Sentiment Extraction from Stock Message Boards

The Das and Chen Paper

Nicholas Waltner

University of Washington

Linguistics 575

Tuesday 6th May, 2014

Paper

MANAGEMENT SCIENCE
Vol. 53, No. 9, September 2007, pp. 1375–1388
ISSN 0025-1909 | EISSN 1526-5501 | 07 | 5309 | 1375

informs

DOI 10.1287/mnsc.1070.0704

Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web

Sanjiv R. Das
Department of Finance, Leavey School of Business, Santa Clara University,
Santa Clara, California 95053, srdas@scu.edu

Mike Y. Chen
Ludic Labs, San Mateo, California 94401, mike@ludic-lab.com

General Factoids

- Das is an ex-Wall Streeter and a finance Ph.D. from NYU. http://algo.scu.edu/ sanjivdas/
- Mike Chen is a computer science Ph.D. from the U of C, Berkeley.
- Approach this NLP task from a different perspective on NLP than other papers discussed in this course.
- Leverage Das's finance background to test a number of sentiment hypotheses using financial market data.

Task

- Focus on stock message boards for technology stocks, where there is a lot of chatter.
- Classify each message as either buy, hold or sell (+1,0,-1).
- Aggregative individual stock sentiment into a sentiment index on the Morgan Stanely High-Tech Stock Index (MSH).
- Using this index they then look for relationships in stock price levels and change in prices.
- Further look at the relationships between changes in sentiment, message agreement, message volumes, trading volumes and stock price volatilities.

