Analyzing the Analyzer

An exercise in Natural Language Processing

Paul Miller
General Assembly DSI
Jan 8, 2021

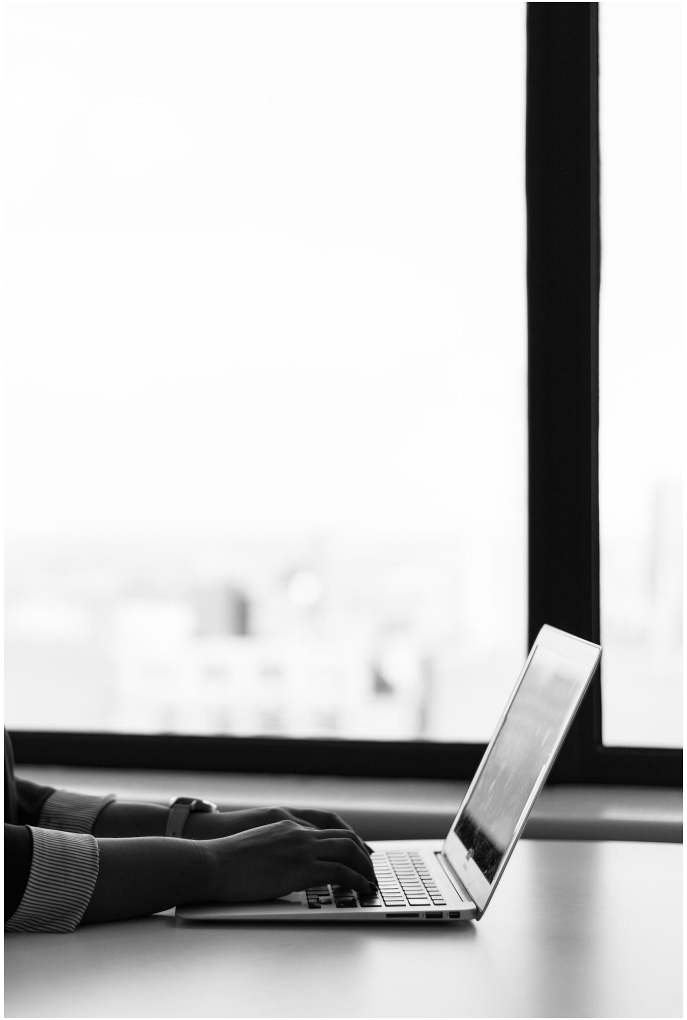


Table of contents

01 Feelings!

Is it healthy to be alone?

02 reddit

Why these two?

03 Classifiers

Run or cook?

04 More Feelings!

Sentiment Analyzers

05 Analyzer thoughts

Buyer beware

Feelings!

Reddits

Classifiers

More Feelings!

Analyzer Thoughts

Is Being Alone A Bad Thing?

Pros:

- Recharge
- More Productive
- Boosts Creativity

Cons:

- Critical Inner Voice
- Depression
- Not Healthy

And the **NLTK SentimentIntensityAnalyzer** says so:

```
sent.polarity_scores("alone")

{'neg': 1.0, 'neu': 0.0, 'pos': 0.0, 'compound': -0.25}
```

https://www.psychalive.org/being-alone/

Photo by https://www.pexels.com/@quang-nguyen-vinh-222549

Feelings!

Reddits

Classifiers

More Feelings!

Analyzer Thoughts

Pick two Reddits to classify

reddit r/AskCulinary

reddit r/running

- Lots of text
- Problem statement:
  - If being alone is negative...
  - ... then Running subreddit -> lower Sentiment scores

Feelings!

Reddits

Classifiers

More Feelings!

Analyzer Thoughts

Classifiers

❖ **Models:**

1. Pipeline with:
*** CountVectorizer (transformer)**
*** Multinomial Naive Bayes (estimator)**

Training model score: 0.992
Testing model score: 0.991

2. Pipeline with:
*** TfidfVectorizer (transformer)**
*** KNeighborsClassifier (estimator)**

Training model score: 1.0
Testing model score: 0.985

3. Pipeline with:
*** CountVectorizer (transformer)**
*** Random Forest (estimator)**

Training model score: 0.946
Testing model score: 0.939

So if classifiers are this good, can’t wait to see how Sentiment Analyzers do!

Feelings!

Reddits

Classifiers

More Feelings!

Analyzer Thoughts

Sentiment Analyzers

❖ **NLTK VADER Sentiment Analyzers**

- Vader_lexicon.txt
- 7500 terms
- If term not in lexicon, return neutral score.
  (run, running, cook, bake, family, together)

Term	Mean Rating	Human Ratings
friend	2.2	[2, 2, 3, 2, 3, 3, 1, 2, 2, 2]
alone	-1.0	[-2, -1, -1, -1, -2, -1, 0, -1, -1, 0]

Humans rate terms on scale of sentiment intensity
-4 most negative
+4 most positive

Feelings!

