NYC DATA SCIENCE ACADEMY

Scraping Alpha$^{\alpha}$

STOCK PREDICTION THROUGH SENTIMENT ANALYSIS TECHNIQUES

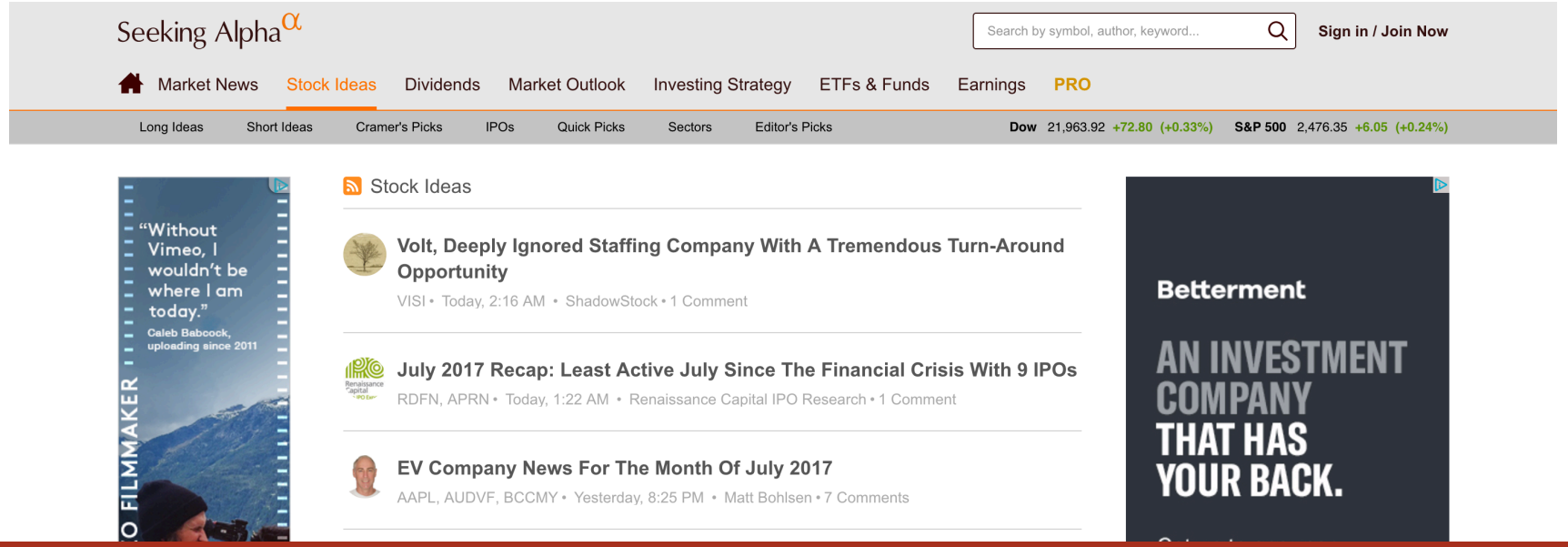
AARSH SACHDEVA

Table of Contents

Seeking Alpha

Seeking Alpha is crowd-sourced content service website for financial markets. Published articles and research provide financial commentary and analysis on a wide-variety of topics in the world of finance, such as publicly traded equities, credit, ETFs, investment strategies, company news, and current events.

I chose to scrape stock idea articles



Sentiment Analysis

Sentiment Classification

Typical sentiment analysis attempts to predict the writer's orientation towards an object by analyzing the words they use to describe it.

For example:

- Positive Sentiment: "…zany characters and richly applied satire, and some great plot twists"
- Negative Sentiment: "It was pathetic. The worst part about it was the boxing scenes…"

The goal of this analysis is to assign an article a classification of "high" or "low" that indicates whether the article is successful at predicting positive returns.

