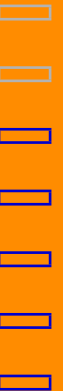


Yahoo for Amazon: Language Algorithms for Extracting Market Sentiment from Stock Message Boards

Sanjiv Das & Mike Chen

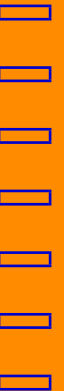
UC Berkeley & Santa Clara University

and other work with Asis Martinez-Jerez, Danny Tom, Peter Tufano & Jason Waddle



Outline

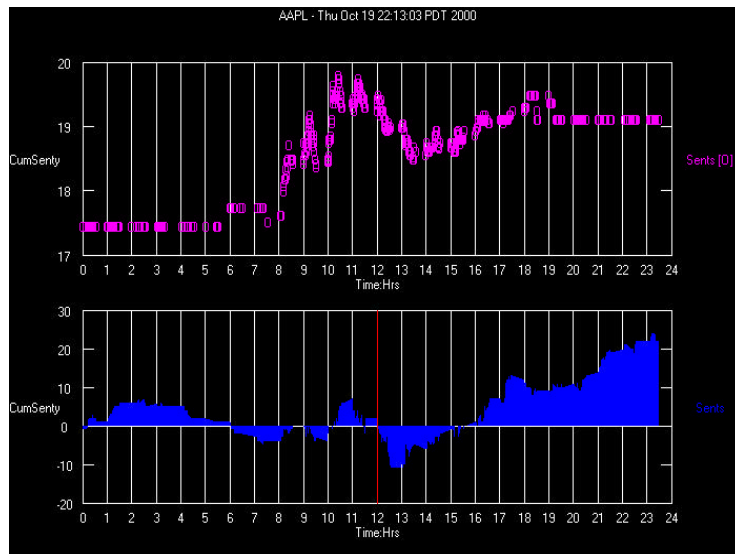
- Motivation
- Statistical Language Processing
- Classification Models
- Results from Amazon's message board.
- Results from a micro analysis of Apple Computer
- Preliminary results from a more extensive dataset.



Stock Market Sentiment Index

[AAPL](#) [AMZN](#) [CDNW](#) [DELL](#) [EBAY](#) [ITWO](#)

Ticker:

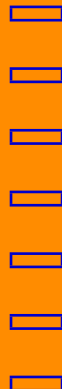


[Easy Plot](#)

by [Sanjiv Das](#) and [Mike Chen](#)

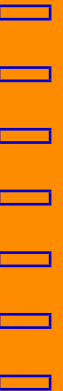


3/65



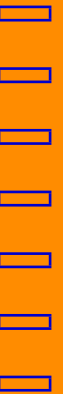
Motivation

- Understand “e-information”.
- The role of small investors in the information generating process.
- Understand the impact of (a) Quicker transmission of information, (b) Broader dissemination of news, and (c) New(s) channels.
- Market efficiency (noise vs signals).
- e-info and volatility.
- Exploit the biggest dynamic data set in the world.
- Process emotive/behavioral content in financial markets.





“Language is itself the collective art of expression, a summary of thousands upon thousands of individual intuitions. The individual gets lost in the collective creation, but his personal expression has left some trace in a certain give and flexibility that are inherent in all collective works of the human spirit” – Edward Sapir





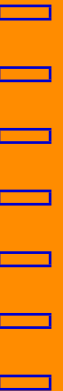
Is there a role for small investors?

- 15% of NASDAQ volume comes from day-trading by small investors.
- In corporate 401(k) plans, web-based trading has had huge impact: trading frequency doubles, portfolio turnover up 50%.
- The cost of trading is falling rapidly.
- Information acquisition is even cheaper.
- E.g: Amazon: 70K messages end 1998, 250,000 end 1999.
- There are 8,000 message boards.
- 5-10 major message board providers [Yahoo, Motley Fool, Silicon Investor, Raging Bull].
- Active boards: message posting 3-5m:24/7 [every 3-5 minutes].



Complexities of Message Flow

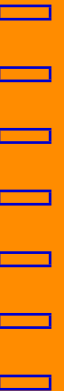
- Fundamental(s) information (publication of quantitative analysis).
- Market sentiment (opinions and “inside” information).
- Manipulative behavior (Emulex declines 62% on false posting) .
- Over and under reactions .
- Sources: investors, corporates, and regulators (physical scrutiny).
- Unusually heterogenous information sources.



Emulex

"Since the events of last Friday [August 25, 2000], we have noted increasingly widespread concern over the vulnerability of the financial markets due to fraudulent acts of this nature. Clearly, the huge impact in our stock that resulted from this incident has demonstrated the high regard and trust that the public has for financial news services. We believe that the public's trust is well founded and the stability of financial markets worldwide remains critically dependent on the continued role in disseminating accurate information. While there will always be challenges in completely safeguarding the integrity of electronic information, in the aftermath of this incident, we have been gratified by the promise of increased vigilance and scrutiny by financial news services worldwide.

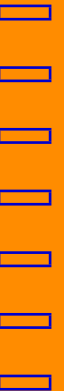
– Paul Folino, President and CEO, Emulex Corp, after the SEC apprehended the message poster, who claimed that Folino had resigned and earnings were down.



The Impact

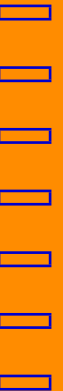
Message board activity impacts:

- Trading behavior (individual and herd)
- Information flow
- Regulation
- Market volatility
- Institutional design
- Market efficiency



Statistical Language Processing

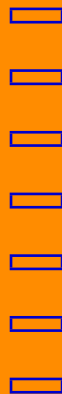
- Classification problems (what search engines do in their indexing algorithms)[IBM's "Clever" project].
- Parts of speech parsing (more complex)[Mainstream natural language work].
- Sentiment parsing (for emotive content) is more complicated ["Rage factors"].
- Other settings: political opinion polls, consumer research, medical studies, economic news classification, editorial analysis.
- Language dependence vs independence.



Board Volumes - Motley Fool

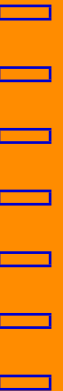


Ticker Symbol	Last Message Date	Number of Messages
DELL	May 9 1999	33382
IOM	May 9 1999	21142
AMZN	May 9 1999	19742
TDFX	May 9 1999	17200
AOL	May 9 1999	11540
CMGI	May 9 1999	10878
ATHM	May 9 1999	9823
MSFT	May 9 1999	9412
THW	May 9 1999	7428
CKO	May 9 1999	6760
BRK.A	May 9 1999	6354
INTC	May 9 1999	6111
CPQ	May 9 1999	5353
YHOO	May 8 1999	5127
CSCO	May 9 1999	5079



Other Work

- Wysocki 1998: Looks at message volume.
 - Variation in message posting volume is a function of firm characteristics.
 - Changes in daily posting volume are related to news and earnings announcements, in a statistically significant way.
 - Posting volume is highest for firms with extreme past returns, low book to market values, high price-earnings ratios, high analyst following and low institutional holdings.
- Bagnoli, Beneish and Watts (1999) examined the predictive validity of whisper forecasts, and found them to be superior to those of First Call analysts. A trading strategy based on whispers yielded better than risk-adjusted profits.
- Visit www.whispernumber.com.









**THE WALL updated for you every nanosecond,
more or less STREET JOURNAL.**

Home | Sign In | About | Press Room | Contact Us | FAQ
Symbol lookup

Detailed Whisper Blast - Enter Symbol: go

WHISPER NUMBERS >> TOP 5

Investor Expectations for Corporate Earnings

Click on Company Name or Symbol for detailed information.