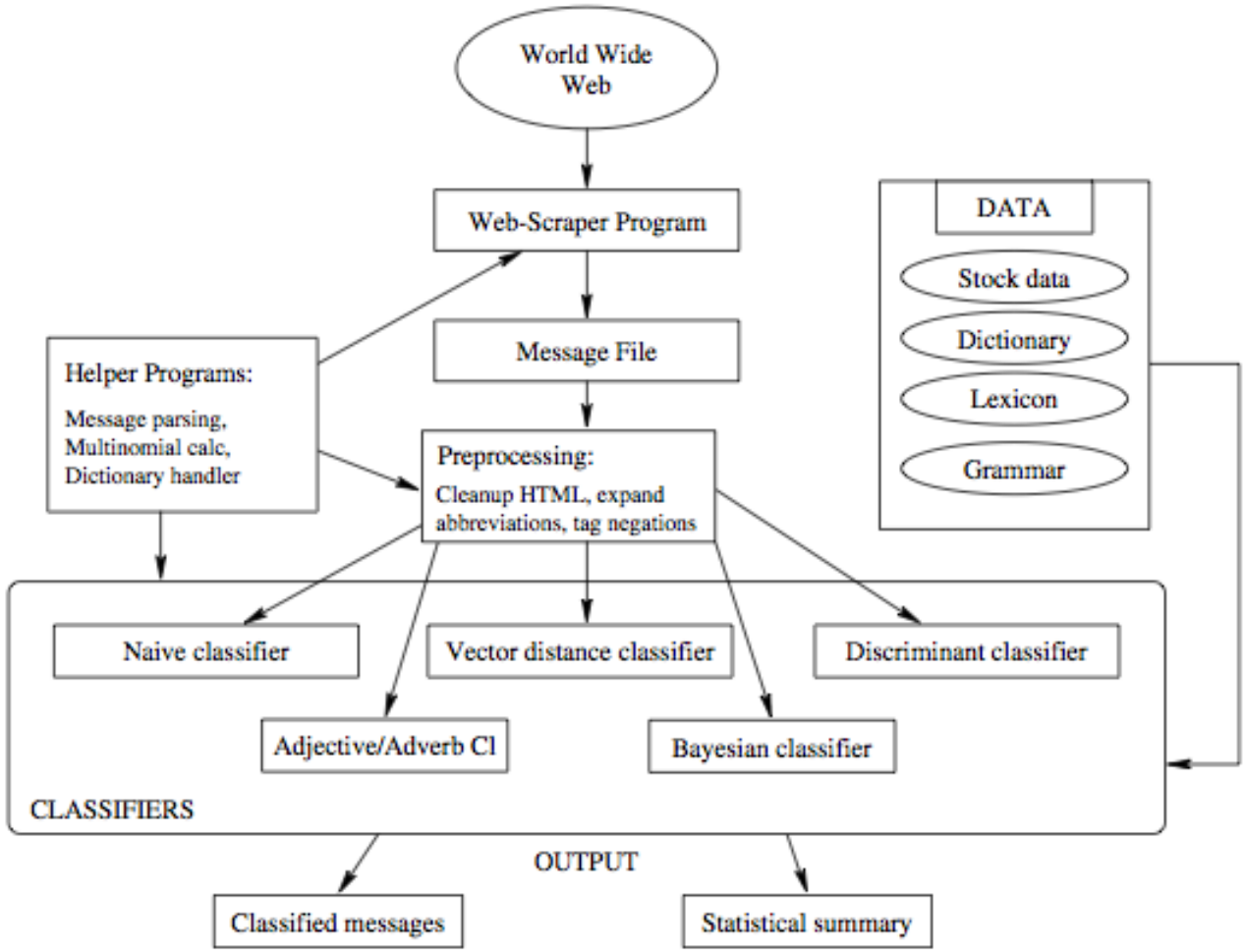
Data Sets

Das and Chen focused on stock market messages boards in a pre-Twitter era.

- Training: In-sample 374 messages.
- Test: Out-of-sample 913 message.
- Live Test: Out-of-sample 50,952 total messages.
- Choose smaller sizes to avoid over-fitting.
- Developed their own corpus using their own annotation arriving at a 72.46% agreement rate between their two annotators.

End-to-End Model

Figure 1 **Schematic of the Algorithms and System Design Used for Sentiment Extraction**



Pre-Processing

They employ three supplementary databases:

- They use CUVOLAD (Computer Usable Version of the Oxford Advanced Learner's Dictionary) to determined POS.
- Developed a lexicon of positive and negative words using discriminant analysis.
- Developed a grammar for the messages, but were not very clear about what they did.
- The used some pre-processing to deal with contractions and negation.

Classification

They employ five classifiers to extract sentiment:

- Naive Classifier: Counting of "buy" and "sell" words using GI and something else.
- Vector Distance Classifier: Simply a Vector Space Model to calculate cosine distances among the messages.
- Discriminant-Based Classifier: Use discriminant analysis, which is popular in the financial econometrics field, to determine which works are more meaningful.
- Adjective-Adverb Phrase Classifier. Score sentiment only on triplets containing an adjective or adverb with the two following words typically noun phrases.
- Basyesian Classifier. Provides simple probabilities of of being buy, hold or sell.

Voting

They then use a Voting Method between the five classifiers to determine polarity.

- Three of the methods must agree on message polarity to establish a simple majority.
