Natural Language Processing

Edward Tjoe

Problem Statement: Model for Target Marketing

Coach Ella Advisors requests the creation of a machine learning-based algorithm to identify the ways they can **maximize the efficiency of their marketing spend based on their target audience**.

As a newly created firm, they wish for us to **minimize wasteful spending** as they are unable to accept clients for or provide investment advice. However, they would also like for us to find the **most efficient way to identify their potential clients** as well.

Target Metrics:

Accuracy: Maximize identification of target and non-target audience

Sensitivity: Minimize missed target audience

Specificity: Minimize wasteful spending on non-target audience

Precision: Maximize efficient spending on target audience

Data Collection: Reddit

API: PushAPI.io

Subreddits Selected:

Target: r/personalfinance [1138 Obs]

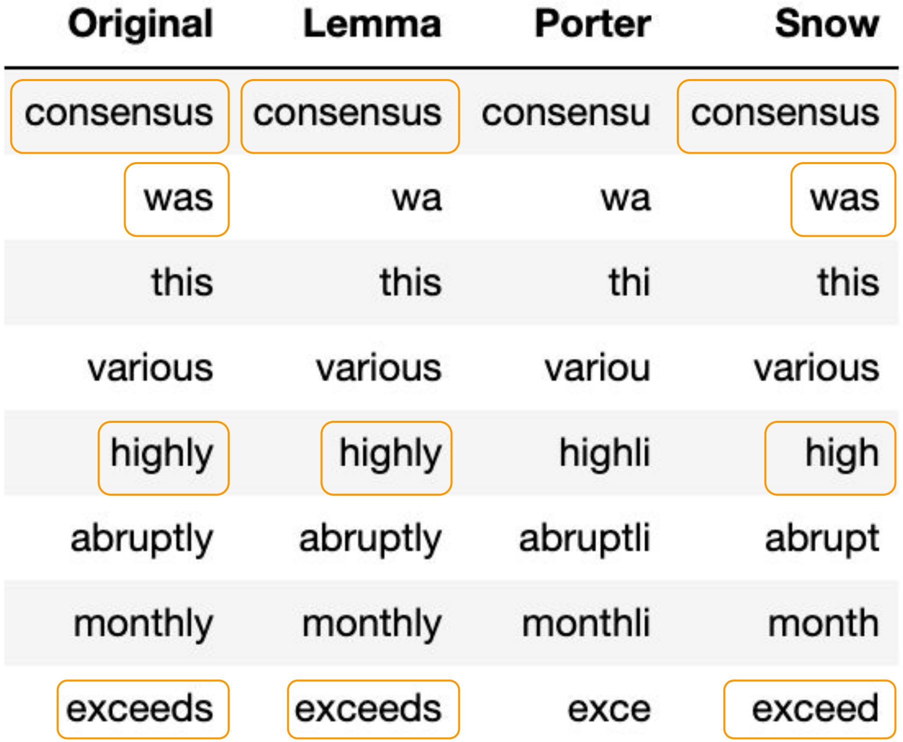
‘Learn about budgeting, saving, getting out of debt, credit, investing, and retirement planning.’

Avoid: r/investing [1113 Obs]

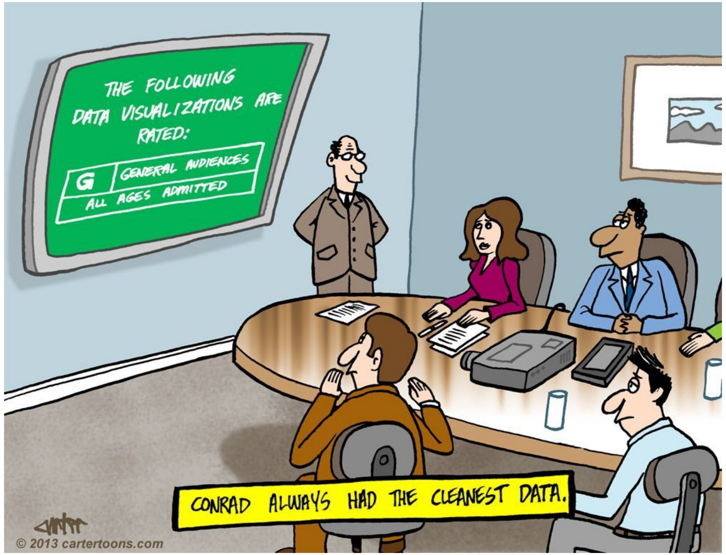
‘Lose money with friends!’

