News Classification Model

Team Birch
Ayan Maiti, Hazel Mitchley, Joshua Schroeder

https://github.com/HJMitchley/News-Classification-Project

OVERVIEW

Objective:
Build a model that categorizes news articles based on their headlines & subheadlines.

Stakeholders:
3rd party websites providing syndicated news content, with recommendation algorithms based on article categories.

Not all sources provide categories for their articles, and/or exact labels may vary across different websites. A classification model that is domestic to the 3rd party website can obviate these problems

Key Performance Indicators

Main KPI: Overall accuracy

For recommendation algorithms, the aim is not to recommend every possible instance of category X (recall), but rather to ensure that if an article is recommended on the basis of being category X, it actually is category X

Secondary KPI: F1 score for each category

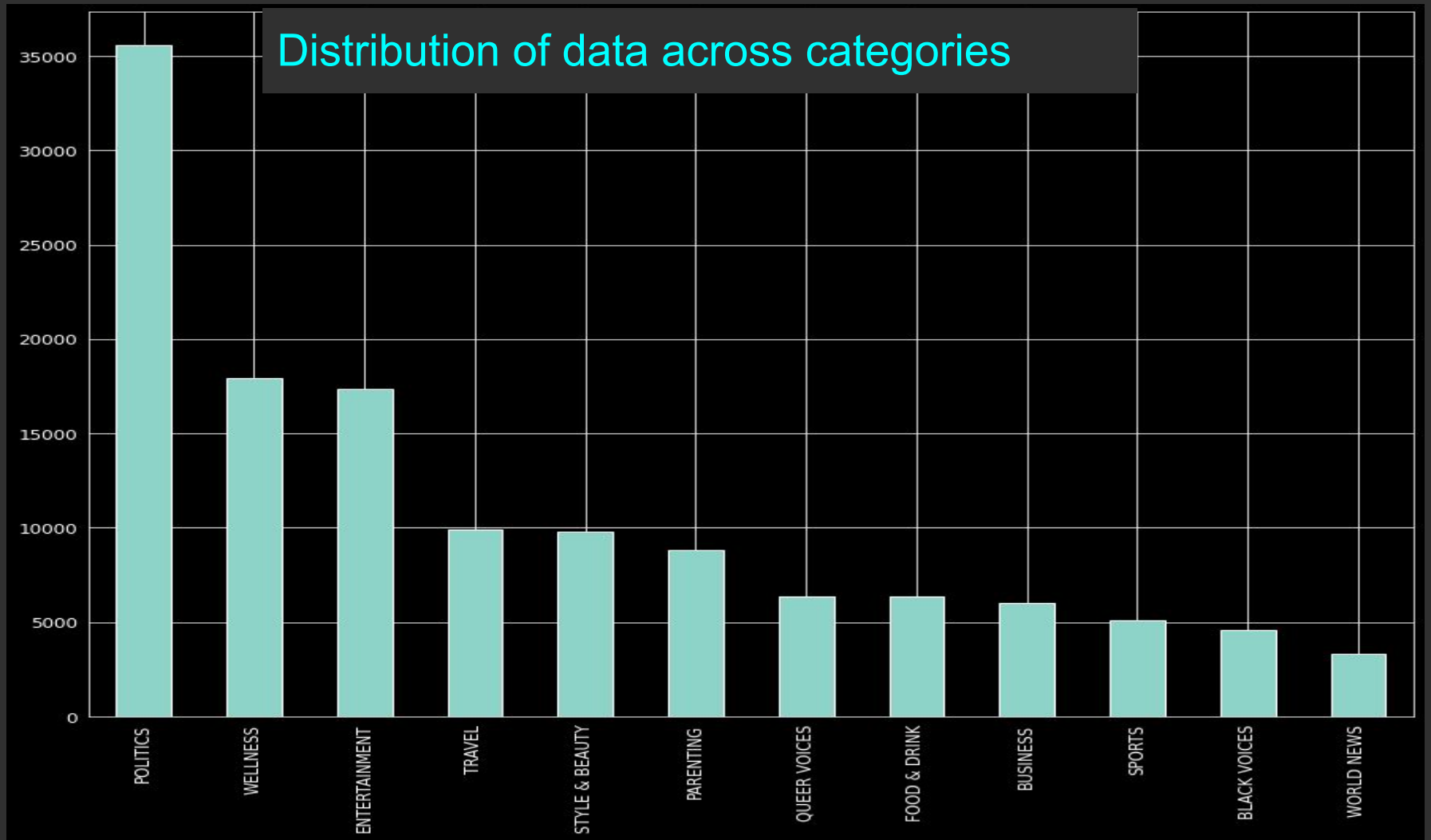
We use this metric to understand what categories the model does better and worse on

