



Data Glacier

Your Deep Learning Partner

News Classifier EDA

Chow Jun Wei and Esraa Sultan

Group name: Jun and Esraa

24 November 2021

Required Team member details

Name	Chow Jun Wei	Esraa Sultan
Email	Jun.chow.18@ucl.ac.uk	Alzahrani.Esraa1@gmail.com
Country	Brunei Darussalam	Saudi Arabia
College/Company	NIL	Esri Saudi Arabia
Specialization	Data Science, NLP, Data Analyst	Data Science

Problem Definition

- To classify a (group of) sentence(s) into their corresponding category. We have 20 categories.

Comp.graph hics	Comp.os.ms- windows.m isc	Comp.sys.ibm.pc.hardw are	Comp.sys.m ac.hardwar e
Comp.wind ows.x	Rec.autos	Rec.motorc ycles	Rec.sport.b aseball
Rec.sport.h ockey	Sci.crypt	Sci.electron ics	Sci.med
Sci.space	Misc.forsale	Talk.politics .misc	Talk.politics .guns
Talk.politics .mideast	Talk.religion .misc	Alt.atheism	Soc.religion .christian

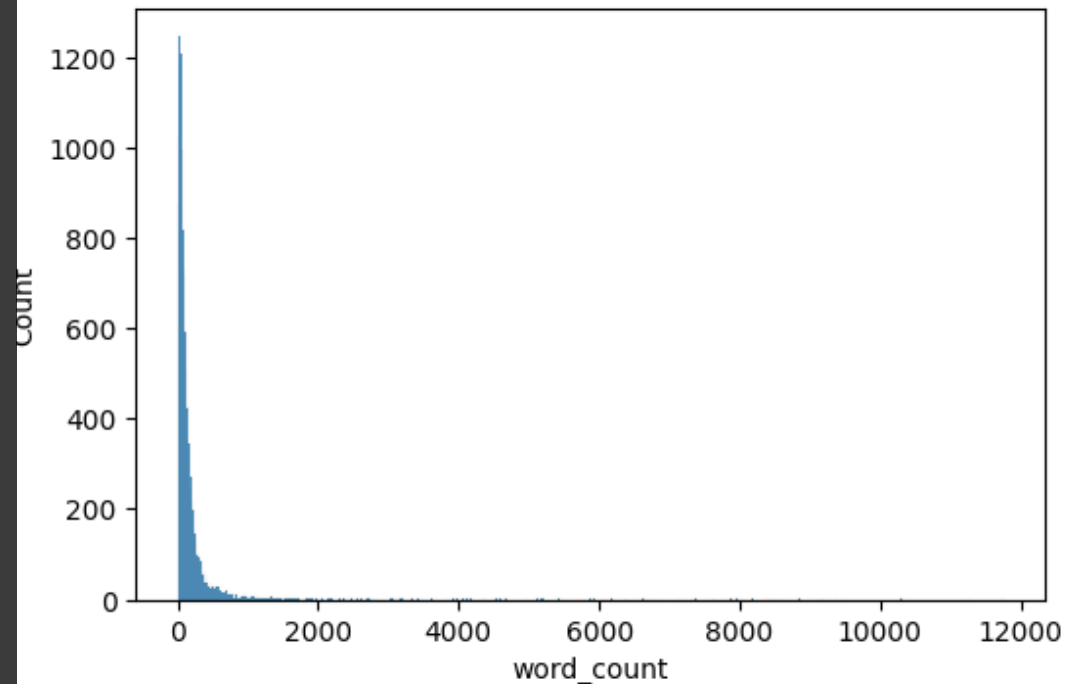
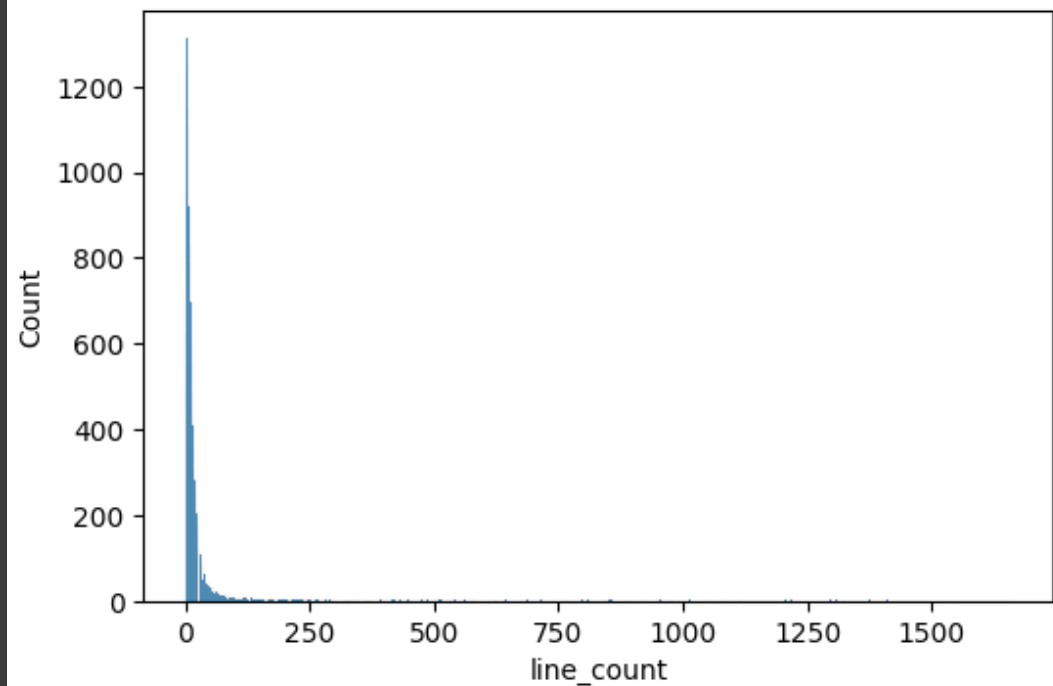
Purpose of this presentation

