Classification Project

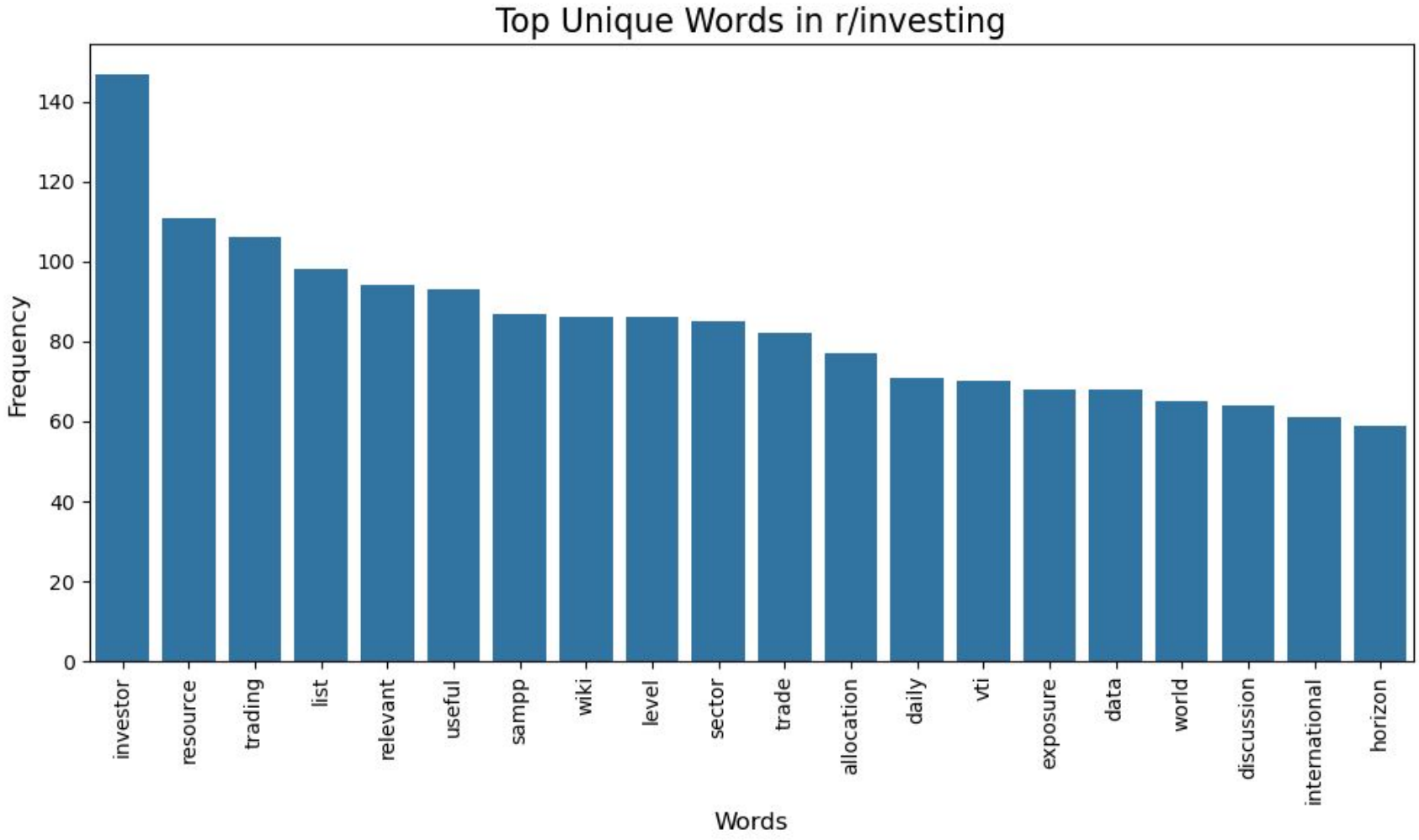
r/investing vs r/personalfinance

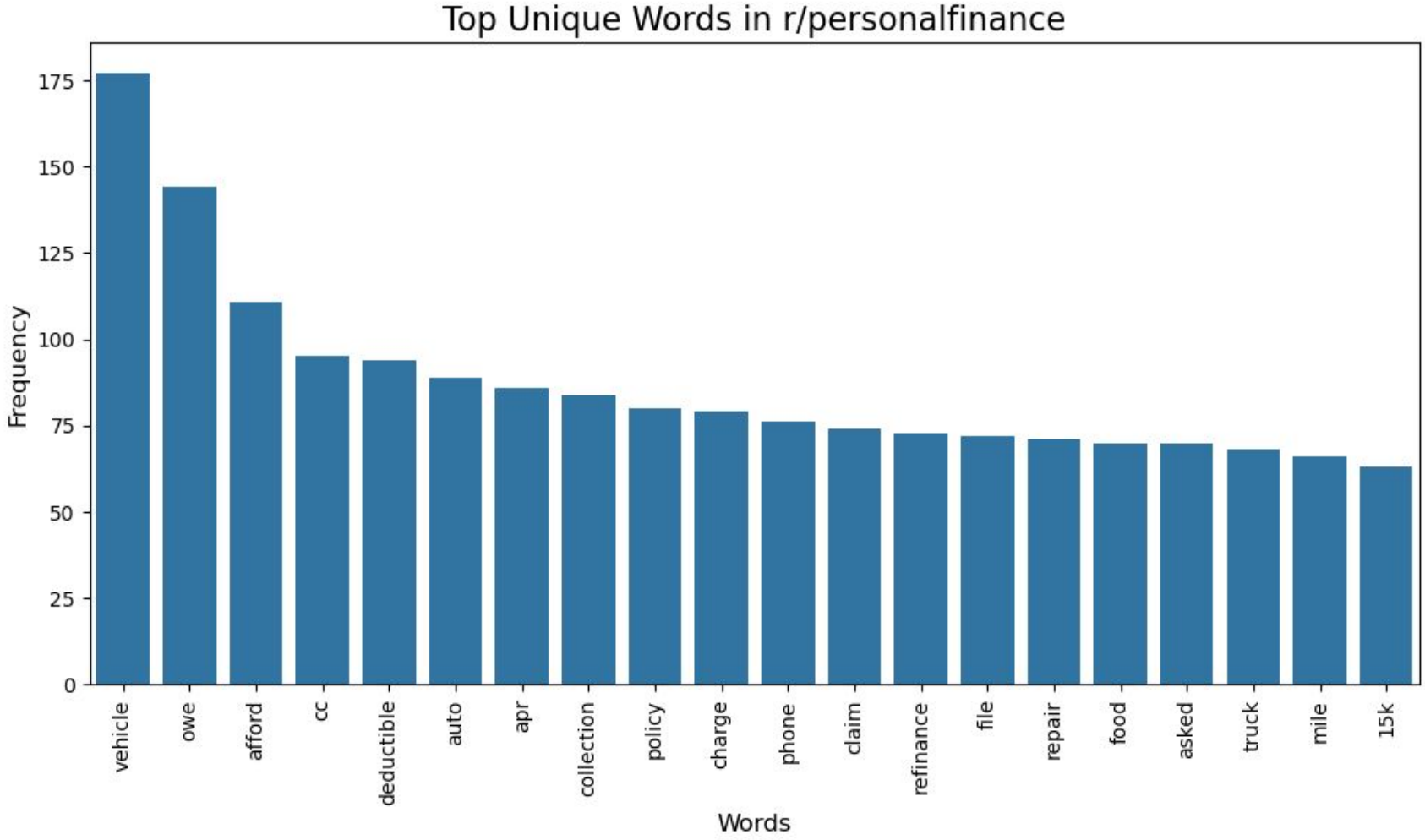
October 12, 2024

Problem Statement

Develop a **classification model** that is able to distinguish between two types of users based on their Reddit posts:

- Users who are experienced investors (professionals)
- Users who are interested in finance but are not heavily investing yet (amateurs).


Top Unique Words in r/investing
Frequency
0
20
40
60
80
100
120
140
investor
resource
trading
list
relevant
useful
sampp
wiki
level
sector
trade
allocation
daily
vti
exposure
data
world
discussion
international
horizon
Words


Top Unique Words in r/personalfinance
Frequency
0
25
50
75
100
125
150
175
vehicle
owe
afford
cc
deductible
auto
apr
collection
policy
charge
phone
claim
refinance
file
repair
food
asked
truck
mile
15k
Words

