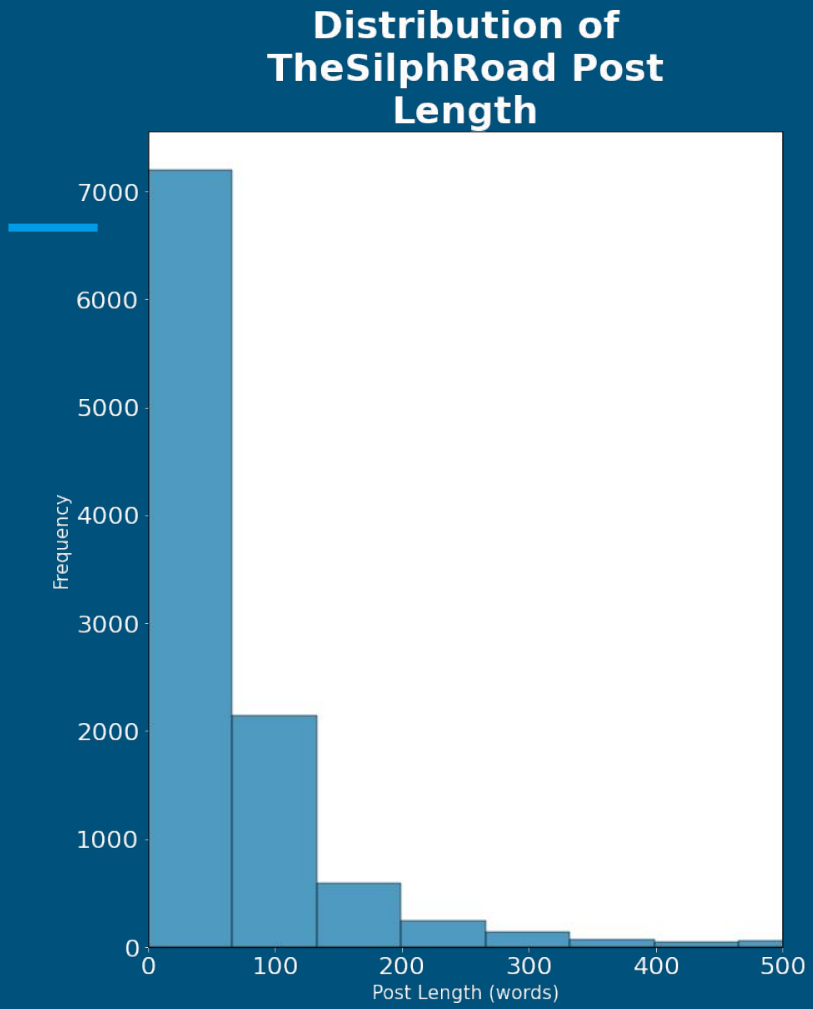
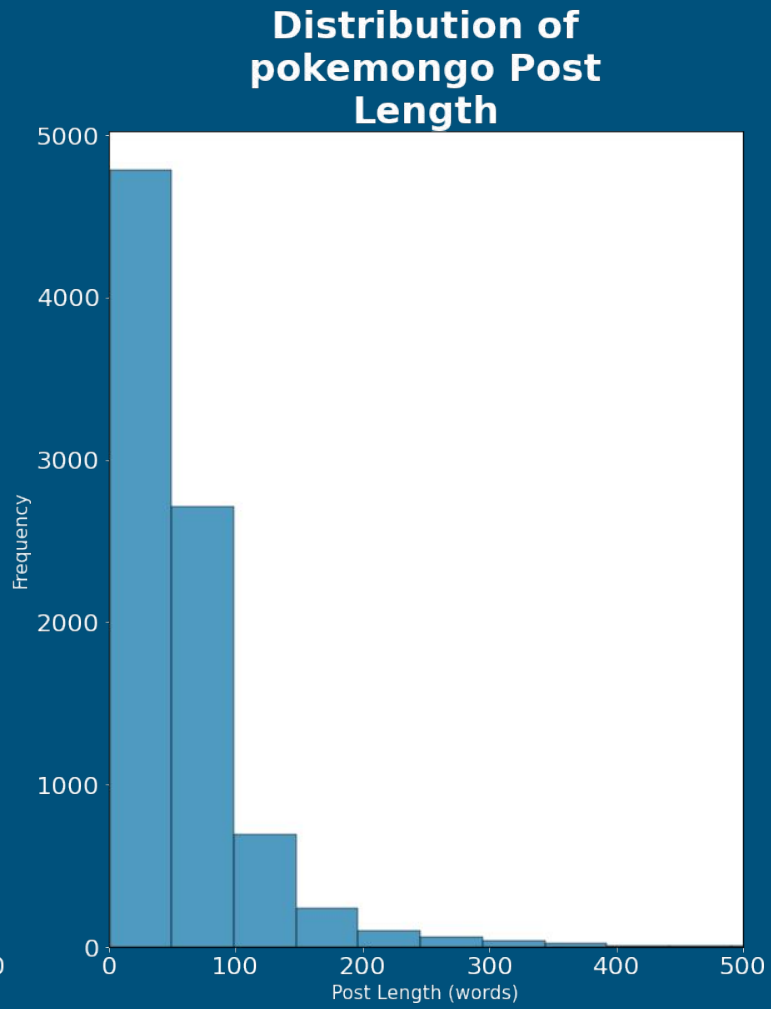
I Want To Be the Very Best Like No One Ever Was