The Dataset

- Misra, Rishabh. "News Category Dataset." arXiv preprint arXiv:2209.11429 (2022). https://www.kaggle.com/datasets/rmisra/news-category-dataset
- Approximately 210K articles headline and subheadlines
- 42 categories

Reduced dataset:

- Picked 12 top categories (131 023 data points)
- Train ~ test split: 80~20
- Train_train ~ validation split: 80~20
- Preprocessing: lemmatization, removing punctuation and stop words


Distribution of data across categories
35000
30000
25000
20000
15000
10000
5000
0
POLITICS
WELLNESS
ENTERTAINMENT
TRAVEL
STYLE & BEAUTY
PARENTING
QUEER VOICES
FOOD & DRINK
BUSINESS
SPORTS
BLACK VOICES
WORLD NEWS

Models

- **Boosted decision tree**
  Validation set accuracy: 61% Weighted avg F1 Score: 70%

- **Logistic regression model**
  Validation set accuracy: 79% Weighted avg F1 Score: 78%

- **Convolutional Neural Network**
  Validation set accuracy: 82% Weighted avg F1 Score: 82%

CNN model basic setup

- Convolutional neural network:
  - Words are mapped to numbers with a dictionary; inputs are sequences padded with zeros to be equal length
  - Embedding layer converts words into n-dimensional vectors
  - Convolutional layer scans k-length phrases for any of m features (ReLU activation)
  - Max pooling over the convolutional layer returns 1 output for each of the m features
  - Dense softmax layer converts the m features into probabilities for the 12 classes
- Loss function was binary cross-entropy
- Number of epochs to train chosen with a holdout validation set