Bayes Theorem

Bayes Theorem is based upon the laws of conditional probability

$$P(A|B)P(B) = P(B|A)P(A)$$

Relative size	Case B	Case $\bar{B}$	Total
Condition A	w	x	$w+x$
Condition $\bar{A}$	y	z	$y+z$
Total	$w+y$	$x+z$	$w+x+y+z$

$$P(A|B) \times P(B) = \frac{w}{w+y} \times \frac{w+y}{w+x+y+z} = \frac{w}{w+x+y+z}$$

$$P(B|A) \times P(A) = \frac{w}{w+x} \times \frac{w+x}{w+x+y+z} = \frac{w}{w+x+y+z}$$

Naïve Bayes Theorem

A Naive Bayes probabilistic classifier attempts to predict which category a piece of text belongs to by returning the category of maximum likelihood given the words in the text

$$Predicted\ Class = argmax_{class\ \in Possible\ Classes}\ P(class|article) = \frac{P(article|class)P(class)}{P(article)}$$

Since P(article) will be the same across all calculations for every class, we can ignore this term

Naïve Bayes Theorem

An article can be represented by the words in it, that is:

$$P(article|class) = P(word_1, word_2, word_3|class) = P(word_1\ |class)P(word_2, |class)P(word_3\ |class)$$

Note: In implementation, the logarithm of each side will be taken and probabilities will be summed

This theorem is called "naïve" because it assumes that the probability of a word being associated with a certain class is independent other words' probabilities of being associated with that class

The probability of a word belonging to a class is calculated as follows:

$$\hat{P}(word_i|class) = \frac{count(word_i|class) + 1}{(\sum_{w \in V} count(word|class)) + |V|}$$

Here, V is the full range of words across all classes

Binary Naïve Bayes Theorem: Example

The following sentences constitute our training data and have already been classified:

- Negative: “it was pathetic the worst part was the boxing scenes”
- Negative: “no plot twists or great scenes”
- Positive: “and satire and great plot twists”
- Positive: ”great scenes and great film”

Binary Naïve Bayes Theorem will first reduce each sentence to remove duplicates within each sentence:

- Negative: “it was pathetic the worst part boxing scenes”
- Negative: “no plot twists or great scenes”
- Positive: “and satire great plot twists”
- Positive: ”great scenes film”

This training data will then be used to classify the following test sentence:

- “great plot with awesome boxing scenes”

Binary Naïve Bayes Theorem: Example

By our training set, we know the following prior probabilities are true:

- P(-) = 2/4 = 1/2
- P(+) = 2/4 = 1/2

We will then calculate the likelihood of each word in the test sentence belonging to each class:

- $P(\text{“}great\text{”}|-) = \frac{1+1}{14+22}$ $\quad P(\text{“}great\text{”}|+) = \frac{2+1}{8+22}$
- $P(\text{“}plot\text{”}|-) = \frac{1+1}{14+22}$ $\quad P(\text{“}plot\text{”}|+) = \frac{1+1}{8+22}$
- $P(\text{“}boxing\text{”}|-) = \frac{1+1}{14+22}$ $\quad P(\text{“}boxing\text{”}|+) = \frac{0+1}{8+22}$
- $P(\text{“}scenes\text{”}|-) = \frac{2+1}{14+22}$ $\quad P(\text{“}scenes\text{”}|+) = \frac{1+1}{8+22}$

Note: we ignore words in the test sentence that don’t appear in the training data

Binary Naïve Bayes Theorem: Example

The probability of the test sentence belonging to a certain class can be calculated as:

- $P(sentence|-) = \frac{1}{2} * \frac{2*2*2*3}{36^4} = 7.1 * 10^{-6}$
- $P(sentence|+) = \frac{1}{2} * \frac{3*2*1*2}{30^4} = 7.4 * 10^{-6}$

Thus, the sentence would be classified as positive (which is correct).

Implementation

Implementation Process

Scrape Seeking Alpha for long/short idea articles

Chunk off a specific time frame of articles for the training set and a specific frame for the test set. In this analysis, a three year period between 08/2012 and 08/2015 was used for training and a two year period between 08/2015 and 08/2017 was used for testing

Sample a certain number of articles within each time frame. In this analysis, 500 articles were used in training and 200 were used in testing.

