Predicting Financial Market Direction Using Social Media Data

Benedict Augustine[1], Baodan Zhang[1], Dr. Yuri G Balasanov[2,*], Dr. Sema Barlas[3,*]

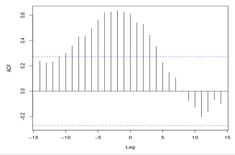
1.Candidate for MS in Analytics, Graham School, University of Chicago 2.Supervisor and Faculty of MSc in Analytics program 3.Director of MSc in Analytics program

Research Objective

- Examine whether relationship exists between Sentiment Indicators and Market Direction
- Validate whether these Sentiment Indicators are leading and adds dimension to current Technical Indicators
- Better predict market direction using this additional information

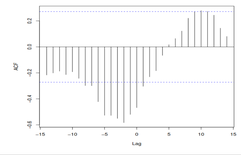
Interesting Findings

Positive Articles Ratio vs Russell Index



Positive sentiment affects the market much slower than negative sentiment. Figure above shows the ccf plot of latest positive articles ratio vs Russell and it shows a positive significant correlation. It also shows that the effect of positive ratio rises gradually. This could mean that Russell adjusts much slower to this positive ratio indicator and that this indicator has both leading and lagging effects.

Negative Articles Ratio vs Russell Index



The negative articles ratio on the other hand, as shown in the figure above, has, as expected, a negative significant correlation from lag -6 to 0. The ccf also shows that unlike positive ratio, the negative articles ratio effect rises sharply but dies out pretty quickly after day 0. This could mean that Russell adjusts quite quickly to negative sentiment and that the market price reflects the current sentiment.

Preprocessing

WORD	TAG	CHUNK	LEMMA	ROLE
Gilead	NNP	NP	gilead	–
Sciences	NNP	NP ^	sciences	–
–	:	–	–	–
a	DT	NP	a	–
Rare	JJ	NP ^	rare	–
Undervalued	NNP	NP ^	undervalued	–
Company	NNP	NP ^	company	–
In	IN	PP	in	–
A	DT	NP	a	–
Frothy	JJ	NP ^	frothy	–
Biotech	NNP	NP ^	biotech	–
Market	NNP	NP ^	market	–

Before Preprocessing

WORD	TAG	CHUNK	LEMMA	ROLE
Gilead	NNP	NP	gilead	SBJ
Sciences	NNPS	NP ^	sciences	SBJ
is	VBZ	–	be	–
a	DT	NP	a	OBJ
Rare	JJ	NP ^	rare	OBJ
Undervalued	NNP	NP ^	undervalued	OBJ
Company	NNP	NP ^	company	OBJ
In	IN	PP	in	–
A	DT	NP	a	–
Frothy	JJ	NP ^	frothy	–
Biotech	NNP	NP ^	biotech	–
Market	NNP	NP ^	market	–

After Preprocessing

Classification Example

Example: Complicated Scenario

WORD	TAG	CHUNK	ROLE	LEMMA
Exelon		NP	–	exelon
's		–	–	's
cost	NN	NP	SBJ	cost
is	VBZ	VP	–	be
n't	RB	VP ^	–	n't
reduced	VBN	VP ^	–	reduce
but	CC	–	–	but
Excelon	NNP	NP	SBJ	excelon
could	MD	VP	–	can
Be	VB	VP ^	–	be
Considered	VBN	VP ^	–	consider
For	IN	PP	–	for
Income	NN	NP	–	income
And	CC	NP ^	–	and
Long-Term	NNP	NP ^	–	long-term
Growth	NNP	NP ^	–	growth

