Subreddit Classification ML Model

by Syed Saamir Shamsie

The Problem

Investors struggle to distinguish cryptocurrency from equity forums, leading to imprecise market trend predictions and misinformed investment decisions

01

Workflow

Project Roadmap

Data Collection

- Extract data from two subreddit forums, 'wallstreetbets' and 'CryptoCurrency', using the Reddit API

EDA

- Preprocess and clean the text data
- Engineer new features like text length, word count, and composite sentiment scores

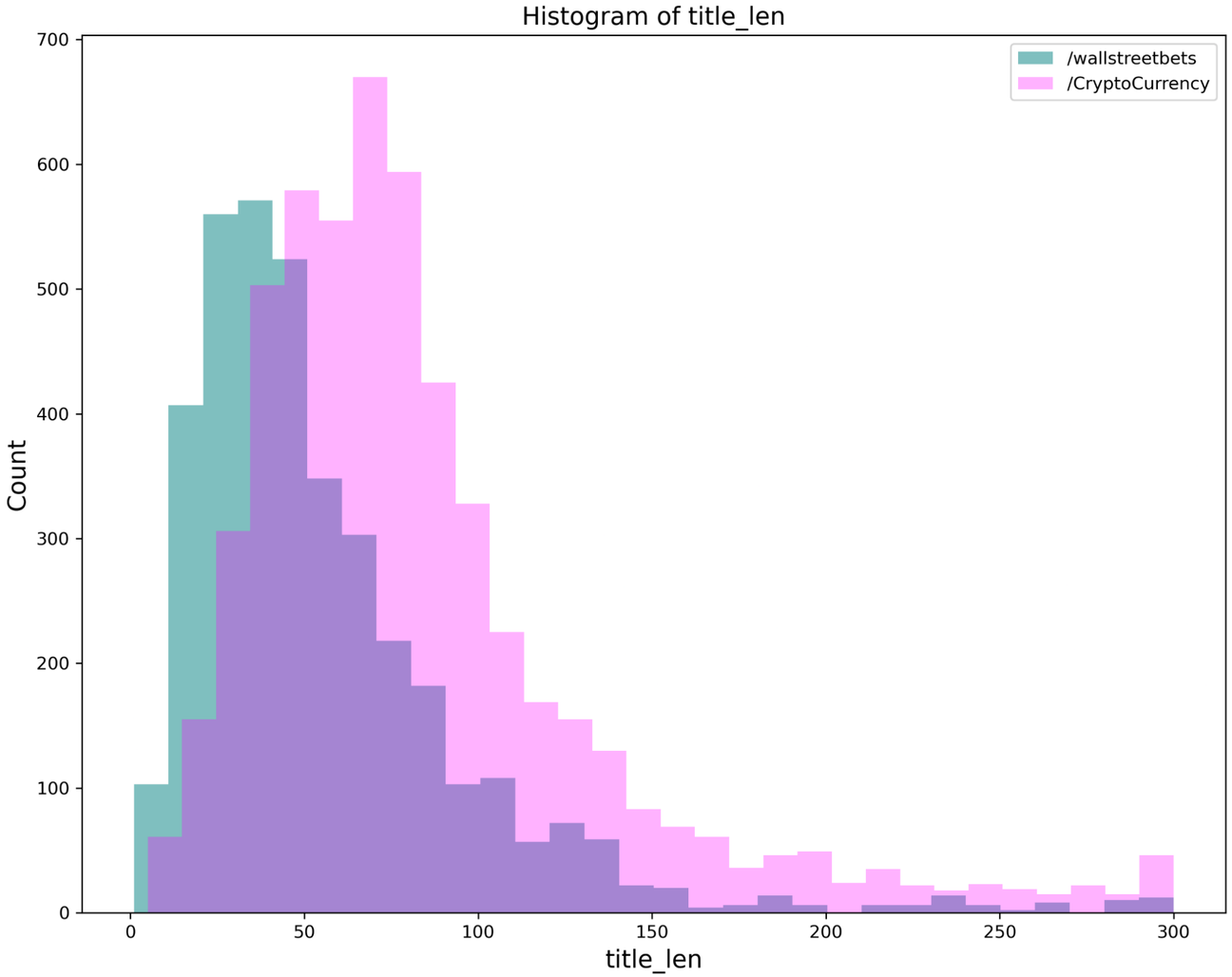
Modeling

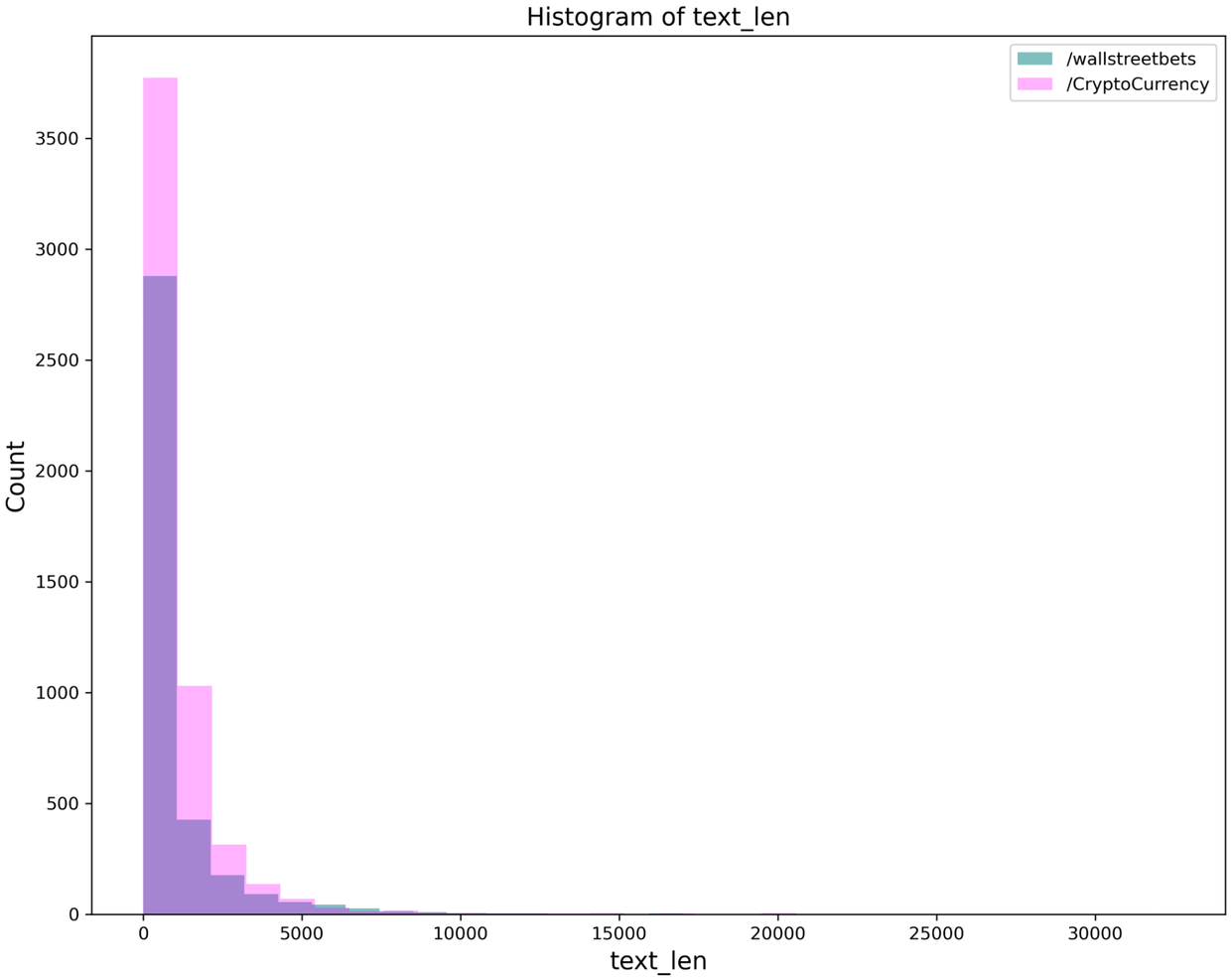
- Construct, optimize, and evaluate two different classification models.

