Data Society Technical Exercise

Daniel Chen

Data

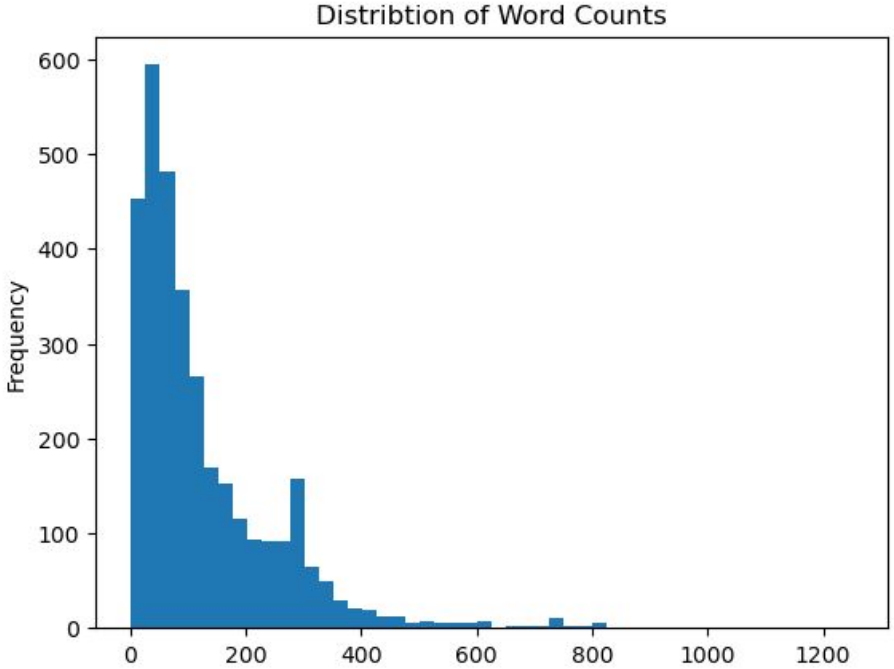
-Public comments about removing EPA regulations that restricted use of certain waterways as disposal sites specifically for mining

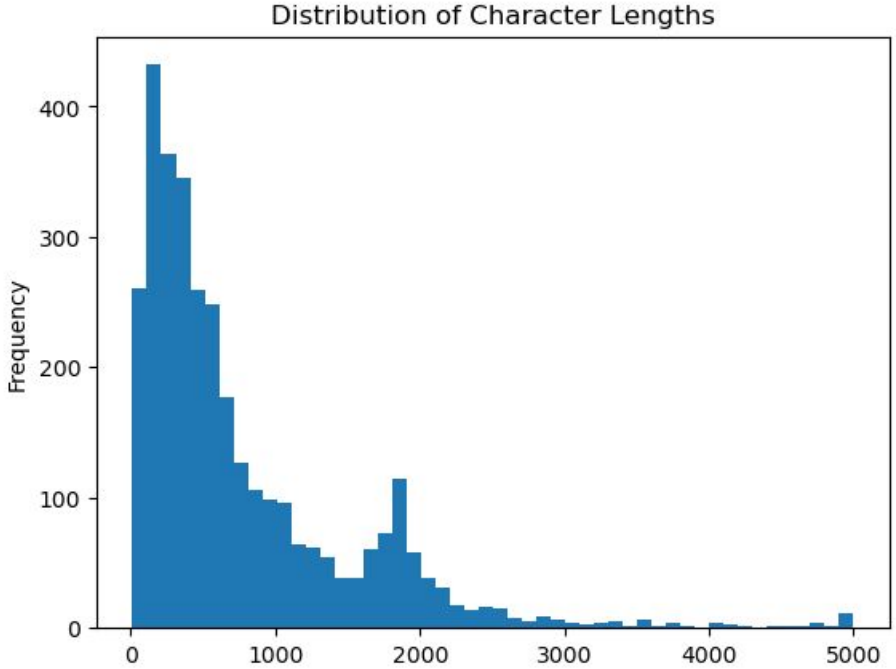
Dataset -

~3k comments

Average word count - 128

Average Character Length - 783


Distribtion of Word Counts
Frequency
0
100
200
300
400
500
600
0
200
400
600
800
1000
1200

