SUBREDDIT CLASSIFICATION: CLASSIFYING INVESTORS

PROBLEM STATEMENT

- Position: Data Scientist for financial advisory company
- Problem Statement: Classify lower net worth investors who require basic savings/investment plans to higher net worth individuals who prefer high growth investment plans.
- Desired output: We aim to use this model to classify individuals on other forums so that our financial advisors can reach out to them with a suitable financial plan.

DATA SETS

r/povertyfinance

- 1m subscribers
- Financially challenged individuals - low risk low growth investment/savings plans

r/wallstreetbets

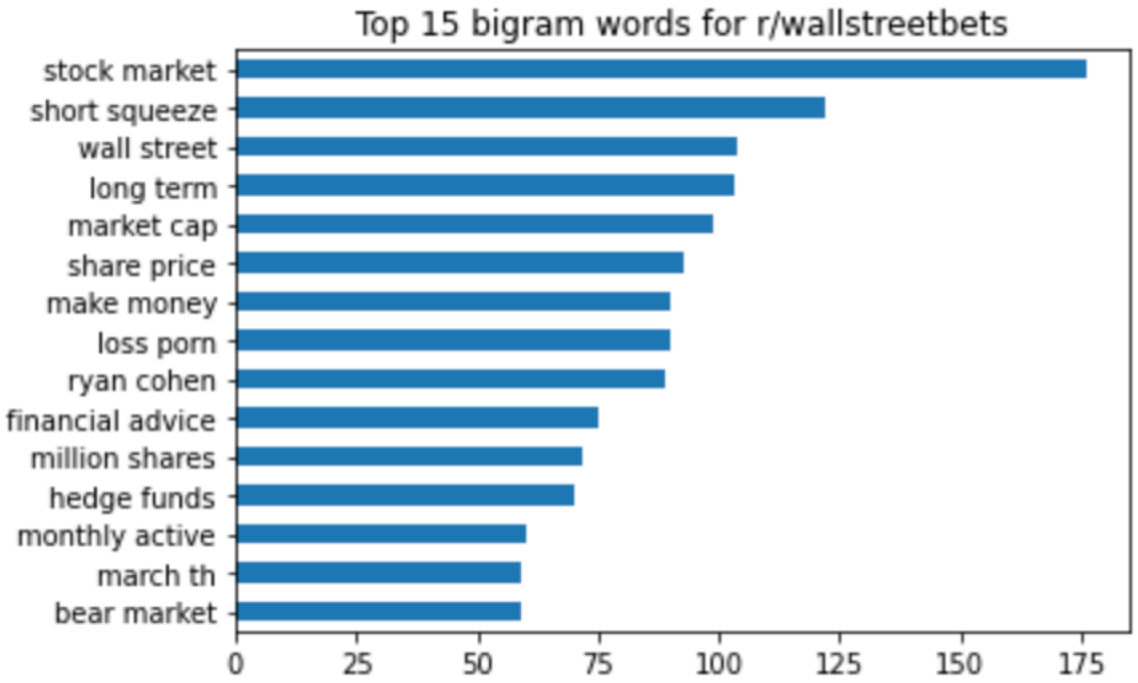
- 11.8m subscribers
- Risk loving investors – High risk high growth investment plans

