News Category Dataset

Rishabh Misra
UC San Diego
r1misra@eng.ucsd.edu

Abstract. People rely on daily news to know what is happening around the world. In today's world, when the proliferation of fake news is rampant, having a large-scale and high-quality source of authentic news articles with the published category information would be valuable to learning authentic news' Natural Language syntax and semantics. We present a *News Category Dataset* that contains around 200K news headlines from the year 2012 to 2018 obtained from HuffPost. To make it more useful, we have included the source links of the news articles so that more data can be extracted as needed. In this paper, we describe various details about the dataset and potential use cases where it can be used.

1. Motivation

Journalistic trends across the world can provide great insights into what is currently happening in society. The key part of uncovering such trends is the meta-data around what the news articles are about and when they were published. Thus, in the "News Category Dataset" we describe in this paper, we explicitly note the category information and publication date for each news article. In the technical aspect, the dataset with news category information can help folks do various Natural Language tasks like Semantic Tagging, Named Entity Recognition (NER), Word Sense Disambiguation, etc..

2. News Category Dataset

We present[1] a large-scale and high-quality News dataset whose unique contribution is the availability of category information with each article. Table 1 notes all the categories available and the corresponding number of articles from it in the dataset. We have a total

[1] Dataset is available at https://rishabhmisra.github.io/publications/

of 202,372 records in the dataset that subsumes news articles published between 2012 and 2018.

News Category	# of Articles
POLITICS	32739
WELLNESS	17827
ENTERTAINMENT	16058
TRAVEL	9887
STYLE & BEAUTY	9649
PARENTING	8677
HEALTHY LIVING	6694
QUEER VOICES	6314
FOOD & DRINK	6226
BUSINESS	5937
COMEDY	5175
SPORTS	4884
BLACK VOICES	4528
HOME & LIVING	4195
PARENTS	3955
THE WORLDPOST	3664
WEDDINGS	3651
WOMEN	3490
IMPACT	3459
DIVORCE	3426
CRIME	3405
MEDIA	2815
WEIRD NEWS	2670
GREEN	2622
WORLDPOST	2579
RELIGION	2556
STYLE	2254
SCIENCE	2178

WORLD NEWS	2177
TASTE	2096
TECH	2082
MONEY	1707
ARTS	1509
FIFTY	1401
GOOD NEWS	1398
ARTS & CULTURE	1339
ENVIRONMENT	1323
COLLEGE	1144
LATINO VOICES	1129
CULTURE & ARTS	1030
EDUCATION	1004

Table 1: Article count of various news categories in the dataset.

Each record in the dataset consists of the following attributes:

- `category`: category in which the article was published.
- `headline`: the headline of the news article.
- `authors`: list of authors who contributed to the article.
- `link`: link to the original news article.
- `short_description`: Abstract of the news article.
- `date`: publication date of the article.

We include article links corresponding to each headline so that more text regarding the news can be extracted as needed by any Machine Learning task.

3. Data Curation Method

We make use of open-source tools like BeautifulSoup, Selenium, and Chrome Driver to curate the dataset. For collecting data from Huffington Post, we use Huffington Post's archive link as the base. In all the articles presented, we extract their link, category, headline, abstract, authors, and date published using BeautifulSoup API. Once that is done on one page, we simulate a button click action using Selenium to go to the next page and repeat the process. Among all the news categories present on Huffington Post, there were a couple of categories that had a very low article count. In order to publish a high-quality dataset, we remove all the articles from categories that had less than 1000

articles. This is to ensure that any Machine Learning model training on this data is not affected by the data skew. Since the headline text comes from a professional news website and has reasonably good quality (no misspellings, abbreviations, etc.), we did not do any additional pre-processing on it.

4. Reading the Data

Once you download the dataset, you can use the following code snippet to read the data for your machine learning methods:

```
import json

def parse_data(file):
    for l in open(file,'r'):
        yield json.loads(l)

data = list(parse_data('./News_Category_Dataset.json'))
```

5. Exploratory Data Analysis

As a basic exploration, we visualize the distribution of news articles from various years in Figure 1. Please note that since the dataset is being published in mid-2018, we have significantly fewer articles from 2018.

