Cleaning the dataset:

Removing all non-English languages: We used a library to check whether language is in English and dropped all who weren’t. We only want to count English words.

Cleaning up html tags: Used a regex and pandas replace method to find and remove all html tags.

Removing punctuation: removed punctuation to count the words more precisely

Metadata usage:

There is a label that marks the article reliable or not. This will be useful when we classify.