Whisper Numbers
Investor Expectations for Corporate Earnings

IPO Whisperm
Investor Sentiment and Pricing Expectations for IPOs

The G-Report
Whispers for Government, Goods, Global Economics & Greenspan Reports

Calendar
Upcoming Dates for Earnings, IPOs, and G-Reports

Board Watch
Volume, Growth and Decline on the Message Boards for Individual Stocks

Recommendations
The Buy, Sell and Hold Recommendations of Self-Directed Investors

Whisper News
WhisperNumbers in the News



Company	Symbol	Earnings Date	Consensus	Whisper Number
ADVANCED MICRO DEVICES INC	AMD	10/11	0.61	\$ 0.67
INTEL CORP.	INTC	10/17	0.38	\$ 0.40
YAHOO! INC.	YHOO	10/10	0.12	\$ 0.13
OMEGA CORP.	OM	10/12	0.06	\$ 0.10
TEXAS INSTRUMENTS, INC.	TXN	10/18	0.33	\$ 0.35

[more whisper numbers...](#)

IPO WHISPER NUMBERS >> TOP 5

Company	Symbol	Date	Price	IPO Whisper
EXHAUST TECHNOLOGIES INC	XTEC	10/10-10/13	N/A	\$ 41.67
URBAN COOL NETWORK INC	UBN	10/23-10/29	\$ 9.00	\$ 3.97
LA FAYETTE COMMUNITY BANCORP	LCB	10/06-10/09	\$ 10.00	\$ 47.75
KPMG CONSULTING INC	KGIN	10/16-10/20	\$ 7.75	\$ 50.00
ADVANCED SWITCHING COMMUNICATIONS INC	ASCX	10/04-10/06	\$ 14.00	\$ 72.50

[more IPO whisperm...](#)

MESSAGE BOARD MOVERS >> TOP 5

Rank	Company	Symbol	This Week	Last Week	% Growth
1	AMERICAN MANAGEMENT SYSTEMS, INC.	AMSY	15550	4104	278.90%
2	ARTEST CORP	ARTE	7462	2030	267.59%
3	ADVOCAT INC.	AVC	10053	3924	156.19%
4	DSL NET INC	DSLN	29252	12537	133.33%
5	CROSSROADS SYSTEMS INC	CRDS	28009	12381	126.23%

[more IPO whisperm...](#)

WHISPERNUMBER Home | Sign In | About | Press Room | Disclaimer | Contact Us

Copyright © 2000 Patent Pending, Internet Financial Network, Inc. All Rights Reserved. WhisperNumber.com, Whisper Blast and G-Report are trademarks of Internet Financial Network, Inc.



**THE WALL updated for you every nanosecond,
more or less STREET JOURNAL.**

Sign up & start whispering today!
Click here to register

Infogate
Find out which whisperm were on the money

Psst: What's with whisperm numbers?
CBS Market Watch

WHISPER NEWS
Family Dollar Misses Whisperm by a Nickel


Investors Expect Research in Motion's Losses to Widen, Results to Miss the "RIMM"

More Whisperm News

John Scherr, co-founder of WhisperNumber.com discusses the value of whisperm numbers.
CN24 Financial Network [click here to listen!](#)



14/65



THE WALL updated faster than
you can read this **STREET JOURNAL**.

Home | Sign In | About | Press Room | Contact Us | FAQ |

Detailed Whisper Blast - Enter Symbol : [Symbol lookup](#)

INTEL CORP. — INTC

Username: Password: [Need to register?](#) [Forgot your password?](#)

AUDIO Available (CLICK HERE TO LISTEN)

Whisper Numbers
Investor Expectations for Corporate Earnings

IPO Whispers
Investor Sentiment and Pricing Expectations for IPOs


The G-Report
Whispers for Government, Goods, Global Economics & Greenspan Reports


Calendar
Upcoming Dates for Earnings, IPOs, and G-Reports

Board Watch
Volume, Growth and Decline on the Message Boards for Individual Stocks

Recommendations
The Buy, Sell and Hold Recommendations of Self-Directed Investors

Whisper News
WhisperNumbers in the News

 Find out which whispers were on the money



WHISPERNUMBER Home | Sign In | About | Press Room | Disclaimer | Contact Us

Copyright © 2000 Patent Pending. Internet Financial Network, Inc. All Rights Reserved. WhisperNumber.com, Whisper Blast and G-Report are trademarks of Internet Financial Network, Inc.

WHISPER:

QTR ENDING	EARNINGS DATE	CONSENSUS NUMBER	ACTUAL NUMBER	WHISPER NUMBER	TOTAL WHISPERS	24hr MOVEMENT	NEXT Q EARNINGS DATE	NEXT Q CONSENSUS
9/2000	10/17/2000	0.38	N/A	0.40	42	N/A	N/A	0.43

ENTER YOUR WHISPER :

COUNTDOWN TO EARNINGS DATE... 16 days remaining

[VIEW DETAILED WHISPER BLAST ON THIS STOCK](#)

SEARCH DATABASE: Symbol: [Symbol lookup](#)

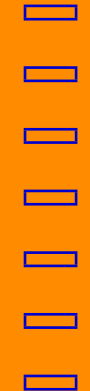
SEARCH BY INDUSTRY SECTOR:

DELAYED QUOTES:


Last: 41.56	Change: -2.88	Open: 43.81	High: 43.88	Low: 41.38	Volume: 72985600
	Percent Change: -6.92%	Yield: N/A	P/E Ratio: N/A	52 Week Range: N/A	


RECENT NEWS:

- 09/28/00 - 05:01 AM [Investors Expect Research in Motion's Losses to Widen, Results to Miss the "RIMM"](#)
- 09/22/00 - 04:18 PM [Markets Storm Back From Early Losses](#)
- 09/22/00 - 10:34 AM [Intel Revenue Warning KO's Markets](#)
- 09/21/00 - 05:20 PM [Weak European Demand Hits Intel's Top Line Growth](#)
- 09/20/00 - 04:44 PM [Dow Jones Takes Investors on a Roller Coaster Ride](#)
- 09/18/00 - 05:21 PM [Profitability Questions Take a Toll on the Dow and Nasdaq](#)
- 09/13/00 - 04:46 PM [Techs Take a Toll on Dow Jones](#)
- 09/11/00 - 06:02 PM [Techs Tumble on Earnings Jitters](#)
- 09/08/00 - 05:10 PM [Tech Meltdown Drags Nasdaq into Triple Digit Losses](#)
- 09/06/00 - 04:37 PM [Nasdaq Can't Shake That Sinking Feeling](#)









THE WALL why read about it tomorrow
if it's happening now **STREET JOURNAL.**

Home | Sign In | About | Press Room | Contact Us | FAQ
Symbol lookup

Detailed Whisper Blast - Enter Symbol:

Username:

Password:

[Need to register?](#)

[Forgot your password?](#)

Whisper Numbers
Investor Expectations for Corporate Earnings

IPO Whispers
Investor Sentiment and Pricing Expectations for IPOs


The G-Report
Whispers for Government, Goods, Global Economics & Greenspan Reports

Calendar
Upcoming Dates for Earnings, IPOs, and G-Reports

Board Watch
Volume, Growth and Decline on the Message Boards for Individual Stocks

Recommendations
The Buy, Sell and Hold Recommendations of Self-Directed Investors

Whisper News
WhisperNumbers in the News



Find out which whispers were on the money

MONITOR THE MESSAGE BOARD SENTIMENT INDICATORS

Individual investors' sentiments undoubtedly continue to exert a strong influence on stock price movements. If you noticed the surge in volume on CMGI's Yahoo! message board back in September 1998, you may have picked up on the growth of one of today's hottest stocks - and you could have jumped in to an early, lucrative investment. Of the 11 million stock-related message boards on the Internet, find out who the next big mover will be, where all of the message board volume is, and which stock will become the next big board winner (or loser!). Monitor it here!