In the training set, calculate returns of the associated stocks for a specified holding period. Then assign a cutoff return that will be used to categorize the article as either “high” or ”low”

Optimize the holding period and cutoff point within the training set to achieve best predictive performance

Use these parameters in the model to predict the category an article will fall into in the test set. Then calculate actual realized returns and compare the results

Training Set: A Snapshot

```
In [47]: testbayes.data
```

Out[47]:

	article	author	datetime	headline	ticker	Words	Return	Target
73752	Biomarin (BMRN) is a company that specialize...	Chimera Research Group	2012-08-02 16:36:57	What's To Come At BioMarin	BMRN	{24-weeks, (, response, guidance, rare, phase,...	-0.00655996	low
73874	The recent drop in Weight Watchers' (NYSE: WTW...	David Trainer	2012-08-07 10:25:41	Risk Gets Lighter, Opportunity Gets Bigger For...	WTW	{contracts, (, And, some, call, WTW, comment, ...	0.0530486	high
73917	We published a full, comprehensive research re...	Saibus Research	2012-08-08 05:31:46	Teva: Undervalued Industry Leading Generic Dru...	TEVA	{view, comprehensive, buy, restored, directly,...	-0.00223602	low
73927	OfficeMax (NYSE: OMX) is starting to show sig...	Shane Blackmon	2012-08-08 10:11:36	The Stars Are Aligning For This Retailer	OMX	{initiating, implemented, view, rather, sold, ...	0.053211	high
73683	One of the earliest influences for our firm's ...	Saibus Research	2012-08-09 12:13:24	Berkshire Hathaway: 8.5% YTD Growth In Per Sha...	BRK.B	{view, buy, 1997-2010, contracts, directly, (,...	0.000118287	high
73509	Sirius XM (NASDAQ: SIRI) has just announced t...	Little Apple	2012-08-22 15:43:42	Sirius XM: Be Prepared For A Major Tree Shake	SIRI	{view, sold, buy, directly, (, unique, increas...	NaN	low
73653	A company with close to $400 billion of annual...	Yale Bock	2012-08-27 11:26:47	Steady Goes The Big Ship - An Update On The Pr...	BP	{facility, view, transform, Santiago, related,...	0.0253555	high