CNN model's performance on the test set

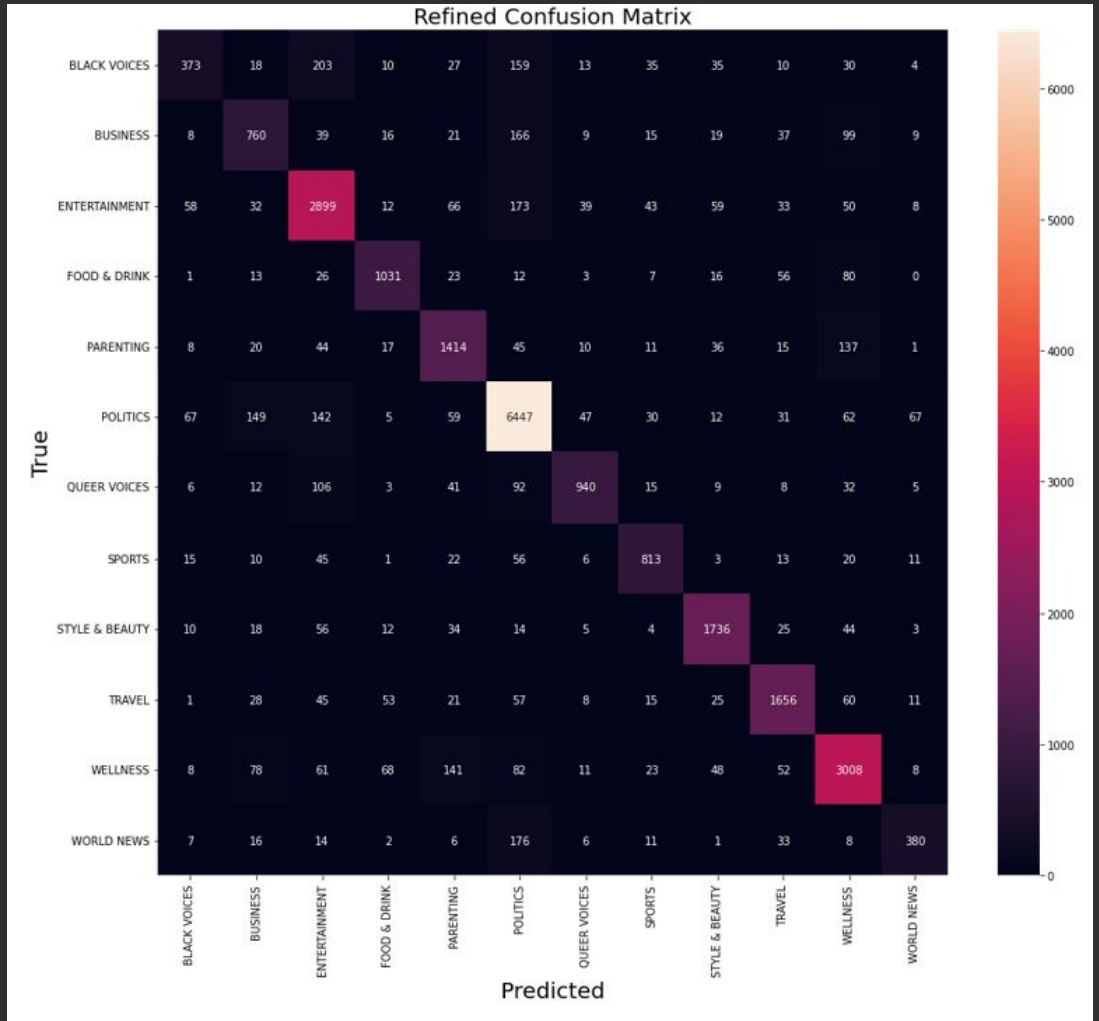
Overall accuracy: 82%

Weighted F1 Score: 82%

Categories with higher representation had higher F1 scores (eg politics, wellness)

Categories with lower representation had lower F1 scores (eg black voices, world news)

	precision	recall	f1-score	support
BLACK VOICES	0.66	0.41	0.50	917
BUSINESS	0.66	0.63	0.65	1198
ENTERTAINMENT	0.79	0.83	0.81	3472
FOOD & DRINK	0.84	0.81	0.83	1268
PARENTING	0.75	0.80	0.78	1758
POLITICS	0.86	0.91	0.88	7118
QUEER VOICES	0.86	0.74	0.79	1269
SPORTS	0.80	0.80	0.80	1015
STYLE & BEAUTY	0.87	0.89	0.88	1961
TRAVEL	0.84	0.84	0.84	1980
WELLNESS	0.83	0.84	0.83	3588
WORLD NEWS	0.75	0.58	0.65	660
micro avg	0.82	0.82	0.82	26204
macro avg	0.79	0.76	0.77	26204
weighted avg	0.82	0.82	0.82	26204
samples avg	0.82	0.82	0.82	26204


Refined Confusion Matrix
BLACK VOICES 373 18 203 10 27 159 13 35 35 10 30 4
BUSINESS 8 760 39 16 21 166 9 15 19 37 99 9
ENTERTAINMENT 58 32 2899 12 66 173 39 43 59 33 50 8
FOOD & DRINK 1 13 26 1031 23 12 3 7 16 56 80 0
PARENTING 8 20 44 17 1414 45 10 11 36 15 137 1
POLITICS 67 149 142 5 59 6447 47 30 12 31 62 67
QUEER VOICES 6 12 106 3 41 92 940 15 9 8 32 5
SPORTS 15 10 45 1 22 56 6 813 3 13 20 11
STYLE & BEAUTY 10 18 56 12 34 14 5 4 1736 25 44 3
TRAVEL 1 28 45 53 21 57 8 15 25 1656 60 11
WELLNESS 8 78 61 68 141 82 11 23 48 52 3008 8
WORLD NEWS 7 16 14 2 6 176 6 11 1 33 8 380
True
BLACK VOICES
BUSINESS
ENTERTAINMENT
FOOD & DRINK
PARENTING
POLITICS
QUEER VOICES
SPORTS
STYLE & BEAUTY
TRAVEL
WELLNESS
WORLD NEWS
Predicted
6000
5000
4000
3000
2000
1000
0

Examples of misclassification errors

1) Multiple labels may be appropriate

   Why So Many Whites Think They Are Discriminated Against. For decades, the GOP has sold their constituents a narrative of white victimhood.

   Label: BLACK VOICES, Model: POLITICS

2) True label is unintuitive – not apparent from headline & short description, even for humans

   Why Some People In America's Salad Bowl Are Eating Junk Food. Even in California's bountiful Central Valley, prices can put healthy fruit and vegetables out of reach.

   Label: POLITICS, Model: FOOD & DRINK

3) The model is just wrong

   Tawana Jackson, Social Worker, On Why She Puts Listerine On Her Feet. And it turns out we were right -- she had tons of advice to dispense. Jackson was well-versed on how to get rid of pimples

   Label: STYLE & BEAUTY, Model: BUSINESS

Ideas for how to improve classification

- More data for under-represented categories?
    - Inherent limitation: small categories with some conceptual overlap with a larger category tended to have the worst F1 scores.
    - perhaps multiple class labels can be simultaneously applicable.
- Explore other neural net architectures, such as RNN’s or transformer models

CNN Model

Hyperparameter tuning/details:

- L1, L2, and L1L2 regularization did not seem to improve model results
- Higher embedding dimension improved results slightly up to 256 dimension representations
- 2, 4, and 6 kernel sizes on the CNN (slightly) outperformed other arrangements of kernels
- Increasing number of convolution feature to about 256 improved results, but further features did not yield higher accuracy
- Lowering learning rate to 0.0002 yielded accuracy improvements and more stability; lower learning rates trained more slowly, but were not more accurate.
- Adding more convolutional or dense layers did not improve accuracy.