- If not, they discard the message.
- Voting reduces the number of messages but increases accuracy.

Metrics

They use four metrics to evaluate their classification results.

- Chi-square test on confusion matrix.
- Ambiguity coefficient = 1-Accuracy. Human agreement was only 72.46%.
- False positive rates.
- Sentiment error. Compare the value of the aggregate sentiment given no classification error versus their classifier. (?).

Test Results

Table 1 **Tests of Various Algorithms**

Algorithm	Accuracy	False possitives	Sentiment error	χ^2	Number of messages
Panel A: In-sample					
NvWtd	92.2460	0.2674	0.5348	64.2581	374
vecDist	49.1979	12.2995	23.5294	6.1179	374
DiscWtd	45.7219	16.0428	35.0267	4.4195	374
AdjAdv	63.3690	7.7540	20.3209	17.4351	374
BayesCl	60.6952	8.2888	14.4385	13.4670	374
Vote	58.2888	7.7540	17.3797	11.1799	374
Vote-d	62.8743	8.6826	17.0659	14.7571	334
Rainbow	97.0430	0.8065	3.2258	75.2281	372
Panel B: Test sample (out-of-sample)					
NvWtd	25.7393	18.2913	6.4622	2.0010	913
vecDist	35.7065	25.8488	5.8050	1.4679	913
DiscWtd	39.1019	25.1917	4.1621	2.6509	913
AdjAdv	31.3253	29.7919	51.3691	0.5711	913
BayesCl	31.5444	19.8248	12.2673	2.0873	913
Vote	30.0110	17.8532	8.3242	2.1955	913
Vote-d	33.1242	20.4517	7.7792	2.3544	797
Rainbow	33.1140	33.0044	38.7061	0.5458	912
Panel C: Test sample (out-of-sample)					
NvWtd	27.6024	20.8000	9.5031	33.8485	50,952
vecDist	38.4224	25.7281	8.5728	103.1038	50,952
DiscWtd	40.6049	25.0530	5.8172	131.8502	50,952
AdjAdv	32.6366	30.2186	54.4434	35.2341	50,952
BayesCl	33.2254	19.9835	14.1525	128.9712	50,952
Vote	31.8614	18.0268	10.2665	136.8215	50,952
Vote-d	35.8050	20.8978	11.0348	138.4210	43,952
Rainbow	35.2335	33.0893	40.0484	24.3460	49,575

