Reddit: Post Classification

Mark Dowicz

Problem Statement

Is it necessary to have two different subreddits for both **rants** and **unpopular-opinions**?

As an avid Reddit user, I've noticed that these two subreddits produce many of the same types of posts. My goal is to produce a classification model designed to predict which subreddit a post came from. The strength of this model will determine if the above subreddits should remain separate, or consider a merge.

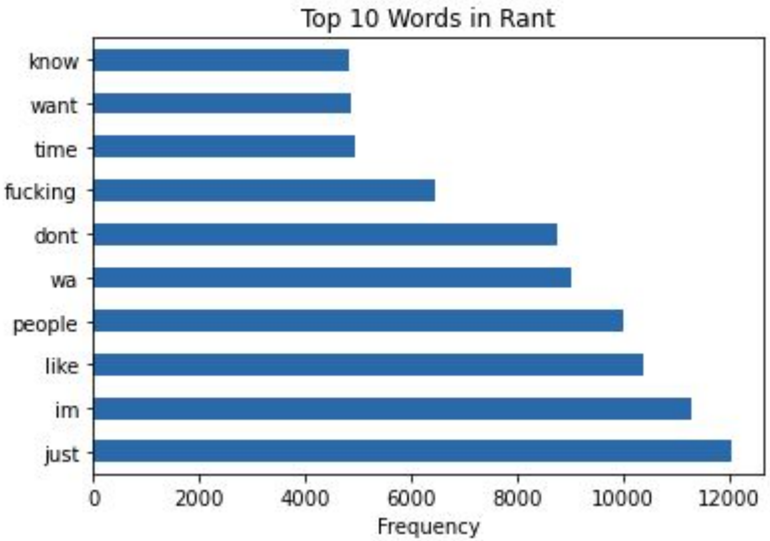
Data Collection & Cleaning

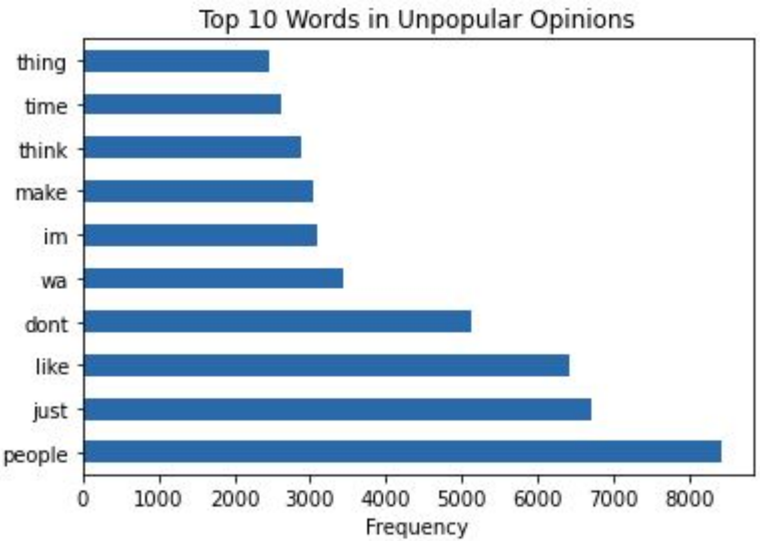
- 15k posts from ‘Rant’ & 15k posts from ‘Unpopular-Opinion’
- Cleaned ‘self_text’
  - Dropped rows with nulls, ‘removed’, or ‘deleted’
  - Removed all special characters
- Created **‘all_text’** and **‘total_word_count’**
  - 25<’total_word_count’<1000
- Created **‘Lem’**
  - Lemmatized each word in ‘all_text’