Jesse Tao

To Catch Them is My Real Test

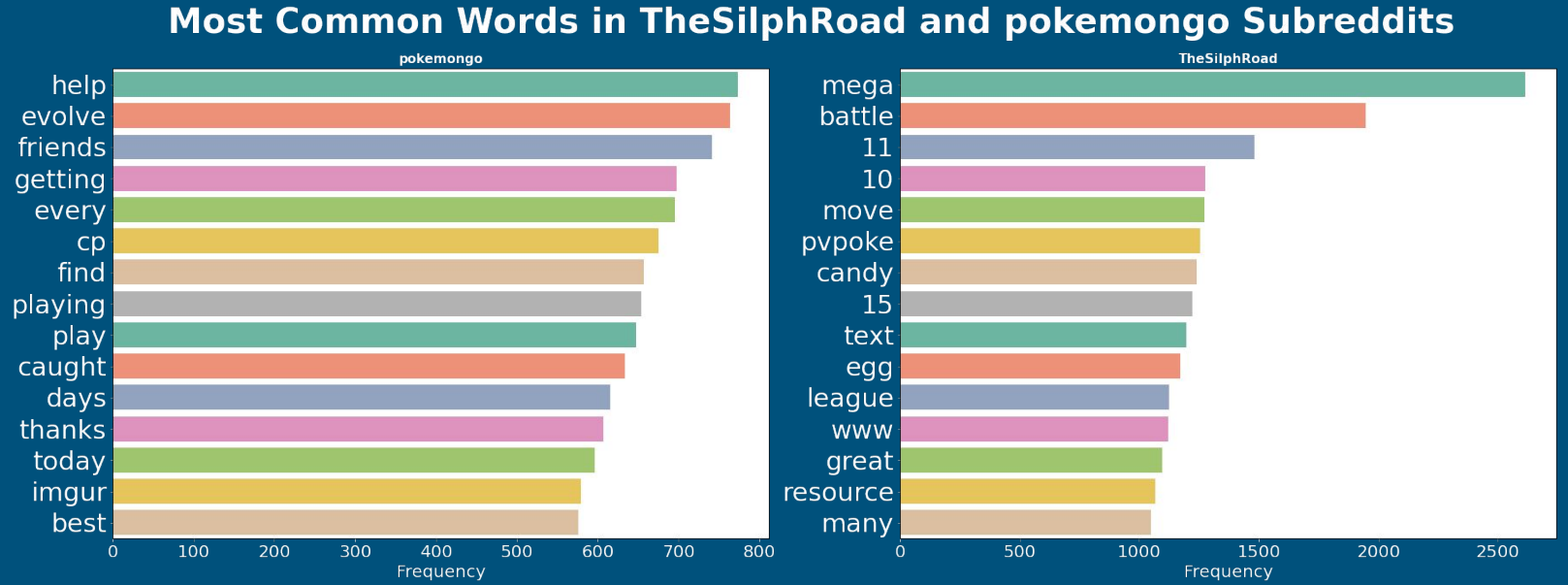
- Posts from r/pokemongo and r/TheSilphRoad using Pushshift API
- Removed posts that were removed or deleted