Data Cleaning: Tokenize, Lemma/Stem

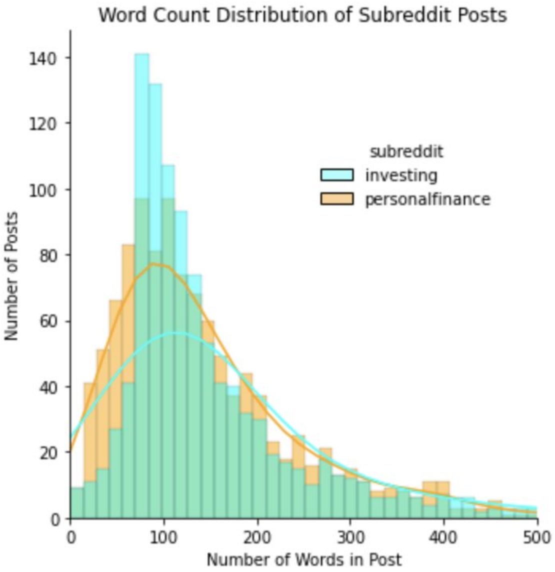
1. Regex:
- URL
- Special characters
- HTML codes
- Digits
2. Remove Stop words
3. Tokenize
4. Lemmatize
5. Porter Stem
6. Snow Stem
7. Compare Results

