Classifying news articles based on headlines & short descriptions

Syndicated news websites often draw content from a range of sources, and use recommendation algorithms to recommend content to consumers based on categories of interest. However, the new outlets which the draw their content from may use different sets of labels for categorizing articles, and/or the labels may not be captured by the third-party website. Having a native model which automatically classifies articles can obviate these problems. We therefore built a model which classifies articles based on their headlines and subheadings.

We used a dataset of approximately 130k headlines and short descriptions, spanning 12 news categories. Reserving 20% of this for our test set, we used the remaining 80% of the data for training and evaluating three possible models (a boosted decision tree model, a logistic regression model, and a convolution neural network). Overall accuracy was chosen as the primary metric for comparing models, as we determined that for the purpose of a recommendation algorithm, accurately classifying most articles was more relevant than correctly identifying as many instances of a particular category as possible. However, we also decided to use the F1 scores of the individual categories as a secondary metric to give us insight into the strengths and weaknesses of the models.

The most successful of the three models was the convolutional neural network, which had an accuracy score of 82%. The average weighted F1 score was also 82%, but by category this score ranged from 50% to 88%. Notably, low F1 scores were associated with categories which were under-represented in the dataset. Training the model on more articles of these categories might therefore lead to better performance. However, when looking at the set of misclassified headlines, it became apparent that there is an inherent limitation, namely that many articles conceptually fall under multiple categories, and so a multi-label system of classification may be more appropriate.