02

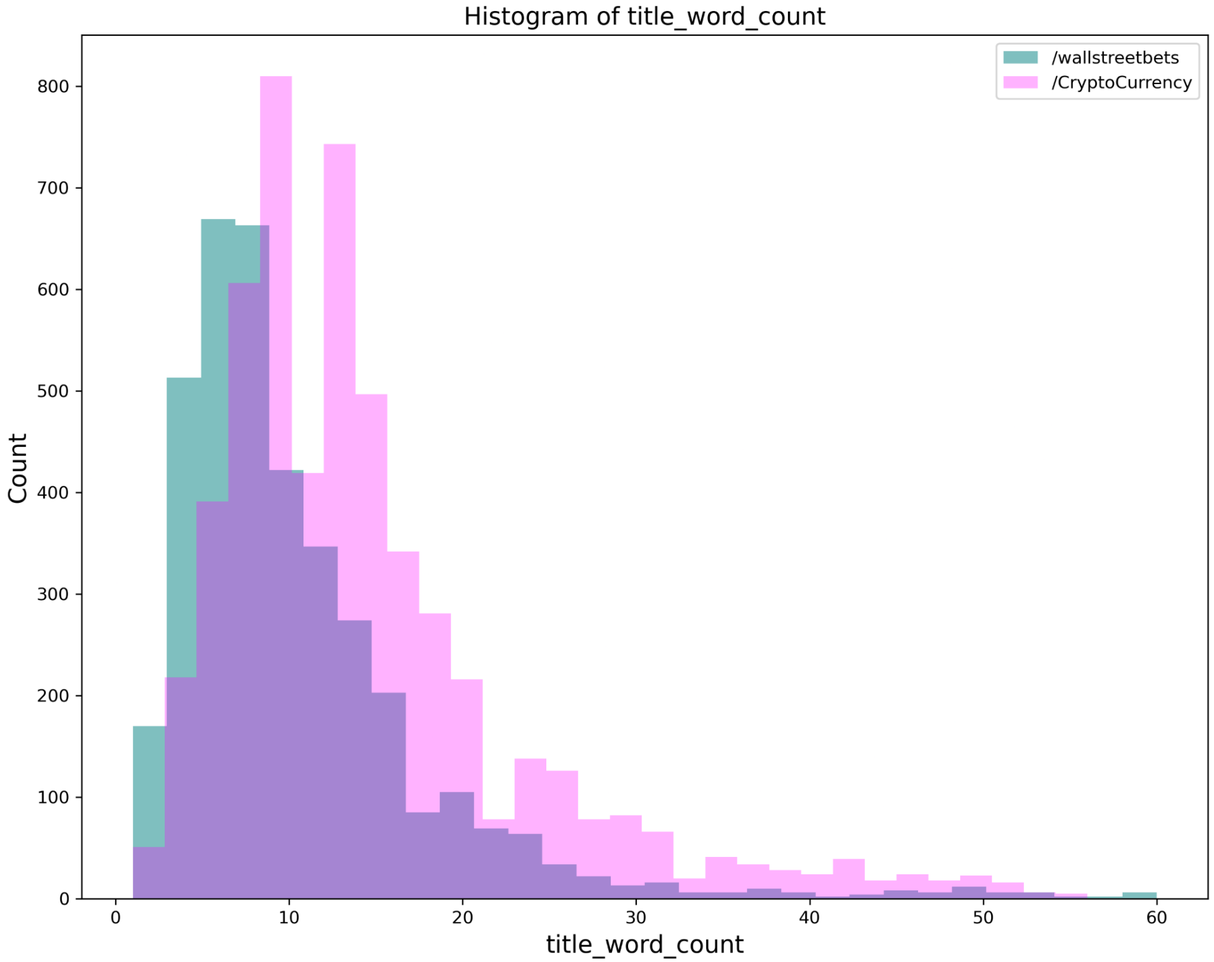
Visuals

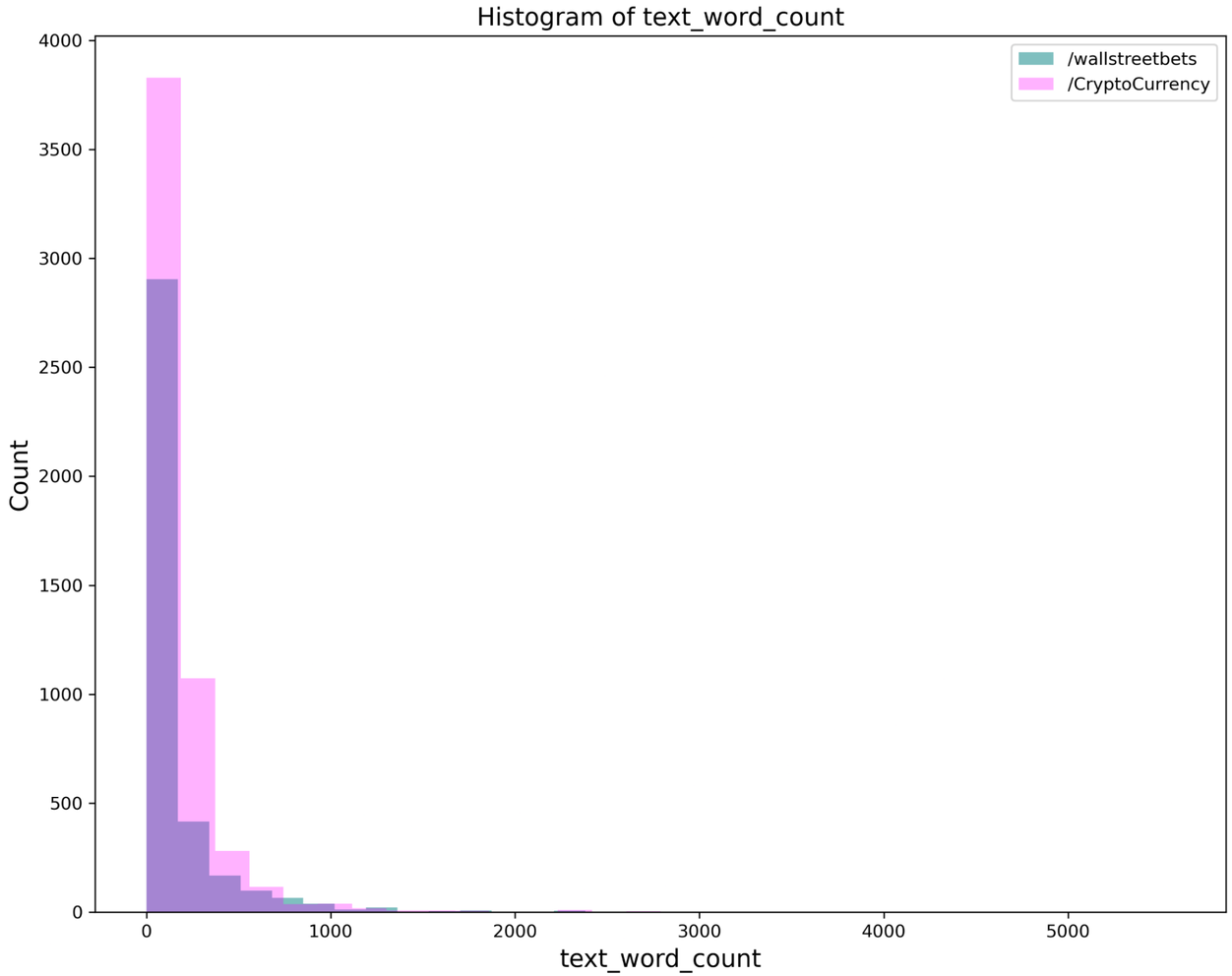
Text Length


Histogram of title_len
/wallstreetbets
/CryptoCurrency
Count
title_len
0
100
200
300
400
500
600
700
0
50
100
150
200
250
300

