What is claimed is:

1. A computer-implemented method comprising:

receiving a set of documents comprising a set of positive labeled documents, each positive labeled document containing or discussing a topic, and a set of unlabeled documents that are not designated as positive or negative for discussing a topic;

processing the set of documents to remove duplicate documents and noise within the documents;

generating a preliminary model for each topic of a set of topics of the set of documents;

for each topic, training the preliminary model to determine an additional set of positive and negative documents based on the set of documents, wherein training comprises in part identifying negative documents from the set of documents using the set of positive labeled documents and the set of unlabeled documents, the negative documents being identified as not containing or discussing the topic; and

generating a customized classification model using the trained preliminary model and the additional set of positive and negative documents, wherein generating the customized classification model includes applying the additional set of positive and negative documents to the preliminary model.

2. The method of claim 1, wherein training the preliminary model further comprises applying a nearest neighbor linking algorithm.

3. The method of claim 1 or 2, further comprising:

testing the customized classification model on a predetermined test set of documents and presenting to the user precision and recall values associated with the testing.

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

providing the customized classification model along with at least a precision and recall value.

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

tuning the preliminary model and the customized classification model using one of triage rules, features selections, user feedback or combinations thereof;

generating a tuned preliminary model and a tuned customized classification model; and

providing the tuned customized classification model along with at least an updated precision or recall value.

6. The method of claim 5, wherein tuning using triage rules comprises:

providing a list of phrases to a user interface; and

applying a selected list of phrases to the set of documents in response to a selection of inclusion or exclusion, wherein a document containing one or more phrases from the list of phrases are indicative of the selected topic.

7. The method of claim 5, wherein tuning using features selection comprises providing a features list to a user interface, the features list further comprising a list of phrases which characterize documents associated with topic, applying a selected list of features to the set of documents in response to a selection of inclusion or exclusion.

8. The method of claim 5, wherein tuning using user feedback comprises providing a selectable dialog box on a user interface to indicate confirmation of a document relating to a topic.

9. The method of claim 7, further comprising weighting the features list based on a ranking.

10. The method of any one of claims 5 to 9, further comprising publishing the tuned customized classification model.

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

validating the set of topics using the set of positive labeled documents and the set of negative documents.

12. The method of claim 11, further comprising determining by a user interface that a topic has sufficient documents for validating.

13. A system comprising:

a classification server including a processor and memory storing instructions that, in response to receiving a set of documents comprising a set of positive labeled documents, each positive labeled document containing or discussing a topic, and a set of unlabeled documents that are not designated as positive or negative for discussing a topic, cause the processor to:

process the set of documents to remove duplicate documents and noise within the documents;

generate a preliminary model for each topic of a set of topics of the set of documents;

for each topic, train the preliminary model to determine an additional set of positive and negative documents based on the set of documents, wherein training comprises in part identifying negative documents from the set of documents using the set of positive labeled documents and the set of unlabeled documents, the negative documents being identified as not containing or discussing the topic; and

generate a customized classification model using the trained preliminary model and the additional set of positive and negative documents, wherein

generating the customized classification model includes applying the additional set of positive and negative documents to the preliminary model.

14. The system of claim 13, wherein the preliminary model applies a nearest neighbor linking algorithm.

15. The system of claim 13 or 14, wherein the memory storing instructions that, in response to receiving the set of documents comprising the set of positive labeled documents and the set of unlabeled documents, cause the processor to:

test the customized classification model on a predetermined test set of documents and presenting to the user the precision and recall values associated with the testing.

16. The system of any one of claims 13 to 15, wherein the memory storing instructions that, in response to receiving the set of documents comprising the set of positive labeled documents and the set of unlabeled documents, cause the processor to:

provide the customized classification model along with at least a precision or recall value.

17. The system of any one of claims 13 to 16, wherein the memory storing instructions that, in response to receiving the set of documents

comprising the set of positive labeled documents and the set of unlabeled documents, cause the processor to:

tune the preliminary model and the customized classification model using one of triage rules, features selections, user feedback or combinations thereof;

generate a tuned preliminary model and a tuned customized classification model; and

provide the tuned customized classification model along with at least the precision and recall value.

18. The system of claim 17, wherein triage rules provides a list of phrases to a user interface and applies a selected list of phrases to the set of documents in response to a selection of inclusion or exclusion, wherein a document containing one or more phrases from the list of phrases are indicative of the selected topic.

19. The system of claim 17, wherein features selections provides a features list to a user interface, the features list further comprises a list of phrases which characterize documents associated with topic, and applies a selected list of features to the set of documents in response to a selection of inclusion or exclusion.

20. The system of claim 17, wherein user feedback provides a selectable dialog box on a user interface to indicate confirmation of a document relating to a topic.

21. The system of claim 19 wherein the features list is weighted based on a ranking.

22. The system of any one of claims 17 to 21, wherein the memory storing instructions that, in response to receiving the set of documents comprising the set of positive labeled documents and the set of unlabeled documents, cause the processor to:

publish the tuned customized classification model.

23. The system of any one of claims 13 to 22, wherein the memory storing instructions that, in response to receiving the set of documents comprising the set of positive labeled documents and the set of unlabeled documents, cause the processor to:

validate the set of topics using the set of positive labeled documents and the set of negative documents.

24. The system of claim 23, wherein the memory storing instructions that, in response to receiving the set of documents comprising the set of positive labeled documents and the set of unlabeled documents, cause the processor to:

determine by a user interface that a topic has sufficient documents for validating.