Rant vs. Unpop-Opinion

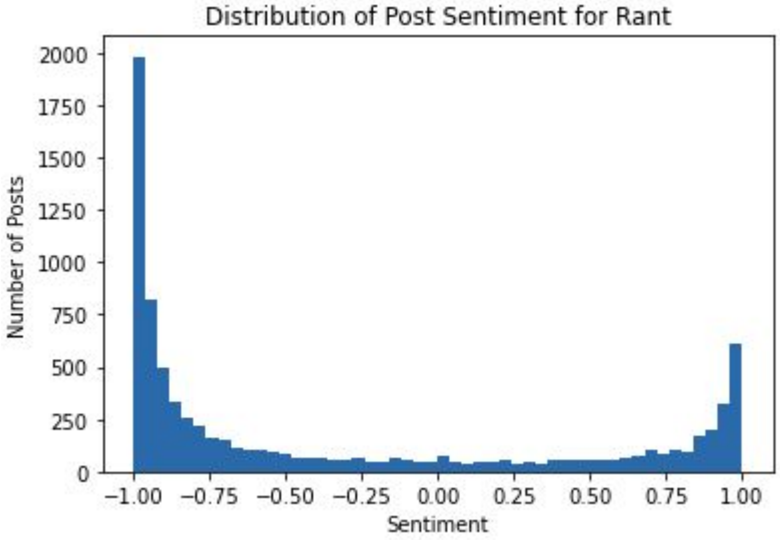
	Rant	Unpop-Opinion
Share	52% (8127)	48% (7488)
Word Count Average	204	133
Word Count Std. Dev.	164	101

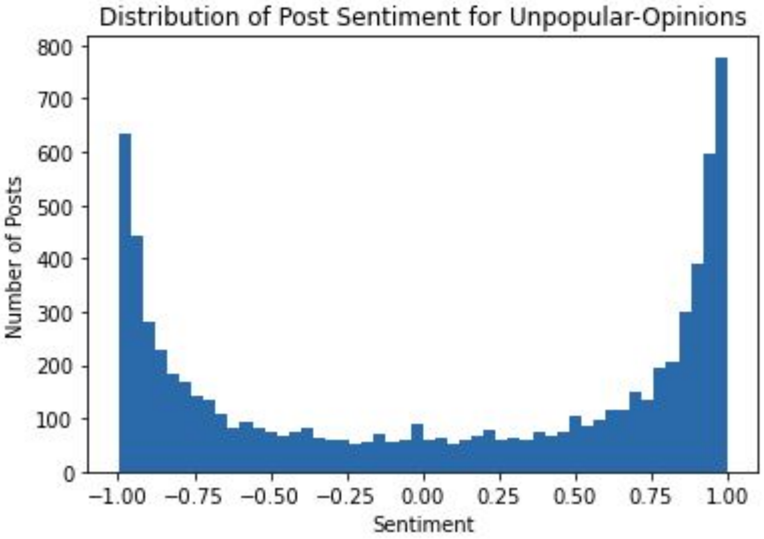
Word Frequency





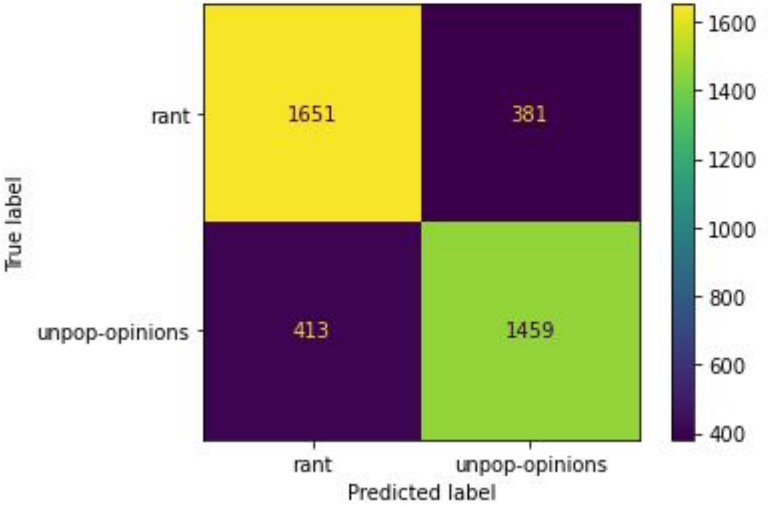
Sentiment Analysis





Best Model

- Dependent Variable: **‘lem’**
- Pipeline --> Grid-Search:
  - CountVectorizer
  - Logistic Regression
- CountVectorizer Params:
  - Stop_words: english
  - Max_features: 2500
  - Min_df: 2
  - Ngram_range: (1, 2)
- Logistic Regression Params:
  - C: 1
  - Penalty: L1
- Train: (0.832), Test: (0.797)



Conclusions

- Both subreddits share very similar frequent words
- In general, Rant posts are more negative
- Model performed equally on each subreddit
    - 79.9% on Rant, 79.2% on Unpop-Opinions
- Each subreddit is distinct and should remain as such
- Issues?
    - Small sample relative to pop.
    - Sample is not random

Thank You!