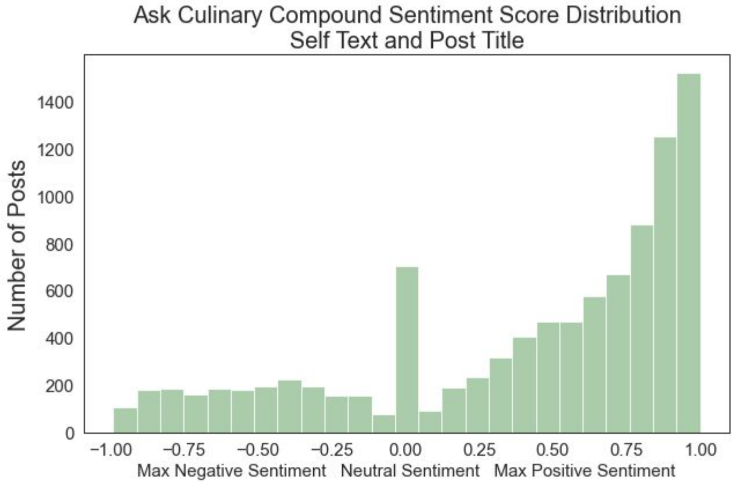
Reddits

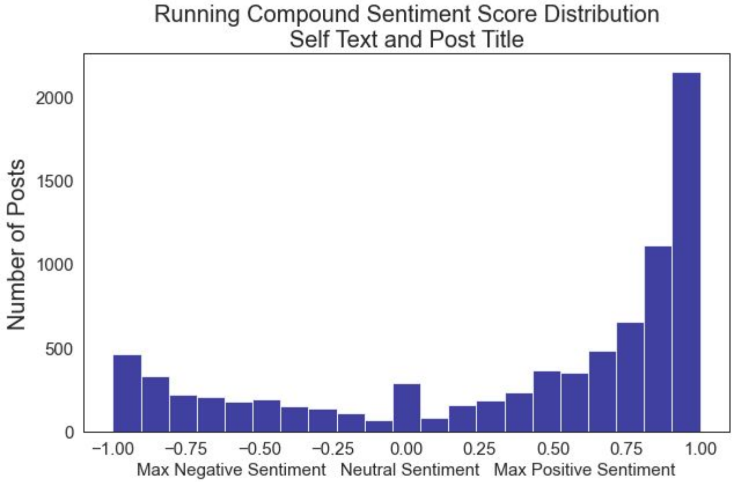
Classifiers

More Feelings!

Analyzer Thoughts

Sentiment Analyzers cont.





Compound Sentiment Scores Normalized -1 to 1

8,000 Ask Culinary posts:

Average Compound Score for all posts: **0.40**

8,000 Running posts:

Average Compound Score for all posts: **0.38**

Running has more positive posts, but a lot more negative to drag it down.

Sentiment Analyzers cont.

❖ **What is missing?**

In **Ask Culinary** 8,000 posts, almost 16,000 vectorized words

- VADER lexicon = 7,500 words
- Only 1,500 of the 16,000 were found in VADER lexicon.

In **Running** 8,000 posts, almost 18,000 vectorized words

- VADER lexicon = 7,500 words
- Only 2,000 of the 16,000 were found in VADER lexicon.

Feelings!

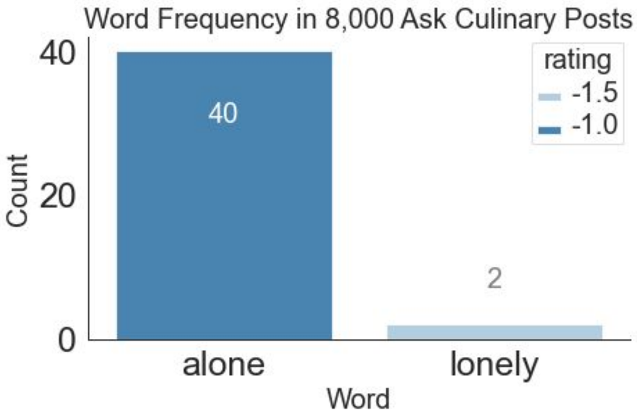
Reddits

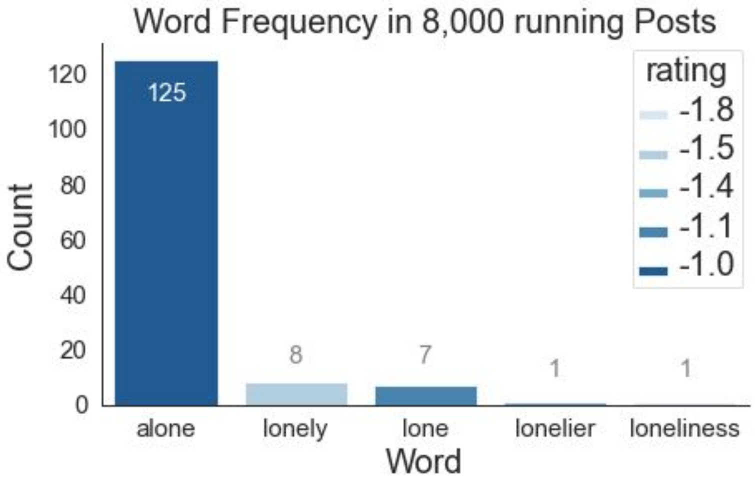
Classifiers

More Feelings!

Analyzer Thoughts

Are runners more lonely?





Feelings!

Reddits

Classifiers

More Feelings!

Analyzer Thoughts

Final Thoughts

- Learn about the tools
  - Stop words!
    - They might impact what you are analyzing (‘fire’, ‘keep’, ‘system’)
  - Lexicon
  - Human ratings
  - Hyperparameters
- Question the results

Analyzing the Analyzer

Thank you!

Paul Miller
General Assembly DSI
Jan 8, 2021