Online Appendix

Empirical Evaluation of an Automated Intraday Stock Recommendation System Incorporating Both Market Data and Textual News

Tomer Geva, Jacob Zahavi
Recanati Business School, Tel Aviv University

Contents

1. List of Companies

#	Ticker	Company
1	ADBE	Adobe Systems Inc
2	AES	Aes Cp Inc
3	AKAM	Akamai Technologies, Inc.
4	AA	Alcoa Inc
5	ALTR	Altera Corporation
6	MO	Altria Group Inc
7	AXP	American Express Inc
8	AIG	American Intl Group Inc
9	AMGN	Amgen
10	BUD	Anheuser Busch
11	AMAT	Applied Materials, Inc.
12	ADSK	Autodesk, Inc.
13	BEAS	BEA Systems, Inc.
14	BA	Boeing Co
15	BRCM	Broadcom Corporation
16	CDNS	Cadence Design Systems, Inc.
17	CDWC	CDW Corporation
18	CHKP	Check Point Software Technologies Ltd.
19	CVX	Chevron Corp
20	CI	Cigna Cp
21	CTAS	Cintas Cp
22	CSCO	Cisco Systems, Inc.
23	CTXS	Citrix Systems, Inc.
24	KO	Coca Cola Co The
25	CTSH	Cognizant Technology Solutions Corporation
26	CL	Colgate Palmolive
27	CMCSA	Comcast Cp
28	CAG	Conagra Food Inc
29	DELL	Dell Inc
30	DD	Du Pont
31	EBAY	Ebay Inc
32	XOM	Exxon Mobile Cp
33	FISV	Fiserv Inc
34	GD	General Dynamics Cp
35	GM	General Motors
36	GOOG	Google Inc.
37	HPQ	Hewlett Packard Co
38	HD	Home Depot Inc
39	INTC	Intel Cp
40	INTU	Intuit Inc.

41	JNJ	Johnson And Johnson Dc
42	JNPR	Juniper Networks, Inc.
43	KLAC	KLA-Tencor Corporation
44	LSI	LSI Logic Cp
45	LRCX	Lam Research Corporation
46	LLTC	Linear Technology Corporation
47	LMT	Lockheed Martin Cp
48	MKC	Mccormick & Co
49	MRK	Merck Co Inc
50	MET	Metlife Inc
51	MSFT	Microsoft Cp
52	MIL	Millipore Cp
53	NSM	National Semiconductor
54	NTAP	Network Appliance, Inc.
55	NUE	Nucor Cp
56	NVDA	Nvidia Corp
57	OMX	Officemax Inc.
58	ORCL	Oracle Corporation
59	PTV	Pactiv Corp
60	PEP	Pepsico Inc
61	PFE	Pfizer Inc
62	PX	Praxair Inc
63	QCOM	Qualcomm Incorporated
64	RSH	Radioshack Corp
65	SNDK	Sandisk Corporation
66	TLAB	Tellabs, Inc.
67	UTX	United Tech
68	VRSN	Verisign, Inc.
69	WMT	Wal Mart Stores
70	DIS	Walt Disney-Disney C
71	XLNX	Xilinx, Inc.
72	YHOO	Yahoo! Inc.

Welch's T-test for difference between our sample of stocks and the rest of the S&P500 index stocks.

	Returns	Sharpe Ratio
Difference between sample mean and remaining S&P500 stocks mean	0.013	0.1
P-Value for Welch's T test	0.57	0.46

H0: Two samples have the same average

2. Detailed Variable List

2.1 General

#	Field	Description	Type
0	id	database id	integer (long)
1	comp_name	company identifier	text
2	daily_time_sec	time of day in seconds	integer (long)
3	is_after_weekend_holiday	1 if the current trading day is after a weekend or holiday	binary
4	is_short_trading_day	1 if the current trading day is a short trading day	binary
5	is_between_930_10	1 if the time is between 9:30 and 10:00	binary
6	is_between_1545_16	1 if the time is between 15:45 and 16:00	binary
7	beta_end_2005	beta value for the end of year 2005 (source CRSP)	decimal
8	shares_outst_sep1	outstanding shares september 1st 2006	integer (long)
9	market_cap_sep1	Market Capital september 1st 2006	integer (long)
10	is_DJI	1 if the companyis included in the DJI 30 index	binary
11	is_tech100	1 if the companyis included in the NASDAQ tech 100 index	binary
12	is_NYSE	1 if the companyis listed in NYSE	binary
13	sector	sector (data from Yahoo finance)	text
14	day_after_dividend	1 if the current trading day is following dividend	binary

2.2 Simple Market Data Transformation

Motivation for including different sets of variable in the prediction model:

Previous literature had already demonstrated that the following types of variables are related to stock returns:

(Past) Returns	Various long term and short term patterns which involve returns are documented in literature e.g., (Rendleman et al., 1982), (Busse and Green, 2002). We expect that this set of information will be useful for data-mining algorithms predicting stock returns.
Trading Volume	Studies such as (Andersen 1996) show that trading volume affects volatility, which in turn, is related to the extent of stock price movements. We therefore expect that this set of information will be useful for data-mining algorithms predicting stock returns.
SPY fund returns	SPY fund data is used to represent the S&P500 index movements, as in (Antweiller and Frank, 2004). We expect that market movements would also affect individual stocks. Thus, this information will be useful for data-mining algorithms predicting stock returns.