Distribution of TheSilphRoad Post Length
Frequency
0
1000
2000
3000
4000
5000
6000
7000
0
100
200
300
400
500
Post Length (words)


Distribution of pokemongo Post Length
Frequency
0
1000
2000
3000
4000
5000
0
100
200
300
400
500
Post Length (words)

Finding Words as Common as Pidgey

More Stopwords, Pidgey was a Bad Example



I Will Travel Across the Land

- After using Count Vectorizer, we have some foreign text
- Used Google Translate to resolve this issue and convert everything to English

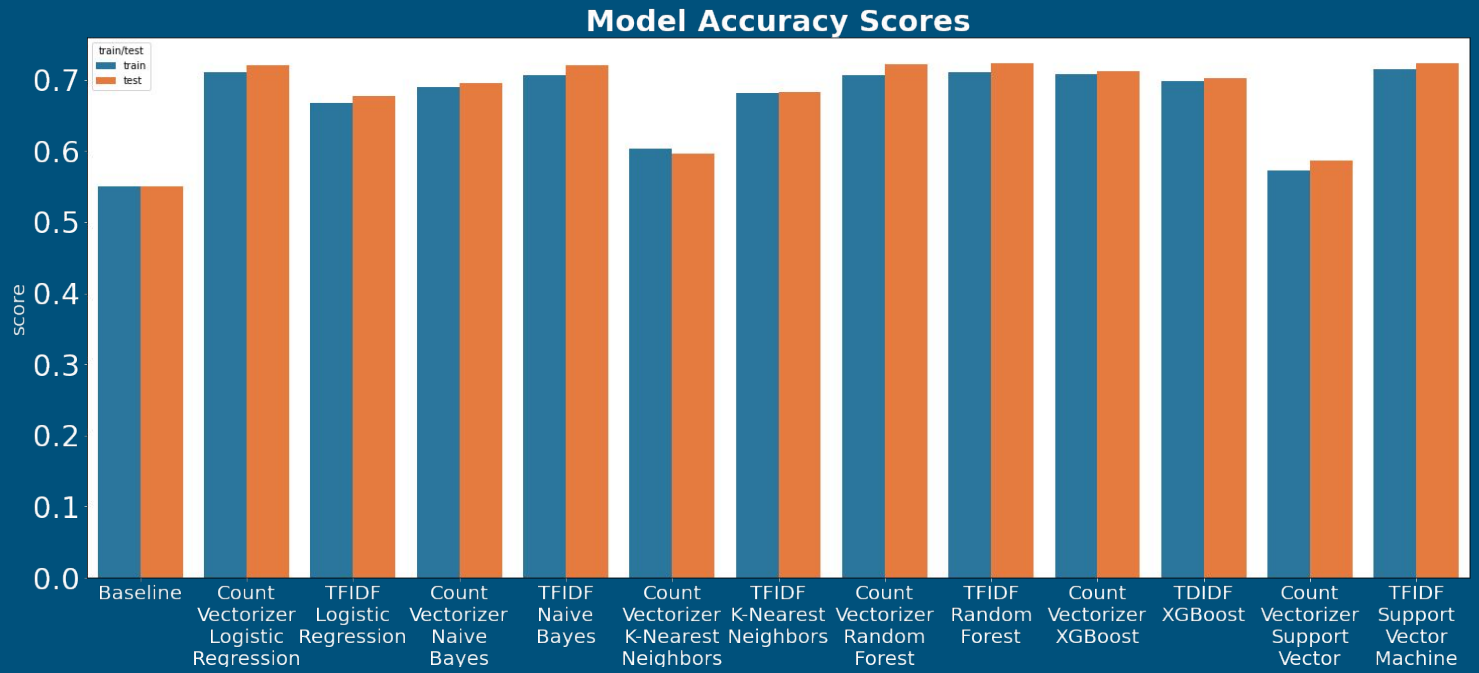
что	шикарно	это	から毎週金曜よる6時55分にお引っ越しすることを記念し	そのうち10月6日	シャトウ	ツ	テレヒ東京系にて放送中のアニメ	ホケットモンスター	特別な内容て開催されます
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Searching Far and Wide