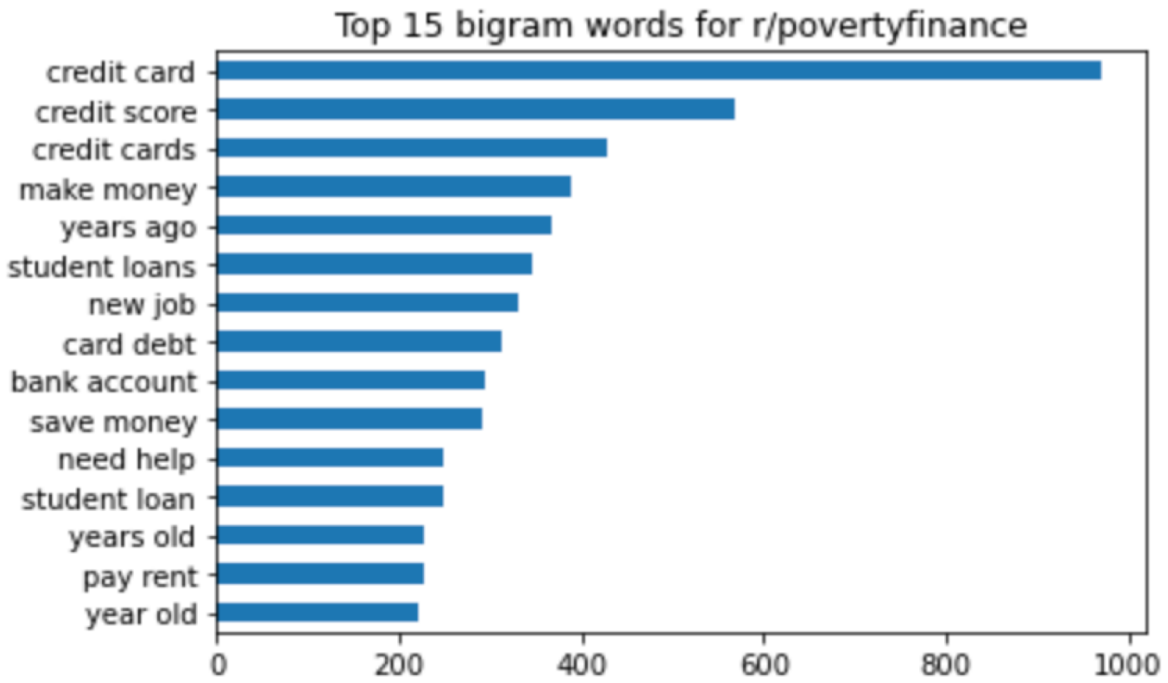
Cleaning

- **Remove duplicates**
- **Remove Unicode, digits, emojis, hyperlinks, double spacing**
- **Tokenize and Lemmatize similar words**