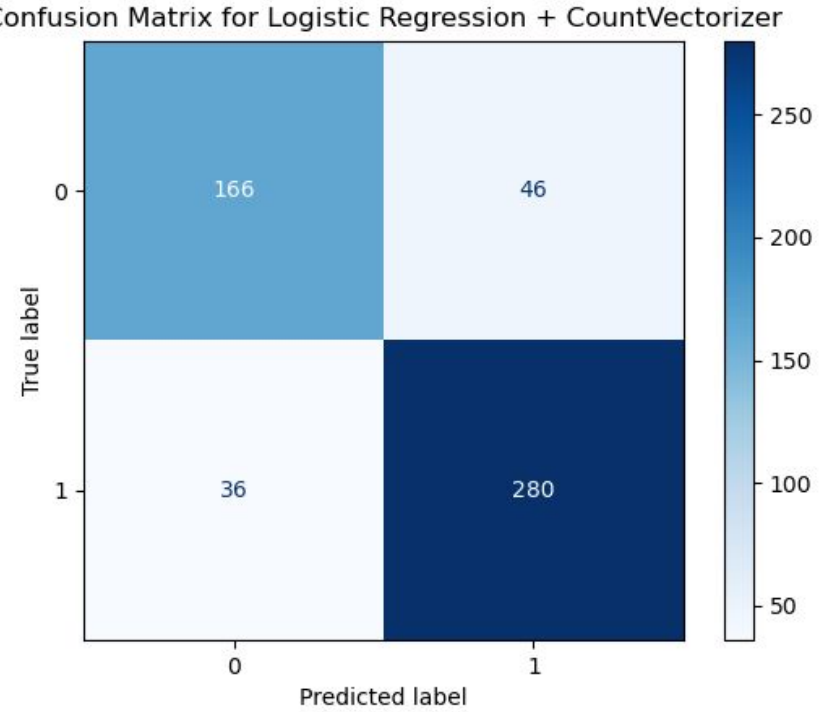
Logistic Regression + CountVectorizer: 0.84