Weekly Top 25 Stock Board Movers
25 Most Active Stocks (by % increase in growth)

Weekly Top 25 Stock Board Losers
25 Least Active Stocks (by lowest % increase in growth)

Weekly Top 25 Stocks by Message Volume
25 Highest Message Board Volume Stocks

SEARCH DATABASE:

Symbol: [Symbol lookup](#)

SEARCH BY INDUSTRY SECTOR:

Select Sector

THIS WEEK'S MESSAGE BOARD MOVERS >> TOP 5

RANKING	COMPANY NAME	SYMBOL	THIS WEEK'S VOLUME	LAST WEEK'S VOLUME	% GROWTH
1	AMERICAN MANAGEMENT SYSTEMS, INC.	AMSY	15550	4104	278.90%
2	ARTEST CORP	ARTE	7462	2030	267.59%
3	ADVOCAT INC.	AVC	10053	3924	156.19%
4	DSL NET INC	DSLN	29252	12537	133.33%
5	CROSSROADS SYSTEMS INC	CRDS	28009	12381	126.23%

Click on Company Name or Symbol for detailed information.

THIS WEEK'S MESSAGE BOARD LOSERS >> TOP 5

RANKING	COMPANY NAME	SYMBOL	THIS WEEK'S VOLUME	LAST WEEK'S VOLUME	% GROWTH
1	TOTALACCESS.COM INC	TXCS.OB	31807	31805	.01%
2	REXALL SUNDOWN INC.	RXSD	21018	21016	.01%
3	SHAMAN PHARMACEUTICALS, INC.	SHMN	20127	20124	.01%
4	INNOVO GROUP, INC.	INNO	17304	17302	.01%
5	ANDOVER NET INC	ANDN	16793	16792	.01%

Click on Company Name or Symbol for detailed information.

THIS WEEK'S MESSAGE BOARD HIGH VOLUME >> TOP 5

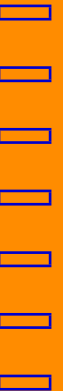
RANKING	COMPANY NAME	SYMBOL	TOTAL VOLUME	LAST WEEK'S VOLUME	% GROWTH
1	AMERICA ONLINE, INC.	AOL	1587581	1583786	.24%
2	DELL COMPUTER CORP.	DELL	1303387	1298912	.34%
3	THE CIRCLE GROUP INTERNET INC	THE	1078168	1075796	.22%



Whisper Numbers

- A “whisper number” is the investor’s expectations for earnings. It is also the investor’s expectations for other financial topics such as an IPO and economic indicators. It continues to gain credibility for its power and accuracy on Wall Street and in the financial community. It continues having a major impact on the direction of a stock’s price after actual earnings are announced as more and more investors place value on whether or not a company met or missed the whisper number.
- “On average our statistics are showing that a stock will decrease in price 74

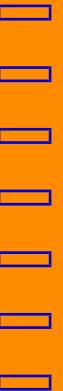
from www.whispernumber.com.





Methodology

- Focus on a few active message boards. e.g. AMZN - interesting history, active debate, steady posting volume.
- Data gathered using a “web-scraper” to crawl the message board. Can download thousands of messages in minutes.
- Parser to establish standard data format.
 - (1) source portal (e.g. Yahoo)
 - (2) stock ticker (e.g. AMZN)
 - (3) message Number (sequence on message board)
 - (4) date
 - (5) message title
 - (6) message body
- 4 months of AMZN (Jan-Apr, 2000), about 24-25K messages.
- Small training set
- Small Testing set (multiple sets).

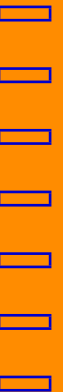




Message Categories

- **3: Buy:** Rating level 3 was assigned if the message indicated positive (i.e buy) sentiment.
- **1: Sell:** Level 1 was assigned if the message indicated negative sentiment.
- **0: Neutral/Spam:** Level 0 applies when the message is neither buy nor sell, and may be simply a neutral message or a nonsense message, i.e. spam.

These classification tags [0,1,3] have no numerical significance, only chosen so as to correspond to some technical features of the algorithm.



Message Example

Yahoo

AMZN

Yer_the_man

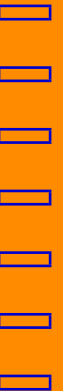
195007

12/19/99 3:30 pm

Is it famous on infamous?

A commodity dumped below cost without profit, I agree. Bezos had a chance to make a profit without sales tax and couldn't do it.

The future looks grim here.<p>



Message Example

Yahoo

AMZN

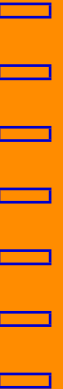
auvernia

195006

12/19/99 3:26 pm

The fact is.....

The value of the company increases because the leader (Bezos) is identified as a commodity with a vision for what the future may hold. He will now be a public figure until the day he dies. That is value.<p>



Message Ambiguity

Yahoo

AMZN

djsbsevenone

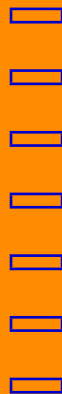
195016

12/19/99 4:01 pm

You're missing this Sonny, the same way the cynics pronounced that "Gone with the Wind" would be a total bust.<p>

This message is harder to classify and may be rated either a 1 (sell) or a 0 (non-informative).

- Human inconsistency $\sim 28\%$.
- Asymptotic ambiguity.





Support Vector Machines (SVMs)

- See Vapnik (1995), Vapnik and Chervonenkis (1964), Chen, Das, Tom and Waddle (1999), Smola and Scholkopf (1998).
- Consider a training data set given by the binary relation

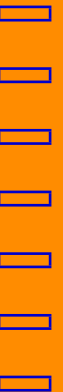
$$\{(x_1, y_1), \dots, (x_n, y_n)\} \subset X \times \mathcal{R}.$$

The set $X \in \mathcal{R}^d$ is the input space and set $Y = \{y_1, \dots, y_n\}$ is a subset of the feature space or categories. We define a function

$$f : x \rightarrow y$$

with the idea that all elements must be mapped from set x into set Y with no more than an ϵ -deviation. A simple linear example of such a model would be

$$f(x_i) = \langle w, x_i \rangle + b, \quad w \in \mathcal{X}, b \in \mathcal{R}.$$



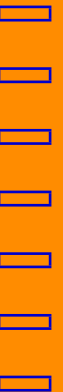
SVMs - Objective Function

The idea in SVM regression is to find the *flattest* w that results in the mapping from $x \rightarrow y$. Thus, we minimize the Euclidean norm of w . We also want to ensure that $|y_i - f(x_i)| \leq \epsilon, \forall i$. The objective function becomes

$$\min \frac{1}{2} \|w\|^2 \quad (1)$$

$$st : y_i - \langle w, x_i \rangle - b \leq \epsilon \quad (2)$$

$$-y_i + \langle w, x_i \rangle + b \leq \epsilon \quad (3)$$





SVM Feasibility

This is a (possibly infeasible) convex optimization problem. Feasibility is obtainable by introducing the slack variables (ξ, ξ^*) . We choose a constant C that scales the degree of infeasibility. The model is then modified to read as

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi + \xi^*) \quad (4)$$

$$st : y_{i-} \langle w, x_i \rangle - b \leq \epsilon + \xi \quad (5)$$

$$-y_{i+} \langle w, x_i \rangle + b \leq \epsilon + \xi^* \quad (6)$$

$$\xi, \xi^* \geq 0. \quad (7)$$

As C increases, the model increases in sensitivity to infeasibility.