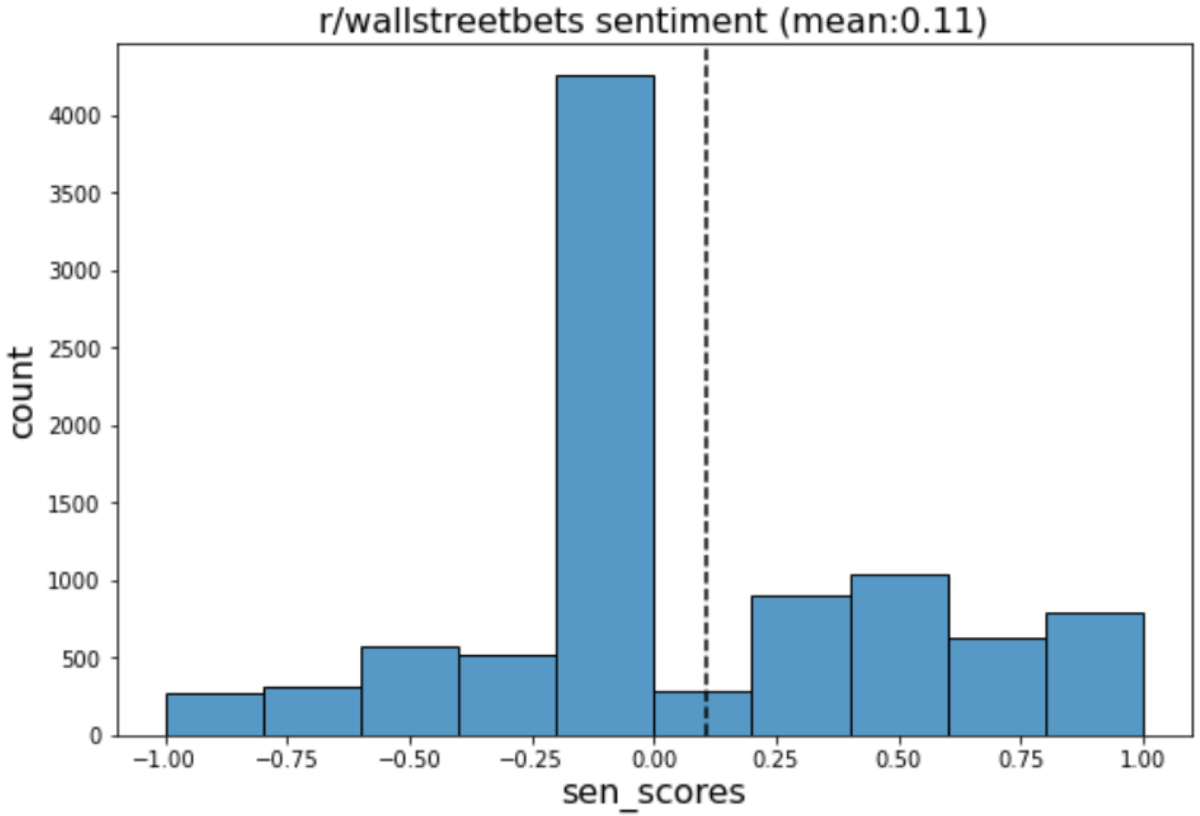
Unigram words

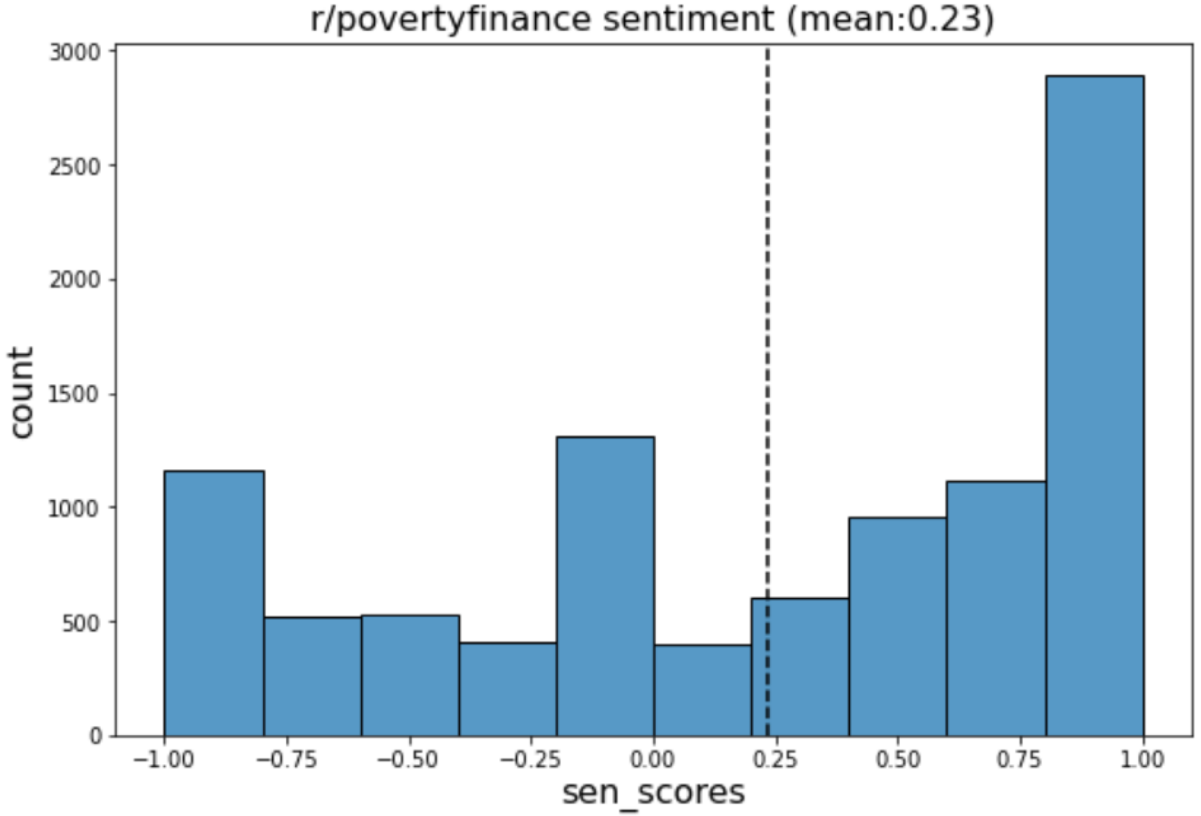
Bigram words





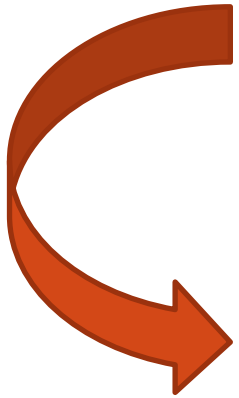
SENTIMENT ANALYSIS





MODELLING

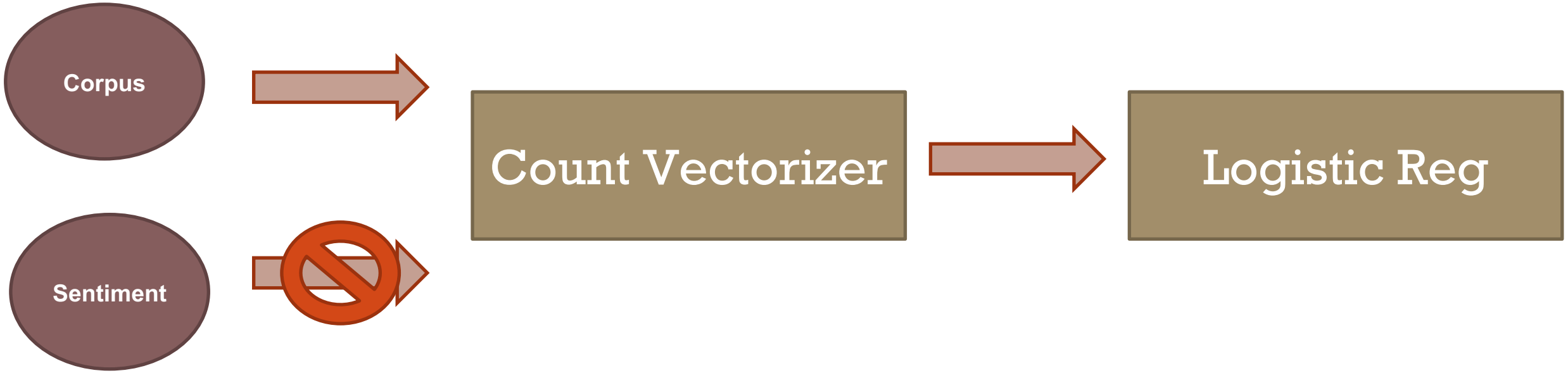
Model	Vectorizer	Train Score	Test Score
Multinomial Naïve Bayes	TFIDF	90.9%	87.8%
	Count	91.5%	90.2%
Logistic Regression	TFIDF	94.1%	91.0%
	Count	97.1%	91.6%
Random Forest	TFIDF	99.8%	86.0%
	Count	99.8%	86.2%



