StoryQ—an Online Environment for Machine Learning of Text Classification

William Finzer[1], Jie Chao[1], Carolyn Rose[2] and Shiyan Jiang[3]
The Concord Consortium[1], Carnegie Mellon University[2] and North Carolina State University[3]

Abstract

The StoryQ environment provides an intuitive graphical user interface for middle and high school students to create features from unstructured text data and train and test classification models using logistic regression. StoryQ runs in a web browser, is free and requires no installation. AI concepts addressed include: features, weights, accuracy, training, bias, error analysis and cross validation. Using the software in conjunction with curriculum currently under development is expected to lead to student understanding of machine learning concepts and workflow; developing the ability to use domain knowledge and basic linguistics to identify, create, analyze, and evaluate features; becoming aware of and appreciating the roles and responsibilities of AI developers;. This paper will consist of an online demo with a brief video walkthrough.

Description of StoryQ

StoryQ is a learning environment in which students engage with text classification as a machine learning problem. It is designed to facilitate understanding through interactive participation with all aspects of the process. Given a training set of labeled texts, learners specify features such as n-grams and parts of speech, and construct features such as specific words or phrases that may occur in a text. Once constructed, features appear in a table along with their frequency of occurrence in each of the classes given in the training dataset.

Students control the model building process by stepping through each iteration of a logistic regression optimization in which weights are shown for each feature and probabilities computed for each text. At any point they can run the iteration forward to completion with computed values for accuracy and kappa, predicted labels for texts, and final weights of features. To analyze results students construct confusion matrix graphs, plots of probability distributions and feature weight distributions.

A novel feature of StoryQ is the interactive display of texts and features made possible by StoryQ's embedding in the Common Online Data Analysis Platform (CODAP). For example, when the learner clicks on a feature in the feature table, all the texts that contain that feature are displayed in a separate area with the feature highlighted and the texts classified according to agreement between original and predicted labels as shown in Figure 1.

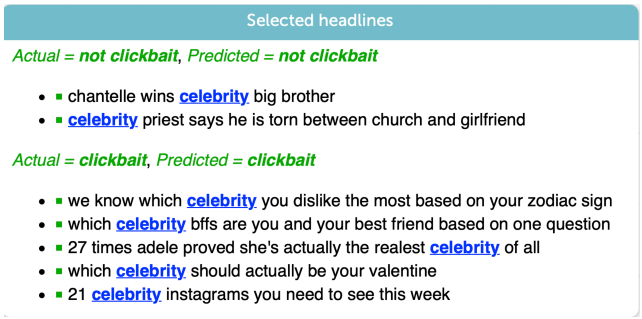

Figure 1. From a training set of headlines classified by whether they are clickbait or not clickbait, the user has built a model with the word "celebrity" as a feature. When clicked all the headlines that contain it are displayed.

Linked selection also applies to points in graphs that represent texts or features. These and other tools allow students to progressively improve their models.

Acknowledgments

This material is based upon work supported by the National Science Foundation under: The Concord Consortium Award ID: 1435470 and The Concord Consortium Award ID: 1949110. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Reference

The Common Online Data Analysis Platform (CODAP), The Concord Consortium, 2020