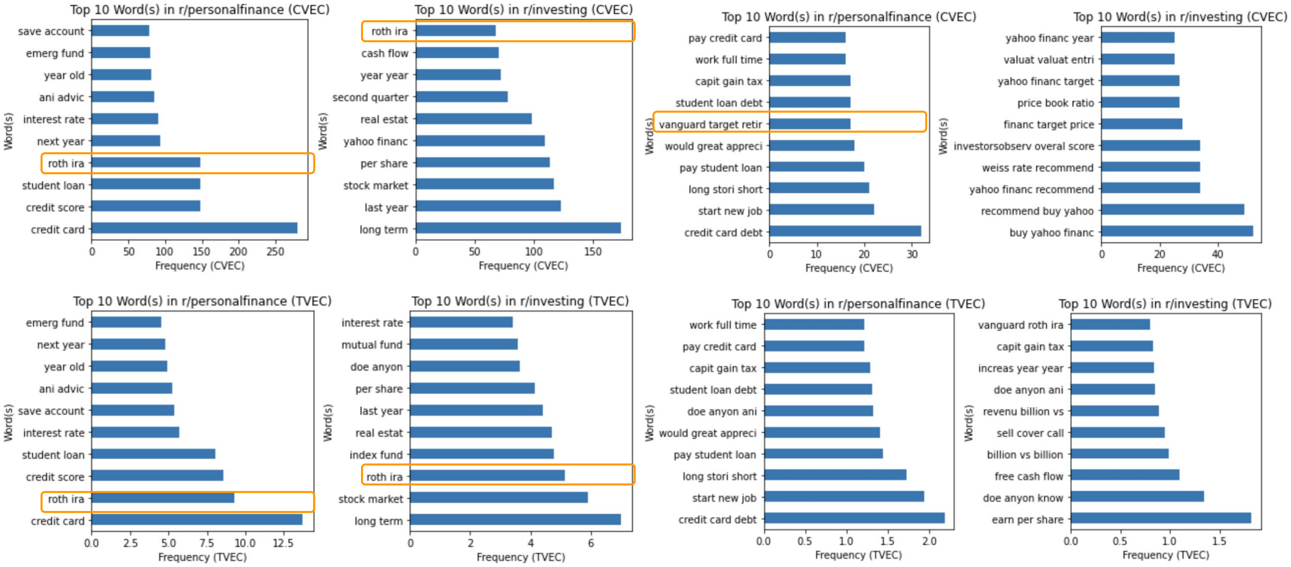


Original	Lemma	Porter	Snow
consensus	consensus	consensu	consensus
was	wa	wa	was
this	this	thi	this
various	various	variou	various
highly	highly	highli	high
abruptly	abruptly	abruptli	abrupt
monthly	monthly	monthli	month
exceeds	exceeds	exce	exceed


THE FOLLOWING DATA VISUALIZATIONS ARE RATED:
G GENERAL AUDIENCES ALL AGES ADMITTED
CONRAD ALWAYS HAD THE CLEANEST DATA.
© 2013 cartertoons.com

Observation: We Write a Similar Length..or Do We?





Unimodal, but sharp increase from previous bin.

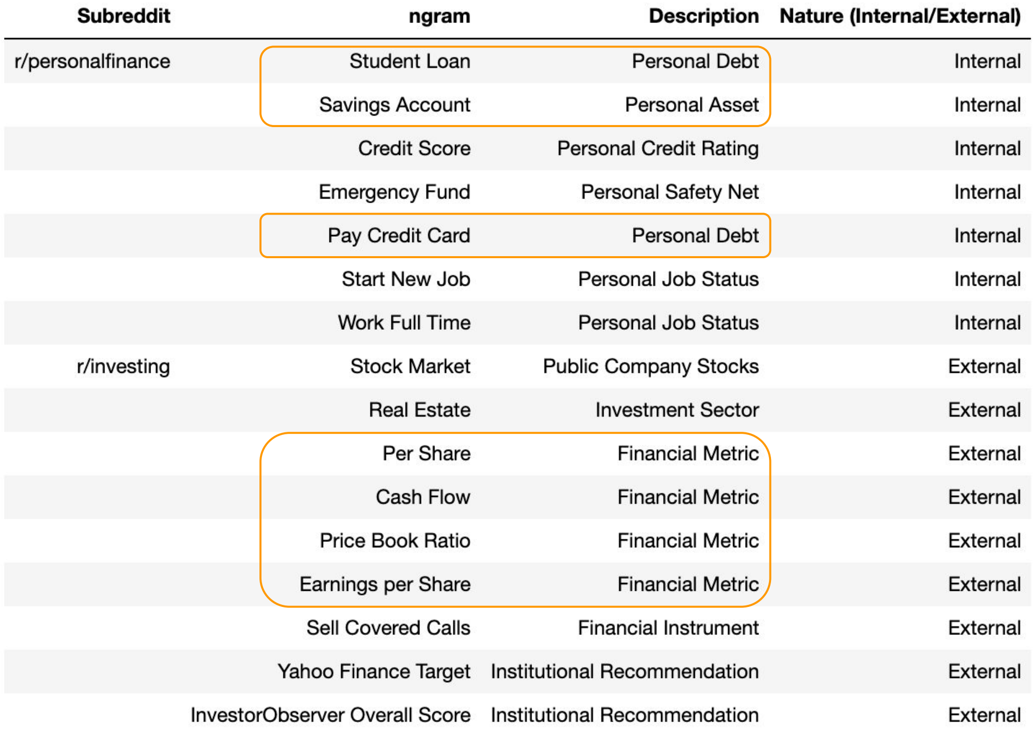
Why? Minimum Word Count!

