December, 2021

Classifying News Articles with NLP

John O’Donnell

Purpose

Categorize new articles for our users

- Context:
  - We acquired NewsBlast, and none of their news articles were labeled
  - Need to categorize these texts into our 40 categories here at HuffPost

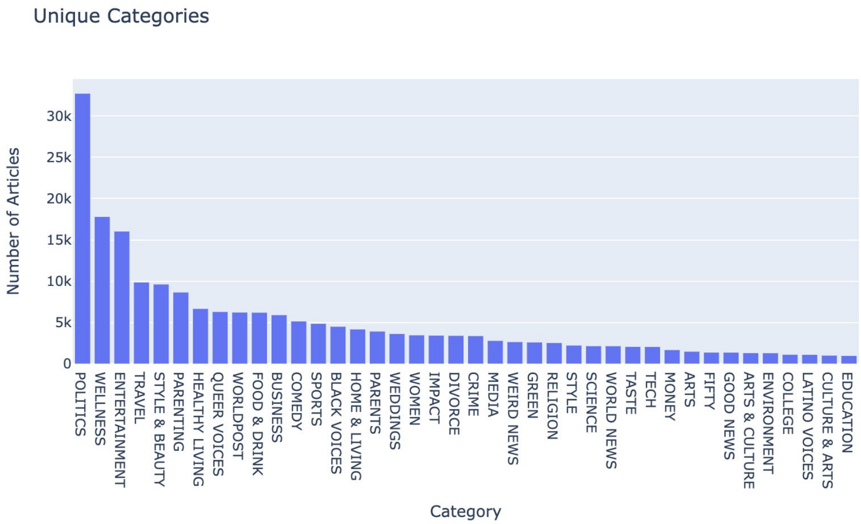
- Given our dataset:
  - Explore high level trends in the data
  - Find relationships between the texts and their categories
  - Generate a model that understands these relationship and can create predictions

- Headwinds:
  - We have 40 different categories, many of which are very similar
  - Many fields are missing, and the data is unstructured

- Solution:
  - Use Natural Language Processing

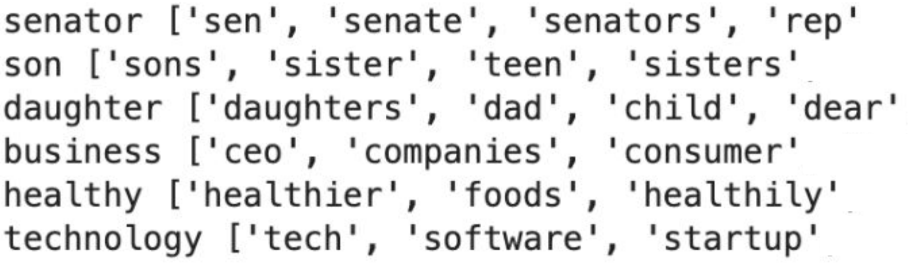
Early Findings

Articles and Text Exploration



- Politics, Wellness, and Entertainment were the most prominent categories in the dataset
- As you can see there are many similar labels
  - Wellness & Healthy Living
  - Style & Style and Beauty
  - Education & College

- A model was built to model the semantic meaning of words



```
senator ['sen', 'senate', 'senators', 'rep'
son ['sons', 'sister', 'teen', 'sisters'
daughter ['daughters', 'dad', 'child', 'dear'
business ['ceo', 'companies', 'consumer'
healthy ['healthier', 'foods', 'healthily'
technology ['tech', 'software', 'startup'
```

Modelling

Two Final Models

- Two models will be shared, each has its own tradeoff

- Model #1 :
  - ~ 78% accurate
  - Simple model, directly interpretable

- Model #2:
  - ~ 80% accurate
  - Complex model, less interpretable

- Recommendation is Model #2:
  - The more complex model is more capable of capturing complex relationships
  - Can still be improved further, less room to grow than model 1
  - Interpretability isn’t very important here, unlikely we will make any business decisions based off this model moving forward

Forward

Next Steps

- Recommendation:
    - Move forward with Model #2
- Next Steps:
    - Continue to tweak Model #2
    - Test out other complex models with different architectures
- Suggestion:
    - Reframe the problem from categories to tags
    - Tags would be more relevant and likely provide more value to our users

Thank You

John O'Donnell