Test Set: A Snapshot

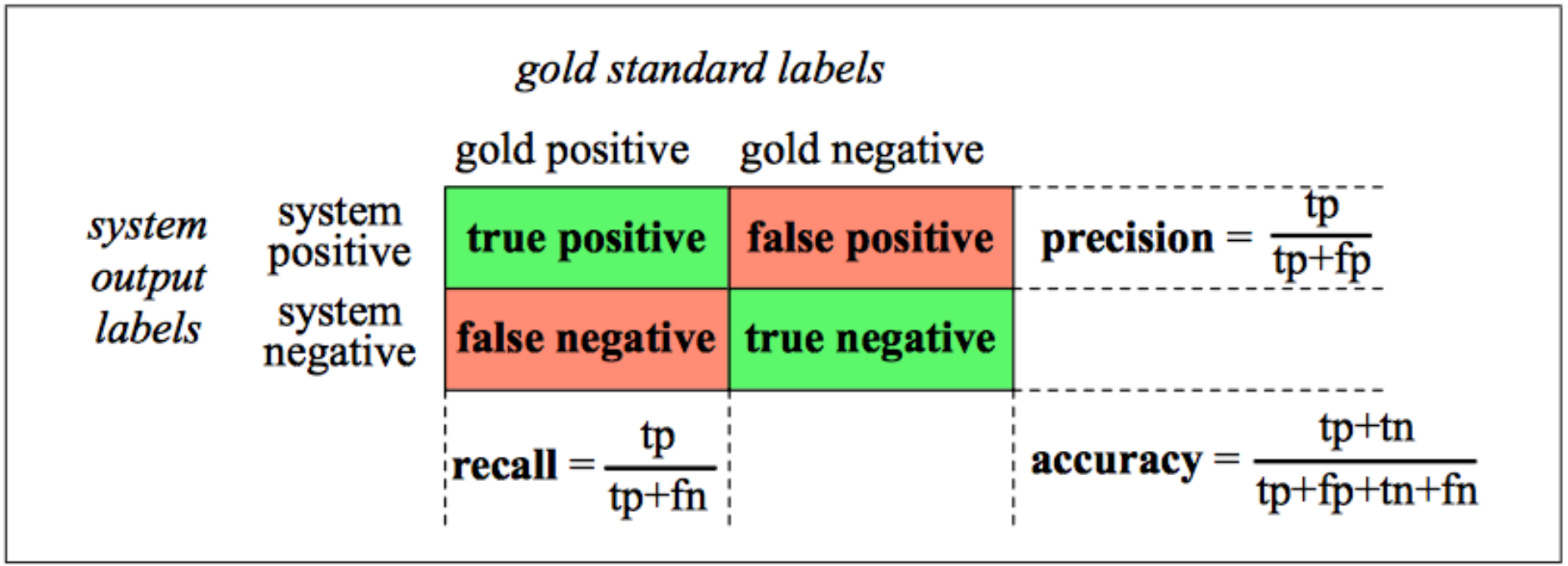
```
In [55]: testbayes.testset
```

Out[55]:

	article	author	datetime	headline	ticker	SentimentPrediction	SentimentPredictionScore	Return	ActualTarget
18231	Summary The so-called "four horsemen" of biote...	Bret Jensen	2015-08-04 11:12:49	Biogen: Ready For A Bounce?	BIIB	high	-3938.278158	-0.157485	low
18261	Summary CAB reported earnings a couple of week...	Josh Arnold	2015-08-04 15:22:56	Cabela's Continues To Struggle	CAB	high	-5129.627411	-0.0892499	low
18268	Summary Bloomberg reported that Nokia could se...	WestEnd511	2015-08-04 16:08:33	Nokia: The Rising Value Of A Map	NOK	low	-3299.903292	-0.14556	low
17687	Summary LGI Homes posted another strong quarte...	ONeil Trader	2015-08-11 05:49:31	LGI Homes: Sharp Rise After Strong Q2, Has Mor...	LGIH	low	-4267.814552	0.0663391	high
17754	Summary Acceptance of OncoSec's technology pla...	Stock Doctor	2015-08-11 15:07:26	OncoSec Medical: Upcoming Catalysts And Big Ph...	ONCS	high	-8059.059122	-0.0306306	low

Evaluation Metrics

Evaluation: Precision, Recall, F-Measure



$$F_1 - Measure: \frac{2*Precision*Recall}{Precision+Recall}$$

Training Set Sensitivity Analysis (100-Article Sample)

F_1-Measure	0%	3%	5%
10-Day	1.0	1.0	1.0
20-Day	1.0	1.0	1.0
60-Day	1.0	1.0	1.0

Accuracy	0%	3%	5%
10-Day	1.0	1.0	1.0
20-Day	1.0	1.0	1.0
60-Day	1.0	1.0	1.0

Precision	0%	3%	5%
10-Day	1.0	1.0	1.0
20-Day	1.0	1.0	1.0
60-Day	1.0	1.0	1.0

Recall	0%	3%	5%
10-Day	1.0	1.0	1.0
20-Day	1.0	1.0	1.0
60-Day	1.0	1.0	1.0

Test Set Performance (20-Day Holding Period, 0% Cutoff)

Precision	Recall	Accuracy	F1-Measure
0.531	0.645	0.570	0.583

Expected 20-Day Return from Long Positions: ***-0.310%***

Expected 20-Day Return from Short Positions: ***0.216%***

Next Steps

Filter out training and test sets by industry (a lot of money was lost in energy/commodities)

Filter out data sets by market cap (we don't want to invest in small-cap or OTC)

Calculate a "Trust-Value" score for reliable authors and incorporate it into the decision to invest

Test out the model across other sources, maybe hop on a Bloomberg and assess professional equity research

Develop a full backtest on Quantopian