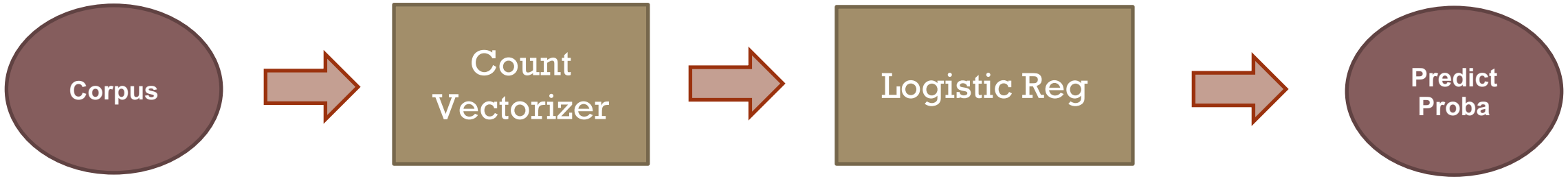
Best Parameters

- **Penalty: L2**
- **C: 1**
- **Max Features: 2000**
- **N_gram range: (1,1)**

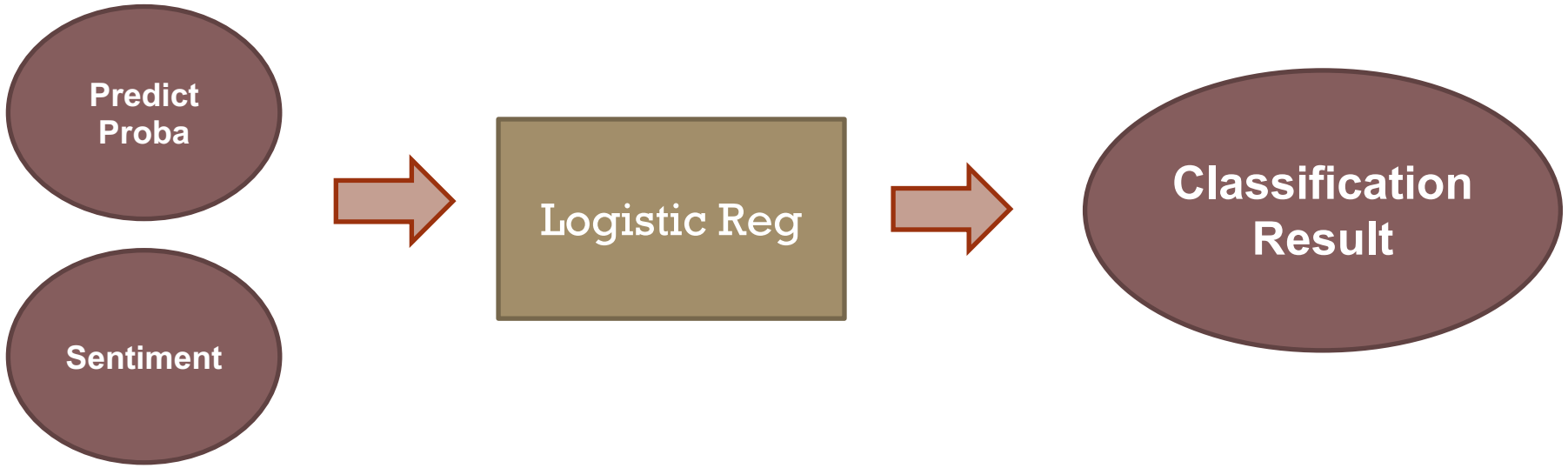
How to include sentiment analysis?



