WE1S uses text classification to categorize news articles into different genres.

In addition to using topic modeling (see M-2), WE1S applies various text classification algorithms to its collections of texts. Whereas topic modeling is an unsupervised method of machine learning, text classification involves the supervised training of an algorithm to categorize texts according to pre-specified groups.

Text classification requires humans to name the categories into which they would like to sort texts and provide samples of texts from each category. These samples are the training data. From this data, the classification algorithm learns salient characteristics of the texts in each category. Then when the algorithm is asked to inspect texts it has not seen before, it is able to sort these texts into the appropriate category based on what it learned from the training data. This method allows researchers to classify large masses of texts into broad categories for further analysis.

Like topic modeling, text classification is probabilistic. Classification algorithms work by building models that predict which category a text is likely to belong to. Researchers assess these models using a variety of metrics designed to help them determine how accurate the predictions of a particular model may be.

There are a wide variety of classification algorithms available, and not all of them are appropriate for every classification problem. WE1S uses two different classification algorithms to perform its classifications. We chose which classification algorithm to use based on experimental comparisons of three of the most commonly used algorithms for text classification.

To answer specific research questions, we used text classification to sort news articles into four categories: articles about the humanities, articles about science, obituaries, and event announcements. (WE1S researchers compiled the training data for each category by classifying sample sets of articles by hand.) The below table lists the average accuracy score for each model over 10 runs, along with the kind of classification algorithm used. An accuracy score of 92% for articles about the humanities, for example, means that in 92 out of 100 cases the classification algorithm correctly predicted if an article was about the humanities or not.

Category	**Algorithm**	**Avg. Accuracy (F1) Score**
Articles about the humanities	Logistic regression	92%
Articles about science	Logistic regression	82%
Obituaries	SVM	81%
Event announcements	SVM	73%

Resources

Recommended readings:

* Sebastiani, Fabrizio, "Machine Learning in Automated Text Categorization" (2002).

* Long, Hoyt, and Richard Jean So, "Literary pattern recognition: Modernism between close reading and machine learning" (2016).

* WE1S Bibliography: Text Classification