Improvements

They use two methods to improve on their initial results:

- Increase the size of the training set without overfitting.
- Screen messages for ambiguity before classifying.
- Use Harvard's GI to build an *optimism score*.

The scores sync with the categories. They then use standard deviation ranges to filter out messages.

Message type	Optimism score	
	Mean	Std. dev.
Buy	0.032	0.075
Hold	0.026	0.069
Sell	0.016	0.071

Improved Sentiment Results

Table 3 **Classification as a Function of Ambiguity: Training Set Size of 913**

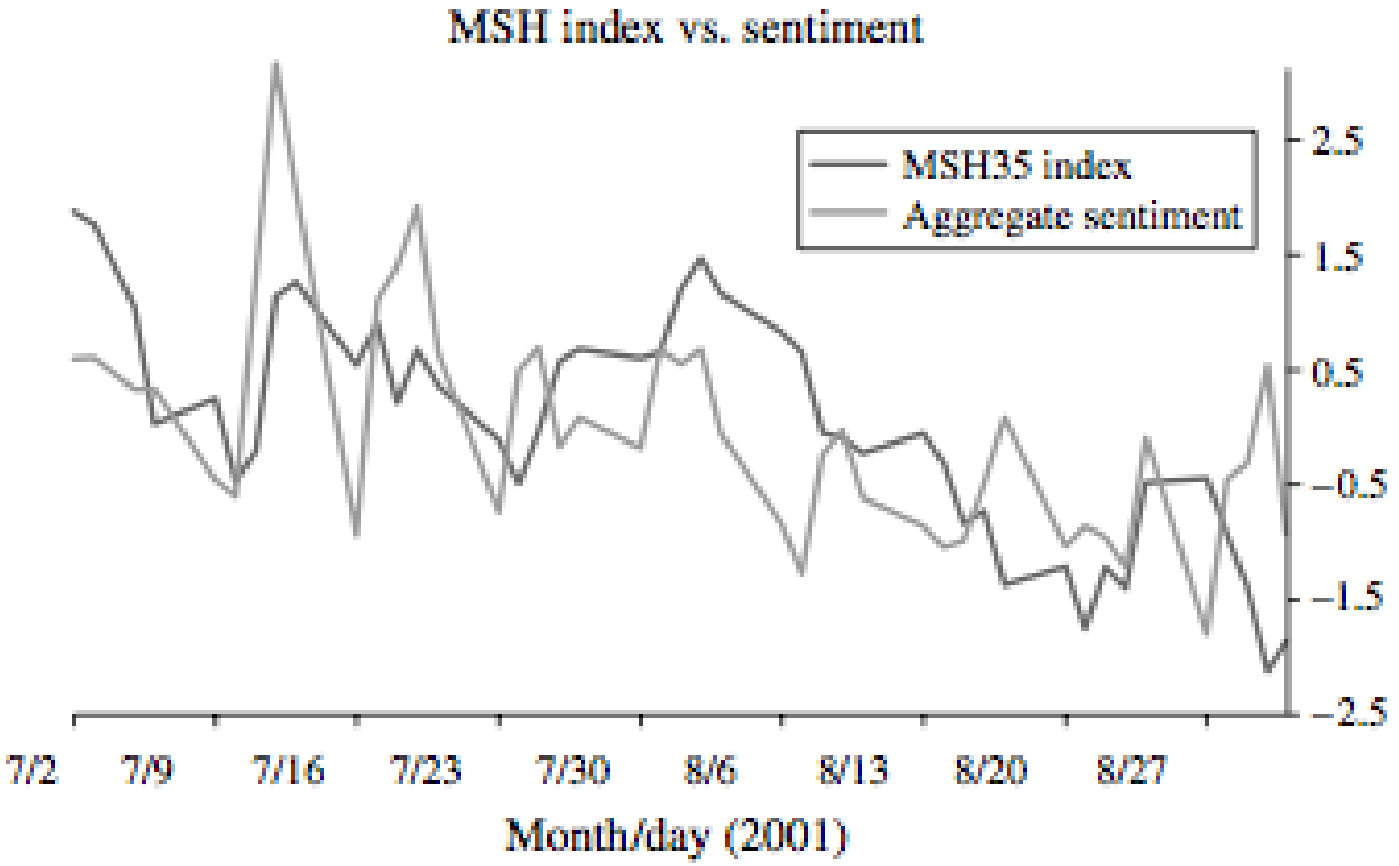
Algorithm	Accuracy	False positives	Sentiment error	χ^2	Number of messages
Panel A: High ambiguity					
NvWtd	45.5016	34.2224	1.8843	16.7918	7,536
vecDist	61.0934	14.2118	10.2309	236.2058	7,536
DiscWtd	54.8832	13.1900	10.0318	204.3926	7,536
AdjAdv	64.1056	33.5987	41.2818	74.0030	7,536
BayesCl	57.4443	12.3275	7.8025	238.8515	7,536
Vote3	53.3838	10.1778	14.3577	242.0414	7,536
Vote3-d	63.2705	12.1534	12.4703	227.2089	6,311
Rainbow	64.8818	32.6479	13.0191	86.8046	7,489
Panel B: Medium ambiguity					
NvWtd	46.9638	30.7494	1.0982	6.2881	1,548
vecDist	64.5995	8.6563	8.6563	69.9800	1,548
DiscWtd	58.1395	8.5917	9.4961	58.2234	1,548
AdjAdv	65.7623	28.4884	41.5375	23.2180	1,548
BayesCl	61.4341	7.7519	8.0749	67.8975	1,548
Vote3	58.3979	6.3307	13.5659	68.8180	1,548
Vote3-d	66.7671	7.3851	11.6805	65.8436	1,327
Rainbow	65.4816	28.9593	10.7304	27.4832	1,547
Panel C: Low ambiguity					
NvWtd	46.5517	25.5172	9.3103	1.9822	290
vecDist	66.8966	3.7931	7.9310	16.9034	290
DiscWtd	57.2414	5.5172	8.2759	11.7723	290
AdjAdv	61.3793	24.1379	40.0000	4.7444	290
BayesCl	64.4828	4.4828	8.2759	15.2331	290
Vote3	63.4483	4.1379	12.4138	15.3550	290
Vote3-d	66.7939	4.5802	11.0687	14.5289	262
Rainbow	67.5862	18.2759	12.0690	9.0446	290

Test Dataset

- Scraped the messages board for the 24 stocks in MSH from July to August 2001.
- Total sample of 145,110 messages.
- Collected until 4PM New York time each for each trading and ignored weekends.
- Individual sentiment indices were incremented by $+1$ for each buy message and by -1 for each sell.
- The data was aggregated on an equally weighted basis to form a MSH sentiment index.

Normalized Indices

Figure 2 **Normalized MSH Index and Aggregate Sentiment, Daily, July–August 2001**



Further Metrics

Four other metrics were constructed for further analysis:

- Index normalization: MSH and the aggregate sentiment index were statistically normalized (subtract mean and divided by standard deviation) to provide unify the scale across individual stocks.
- Disagreement: Tracked this metric over time.

$$\text{DISAG} = \left|1 - \left|\frac{B-S}{B+S}\right|\right|,$$

- Volatility: Defined it as the difference between high and low stock price divided by the average of the open and closing prices.
- Volume: Trading volume in the number of shares per day (should be dollar value instead).

Index Level Results

Ran four regression tests with significant results on level with weak ones on changes.