SVM Modifications

1. We can tune the objective function by introducing cost functions $c(\cdot), c^*(\cdot)$. Then, the objective function becomes

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n [c(\xi) + c^*(\xi^*)]$$

2. We replace the function $[f(x) - y]$ with a “kernel” $K(x, y)$ introducing nonlinearity into the problem. The choice of the kernel is a matter of judgment based on the nature of the application being examined.



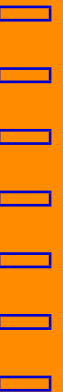


SVM Packages

1. SVM Light: The package in some detail is available from the developers at the following URL: <http://www-ai.informatik.uni-dortmund.de>. SVMlight is an implementation of Vapnik's Support Vector Machine for the problem of pattern recognition. The algorithm has scalable memory requirements and can handle problems with many thousands of support vectors efficiently. The algorithm proceeds by solving a sequence of optimization problems lower-bounding the solution using a form of local search. Based on work by Joachims (1999).
2. University of London SVM: We used the Royal Holloway Department of Computer Science algorithm. The reference URL for this model is <http://svm.dcs.rhnc.ac.uk/>. This SVM allows many different estimation kernels. e.g. The Radial Basis function kernel minimizes the distance between inputs (x) and targets (y) based on

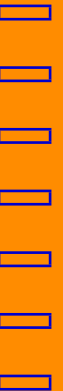
$$f(x, y; \gamma) = \exp(-\gamma|x - y|^2)$$

where γ is a user defined squashing parameter.

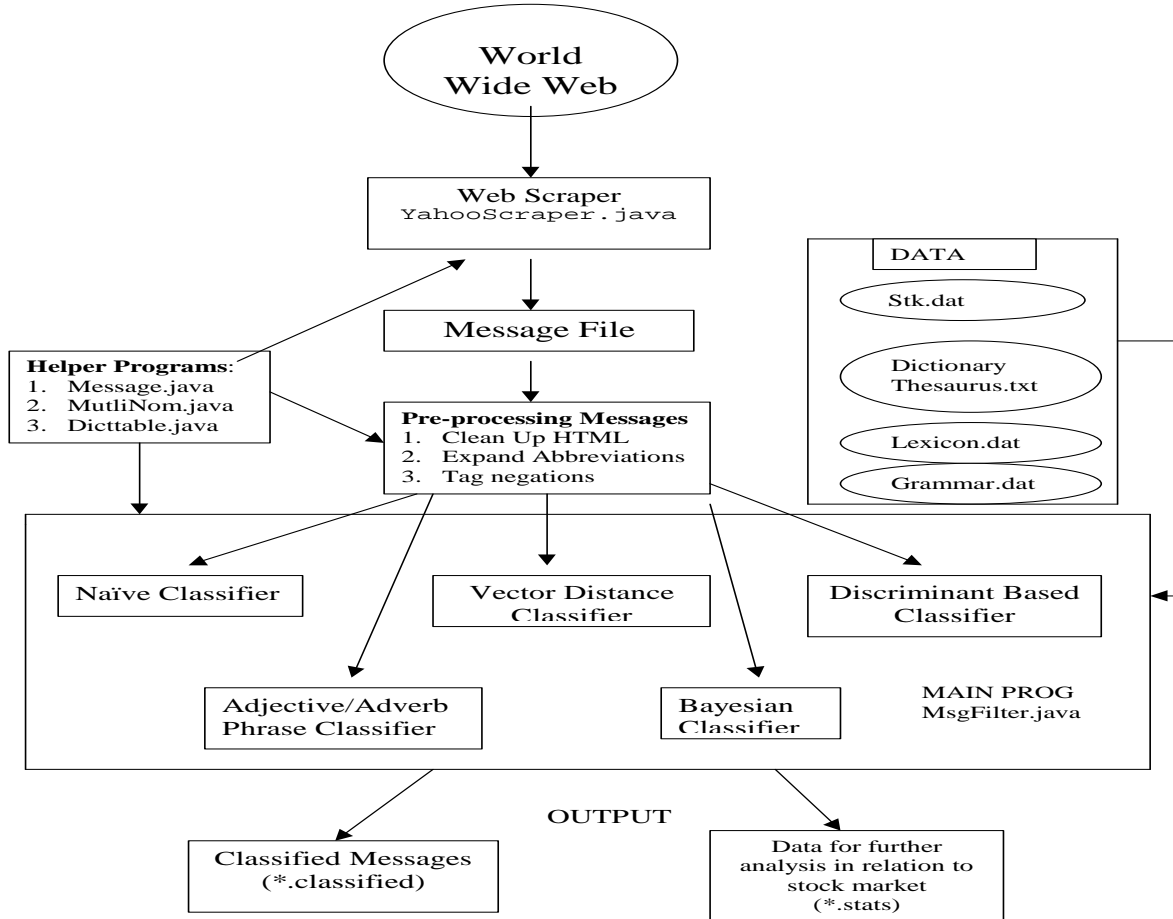


Classification Algorithms

- 5 algorithms, allows voting.
- Non-optimizer dependent, fully analytical, faster than SVMs.
- Some language dependence.
- Parts-of-speech tagging done using CUVOALD (Computer Usable Version of the Oxford Advanced Learner's Dictionary), Birkbeck College, University of London, courtesy of Roger Mitton of the Computer Science Department. It contains about 70,000 words, covers 80-90 percent of the words in a message were found in the dictionary.



SCHEMATIC OF LANGUAGE ALGORITHMS FOR STOCK MESSAGE CLASSIFICATION

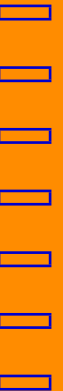


The Lexicon

Words are the heart of any language inference system, and in a specialized domain, this is even more so. In the words of F.C. Bartlett,

“Words ... can indicate the qualitative and relational features of a situation in their general aspect just as directly as, and perhaps even more satisfactorily than, they can describe its particular individuality. This is, in fact, what gives to language its intimate relation to thought processes.”

The text classification model relies on a lexicon of “discriminant” words. This lexicon was designed using domain knowledge and statistical methods.





Constructing the Lexicon

A discriminant function was used to statistically detect which words in the training corpus were good candidates for classifier usage.

The features of the lexicon are as follows:

1. These words are hand-selected based on reading a few thousand messages.
2. The lexicon may be completely user-specified, allowing the methodology to be tailored to individual preference.
3. For each word in the lexicon, we tag it with a “base” value, i.e. the category in which it usually appears.
4. Each word is also “expanded” (reverse stemming).
5. “Negation” counterparts.
6. About 300 base words.



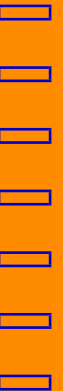
Grammar

- Training corpus, a set of rules.
- Search the grammar for a rule that may be applied to the message. A distance function under a carefully chosen metric is used to identify the applicable rule. Any message is explored for affinity to a rule, takes on the classification character of the rule if thresholds are met.
- It is a “conceptual processor” (Roger Schank). The grammar rules work together to make sense of the “thought bullets” posted to the web. Schank states this particularly well: “People do not usually state all the parts of a given thought that they are trying to communicate because the speaker tries to be brief and leaves out assumed or unessential information. The conceptual processor searches for a given type of information in a sentence or a larger unit of discourse that will fill the needed slot.”



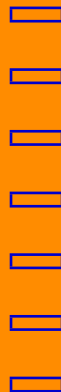
Negation

Whenever a negation word appears in a sentence, it usually causes the meaning of the sentence to be the opposite of that without the negation. For example, the sentence “It is not a bullish market” actually means the opposite of a bull market. Words such as “not”, “never”, “no”, etc., serve to reverse meaning. We handle negation by detecting these words and then tagging the rest of the words in the sentence after the negation word with markers. These markers appear as “--n” concatenated to the end of the word. The lexicon is designed to contain words with negation markers, so that their presence in a message will be correctly interpreted.