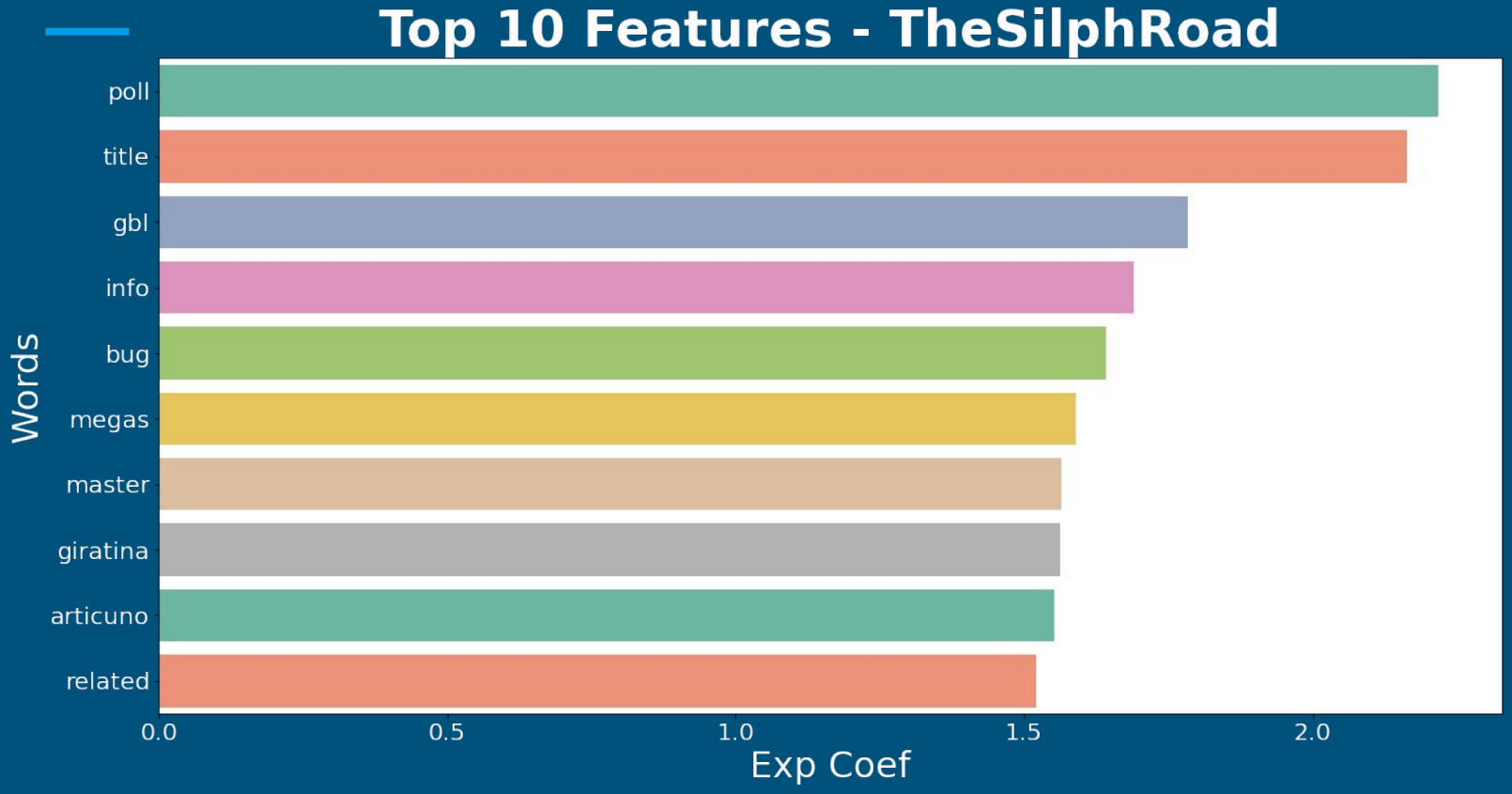
- Created multiple pipelines that used either CountVectorizer or TfidfVectorizer along with a classification model
- Used a custom lemmatization or stemming preprocessor
- Conducted GridSearch over a wide amount of hyperparameters
- Used HalvingRandomSearchCV to optimize for time

These Models to Understand

- Key parameters: https://jesseptao.com/key-parameters/



The Power That's Inside



How do we Catch ‘em All?

- Optimize specificity to distance ourselves from pokemongo subreddit
- Wait for potential more complex models to finish fitting
- Find better models using cloud computing with more than 50 CPU threads
- Build a custom sentiment analyzer
- Implement a custom tokenizer to ignore long strings of numbers

Questions?