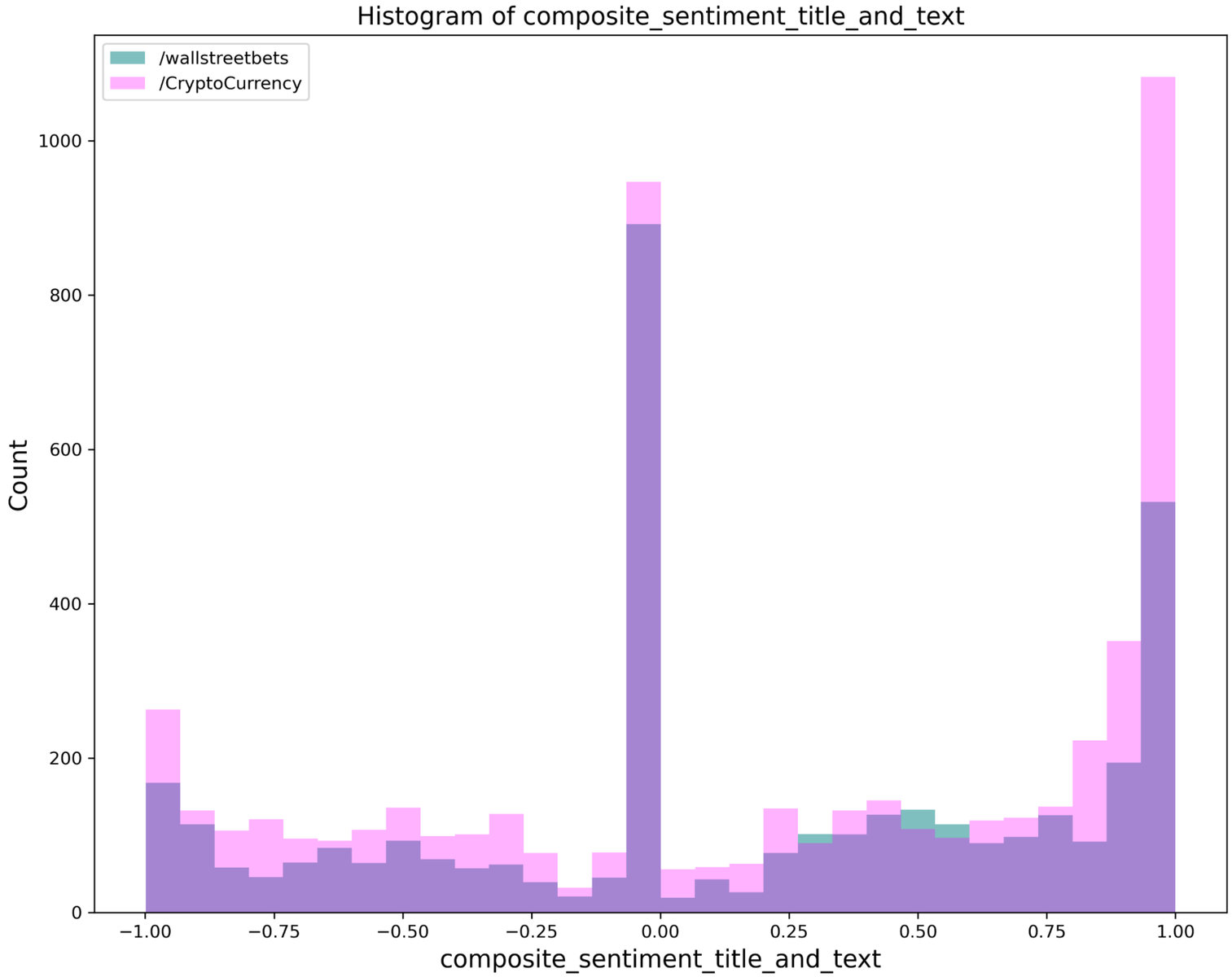

Histogram of text_len
/wallstreetbets
/CryptoCurrency
Count
text_len
0
500
1000
1500
2000
2500
3000
3500
0
5000
10000
15000
20000
25000
30000