Naive Classifier (NC)

- This algorithm is based on a simple word count of positive and negative connotation words.
- If the net number crosses a given threshold, we can classify it as a buy or sell, else it would be treated as neutral.
- Each word is parsed through the lexicon, and assigned a value based on the default value in the lexicon.
- A baseline approach to the classification problem.



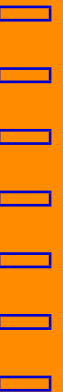


Vector Distance Classifier [VDC]

- This algorithm treats each message as a word vector. Therefore, each hand-tagged message becomes a grammar rule.
- Each message in the test set is then compared to the rule set and is assigned a classification based on which rule comes closest in vector space.
- The angle between the message vector (M) and the vectors in the grammar (G) provides a measure of proximity.

$$\cos(\theta) = \frac{M \cdot G}{|M| \cdot |G|}$$

- Variations on this theme are made possible by using sets of top- n closest rules, rather than only the closest rule.
- Similar to regression analysis.





Discriminant-Based Classifier [DBC]

- Weight lexical items differently.
- Categories $C = \{0, 1, 3\}$.
- Mean score (average number of times term t appears in a message of category i) of each term for each category $= \mu_i$, where i indexes category (we suppress all subscripts for t to simplify exposition here).
- Messages indexed by j . The number of times term t appears in a message j of category i is denoted m_{ij} . Let n_i be the number of times term t appears in category i .

$$F(t) = \frac{\frac{1}{|C|} \sum_{i \neq k} (\mu_i - \mu_k)^2}{\sum_i \frac{1}{n_i} \sum_j (m_{ij} - \mu_i)^2} \quad (8)$$

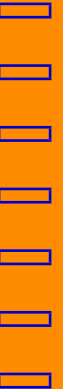
- Ratio of the across-class (class i vs class k) variance to the average of within-class (class $i \in C$) variances.



Discriminant Values

SAMPLE DISCRIMINANT VALUES

bad 0.040507943664639216
hot 0.016124148231134897
hype 0.008943543938332603
improve 0.012395140059803732
joke 0.02689751948279659
jump 0.010691670826157351
killing 0.010691670826157329
killed 0.016037506239236058
lead 0.003745650480005731
leader 0.0031710056164216908
like 0.003745470397428718
long 0.01625037430824596
lose 0.12114219092843743
loss 0.007681269362162742
money 0.15378504322023162
oversell 0.0
overvalue 0.016037506239236197
own 0.0030845538644182426
gold__n 0.0
good__n 0.04846852990132937
grow__n 0.016037506239236058



Adjective-Adverb Phrase Classifier [AAPC]

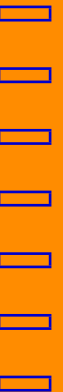
- Using a parts of speech tagger in conjunction with the CUVOALD dictionary, search for noun phrases containing adjectives or adverbs, i.e. in its simplest form, this would be an adjective-noun pair.
- Form a “triplet”, which consists of the adjective or adverb and the two words immediately following it in the message.
- Triplet contains meaningful interpretive information because adjective or adverb add emphasis to the phrase in which they are embedded.
- Submit significant phrases for lexical-grammar analysis, obtaining either a buy or sell token for the phrase.





Bayesian Classifier [BC]

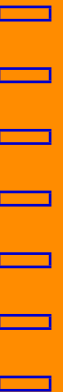
- See Koller and Sahami (1997), Mitchell (1997), Charkrabarti, Dom, Agrawal and Raghavan (1998).
- Based on a term-message-class (t, m, c) model.
- The Bayesian classifier works off word-based probabilities, and is thus indifferent to the structure of the language. Since it is language-independent, it has wide applicability. In particular, the method enables investigation of message boards in other financial markets, where the underlying language may not be English.





Bayesian Classifier - Basic Notation

- The total number of classes is denoted C , such that we have $c_i, i = 1 \dots C$.
- Each message is denoted $m_j, j = 1 \dots M_i, \forall i$.
- M_i is the total number of messages per class.
- The total number of terms is T .
- Lexicon: a set of terms $F = \{t_k\}_{k=1}^T$ that are employed in distinguishing which class each message falls into.
- Use discriminant analysis to determine a good set of terms $\{t_k\}$.

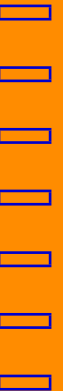




Bayesian Classifier - Definitions

- $n(m, t) \equiv n(m_j, t_k)$: is the total number of times term t_k appears in message m_j .
- $q(m, t) \equiv q(m_j, t_k)$ is an indicator function which is equal to 1 if the term t_k appears in message m_j , else it is equal to zero.
- $n(m)$: is the total number of terms/words in message d including duplicates. Of course, $n(m_j) = \sum_{k=1}^T n(m_j, t_k)$.
- $q(m)$: is the total number of terms/words in message d excluding duplicates. Of course, $q(m_j) = \sum_{k=1}^T q(m_j, t_k)$.
- $q(t)$: is the total number of messages in which term t_k appears across *all* classes. $q(t_k) = \sum_{j=1}^M q(m_j, t_k)$. Here $M = \sum_{i=1}^C M_i$.
- $n(c)$: the total number of terms in all $m \in c$. This is trivially:

$$n(c_i) = \sum_{m_j \in c_i} n(m_j). \quad (9)$$





- $q(c)$: total number of distinct terms in all $m \in c$, which must satisfy $q(c) \leq n(c)$.
- $n(c, t)$: the number of times term t appears in all $m \in c$. This is

$$n(c_i, t_k) = \sum_{m_j \in c_i} n(m_j, t_k) \quad (10)$$

- $\theta(c_i, t_k)$: the probability with which term t appears in all messages m in class c . Thus, a casual interpretation of $\theta(c, t)$ may be the probabilistic proportion of $t \in m \in c$. We write:

$$\theta(c, t) = \frac{\sum_{m_j \in c_i} n(m_j, t_k)}{\sum_{m_j \in c_i} \sum_k n(m_j, t_k)} = \frac{n(c_i, t_k)}{n(c_i)} \quad (11)$$





Bayesian Classifier - Main Concept

- Compute the most probable class c_i given any message m_j . Bayes' Theorem:

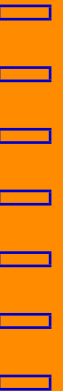
$$\Pr[c_i|m_j] = \frac{\Pr[m_j|c_i] \cdot \Pr[c_i]}{\sum_{i=1}^C \Pr[m_j|c_i] \cdot \Pr[c_i]} \quad (12)$$

- Need reverse conditional probabilities $\Pr[m_j|c_i]$.

$$\Pr[m_j|c_i] = \binom{n(m_j)}{\{n(m_j, t_k)\}} \prod_{k=1}^T \theta(c_i, t_k)^{n(m_j, t_k)} \quad (13)$$

$$= \frac{n(m_j)!}{n(m_j, t_1)! \times n(m_j, t_2)! \times \dots \times n(m_j, t_T)!} \times \prod_{k=1}^T \theta(c_i, t_k)^{n(m_j, t_k)} \quad (14)$$

- Ensure that $\theta(c_i, t_k) \neq 0 \forall c_i, t_k$. This is usually done by employing



Laplace's formula which is

$$\theta(c_i, t_k) = \frac{n(c_i, t_k) + 1}{n(c_i) + K}.$$

where K is the size of the lexicon used by the algorithm.