Step 1

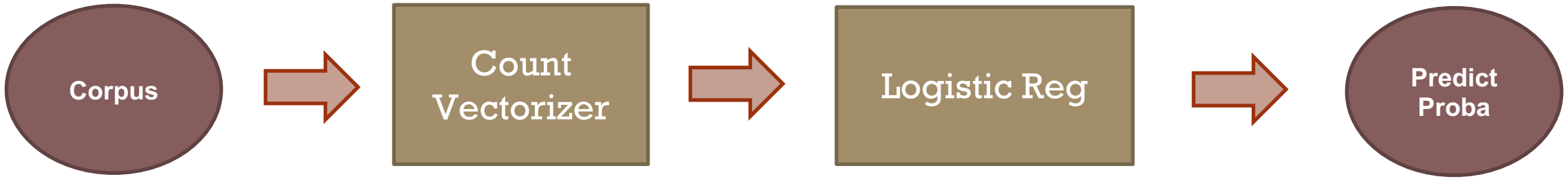

Corpus
Count Vectorizer
Logistic Reg
Predict Proba

Step 2

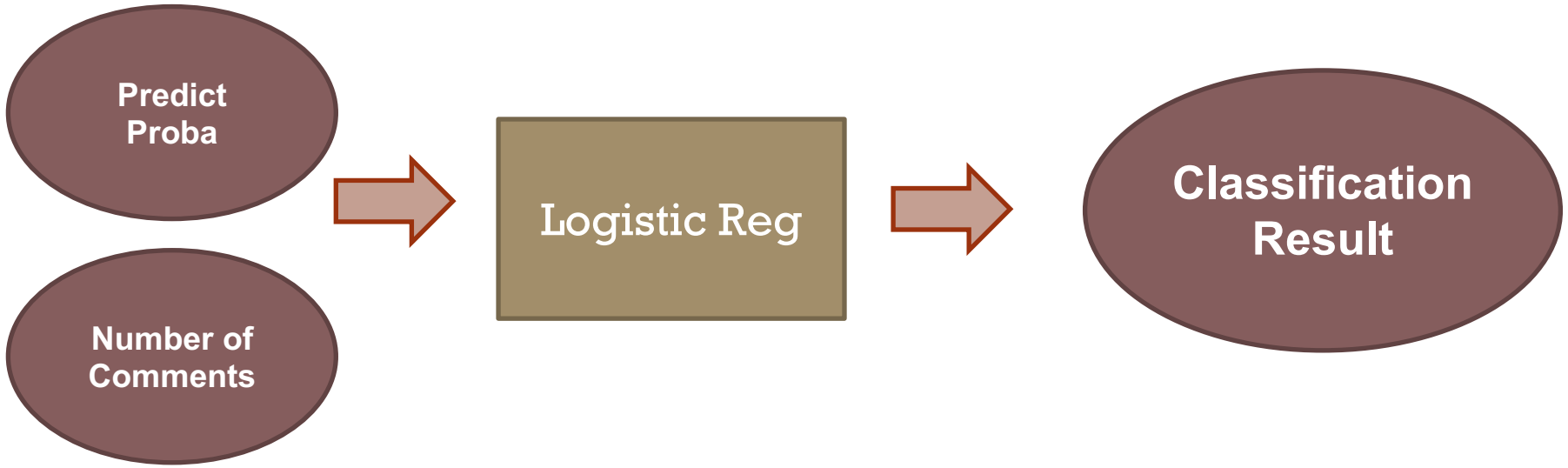

Predict Proba
Sentiment
Logistic Reg
Classification Result

Final Model

Step 1



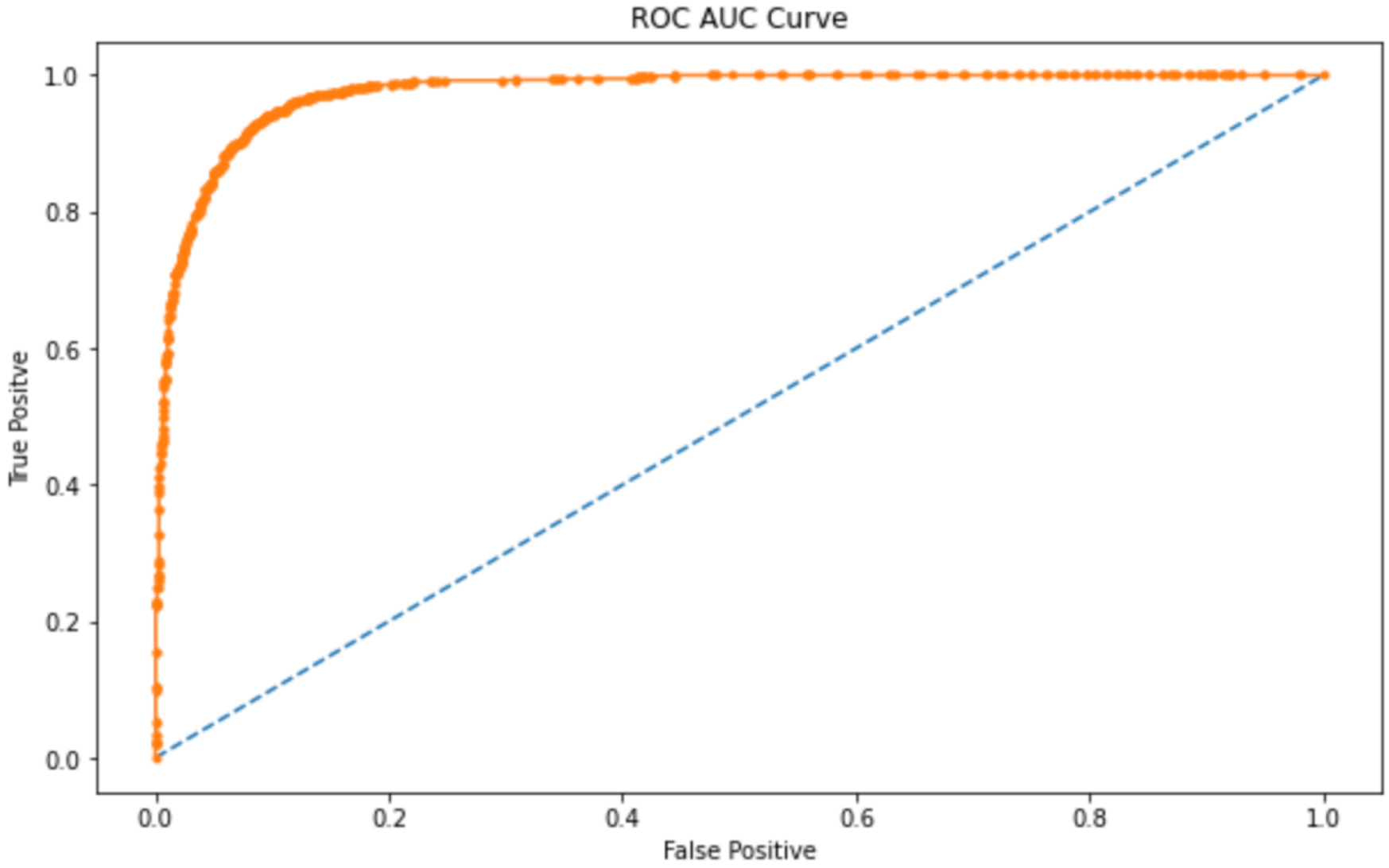
Step 2

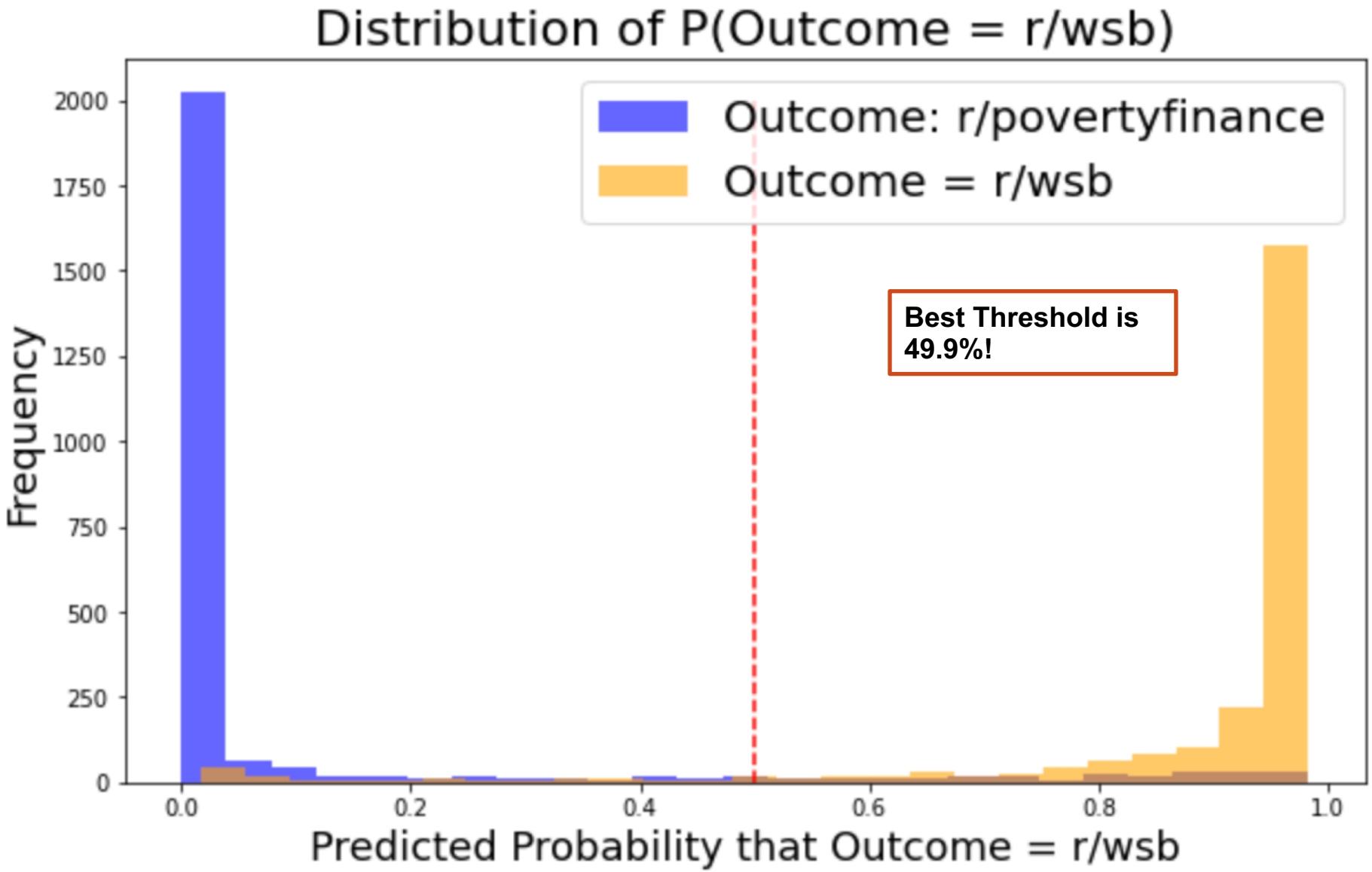


Number of comments

Number of Comments	Mean
r/Povertyfinance	21.46
r/Wallstreetbets	0.48

Financially challenged people posts are a cry for help!


ROC AUC Curve
True Positve
False Positive
0.0
0.2
0.4
0.6
0.8
1.0


Distribution of P(Outcome = r/wsb)
Outcome: r/povertyfinance
Outcome = r/wsb
Best Threshold is 49.9%!
Frequency
2000
1750
1500
1250
1000
750
500
250
0
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability that Outcome = r/wsb

Optimizing Threshold

	Predict r/povertyfinance	**Predict r/wallstreetbets**
