the master AI subsystem is to:

select one of the universal tokenizer or the entropy-based on-demand tokenizer based on the data items; and

tokenize the nodes of the DOM tree using selected one of the universal tokenizer or the entropy-based on-demand tokenizer.

9. The system of claim 8, wherein the feature generation component comprises a universal natural language processing (NLP) feature generator to generate one of universal NLP features or a hierarchy of NLP features using the tokens, wherein the hierarchy of features comprise a high level features representing domain knowledges and a low level features representing NLP characteristics, and wherein the master AI subsystem is to selectively use one of the universal NLP features or the hierarchy of NLP features.

10. The system of claim 9, wherein the hyper-parameter selection component provides a plurality of machine learning algorithms, and wherein the master AI subsystem is to selectively use at least one of the plurality of machine learning algorithms based on the data items, and adjust parameters specifying the at least one of the plurality of machine learning algorithms in a training process using the data items.

11. The system of claim 10, wherein the reference search component provides a plurality of data input sources, and wherein the master AI subsystem is to cross-validate validity of the data items from the plurality of data input sources.

12. The system of claim 11, wherein the standardization component provides a plurality of post-processing methods, and wherein the master AI subsystem is to selectively use one of the plurality of post-processing methods to reformat the data items to an output format.

13. A method for training a machine learning system, the method comprising:
providing a machine learning processing pipeline comprising a plurality of machine learning components to process an input document, wherein each of at least two of the plurality of machine learning components is provided with at least two candidate implementations, wherein the plurality of machine learning component comprise a data balancing component, and wherein the data balancing component comprises at least two of an informative down-sampling implementation, a down-up sampling implementation, or a minority-class-oriented active sampling implementation; and
training the machine learning processing pipeline by selectively deploying the at least two candidate implementations for each of the at least two of the plurality of machine learning components.

14. The method of claim 13, wherein the plurality of machine learning components further comprise a file conversion component, a data grouping component, a domain look-up component, a document parser, a tokenization component, a feature generation component, a hyper-parameter selection component, a reference search component, and a standardization component.

15. The method of claim 14, the method further comprising:
selecting one of the at least two of the informative down-sampling implementation, the down-up sampling implementation, or the minority-class-oriented active sampling implementation based on a test run on the data items in the input document using each of the at least two of the informative down-sampling implementation, the down-up sampling implementation, or the minority-class-oriented active sampling implementation.

16. The method of claim 15, wherein the document parser is to generate a document object model (DOM) tree based on the data items of the input document, and wherein each node of the DOM tree comprises one of a sentence or a paragraph.

17. The method of claim 16, wherein the tokenization component comprises a universal tokenizer and an entropy-based on-demand tokenizer for generating tokens, the method further comprising:

selecting one of the universal tokenizer or the entropy-based on-demand tokenizer based on the data items; and

tokenizing the nodes of the DOM tree using selected one of the universal tokenizer or the entropy-based on-demand tokenizer.

18. The method of claim 17, wherein the feature generation component comprises a universal natural language processing (NLP) feature generator to generate one of universal NLP features or a hierarchy of NLP features using the tokens, wherein the hierarchy of features comprise a high level features representing domain knowledges and a low level features representing NLP characteristics, and wherein the master AI subsystem is to selectively use one of the universal NLP features or the hierarchy of NLP features.

19. The method of claim 18, wherein the hyper-parameter selection component provides a plurality of machine learning algorithms, and wherein the master AI subsystem is to selectively use at least one of the plurality of machine learning algorithms based on the data items, and adjust parameters specifying the at least one of the plurality of machine learning algorithms in a training process using the data items.

20. A machine-readable non-transitory storage media encoded with instructions that, when executed by one or more computers, cause the one or more computer to train a machine learning system, to:

provide a machine learning processing pipeline comprising a plurality of machine learning components to process an input document, wherein each of at least two of the plurality of machine learning components is provided with at least two candidate implementations, wherein the plurality of machine learning component comprise a data balancing component, and wherein the data balancing component comprises at least two of an informative down-sampling implementation, a down-up sampling implementation, or a minority-class-oriented active sampling implementation; and

train the machine learning processing pipeline by selectively deploying the at least two candidate implementations for each of the at least two of the plurality of machine learning components.