Variable List:

#	Field	Description	Type
15	returns1	returns from 1 previous interval	decimal
16	returns2	returns from 2 previous intervals to 1 previous intervals	decimal
17	returns3	returns from 3 previous intervals to 2 previous intervals	decimal
18	returns4	returns from 4 previous intervals to 3 previous intervals	decimal
19	returns5	returns from 5 previous intervals to 4 previous intervals	decimal
20	thr_returns1	1 if returns1 >0	binary
21	thr_returns2	1 if returns2 >0	binary
22	thr_returns3	1 if returns3 >0	binary
23	thr_returns4	1 if returns4 >0	binary
24	thr_returns5	1 if returns5 >0	binary
25	AbsRtn1	absolute value of returns1	decimal
26	AbsRtn2	absolute value of returns2	decimal
27	AbsRtn3	absolute value of returns3	decimal
28	AbsRtn4	absolute value of returns4	decimal
29	AbsRtn5	absolute value of returns5	decimal
30	ThAbsRtn1	1 if AbsRtn1 >0	binary
31	ThAbsRtn2	1 if AbsRtn2 >0	binary
32	ThAbsRtn3	1 if AbsRtn3 >0	binary
33	ThAbsRtn4	1 if AbsRtn4 >0	binary
34	ThAbsRtn5	1 if AbsRtn5 >0	binary
35	RtnTodayOpen	returns from today opening time	decimal
36	RtnPrvClsDySt	returns from previous day close time to today start time	decimal
37	AbsRtnTodayOpen	1 if RtnTodayOpen >0	binary
38	AbsRtnPrvDyCls	1 if RtnPrvDyCls >0	binary
39	AbsRtnPrvClsDySt	1 if RtnPrvClsDySt >0	binary
40	spy_returns1	SPY fund returns from 1 previous interval	decimal
41	spy_returns2	SPY fund returns from 2 previous intervals to 1 previous intervals	decimal
42	spy_returns3	SPY fund returns from 3 previous intervals to 2 previous intervals	decimal
43	spy_returns4	SPY fund returns from 4 previous intervals to 3 previous intervals	decimal
44	spy_returns5	SPY fund returns from 5 previous intervals to 4 previous intervals	decimal
45	thr_spy_returns1	1 if spy_returns1 >0	binary
46	thr_spy_returns2	1 if spy_returns2 >0	binary
47	thr_spy_returns3	1 if spy_returns3 >0	binary
48	thr_spy_returns4	1 if spy_returns4 >0	binary
49	thr_spy_returns5	1 if spy_returns5 >0	binary
50	AbsSpyRtn1	absolute value of spy_returns1	decimal
51	AbsSpyRtn2	absolute value of spy_returns2	decimal
52	AbsSpyRtn3	absolute value of spy_returns3	decimal
53	AbsSpyRtn4	absolute value of spy_returns4	decimal
54	AbsSpyRtn5	absolute value of spy_returns5	decimal
55	RtnSpyTodayOpen	SPY fund returns from today opening time	decimal
56	RtnSpyPrvClsDySt	SPY fund returns from previous day close time to today start time	decimal
57	AbsRtnSpyTodayOpen	1 if RtnSpyTodayOpen >0	binary
58	AbsRtnSpyPrvDyCls	1 if RtnSpyPrvDyCls >0	binary
59	AbsRtnSpyPrvClsDySt	1 if RtnSpyPrvClsDySt >0	binary
60	prev_trades1	trades from 1 previous interval	integer (long)
61	prev_trades2	trades between 2 previous intervals to 1 previous intervals	integer (long)
62	prev_trades3	trades between 3 previous intervals to 2 previous intervals	integer (long)
63	prev_trades4	trades between 4 previous intervals to 3 previous intervals	integer (long)
64	prev_trades5	trades between 5 previous intervals to 4 previous intervals	integer (long)
65	nrm_prev_trades1	normalize prev_trades1 by trading volume per interval during the previous 5 trading days	decimal
66	nrm_prev_trades2	normalize prev_trades2 by trading volume per interval during the previous 5 trading days	decimal
67	nrm_prev_trades3	normalize prev_trades3 by trading volume per interval during the previous 5 trading days	decimal
68	nrm_prev_trades4	normalize prev_trades4 by trading volume per interval during the previous 5 trading days	decimal
69	nrm_prev_trades5	normalize prev_trades5 by trading volume per interval during the previous 5 trading days	decimal
70	today_trades	trades from today opening time	integer (long)
71	TradesPrvDyCls	trades from today opening time	integer (long)
72	TrdPrvDyClsDyStrt	trades from today opening time	integer (long)

2.3 Technical Analysis Indicators

Motivation: Technical Analysis indicators provide additional transformations. Data-mining models may utilize such transformation to capture additional patterns within the data. See, for instance, (Dhar and Chou, 2001) or (Thomas, 2003).