Table 4 **Regressions Relating the Stock Index (MSH) to the Sentiment Index**

Dependent variable	Intercept	MSH_t	$SENTY_t$	R^2
Regressions in levels				
MSH_{t+1}	−0.081	0.793***	0.154*	
t-stat	−1.12	9.28	1.86	0.77
$SENTY_{t+1}$	−0.028	0.100	0.451***	
t-stat	−0.21	0.62	2.90	0.25
Dependent variable	**Intercept**	**ΔMSH_t**	**$\Delta SENTY_t$**	**R^2**
Regressions in changes				
ΔMSH_{t+1}	−0.094	−0.082	0.138*	
t-stat	−1.19	−0.51	1.71	0.07
$\Delta SENTY_{t+1}$	−0.083	−0.485	−0.117	
t-stat	−0.53	−1.51	−0.72	0.08

Stock Level Results

Further their analysis to the 24 individual stocks:

- Although their is positive skew between return and sentiment with significant t-statistics for the SENTY and CH_SENTY variables at 2.08 and 1.66, the models are not statistically significant.
- The r-squarers are 0.0041 and 0.0027, respectively.
- Conclusion: There is likely simply too much noise in the daily sentiment of stocks and their movements.

Further Metric Results

They did, however, find strong correlations between sentiment, disagreement, volumes and volatility:

- Sentiment is inversely related to disagreement, i.e. when disagreement increases, sentiment drops.
- Sentiment is correlated to high message posting levels.
- Message volume and trading volumes are correlated.
- Trading volume and volatility are strongly related.

Table 5 **Relationship of Changes in Volatility and Stock Levels to Changes in Sentiment, Disagreement, Message Volume, and Trading Volume**

Independent variables	ΔVOLY	ΔSTK
Intercept	−0.000	−0.056***
	−0.01	−3.29
ΔSENTY	−0.106	0.059*
	−1.50	1.83
ΔDISAG	0.008	0.001
	0.14	0.03
ΔMSGVOL	0.197***	−0.080***
	3.34	−2.99
ΔTRVOL	0.447***	0.000
	11.84	0.01
R^2	0.20	0.02

Author Conclusions

The authors conclude fivefold:

- Limited understanding of the microstructure of technology stocks.
- Their work can be used to understand the mechanics of herding.
- Their work can be used to monitor market activity.
- Their work can be used by firms to monitor message boards for investor reaction to management actions.
- Sentiment may be applied to test theories in the field of behavioral finance.

Critique

- This is one of many papers written by finance professors dabbling in comp-ling. As such, the comp-ling side of their is generally, and understandably, not state-of-the-art within the comp-ling field.
- Is market microstructure misunderstood, or is the buy, hold, sell paradigm too blunt of a sentiment measurement tool?
- Much has evolved, in comp-ling and, the web and the markets, since they collected their data in 2001.
- Our work shows that sentiment evolves over longer periods than one day.
- Clearly, there is a behavioral effect (Table 5), but volatility seems easier to predict than price movements, which also syncs with our work.

Similar Finance Papers

- Antweiler and Frank. 2002 *Internet stock message boards and stock returns*.
- Antweiler and Frank. 2004 *Is all that talk just noise? The information content of Internet stock message boards*.
- Antweiler and Frank. 2005 *The market impact of news stories*.
- Choi, Laibson and Merick. 2002 *Does the Internet increase trading? Evidence from investor behavior*.
- Boudoukh, Feldman, Kogan and Richardson. 2013 *Which News Moves Stock Prices? A Textual Analysis*.
- Tetlock. 2005 *Giving content to investor sentiment: The role of the media in the stock market*.
- Wysocki. 1998 *Cheap talk on the web: The determinants of postings on stock message boards*.

Future Work

- The work of finance professors on stock market texts has provided a number of insights into investor behavior.
- However, much of their work was done during the infancy of the comp-ling field and exclusively with *shallow techniques* and without more advanced machine learning approaches.
- Further, very little attention has been paid to the role of *emotion* in the financial markets, i.e. buying and selling shares of Apple is just or more emotionally charged as buying a toothbrush or a new car.
- A revaluation of Tetlock's work (15 years of analysis of the Wall Street Journal's *Abreast of the Market* column (1984-95)) with more fine grained sentiment tools developed at Madison Park, may provide deeper insights into the behavioral biased exhibited by investors.