r/investing

- Precision: 0.82
- Recall: 0.78
- F1- score: 0.80

r/personalfinance

- Precision: 0.86
- Recall: 0.89
- F1- score: 0.87



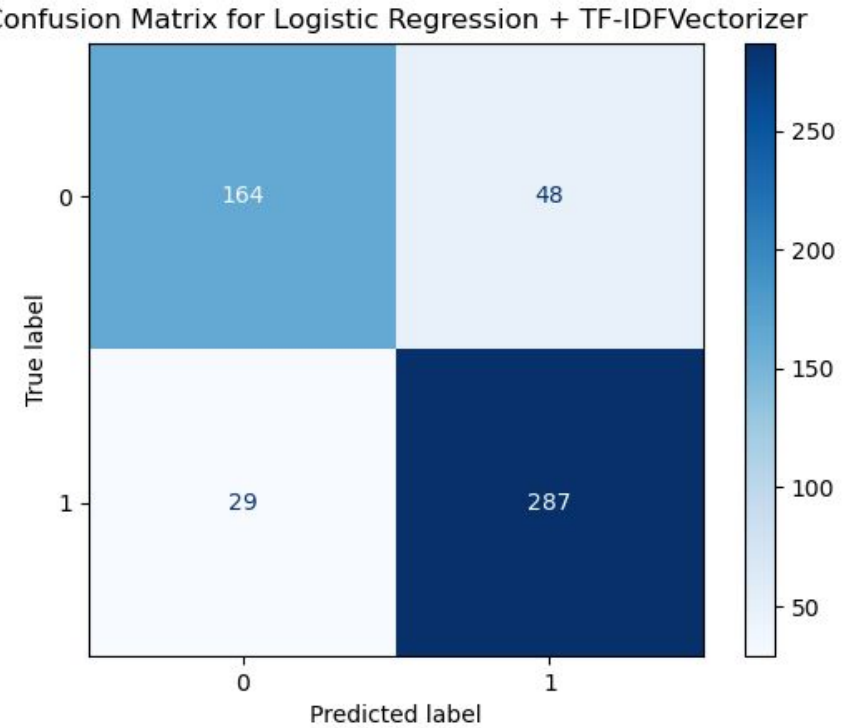
Logistic Regression + TF-IDF: 0.85

r/investing

- Precision: 0.85
- Recall: 0.77
- F1- score: 0.81

r/personalfinance

- Precision: 0.86
- Recall: 0.91
- F1- score: 0.88



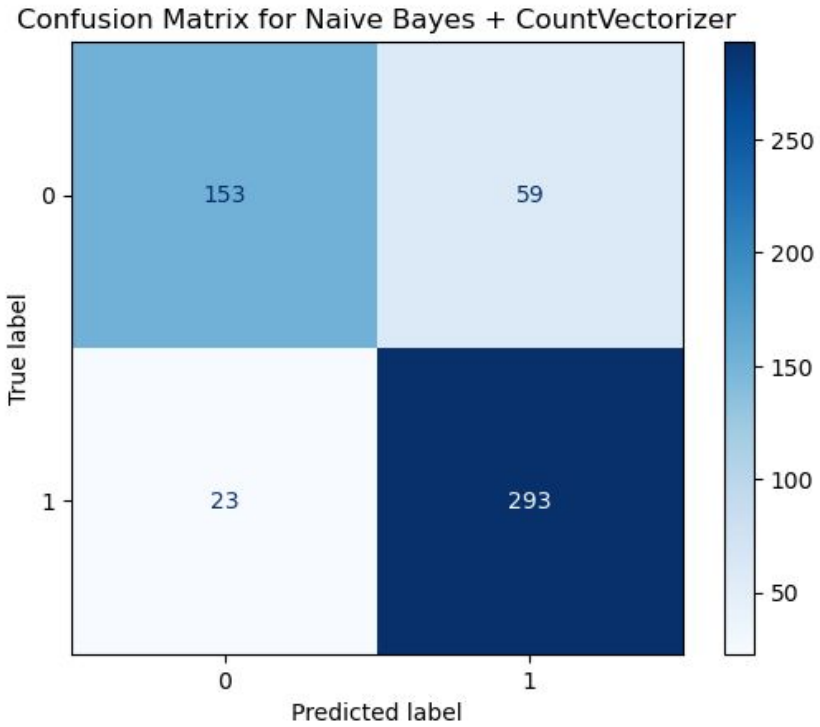
Naive Bayes + CountVectorizer: 0.85

r/investing

- Precision: 0.82
- Recall: 0.81
- F1- score: 0.81

r/personalfinance

- Precision: 0.87
- Recall: 0.88
- F1- score: 0.88



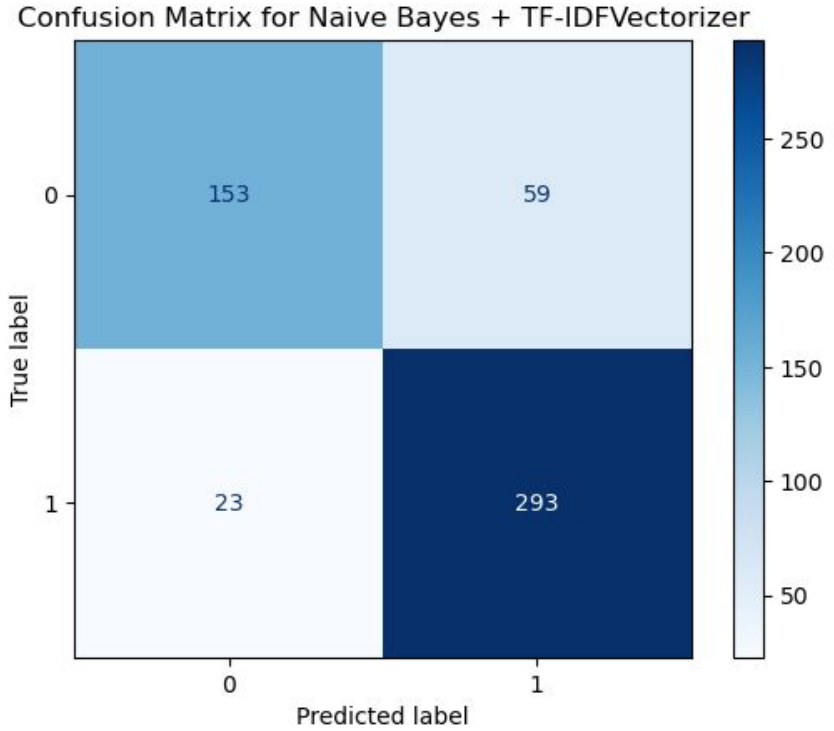
Naive Bayes + TF-IDF: 0.84

r/investing

- Precision: 0.87
- Recall: 0.72
- F1- score: 0.79

r/personalfinance

- Precision: 0.83
- Recall: 0.93
- F1- score: 0.88



Best Model

Logistic Regression with TF-IDF

Thank you!