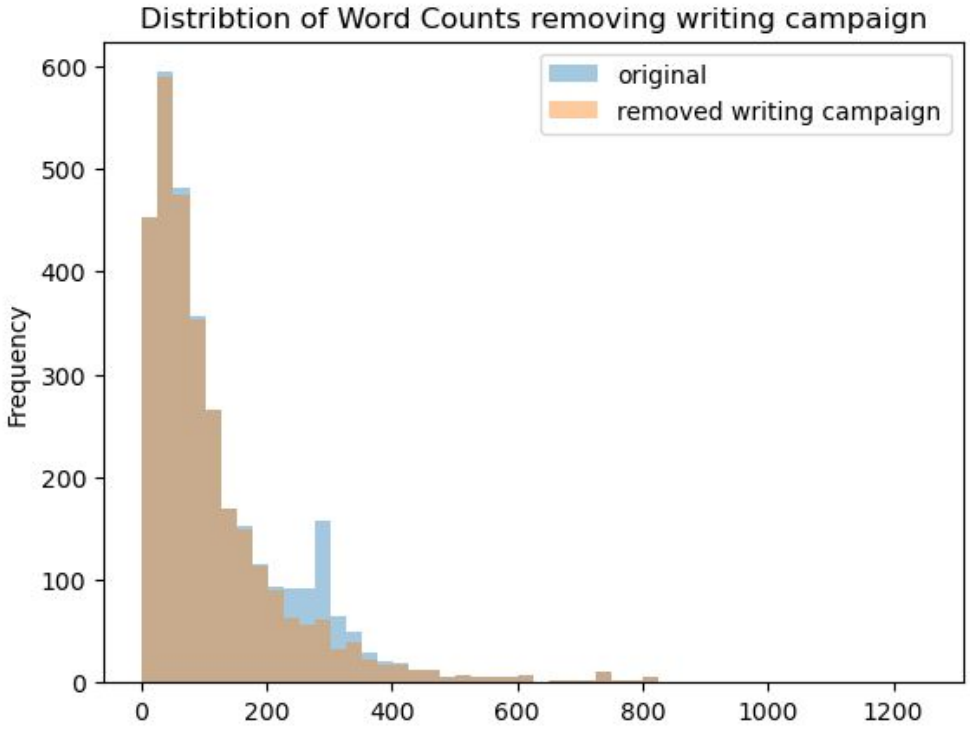

Distribution of Character Lengths
Frequency
0
100
200
300
400
0
1000
2000
3000
4000
5000

Analyzing spikes in both word count and character length

I found **26** identical comments. Using a subset from that comment I was able to find an additional **235** more comments the contain the same general message with personalizations added on the end.

This appears to be a coordinated campaign by an organization to get more people to comment by giving them a common starting message

If we remove those comments from the dataset the word count distribution looks much more like what we would expect. (the comments will be left in for further analysis)