To show how the data looks like, whenever possible. We **do not** yet show any recommendations from this presentation.

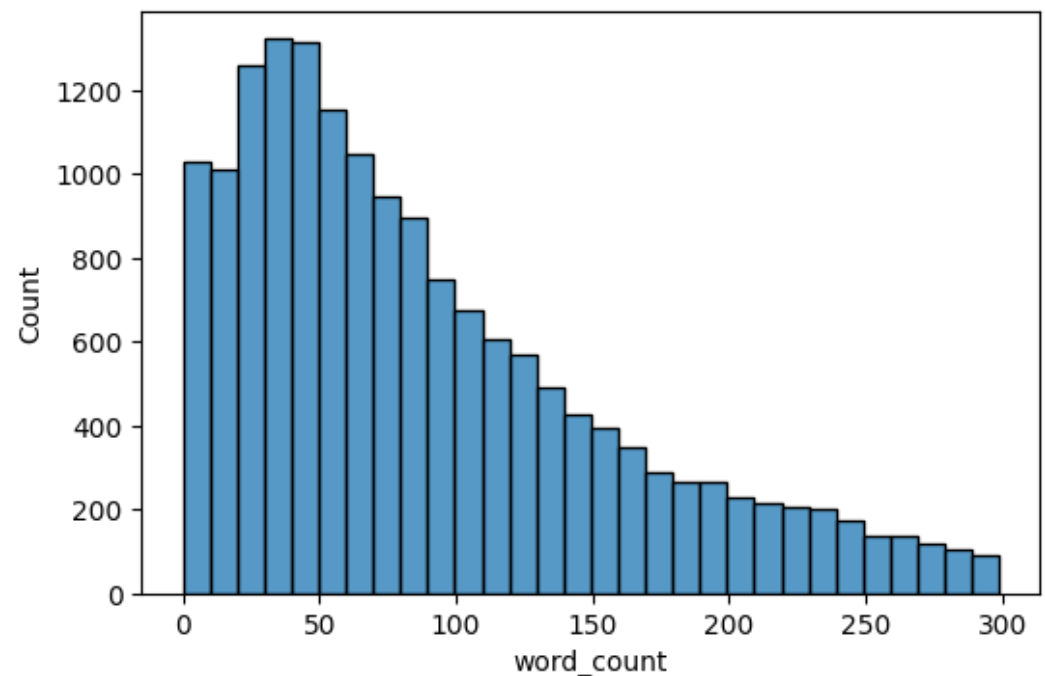
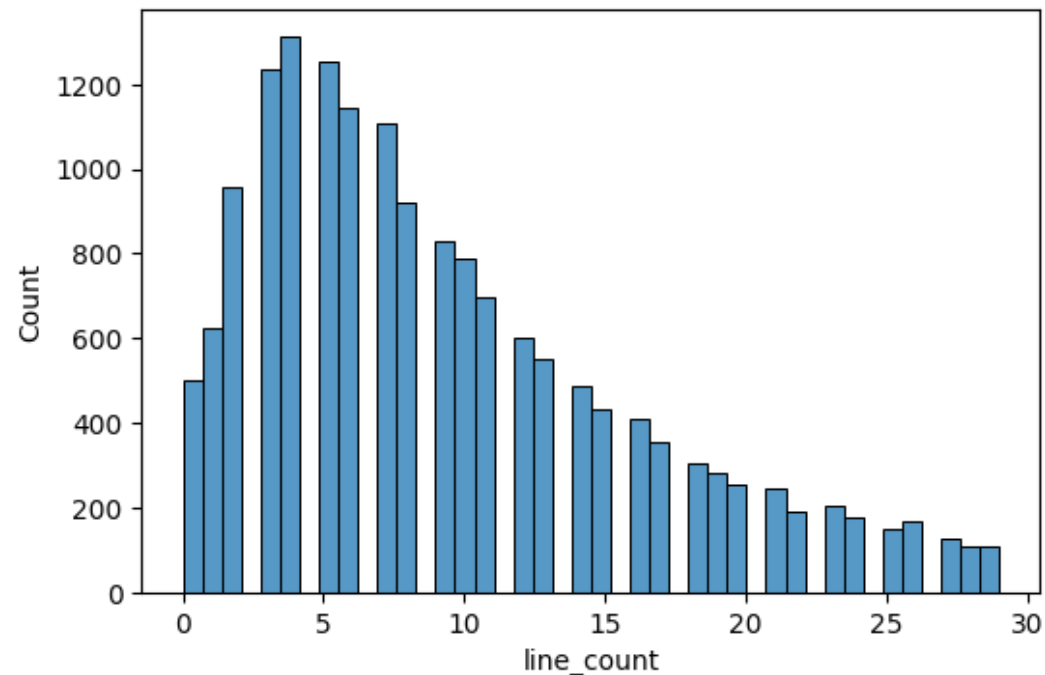
The next few slides
contain information on
before we clean the data