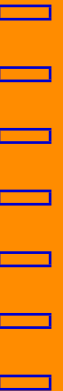
- Computing $\Pr[c_i]$: the proportion of messages classified into class c_i .
- For each message, we get three posterior probabilities, one for each message category. The category with the highest probability is assigned to the message.





Voting Classifier [VC]

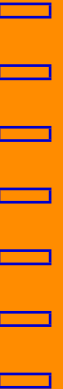
- First, a simple majority rule, i.e. 3 of 5 classifiers should agree on the message type. Second, raise threshold to requiring agreement amongst 4 of the 5 classifiers. Third, unanimity.
- As we go from simple majority to full consensus, the accuracy of classification improves dramatically, but the number of messages classified falls as well.
- Chose simple majority rule



Results

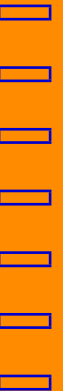
- Benchmark: 100% or human asymptotic agreement level?
- Metric:

$$\text{Accuracy}(\%) = \frac{\text{No of correct classifications}}{\text{No of attempted classifications}}$$



Algorithm Performance

(1) Method No	(2) Classification Method	(3) Correct Messages	(4) Attempted Messages	(5) Percent Accuracy
1	Naive (NC)	33	64	51.56%
2	Vector-Distance (VDC)	27	64	42.19
3	Discriminant (DBC)	32	64	50.00
4	Adj-Noun (AAPC)	33	64	51.56
5	Bayes (BC)	34	64	53.13
6	2-votes	34	64	53.13
7	3-votes	32	52	61.54
8	4-votes	17	20	85.00
9	5-votes	5	5	100.00



Confusion Matrix

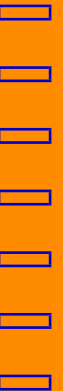
	Null	Sell	Buy
Null	58%	10%	32%
Sell	4%	71%	25%
Buy	40%	20%	40%





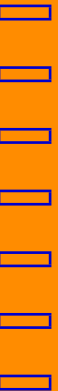
Stock Market Analysis

- Stock returns: computed from the data as $R_t^{daily} = \ln\left(\frac{S_t}{S_{t-1}}\right) \times 260$, i.e. continuously compounded returns, where S_t^{close} represents the closing stock price at date t . The factor 260 annualizes the returns.
- Intra-day returns: given by $R_t^{intra} = \ln\left(\frac{S_t^{close}}{S_t^{open}}\right) \times 260$.
- A proxy for intra-day stock volatility, given as $S_t^{high} - S_t^{low}$.
- Trading volume, provided directly from the data set.
- Message posting volume, equal to the total number of messages posted during the trading day.
- Stock market sentiment (denoted Q), equal to the net of buy and sell messages posted during the day.
- Bull/Bear indicator, given by an indicator variable, equal to 1 if $Q > 0$ and equal to -1 if $Q < 0$.



Descriptive Statistics

	Total	Mean	StdDev
Basic Stock Related Data			
OPEN		66.9032	7.4237
HIGH		69.2449	7.5643
LOW		64.1441	7.6966
CLOSE		66.6364	7.7130
VOLUME	742566000	9643714	6928030
Sentiment Information			
NULL	10892	141.4545	86.3002
BUY	7231	93.9091	53.9898
SELL	6094	79.1429	50.3346
MSGVOL	24217	314.5065	187.7189
SENTY		14.7662	18.9597
BULL/BEAR	41	0.5324	0.8520
Derived Information			
R-DAILY		-1.6624	15.9583
HI-LO		5.1009	2.3212
R-INTRA		-1.1499	12.8512



Correlations

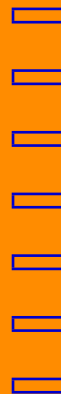


	OPEN	HIGH	LOW	CLOSE	VOL UME	NULL	BUY	SELL	MSG VOL	SENTY	RET- DAILY	HI-LO
HIGH	0.956											
LOW	0.966	0.954										
CLOSE	0.912	0.964	0.951									
VOLUME	0.217	0.358	0.170	0.287								
NULL	0.262	0.328	0.208	0.229	0.700							
BUY	0.302	0.372	0.241	0.273	0.748	0.947						
SELL	0.200	0.257	0.136	0.152	0.687	0.970	0.936					
MSGVOL	0.261	0.327	0.202	0.225	0.722	0.992	0.974	0.983				
SENTY	0.329	0.376	0.327	0.374	0.306	0.124	0.362	0.011	0.164			
R-DAILY	-0.028	0.144	0.092	0.304	0.239	-0.014	0.020	-0.074	-0.021	0.254		
HI-LO	-0.089	0.096	-0.207	-0.011	0.603	0.378	0.411	0.388	0.396	0.142	0.164	
R-INTRA	-0.134	0.090	0.048	0.280	0.159	-0.069	-0.051	-0.102	-0.074	0.126	0.796	0.132



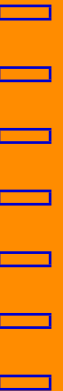
Empirical Results

Dependent Variable	Intercept	Independent Variables		
MSGVOL ($R^2 = 0.52$)	125.96 (4.91)	1.96E-05 VOL (9.03)		
MSGVOL ($R^2 = 0.03$)	290.54 (10.76)	1.62 SENTY (1.13)		
MSGVOL ($R^2 = 0.16$)	151.19 (3.15)	32.02 HI-LO (3.73)		
SENTY ($R^2 = 0.09$)	6.68 (1.87)	8.39E-07 VOL (2.79)		
SENTY ($R^2 = .08$)	15.24 (7.22)	0.50 R-DAILY (2.29)	-0.31 R-INTRA (-1.14)	
SENTY ($R^2 = 0.02$)	8.83 (1.69)	1.16 HI-LO (1.25)		
SENTY(t) ($R^2 = 0.17$)	8.33 (1.96)	4.88E-07 VOL(t-1) (1.59)	8.25E-7 VOL(t) (2.58)	-7.03E-07 VOL(t+1) (-2.28)
SENTY(t) ($R^2 = 0.12$)	14.35 (6.91)	0.04 R-DAILY(t-1) (0.30)	0.25 R-DAILY(t) (1.92)	-0.31 R-DAILY(t+1) (-2.29)



Falling Apple - a situation of gravity

- On September 28, 2000 – Apple Computer announces poorer than expected profits and the stock crashes 50% in one day.
- The volume of messages goes through the roof. Message volume rises 20 times from about 100 a day to 2000 per day.
- The market sentiment index anticipates the sharp drop. There is huge negative sentiment a day before the crash. (Information leakage)
- Trading volume rises by a factor of 10.





[Home](#) - [Yahoo!](#) - [Help](#)

Welcome

Track stocks, view your [bank](#), [brokerage](#), or [credit card](#) accounts, and more! [[Register/Sign In](#)]

[Customize \(Yahoo! ID required\)](#) - [Sign In](#)

Quotes



\$10,000 in prizes to be won! Play the [Yahoo! Investment Challenge](#).

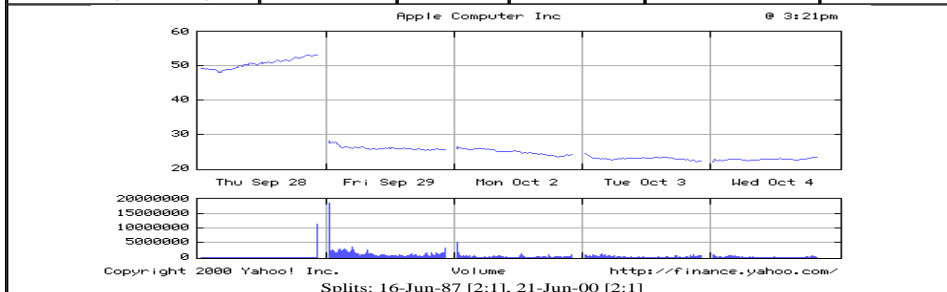
Click to trade or open an account. - [Important Disclaimer](#)

Views: [Basic](#) - [DayWatch](#) - [Performance](#) - [Fundamentals](#) - [Real-time ECN News](#) - [Detailed](#) - [[Create New View](#)]

Wed, October 4 2000 3:38pm ET - U.S. Markets close in 22 minutes.

