

Data Appendix: Discrimination of Real vs. Fake News

Analysis Data File: headlines_clean.csv

Unit of observation:

Each row represents a single news article headline.

Variable: title

Description:

The headline text of the news article.

Type:

Text (string)

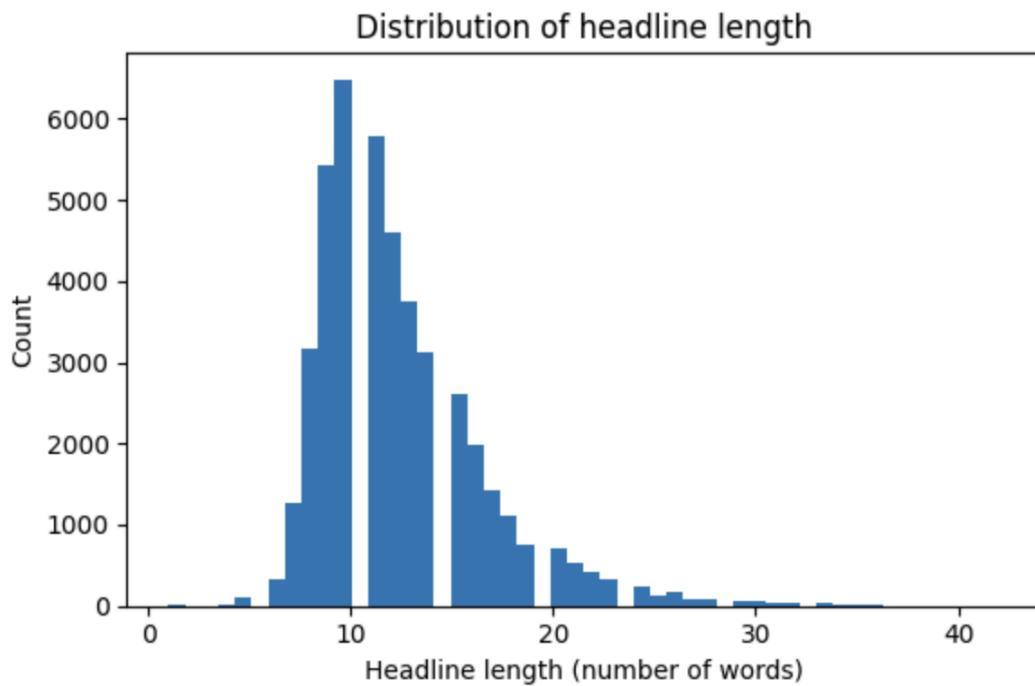
Summary statistics:

Because this variable is text, we summarize it using the number of words in each headline. The table below reports the number of observations, the mean, median, minimum, and maximum headline length (in words).

```
count      44898.000000
mean        12.453472
std         4.111476
min         1.000000
25%        10.000000
50%        11.000000
75%        14.000000
max         42.000000
Name: title_length, dtype: float64
```

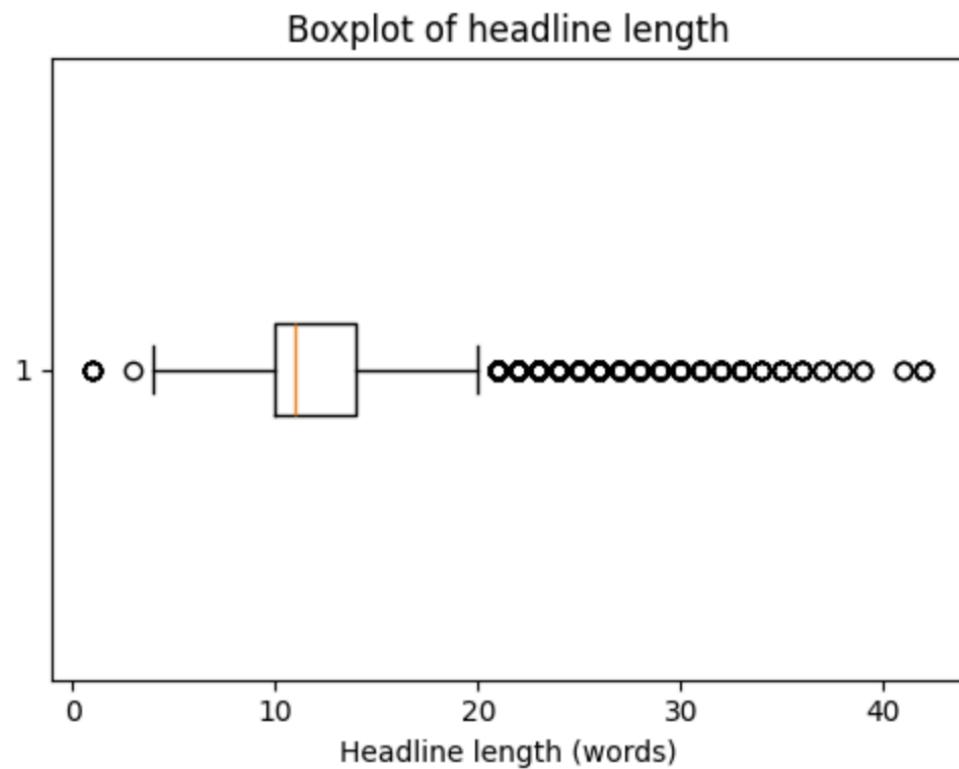
Distribution of headline length:

The histogram shows the distribution of headline lengths across all articles. Most headlines are relatively short, with a smaller number of very long headlines forming a right-skewed tail. This indicates that while typical headlines contain a moderate number of words, a few unusually long headlines exist in the dataset.



Boxplot of headline length:

The boxplot summarizes the spread of headline lengths and highlights the presence of outliers. The median headline length lies near the center of the distribution, and several long-headline outliers are visible, confirming the right-skew observed in the histogram.



Variable: label

Description:

Indicator of whether the headline is real or fake.

Coding scheme:

0 = real

1 = fake

Type:

Categorical (binary)

Summary statistics:

The table below reports the number and proportion of real and fake headlines in the dataset.

```
label
1    0.522985
0    0.477015
Name: proportion, dtype: float64
```

Distribution of labels:

The bar chart shows the number of real and fake headlines in the analysis dataset. The two classes are similar in size, indicating that the dataset is relatively balanced between real and fake headlines.