Line number count and word count

Most contains only a few sentences;
but some are very long

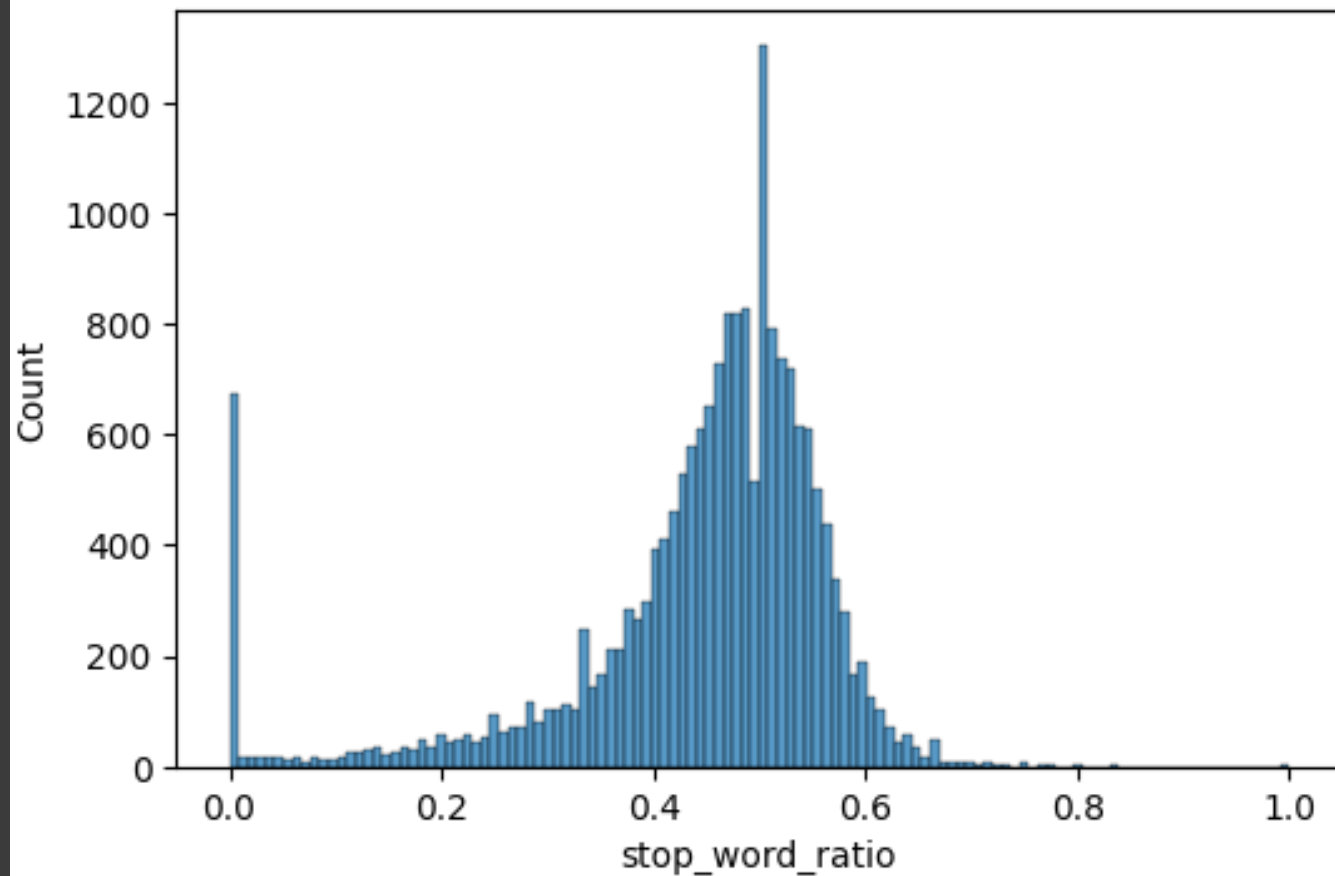


Zoomed in to small range



Stop words per word count ratio

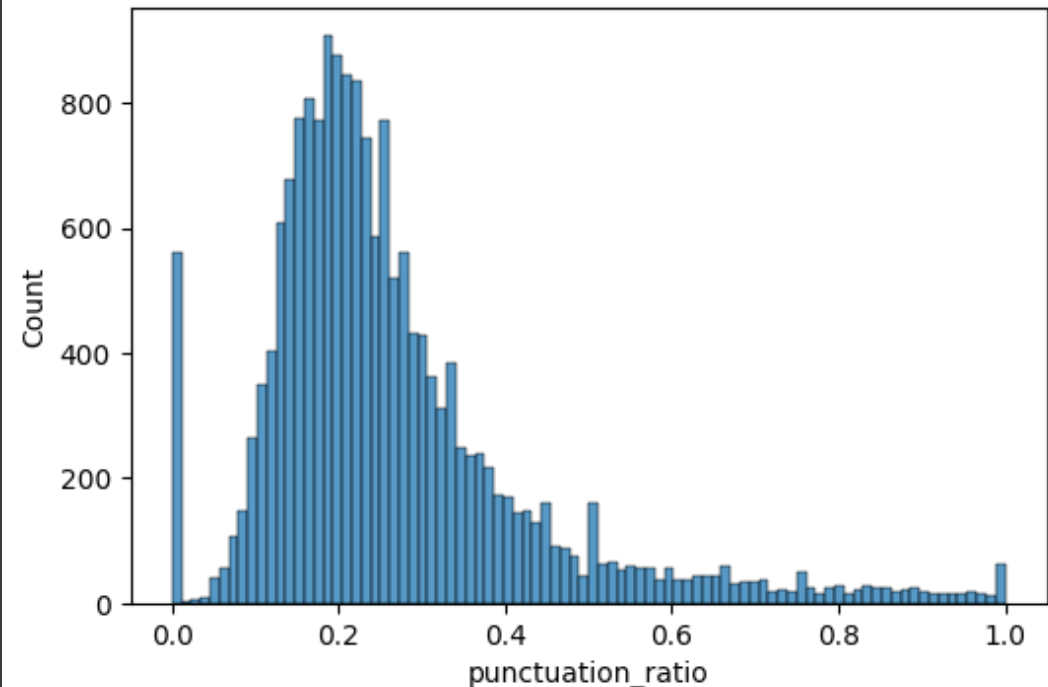
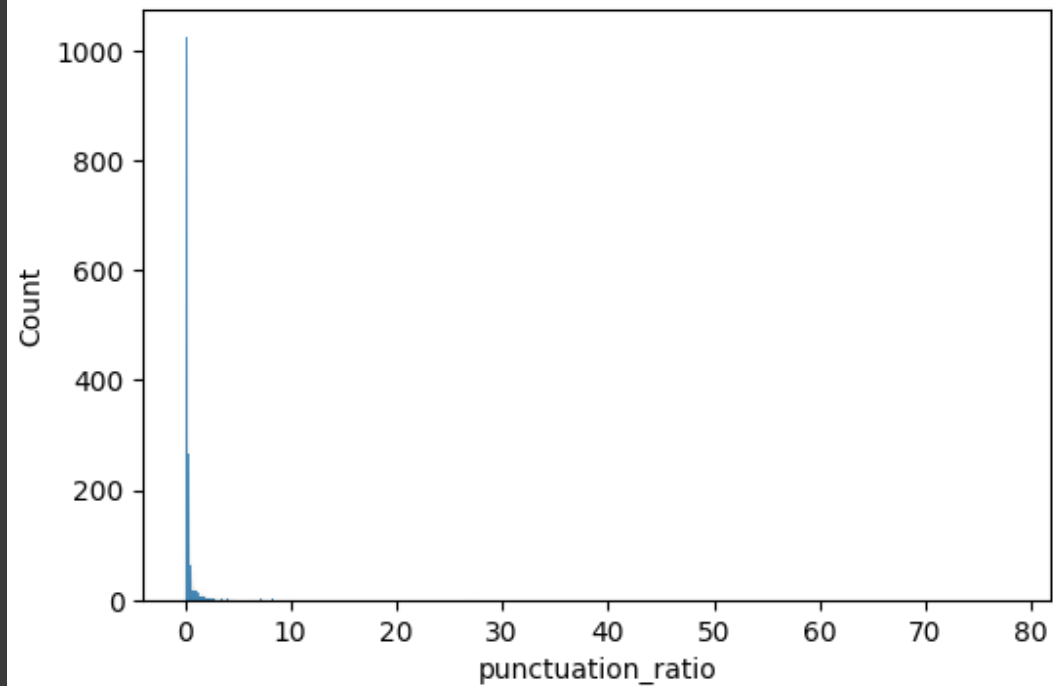
On average, half of the words per article are considered stop words



```
Median: 0.4745762711864407  
Mean: 0.443714166539526  
Std: 0.13075392883348091
```


Punctuation to Word Ratio

There contains a long-tailed distribution, but mostly followed Poisson distribution where it's not at the tail.



What changes after cleaning

Not much changes are noticed after cleaning. The statistics does shift a bit (by 0,002) but that's almost negligible.

Unigrams, Bigrams, Trigrams

The actual table is too huge to put here. We would summarize them in the next three slides.

Unigrams

- Top 3 common words: **don't, think, know.**
- Top 3 uncommon words: **image, space, god.**
- Conclusion: Not much information extracted from unigrams to distinguish between categories.

Bigrams

- Top 3 common words: **don't think, don't know, does know.**
- Uncommon words gives better indications here. Examples: **space station, human rights, medical newsletter, image processing, hard disk, etc.**
- Can separate major categories (between science and politics, for example) but not really for sub-categories (sci.crypt vs sci.electronics).

Trigrams

- Top 3 common: not available.
- Uncommon: emails, phone numbers, separators (====, -----) occupy most top-k trigrams. Useful ones includes **linked allocation unit** for example.
- They are not necessary the most useful separator between categories. Category 10 contains numbers for their top-k trigrams such as 0 0 0, 0 1 1, 2 2 2, that are not meaningful.

Recommended Models (Technical User)

- Any Neural Net models shall do good.
- One suggests the use of AWD-LSTM defaults of fastai NLP.

Thank You