APPLE COMP INC (NasdaqNM:AAPL) - More Info: [News](#) . [Msgs](#) . [Profile](#) . [Research](#) . [Insider](#) . [Options](#)

Last Trade 3:17PM - 23 1/2	Change +1 3/16 (+5.32%)		Prev Cls 22 5/16	Volume 23,023,700	Div Date Jun 20
Day's Range 21 7/8 - 23 9/16	Bid 23 1/2	Ask 23 9/16	Open 22 3/8	Avg Vol 8,210,772	Ex-Div Jun 21
52-week Range 22 3/16 - 75 3/16	Earn/Shr 2.03	P/E 10.99	Mkt Cap 7.637B	Div/Shr N/A	Yield N/A



[1d](#) | [5d](#) | [3m](#) | [1y](#) | [2y](#) | [5y](#) | [max](#) Other: [historical quotes](#) | [small chart](#)

Basic | [Moving Avg](#) | Compare AAPL vs. S&P Nasdaq Dow

[Add to My Portfolio](#) - [Set Alert](#)

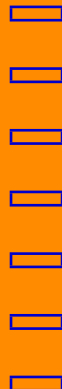
[Non-Tables Version](#) - [Download Spreadsheet](#)

Quotes delayed 15 minutes for Nasdaq, 20 minutes otherwise.

[Customize Finance \(Yahoo! ID required\)](#) - [COOL JOBS @ YAHOO!](#) - [Yahoo! Finance Home](#)

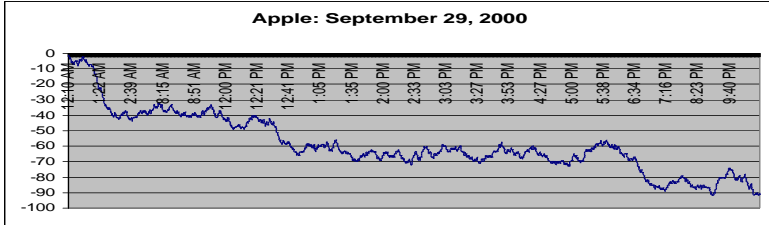
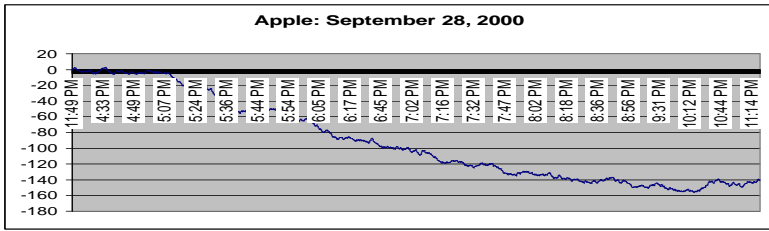
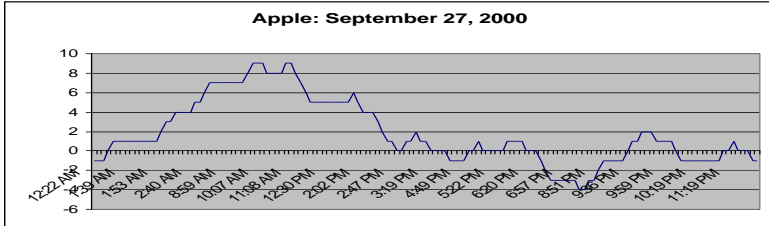
Recent News

[Customize News](#)

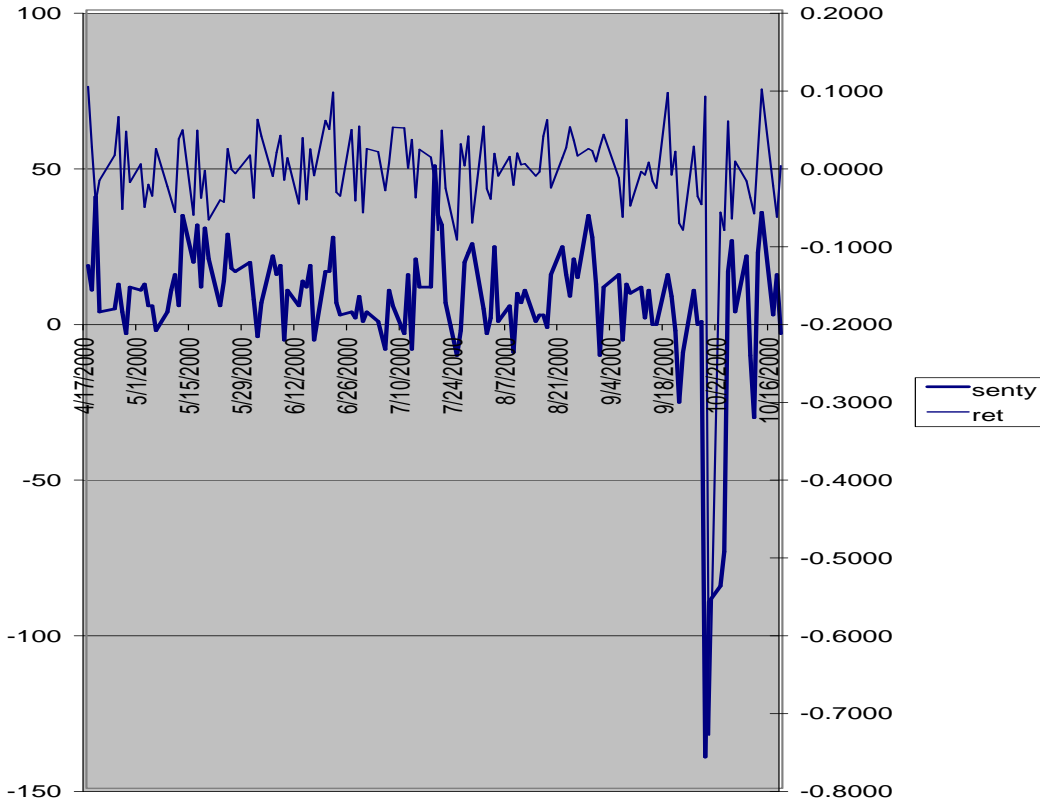




Apple Computer's Sentiment Index
(before, during and after announcing an earnings drop on September 28, 2000 at 4pm).

