Overall Distribution of dataset w/o removing N.A

Bias: 443

BS: 11492

Conspiracy: 430

Fake: 19

Hate: 246

Junksci: 102

Satire: 146

State News: 121

No. of NAs by column:

Author:

Questions:

1. How to remove special punctuation from thread title column?

Check tidyvserse.

1. How to create an ensemble model involving text (NLP) and the other features/analytic methods?
2. Do you know of any efficient webscraping methods? Python script, BeautifulSoup
3. Is it appropriate to simply remove records with NA in them?

(try to build a model without the text features?).

1. What is the appropriate format for the date type columns (“published”, “crawled”, etc.)

Just store as timestamp format.

1. How to predict rare classes? [refer to dist. of dep. variable]
2. What should be done with the social networking features (since they’re 0 for most of the observations).

Check if significant or not.

TODO;

Filter the data and remove NA values [done]