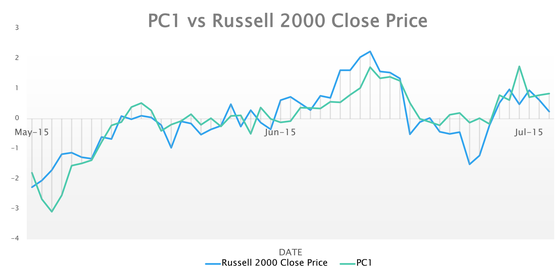
✕ Cost ⟹

✓ Cost is reduced ⟹

✕ Cost is not reduced ⟹

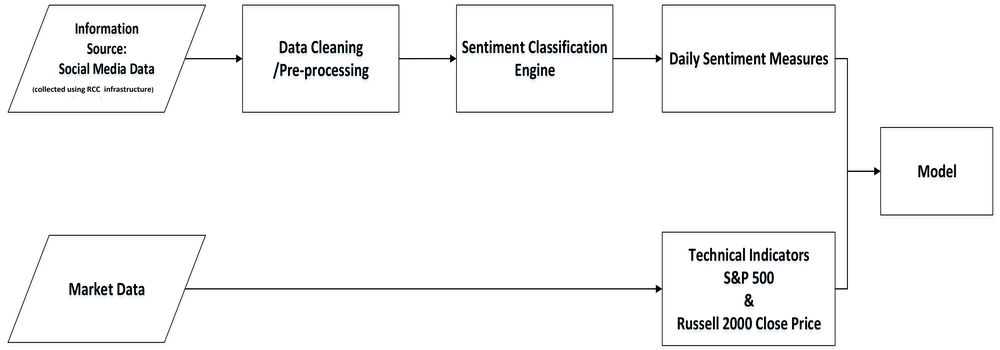
Research Result

Leading relationship exists between Sentiment and Russell 2000



Was able to find a strong, leading relationship between the Sentiment Indicators and Russell 2000 Index. This chart is a plot of first principal component of Sentiment Indicators vs Russell close. It is important to note that almost all principal components were significant, i.e. PCA could not drastically reduce the dimensionality of data.

Methodology



Sentiment Dictionary

- Standard publicly available dictionaries e.g Bill McDonald's , Harvard Inquirer etc.
- Machine Learning to build dictionary
- Manually further customize

Association rule discovery

Rules	Support	Lift
{AWESOME} => {1}	0.011289867	1.773717146

Rules	Support	Lift
{lamest} => {0}	0.000141123	2.292461986
{hate,lamest} => {0}	0.000141123	2.292461986

Findings

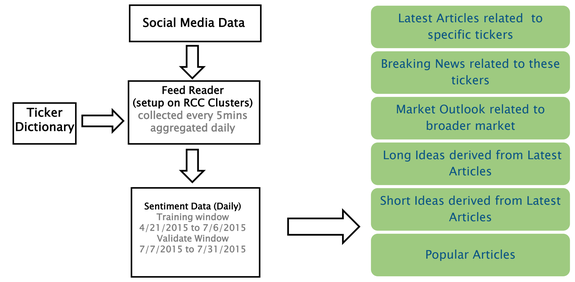
Relationship Hypothesis

- Sentiment Indicators did not exhibit strong (short term) relationship with S&P 500
- Strong, leading, short term relationship with the Russell 2000 Index

Predict Market Direction Hypothesis

- Able to predict the direction of Russell 2000 for next trading day with 80% accuracy; 30% better than just using Technical Indicators

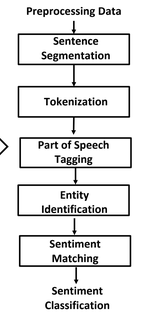
Information Source



Sentiment Classification Engine

- Different from Standard dictionary based classifiers
- Identifies Entity
- Matches Sentiment with the Entity
- Accuracy rate was 84% vs Standard Probabilistic approach was 51%

Sentiment Classification Engine



Conclusion & Further Improvements

- Additional data needed
- Weight the sentiment measures based on stock weights in the index.
- Adjust for Seasonality
- Remove lagging effect from leading information
- Use time warping techniques to give different weights to different periods
- Try additional machine learning techniques