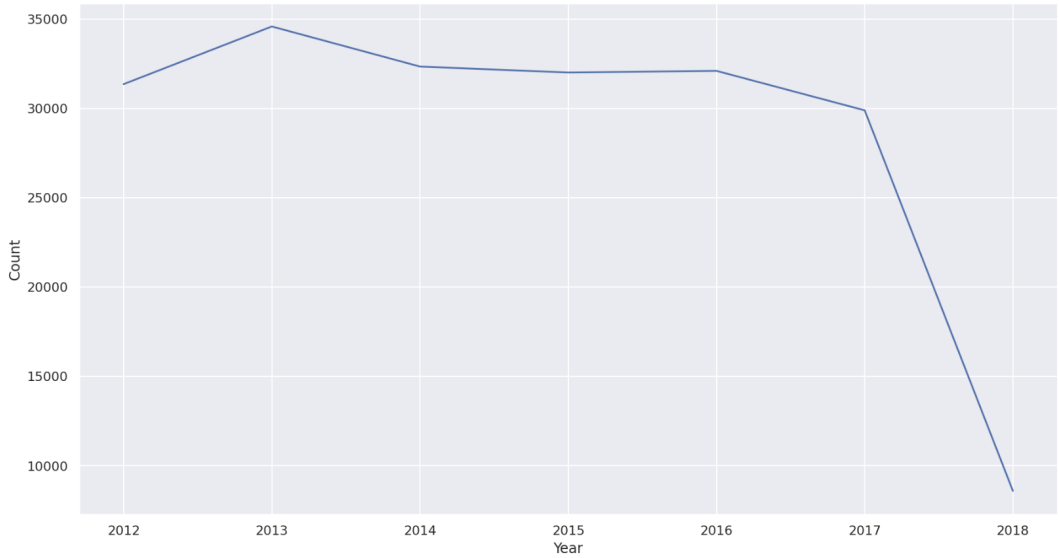

Figure 1: Number of articles published each year.

In figures 2 and 3, we showcase the average number of words present in the headline and short description of articles present in each category, respectively. Although differences in the headline length are not much, short description lengths have more variation.

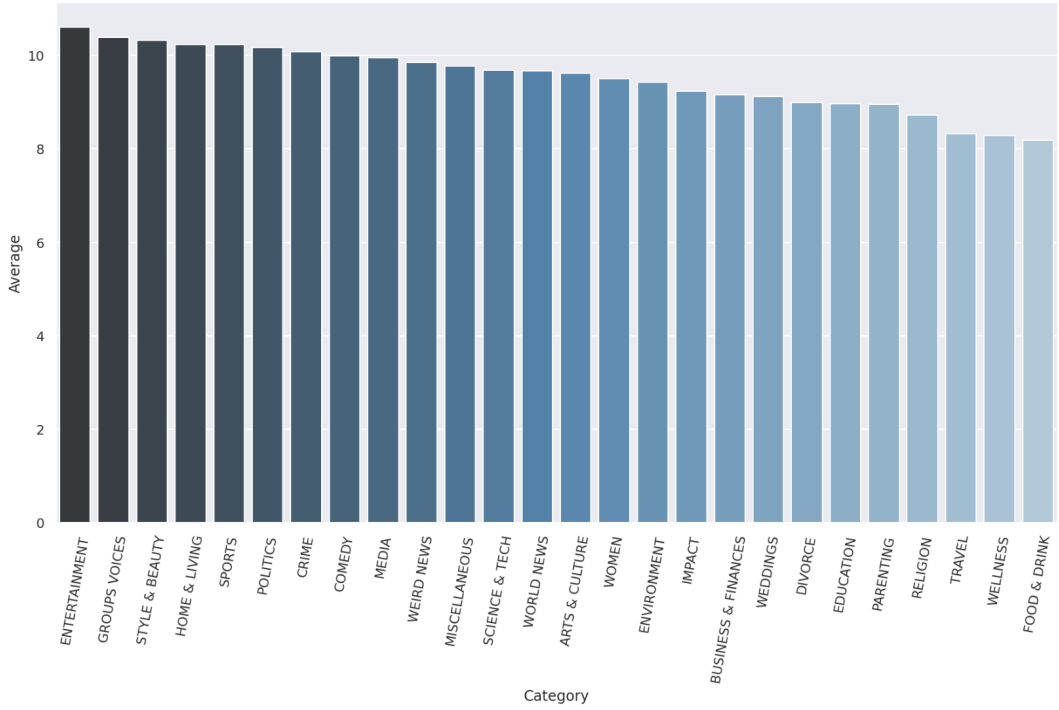

Figure 2: Average headline length for articles under specific categories.

