Project 3 : NLP

DSI-1128

Atigon Hongchumpol / Jan 20, 2023

Problem Statement

- Cats & Dogs photo caption contest
- The image processing team ask for help

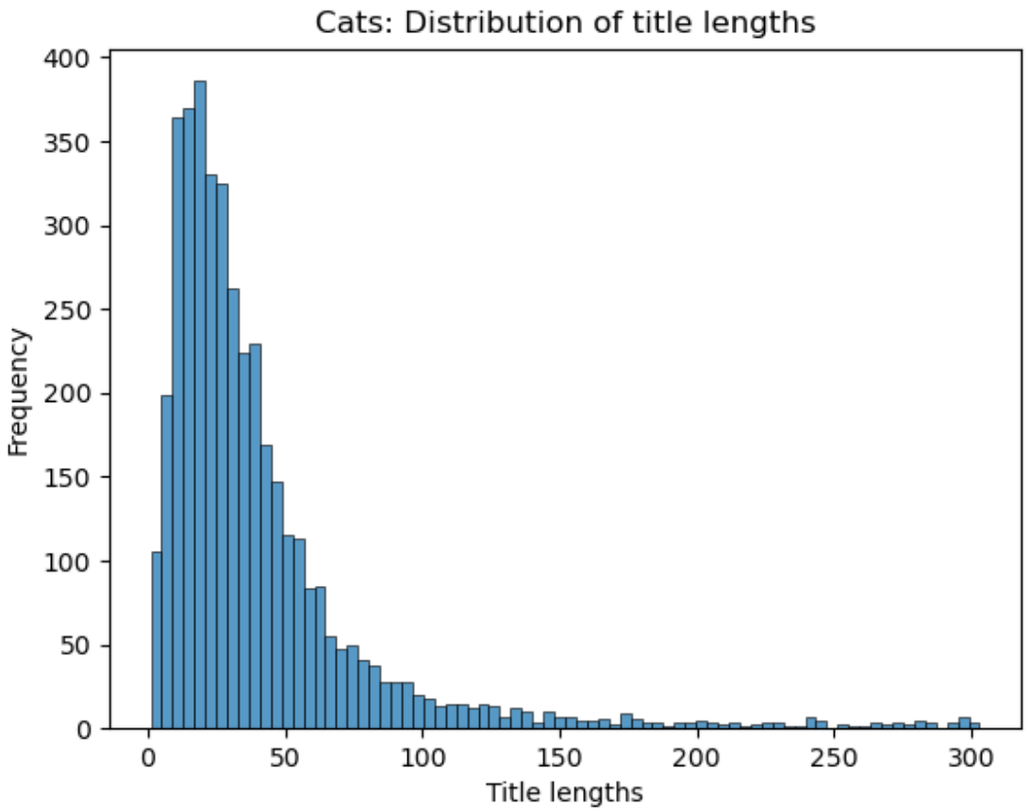
Data collection & Cleaning

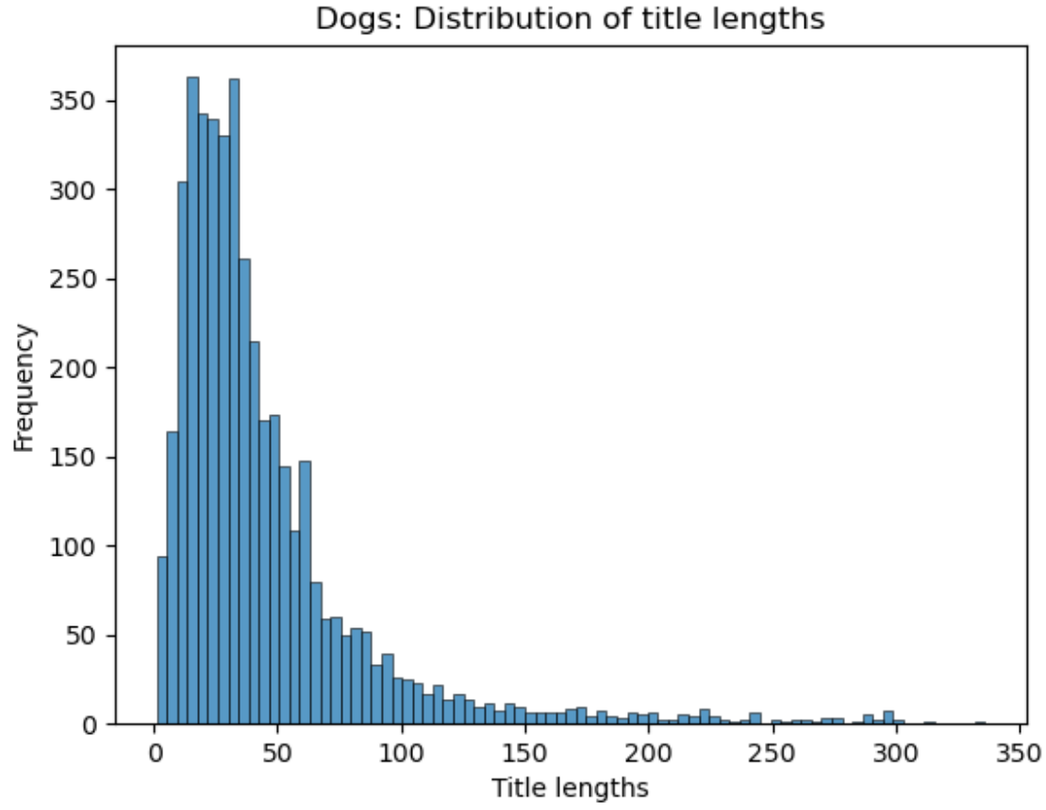
- Pushshift
- r/cat, r/dog
- 5,000 + 5,000 = 10,000 posts

- Remove duplicates title
- Remove incorrect category

EDA

Title lengths distribution





EDA

Most 5 common words using stemming

- CATS

1. cat
2. love
3. like
4. kitten
5. cute

- DOGS

1. dog
2. love
3. **help**
4. puppi
5. **breed**

EDA

Top 10 common emoji

- CATS

❤	84
🥰	62
😍	59
😂	57
😻	50
🤣	34
🐈	31
😺	29
🥹	25
😹	22

- DOGS

❤	67
🥰	45
🐶	43
😍	43
😂	40
🐕	21
🐾	20
🥹	18
💕	17
♥	14

EDA

Sentimental Analysis using compound score

- CATS

0.1528

- DOGS

0.1716

Cats < Dogs
12%

Modeling

- Baseline accuracy 50%
- (CVEC, TF-IDF) + (RF, LR, KNN, NB, ADA)
- STACK

Evaluation

	Train score	Test score	Recall	F-1
CVEC + RF	0.98	0.76	0.82	0.78
TFIDF + LR	0.91	0.78	0.84	0.79
CVEC + NB	0.82	0.77	0.77	0.77
STACK (RF+LR+NB)	0.95	**0.79**	0.84	**0.80**
CVEC+ADA	0.73	0.72	**0.96**	0.78

Summary

- Weak prediction 79%
- AdaBoost recall 0.96

Recommendation

- Bigger dataset
- New word remove technique
- Use another source of

Thank you