UnresolvedMysteries vs FanTheories Subreddits

Binita Patel
Data Scientist

TABLE OF CONTENTS

01
Data Science Problem

Data Science Problem

- Co-worker is about to quit job
- Goes through some of the subreddits and messes them up
- Build a classification model
- Focus on two subreddits : UnresolvedMysteries and FanTheories.

02
Procedure / Methodolgy

Procedure/Methodology

Data Cleaning

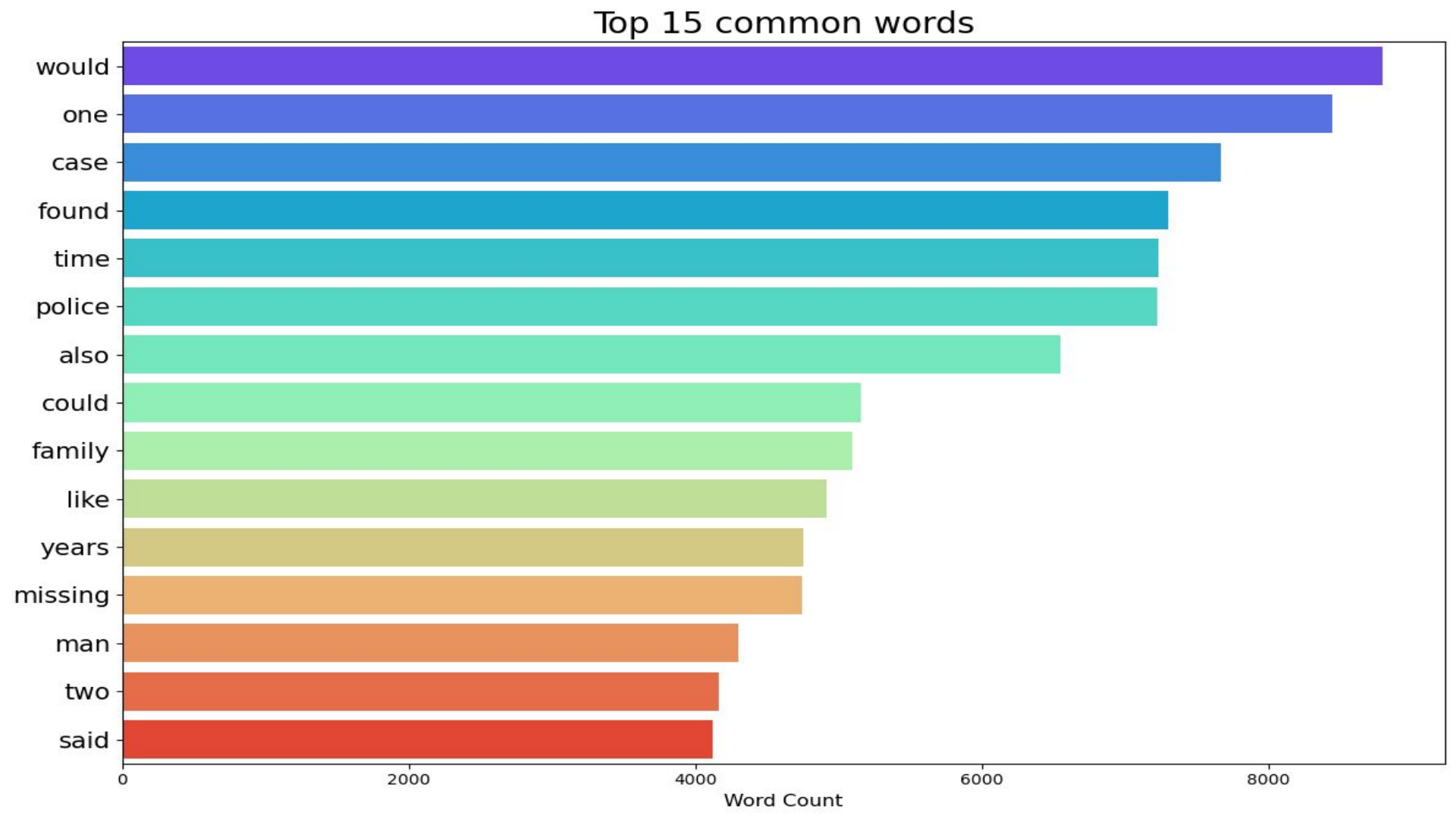
Cleaned dataset by removing Nulls

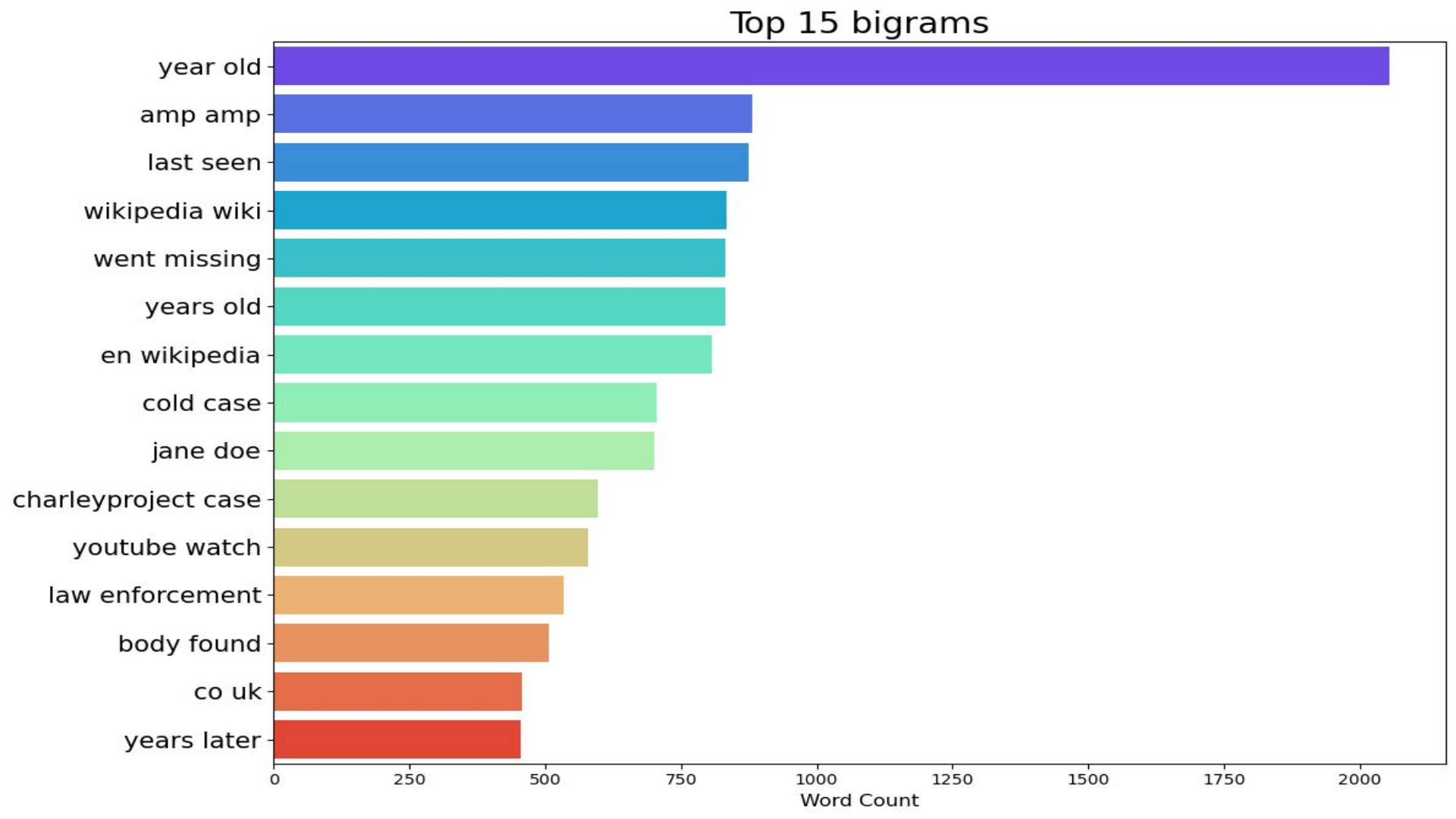
EDA

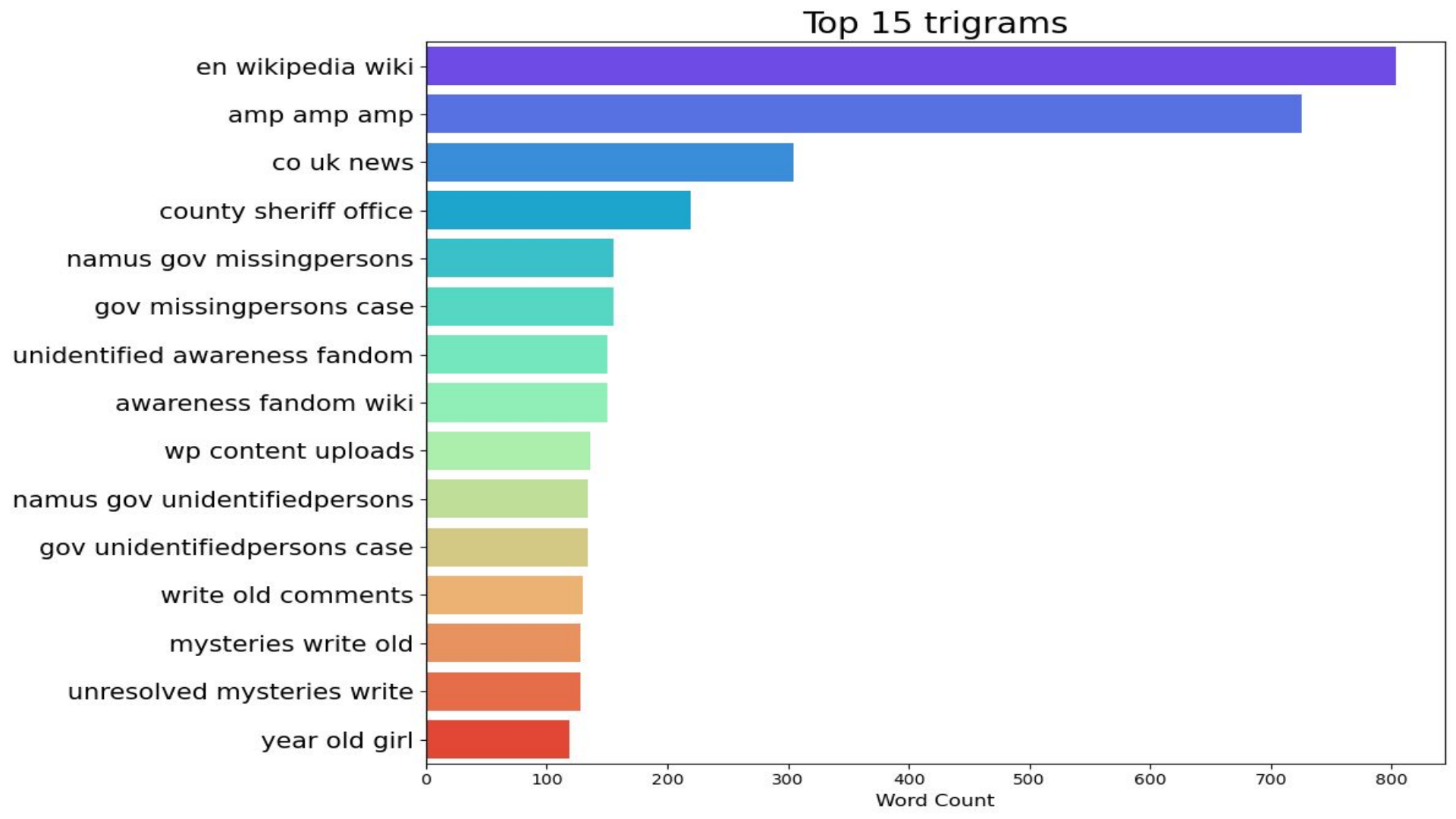
Looked at top words and did sentiment analysis