Word Count


Histogram of title_word_count
/wallstreetbets
/CryptoCurrency
Count
title_word_count
0
100
200
300
400
500
600
700
800
0
10
20
30
40
50
60


Histogram of text_word_count
/wallstreetbets
/CryptoCurrency
Count
text_word_count
0
500
1000
1500
2000
2500
3000
3500
4000
0
1000
2000
3000
4000
5000

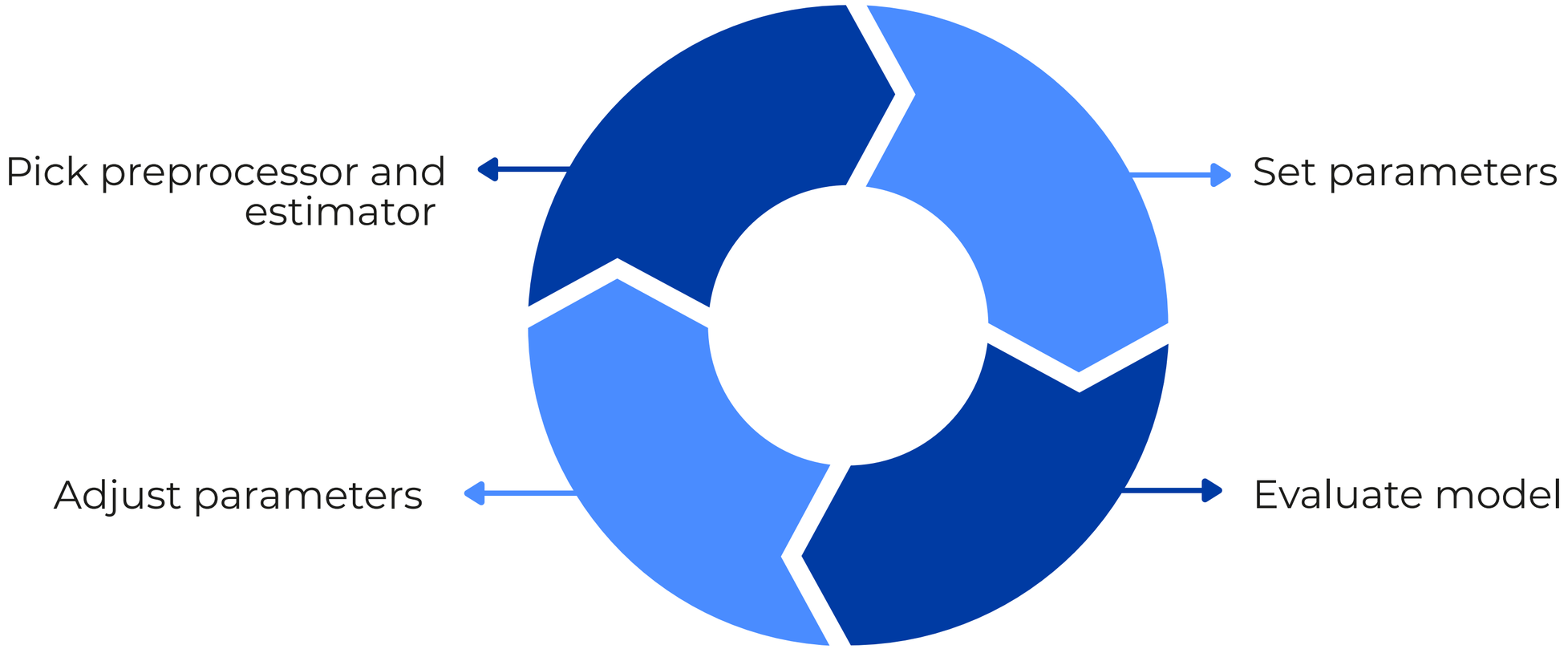
Sentiment Analysis



03

Modeling

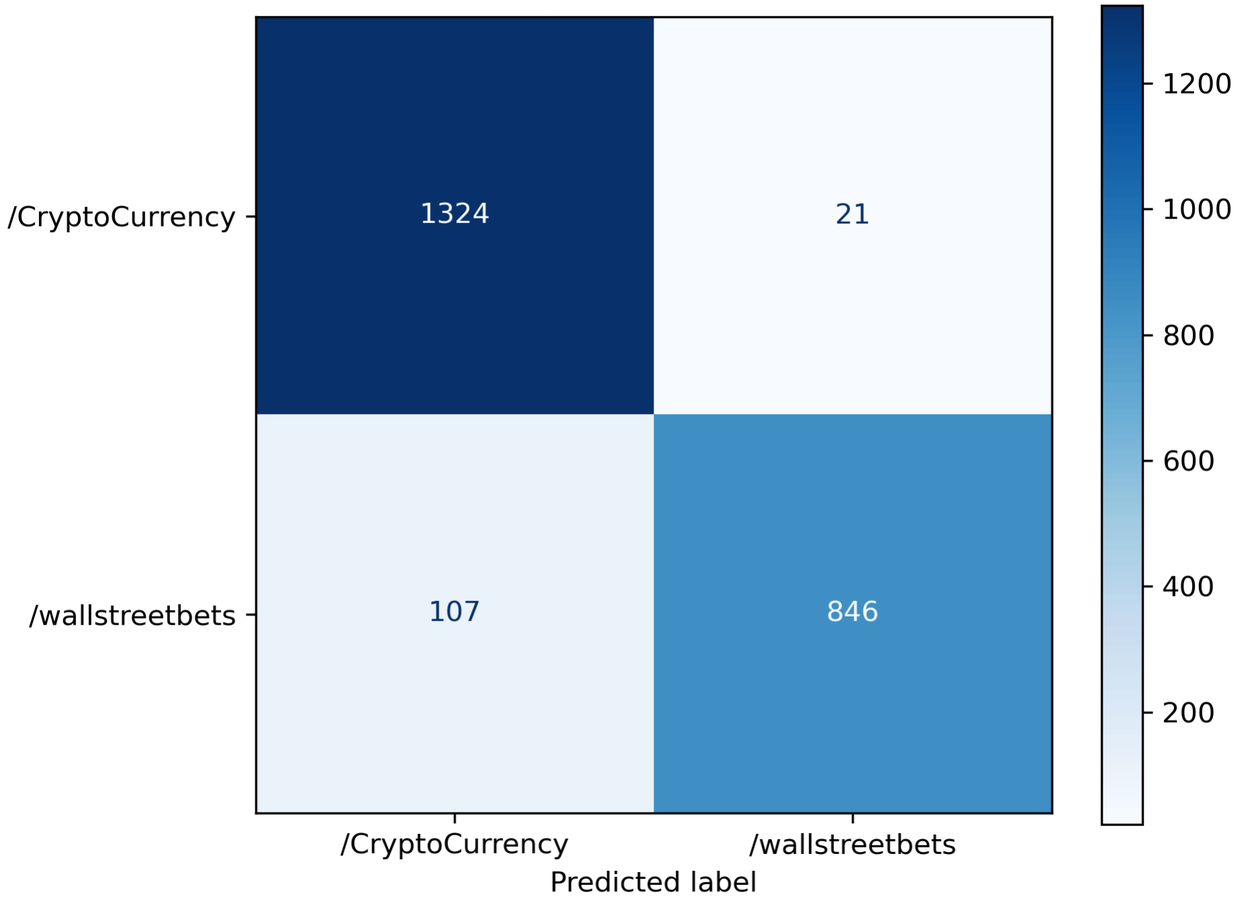
Modeling Flow



Model 1

Preprocessor: CountVectorizer
Estimator: NaiveBayes

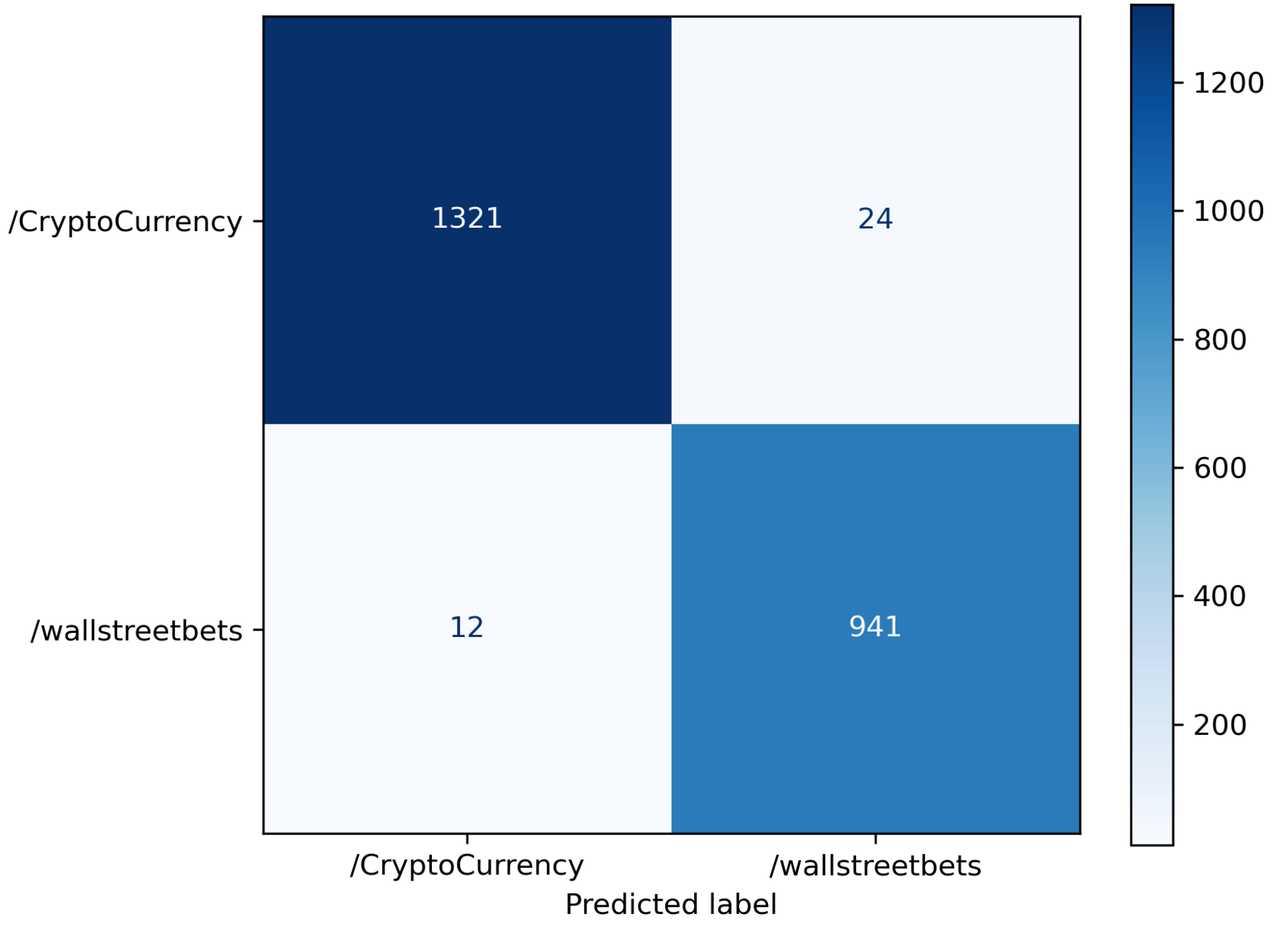
Parameters:
max df: 0.5
min df: 2
max features: 10,000
n-gram range: (1,1)
stop words: 'english'
strip accents: ascii
alpha: 0.12575



Model 2

Preprocessor: TFIDF
Estimator: SVC

Parameters:
SVC C: 2.50075
SVC Kernel: 'rbf'
max features: 4,000
max df: 0.7
min df: 2



Results

Model 1

Train 95.5%

Test 94.4%

Model 2

99.9% Train

98.4% Test

Thanks

Do you have any questions?