Actual r/Povertyfinance	True Negative (+ $1000) - low capital investor sign basic plan	False Positive (+ $500) - low capital investor sign high growth plan
Actual r/wallstreetbets	False Negative (+ $250) - Risk loving investor sign basic plan	True Positive (+ $ 2000) - Risk loving investor sign high growth plan

Summary of Final Model – 'Cvec Logreg + Logreg'

New Threshold: 7%	Predict r/povertyfinance	Predict r/wallstreetbets
Actual r/Povertyfinance	2082 (43%)	415(8%)
Actual r/wallstreetbets	48 (1%)	2325 (48%)

Metrics	Score
Train Score	92.4%
Test Score	90.4%
ROC AUC Score	90.6%
Precision	84.8%
Recall	97.9%
F_score	90.9%

Summary of Findings

Findings	Trend	Words
Financially Challenged Individuals	Posts are more desperate and usually seek help	Help, Need, Credit
	Terms describing basic needs	Food, Car
Individuals seeking high growth high risk investment strategies	Posts are usually more analytical	Short Squeeze, bull/bear market
	More Financial Lingo/Jargon	Moon, Apes

Lessons Learnt

- Training on 10,000 data points makes your model much more effective
- EDA is crucial in ensuring that errors are cleared out

Future Work

- Use a sentiment dictionary more relevant for finance
- Expand ngram_words to investigate trigrams
- Use other models such as knn, boosting
- Use other tools such as word2vec and stemming

SOURCES

1. https://www.forbes.com/sites/georgeschultze/2021/06/15/are-the-apes-now-running-wall-street/?sh=36860ee35e88
2. https://www.fidelity.com/Learning-center/trading-investing/ukraine-russia
3. https://www.vox.com/the-goods/22249458/gamestop-stock-wallstreetbets-reddit-citron
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5147524/
5. https://www.codeproject.com/Articles/5269447/Pros-and-Cons-of-NLTK-Sentiment-Analysis-with-VADE