Why? ‘Effort’

Roth IRA:

1. Retirement Fund
2. Funded with after-tax money (Tax-free)

Observation: But We Talk About Different Stuff Too



Subreddit	ngram	Description	Nature (Internal/External)
r/personalfinance	Student Loan	Personal Debt	Internal
	Savings Account	Personal Asset	Internal
	Credit Score	Personal Credit Rating	Internal
	Emergency Fund	Personal Safety Net	Internal
	Pay Credit Card	Personal Debt	Internal
	Start New Job	Personal Job Status	Internal
	Work Full Time	Personal Job Status	Internal
r/investing	Stock Market	Public Company Stocks	External
	Real Estate	Investment Sector	External
	Per Share	Financial Metric	External
	Cash Flow	Financial Metric	External
	Price Book Ratio	Financial Metric	External
	Earnings per Share	Financial Metric	External
	Sell Covered Calls	Financial Instrument	External
	Yahoo Finance Target	Institutional Recommendation	External
	InvestorObserver Overall Score	Institutional Recommendation	External

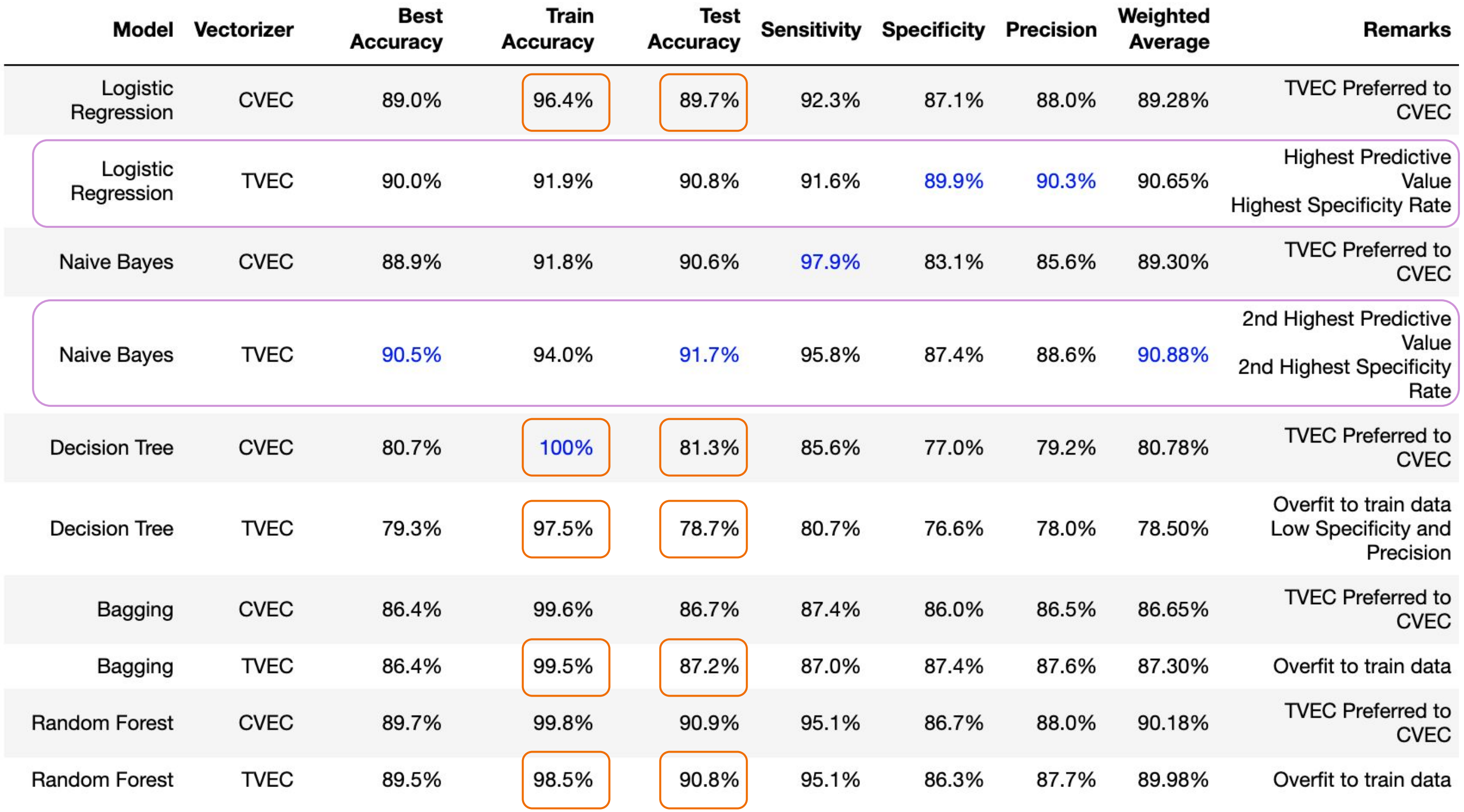
Posts have different nature:

- Internal vs External Focus

- Personal vs Impersonal

- Conservative vs Opportunistic

Modeling: Logit, NB, DTree, Bag, RF



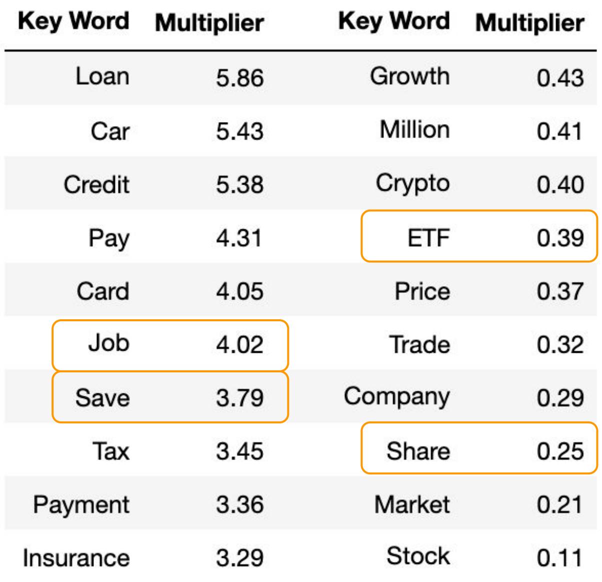
Model	Vectorizer	Best Accuracy	Train Accuracy	Test Accuracy	Sensitivity	Specificity	Precision	Weighted Average	Remarks
Logistic Regression	CVEC	89.0%	96.4%	89.7%	92.3%	87.1%	88.0%	89.28%	TVEC Preferred to CVEC
Logistic Regression	TVEC	90.0%	91.9%	90.8%	91.6%	89.9%	90.3%	90.65%	Highest Predictive Value Highest Specificity Rate
Naive Bayes	CVEC	88.9%	91.8%	90.6%	97.9%	83.1%	85.6%	89.30%	TVEC Preferred to CVEC
Naive Bayes	TVEC	90.5%	94.0%	91.7%	95.8%	87.4%	88.6%	90.88%	2nd Highest Predictive Value 2nd Highest Specificity Rate
Decision Tree	CVEC	80.7%	100%	81.3%	85.6%	77.0%	79.2%	80.78%	TVEC Preferred to CVEC
Decision Tree	TVEC	79.3%	97.5%	78.7%	80.7%	76.6%	78.0%	78.50%	Overfit to train data Low Specificity and Precision
Bagging	CVEC	86.4%	99.6%	86.7%	87.4%	86.0%	86.5%	86.65%	TVEC Preferred to CVEC
Bagging	TVEC	86.4%	99.5%	87.2%	87.0%	87.4%	87.6%	87.30%	Overfit to train data
Random Forest	CVEC	89.7%	99.8%	90.9%	95.1%	86.7%	88.0%	90.18%	TVEC Preferred to CVEC
Random Forest	TVEC	89.5%	98.5%	90.8%	95.1%	86.3%	87.7%	89.98%	Overfit to train data

Hyperparameters tuned for overall best model with slight focus on Specificity:

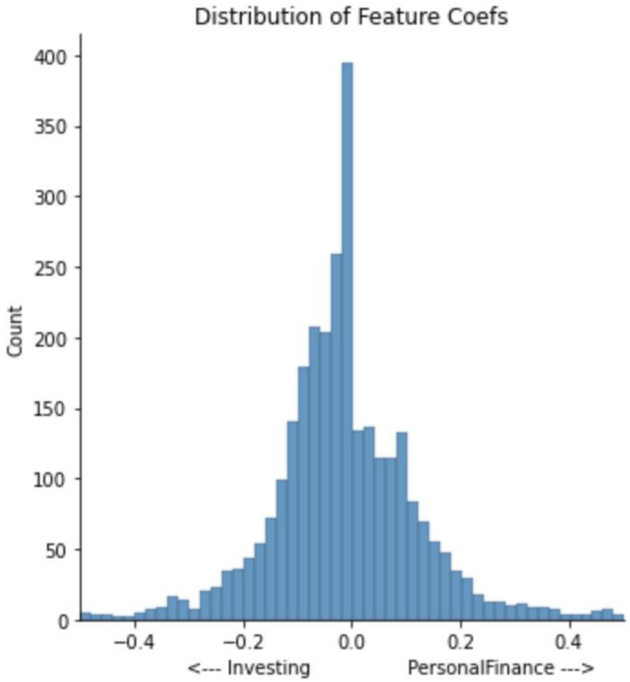
- Max features
- Ngram range
- Max depth (trees)
- Max samples
- N Estimators

1. TF-IDF better than CountVectorizer
   - Penalty on common words
   - Greater weight on focus words
2. Overfitting to train data
   - Large drop from train to test accuracy
   - Likely to perform even worse to blind data

Feature Analysis: Feature Importance and Distribution



Key Word	Multiplier	Key Word	Multiplier
Loan	5.86	Growth	0.43
Car	5.43	Million	0.41
Credit	5.38	Crypto	0.40
Pay	4.31	ETF	0.39
Card	4.05	Price	0.37
Job	4.02	Trade	0.32
Save	3.79	Company	0.29
Tax	3.45	Share	0.25
Payment	3.36	Market	0.21
Insurance	3.29	Stock	0.11



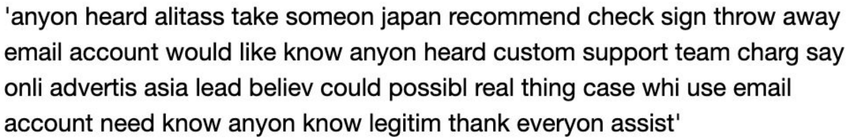
- Multipliers after applying exponential
- ‘Loan’: 1 unit increase translates to 5.86x likelihood to originate from r/personalfinance

- High peak at 0.0 → Many weak features
- Higher area in left of 0.0 → Stronger focus words for r/investing

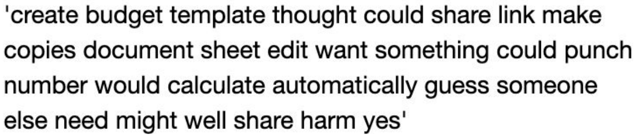
Paying credit card and saving to buy company shares → 6.37x

Misclassification Analysis

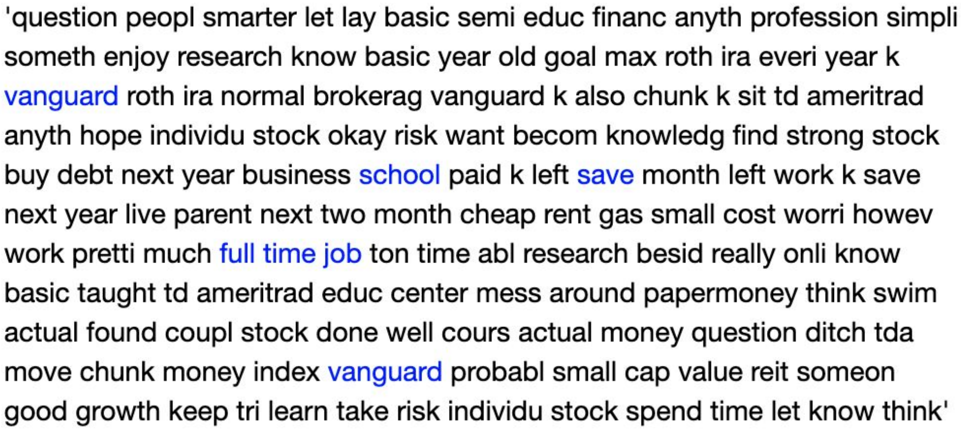
1. Lack of contextual understanding
2. Small number of words (29 Words)
3. Cross-related post topics
4. Focus words