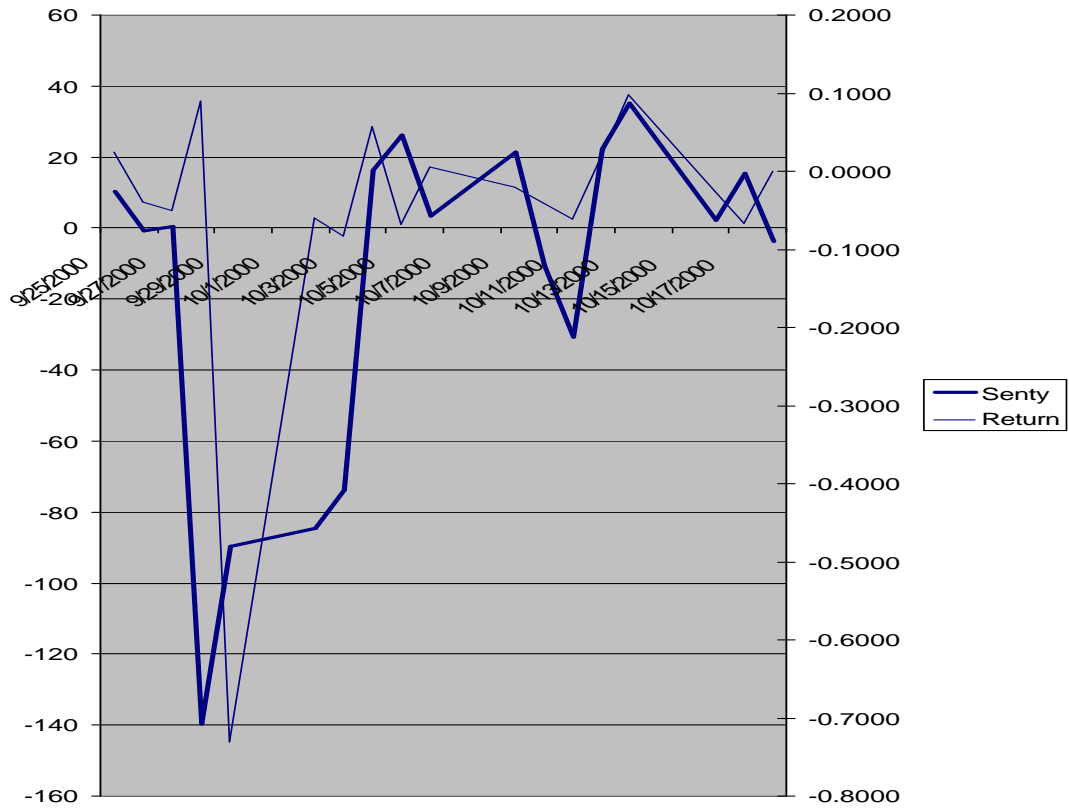


Apple Computer (14-Apr to 18-Oct, 2000)





Apple Computer (25-Sep to 18-Oct, 2000)



Apple Computer Inc

Basic Statistics

Statistic	Mean	Std. Dev	Min	Max
Closing Price	49.63	10.77	19.44	63.56
Trading Volume	7.27M	12.61M	1.26M	133.00M
Sentiment Index	6.08	22.70	-140	+50
Daily Return	-0.79	7.71	-73.12	10.19
NASDAQ Return	-0.04	2.73	-6.12	7.64



Apple Computer Inc



Sentiment and Returns

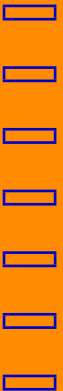
	Entire 130 days		First 100 days	
	Estimate	T-stat	Estimate	T-stat
Intercept	-0.0194	-3.11	0.0034	0.56
SENTY(t)	0.0005	1.60	0.0004	1.26
SENTY(t-1)	0.0013	4.35	-0.0007	-2.00
R^2	0.25		0.05	

Sentiment and Volatility

$$RET_t = \alpha + \beta NASDRET_t + \epsilon_t$$

$$ARCHVAR_t = a_0 + a_1 \epsilon_{t-1}^2 + a_2 SENTRY_{t-1}^2$$

	Entire 130 days		First 100 days	
	Estimate	T-stat	Estimate	T-stat
α	-0.1479	-0.44	0.0449	0.16
β	1.0265	8.87	1.0448	10.21
a_0	8.9572	8.09	7.7827	5.02
a_1	-0.0307	-0.53	-0.1277	-1.83
a_2	0.0186	6.18	0.0086	1.50



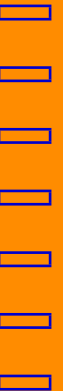
A Four Stock Analysis

Choice of the stocks

	Message Volume	
News Flow	High	Low
High	AMZN	DAL
Low	GMGC	GWRX

Other Criteria:

- Market cap (size)
- Institutional holdings
- Day trading volume





Stock Return relationships

Is sentiment related to returns?

Only contemporaneously, ...

- $STKRET[t] = -0.0075 + 0.0004^{**} SENTRY[t]$
- $STKRET[t] = -0.00016 + -6.4E-05 SENTRY[t-1]$
- $SENTRY[t] = 14.32 + 15.84 STKRET[t-1]$
- Similar results for individual stocks also.
- Exception: posters in GWRX show a statistically negative relationship of STKRET regressed on lagged SENTRY.



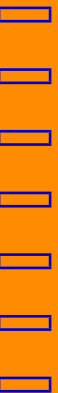
The Correlation of Posting Volume across 4 boards

- Strong Posting Volume correlations, even after removing days on which there were no messages on any board.

	TMF	Yahoo	SillInv
Yahoo	0.547		
SillInv	0.794	0.532	
RBull	0.479	0.434	0.293

- Posting Volume and Volatility

	NBRTOT	HISTOVOL	HILOPCTG	BREAKRUN
HISTOVOL	0.184			
HILOPCTG	0.188	0.460		
BREAKRUN	0.262	-0.507	-0.305	
NBRJUMPP	0.401	-0.071	0.389	0.266





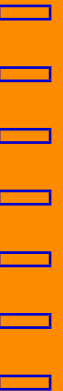
Message Volume and Volatility

- Message Volume is a function of market uncertainty

$$\text{NBRTOT}[t] = -242 + 394.06^{**} \text{DISTOT}[t] + 3.34^{**} \text{BREAKRUN}[t] + 39.35^{**} \text{NBRJUMPP}[t]$$

- Note: DISTOT is a disagreement index which is

$$\text{DISTOT} = \left| \frac{|\text{SENTY}|}{\text{No of msgs}} - 1 \right|$$



Volatility and Information

- Intraday volatility is related to public news announcements:

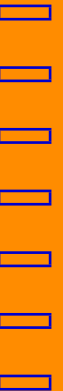
$$\text{HILOPCTG}(t) = 0.0566 + 1.59\text{E-}05 \text{NBRTOT}(t) + 0.0055^{**} \text{NEWSNBR}(t) + 3.07\text{E-}06 \text{NEWSLGTH}(t)$$

- Market direction changes are related to message board activity:

$$\text{BREAKRUN}(t) = 25.45 + 0.016^{**} \text{NBRTOT}(t) - 0.226 \text{NEWSNBR}(t) - 0.0002 \text{NEWSLGTH}(t)$$

- Jumps or market gaps are related to message board volume:

$$\text{NBRJUMPP}(t) = 0.38 + 0.0035^{**} \text{NBRTOT}(t) + 0.063 \text{NEWSNBR}(t) - 0.0001 \text{NEWSLGTH}(t)$$





Surprise and Sentiment

- The number of jumps is contemporaneously related to the absolute value of the sentiment index:

$$\text{NBRJUMPP}(t) = 0.77^{**} + 0.025^{**} \text{ ABSSENT}(t)$$

- The relationship remains even when public news is added as a control:

$$\text{NBRJUMPP}(t) = 0.39^{**} + 0.15^{**} \text{ ABSSENT}(t) + 0.102^{**} \text{ NEWSNBR}(t)$$

- Jumps are predicted by lagged sentiment, but not by lagged news volume:

$$\text{NBRJUMPP}(t) = 0.77^{**} + 0.20^{**} \text{ ABSSENT}(t-1) + 0.01 \text{ NEWSNBR}(t-1)$$



Summary

- Algorithms to parse stock market sentiment are feasible.
- Sentiment volume and trading volume are related.
- Sentiment tracks market returns and volatility.
- Sentiment is more focussed when external news is also present.
- Sentiment appears to predict jumps.
- Sentiment is related to stock return and volatility more closely on a minute-by-minute time scale.