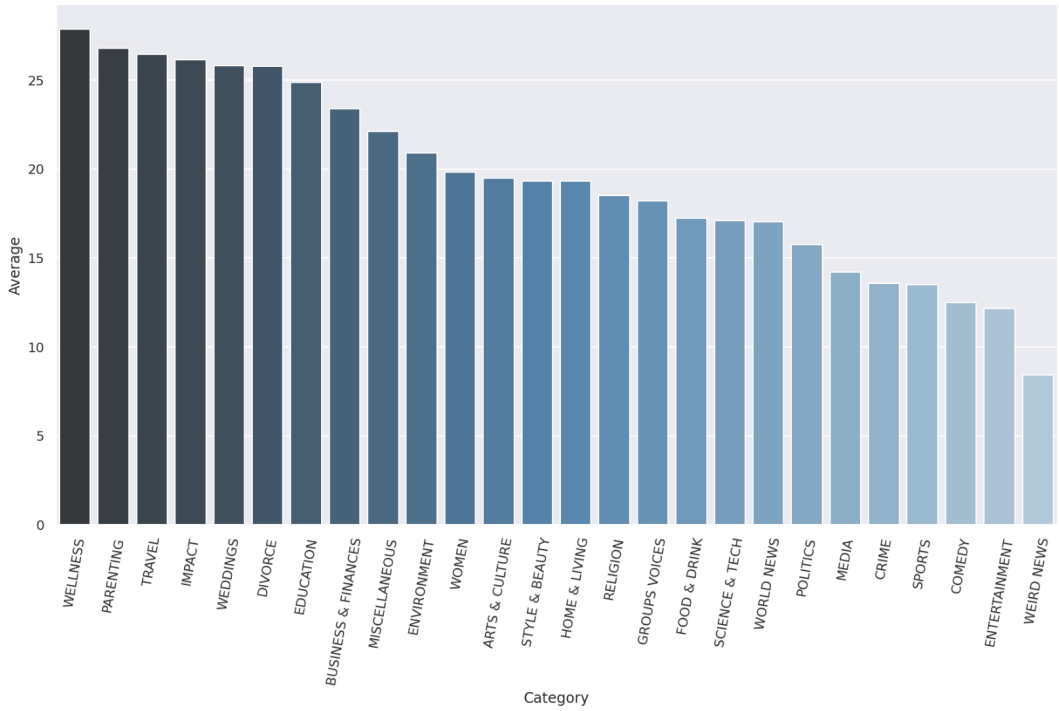

Figure 3: Average abstract length for articles under specific categories.

Lastly, we analyze the distribution of articles published under different categories for two different years 2013 and 2017 in Figure 4. We see a fascinating insight: in 2013, the focus of reporting was a lot on Wellness, parenting, and beauty whereas, in 2017 the focus shifted to Politics, World News, and Entertainment. This clearly shows the shift in what's happening around the world over the years.

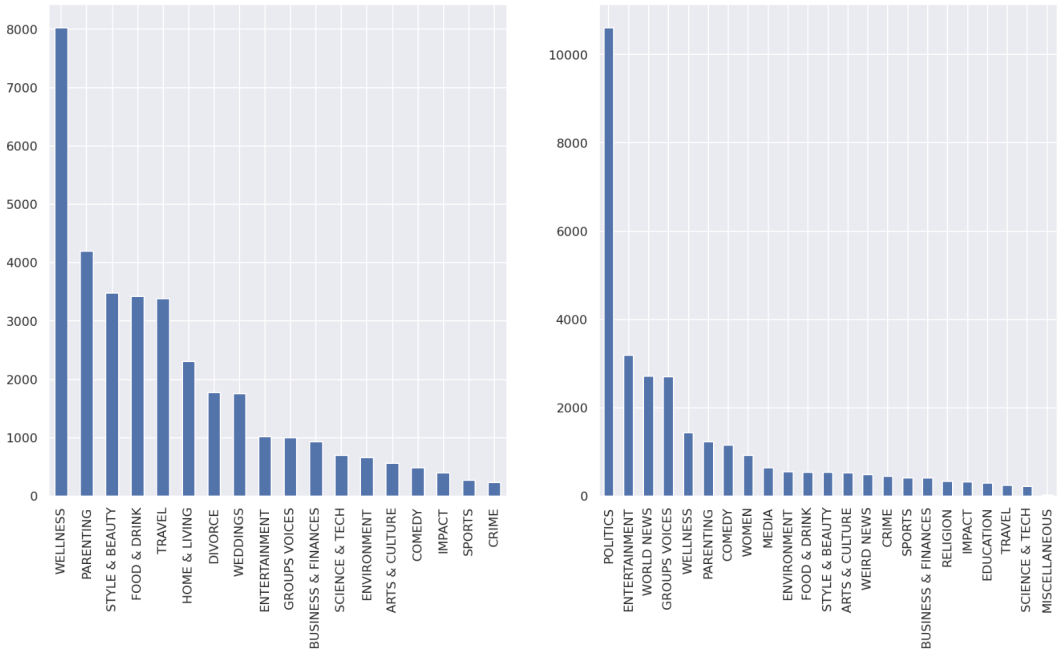

Figure 4: On the left is the distribution of news articles published in 2013. On the right is the distribution of news articles published in 2017.

6. Potential Use Cases

Apart from the evident news category classification task, the "News Category Dataset" can be used for other Natural Language Processing tasks that involve understanding the syntax and semantics of specific categories. The practical value of such methods involves tagging untracked news articles on the web. Furthermore, it provides a good source for doing comparative studies of articles across different categories in terms of writing style, presentation, etc. Another potential use case for this dataset is for training a Machine Learning model to tag specific parts of the text with their tone based on the type of language used.

References

[1] Misra, Rishabh "News Headlines Dataset For Sarcasm Detection," *DOI: 10.13140/RG.2.2.16182.40004*.