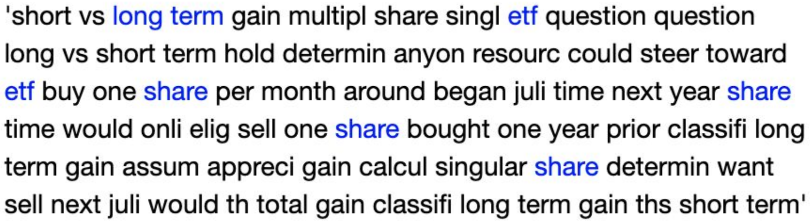
'anyon heard alitass take someon japan recommend check sign throw away email account would like know anyon heard custom support team charg say onli advertis asia lead believ could possibl real thing case whi use email account need know anyon know legitim thank everyon assist'



'create budget template thought could share link make copies document sheet edit want something could punch number would calculate automatically guess someone else need might well share harm yes'



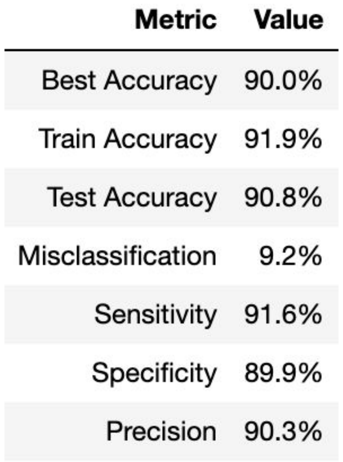
'question peopl smarter let lay basic semi educ financ anyth profession simpli someth enjoy research know basic year old goal max roth ira everi year k vanguard roth ira normal brokerag vanguard k also chunk k sit td ameritrad anyth hope individu stock okay risk want becom knowledg find strong stock buy debt next year business school paid k left save month left work k save next year live parent next two month cheap rent gas small cost worri howev work pretti much full time job ton time abl research besid really onli know basic taught td ameritrad educ center mess around papermoney think swim actual found coupl stock done well cours actual money question ditch tda move chunk money index vanguard probabl small cap value reit someon good growth keep tri learn take risk individu stock spend time let know think'



'short vs long term gain multipl share singl etf question question long vs short term hold determin anyon resourc could steer toward etf buy one share per month around began juli time next year share time would onli elig sell one share bought one year prior classifi long term gain assum appreci gain calcul singular share determin want sell next juli would th total gain classifi long term gain ths short term'

Conclusion

- Snow Stemmer performs better than Lemma, Porter, Original
- TF-IDF Vectorizer performs better than Countvectorizer
- Caution on model overfitting.
- Misclassification happens.
- Problem Statement:
  - Maximize the efficiency of their marketing spend
  - Minimize wasteful spending



Metric	Value
Best Accuracy	90.0%
Train Accuracy	91.9%
Test Accuracy	90.8%
Misclassification	9.2%
Sensitivity	91.6%
Specificity	89.9%
Precision	90.3%

Recommendation: TF-IDF Logistic Regression Model

1. Expand data collection from other sources (not just Reddit).
2. Identifying more stop words to reduce noise in our data.
3. Creating a dictionary to process words (stemming) more appropriately.
4. Obtain greater computing power or more time to process a greater number of hyperparameters.