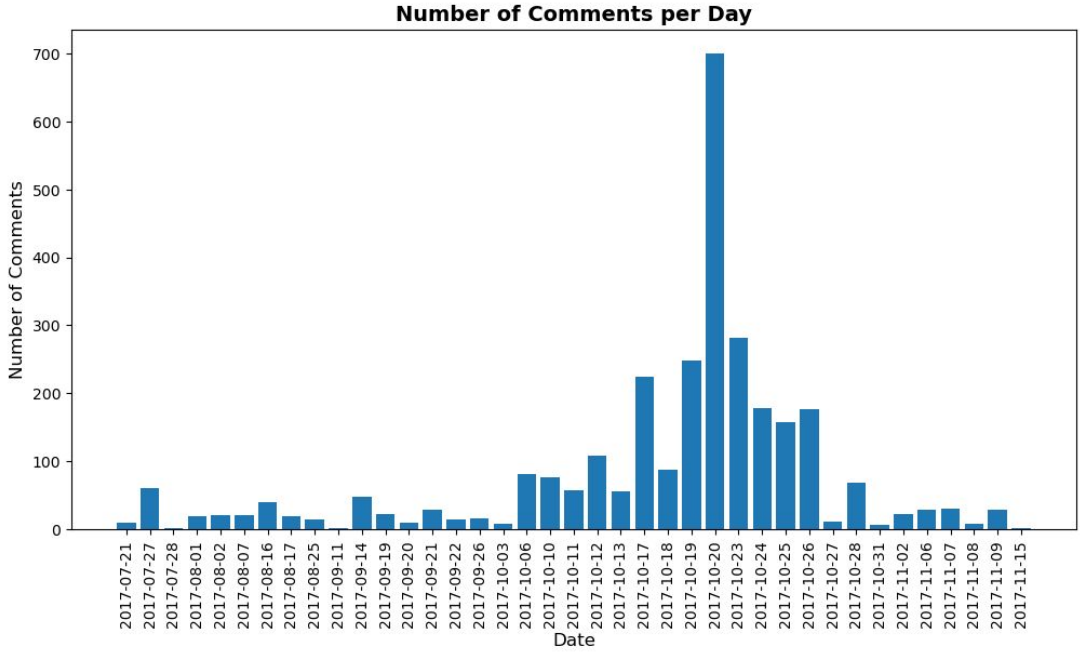
Word Cloud

The words that seem to jump out the most are Bristol Bay, Pebble Mine, Clean Water Alaska, Salmon, protect and opposed

Comment Posted Dates

Three is a spike of comments on 10/20/2017, but no common theme was found yet.

There could have been some news program or article that publicized that they were open for comment but no specific example was found.

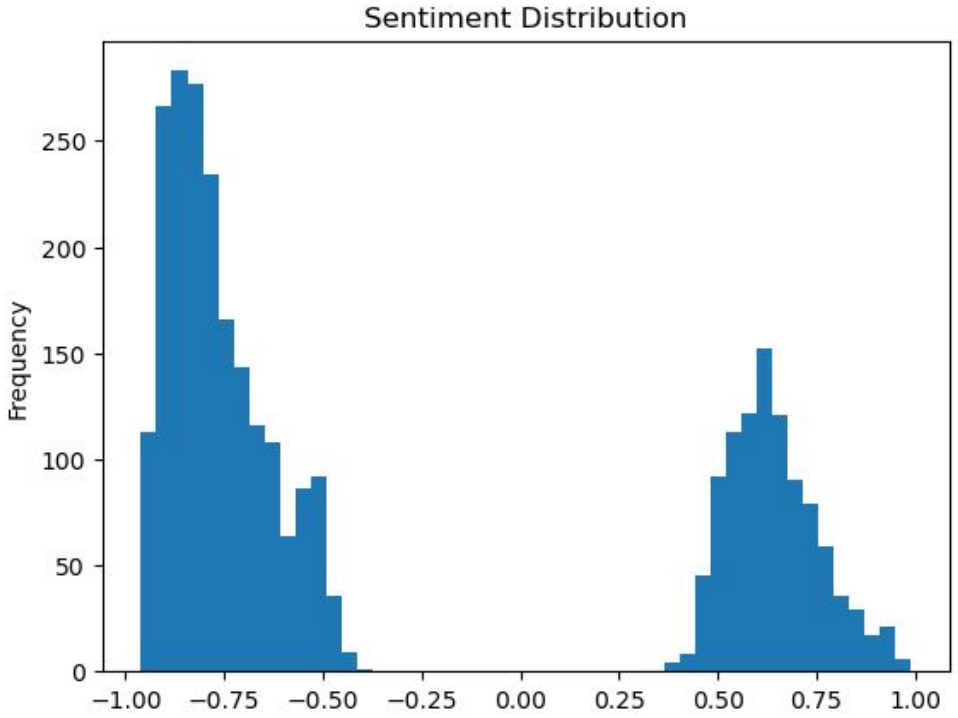


Sentiment Analysis

model="cardiffnlp/twitter-roberta-base-sentiment-latest"