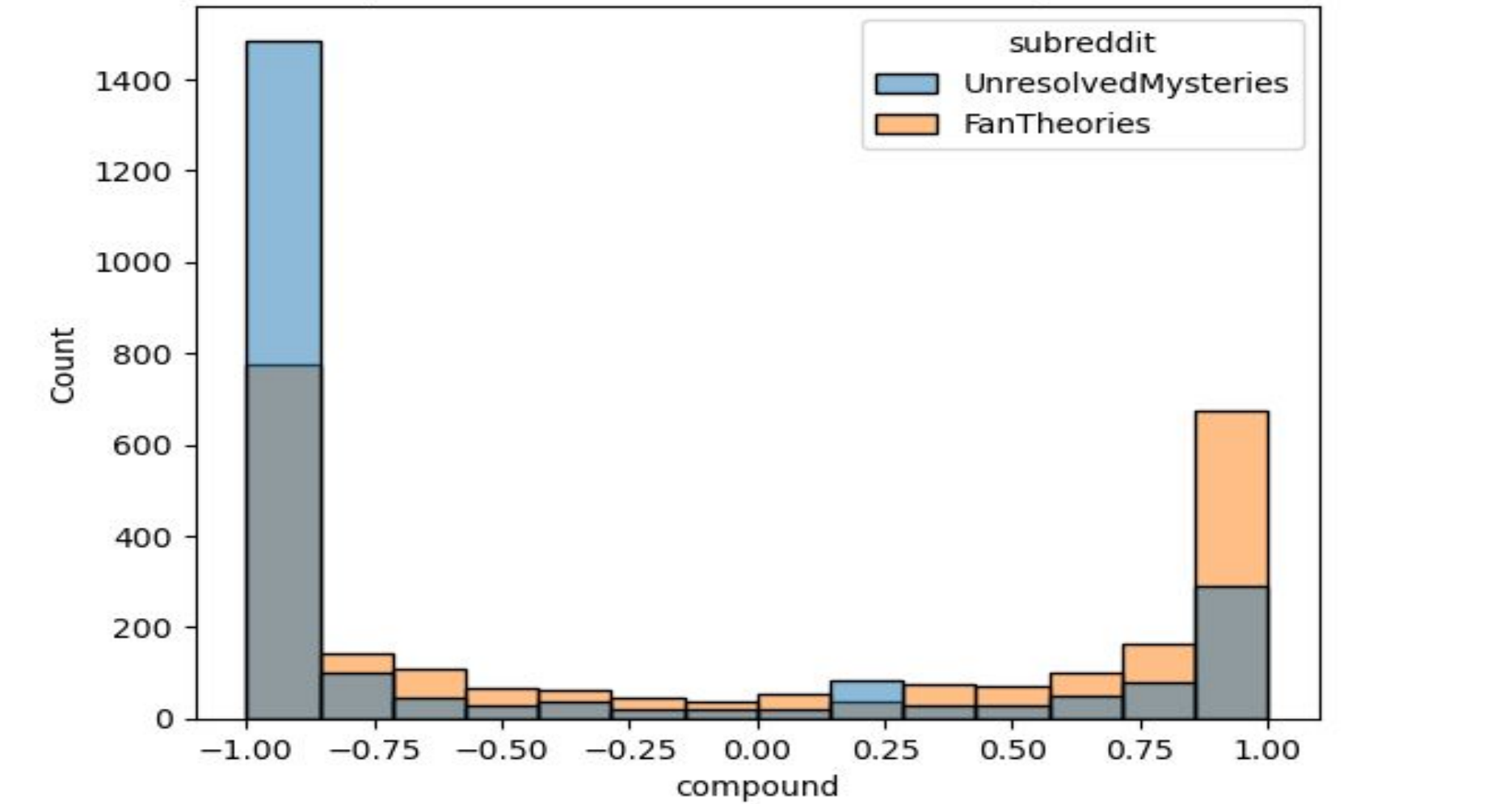
Modeling

Did preprocessing and ran 6 models


Top 15 common words
would
one
case
found
time
police
also
could
family
like
years
missing
man
two
said
0
2000
4000
6000
8000
Word Count


Top 15 bigrams
year old
amp amp
last seen
wikipedia wiki
went missing
years old
en wikipedia
cold case
jane doe
charleyproject case
youtube watch
law enforcement
body found
co uk
years later
0
250
500
750
1000
1250
1500
1750
2000
Word Count


Top 15 trigrams
en wikipedia wiki
amp amp amp
co uk news
county sheriff office
namus gov missingpersons
gov missingpersons case
unidentified awareness fandom
awareness fandom wiki
wp content uploads
namus gov unidentifiedpersons
gov unidentifiedpersons case
write old comments
mysteries write old
unresolved mysteries write
year old girl
0
100
200
300
400
500
600
700
800
Word Count


Distribution of Compound Score between UnresolvedMysteries and FanTheories
subreddit
UnresolvedMysteries
FanTheories
Count
1400
1200
1000
800
600
400
200
0
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
compound

03
Primary Findings

Models

Num	Vectorizer	Model	Best Score	Train Score	Test Score (Accuracy)	Best Parameters
1	CountVectorizer	Multinomial Naïve Bayes	0.9873	0.9969	0.9788	{'cvec__max_features': None, 'cvec__min_df': 2, 'cvec__ngram_range': (2, 2)}
2	**TfidfVectorizer**	**Multinomial Naïve Bayes**	**0.9873**	**0.9975**	**0.9797**	**{'tvec__max_features': None, 'tvec__min_df': 2, 'tvec__ngram_range': (2, 2)}**
3	CountVectorizer	Random Forest Classifier	0.9735	1.0	0.9755	{'rf__max_depth': None, 'rf__min_samples_leaf': 1, 'rf__n_estimators': 150}
4	TfidfVectorizer	Random Forest Classifier	0.9740	1.0	0.9772	{'rf__max_depth': None, 'rf__min_samples_leaf': 1, 'rf__n_estimators': 500}

04
Next Steps / Conclusions

Next Steps

Reddits

Look at other similar reddits

Misplaced button

Misplaced Subreddit Post button

Post titles

Use post titles instead of post content

THANKS

Do you have any questions?

Binita Patel

binita.patel.bp@gmail.com