From the model we can see comments skew negative. And when they are negative the mode has a greater degree of confidence in that precision compared to positive

This model also allows for Neutral state, which 27% of the comments fall into. This is unexpected and required to look more closely at those comments and choice of model



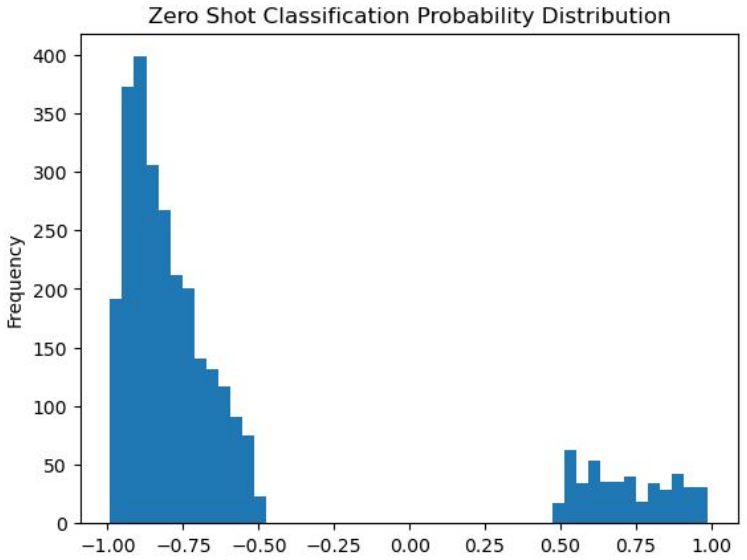
Label	%
Negatve	66.7
Neutral	26.9
Positive	6.35%

Zero-Shot Classification model

model="facebook/bart-large-mnli"

candidate_labels = ["for", "against"]

With his model we now classify all the previously neutral targets. We still see high confidence in Against than we do in For



Label	%
Against	84.5%
For	15.5%

Comparison of Methods

Zero Shot Classification

Sentiment Model

	Against	For
Negative	95.24	4.76
Neutral	71.06	28.94
Positive	29.1	70.9

Given Negative Sentiment, 95% odds Zero-shot agreed

Given Positive Sentiment, 71% odds Zero-shot agreed

Topic Modeling

Performed topic modeling using LDA on Against and For subsets of comment data but didn't get much interesting in results.

Against Topics:

Topic 1: bristol bay proposed salmon determination pebble jobs wild alaska fisheries

Topic 2:salmon bay bristol mining area pebble people jobs world sockeye

Topic 3:epa mining water area clean pebble bay protect environment bristol

For Topics:

Topic 1:epa pebble determination proposed bristol bay mining clean alaska process

Topic 2:salmon bay bristol protect wild world alaska proposed fishery water

Topic 3:pebble alaska project mining process jobs proposed people state epa

Summarizing random subset of comments for and against to get theme

For:
[{'summary_text': " Pro Pebble Mine is a good thing for all Alaskans, and especially for the economy of our struggling burrough. We can't just rely on drilling for oil, or salmon fishing for an economy. We need to expand our mineral resources also. Pro Pebble will make America great Again ."}]

Against :
[{'summary_text': ' I oppose Pebble Mine and voice my support for sustainable jobs in Alaska, traditional ways of life for indigenous people, and the protection of habitat for the largest Pacific salmon runs in the world . Please retain protections for Bristol Bay from the proposed Pebble Mine, which poses significant environmental risk to a highly-productive and sustainable fishery .'}]

In Summary

The proposal is very unpopular with approximately 85% of comments against. With arguments cited as the environment, salmon population, and clean water act

Comments for the proposal revolve around increase in jobs, and anger against too many regulations.