Transformation are defined in (Thomas, 2003)

#	Field	Description	Type
73	smpl_avg10	price simple average - last 10 intervals	decimal
74	exp_avg10	price exponential average - last 10 intervals	decimal
75	RSI_10	Relative strength index - last 10 intervals	decimal
76	stdev_10	Standard Deviation of price - last 10 intervals	decimal
77	NROC_10	NROC normalized n-day rate of change - last 10 intervals	decimal
78	OBV_10	OBV: On Balance Volume - last 10 intervals	decimal
79	CLV_10	Close Location Value (CLV) - last 10 intervals	decimal
80	ADL_10	The accumulation/distribution line (ADL) - last 10 intervals	decimal
81	CMF_10	Chaikin Money Flow (CMF) - last 10 intervals	decimal
82	percent_K10	K% - stochastic oscillators - last 10 intervals	decimal
83	smpl_avg50_ratio	price simple average - last 50 intervals	decimal
84	exp_avg50_ratio	price exponential average - last 50 intervals	decimal
85	RSI_50	Relative strength index - last 50 intervals	decimal
86	stdev_50	Standard Deviation of price - last 50 intervals	decimal
87	NROC_50	NROC normalized n-day rate of change - last 50 intervals	decimal
88	OBV_50	OBV: On Balance Volume - last 50 intervals	decimal
89	CLV_50	Close Location Value (CLV) - last 50 intervals	decimal
90	ADL_50	The accumulation/distribution line (ADL) - last 50 intervals	decimal
91	CMF_50	Chaikin Money Flow (CMF) - last 50 intervals	decimal
92	percent_K50	K% - stochastic oscillators - last 50 intervals	decimal
93	smpl_avg200_ratio	price simple average - last 200 intervals	decimal
94	exp_avg200_ratio	price exponential average - last 200 intervals	decimal
95	RSI_200	Relative strength index - last 200 intervals	decimal
96	stdev_200	Standard Deviation of price - last 200 intervals	decimal
97	NROC_200	NROC normalized n-day rate of change - last 200 intervals	decimal
98	OBV_200	OBV: On Balance Volume - last 200 intervals	decimal
99	CLV_200	Close Location Value (CLV) - last 200 intervals	decimal
100	ADL_200	The accumulation/distribution line (ADL) - last 200 intervals	decimal
101	CMF_200	Chaikin Money Flow (CMF) - last 200 intervals	decimal
102	percent_K200	K% - stochastic oscillators - last 200 intervals	decimal

2.4 Piecewise Linear Representation

Motivation: Provides compact representation of trends within time series data. Trends representation in time series data have been employed in related studies such as (Fung et al. 2003).

#	Field	Description	Type
103	Coefficient	The coefficient of the current linear segment	decimal
104	Duration	The duration of the current linear segment	decimal
105	Coefficient_prev1	The coefficient of the previous linear segment	decimal
106	Duration_prev1	The duration of the previous linear segment	decimal
107	Coefficient_prev2	The coefficient of the linear segment which is 2 segments before the current segment	decimal
108	Duration_prev2	The duration of the linear segment which is 2 segments before the current segment	decimal
109	Coefficient_prev3	The coefficient of the linear segment which is 3 segments before the current segment	decimal
110	Duration_prev3	The duration of the linear segment which is 3 segments before the current segment	decimal
111	Coefficient_prev4	The coefficient of the linear segment which is 4 segments before the current segment	decimal
112	Duration_prev4	The duration of the linear segment which is 4 segments before the current segment	decimal

2.5 Simple News Item Counts

#	Field	Description	Type
113	current_interval_genera	number of news items during the current interval (general count)	integer
114	current_interval_RTRS	number of news items during the current interval (source: PRN Reuters)	integer
115	current_interval_PRN	number of news items during the current interval (source: PRN Newswire)	integer
116	current_interval_BSW	number of news items during the current interval (source: Business Wire)	integer
117	current_interval_EOL	number of news items during the current interval (source: Edgar Online)	integer
118	current_interval_MKW	number of news items during the current interval (source: MarketWatch)	integer
119	current_interval_BREAK	number of "Breaking News" news items during the current interval	integer
120	prev1_interval_general	number of news items during the previous interval (general count)	integer
121	prev1_interval_RTRS	number of news items during the previous interval (source: PRN Reuters)	integer
122	prev1_interval_PRN	number of news items during the previous interval (source: PRN Newswire)	integer
123	prev1_interval_BSW	number of news items during the previous interval (source: Business Wire)	integer
124	prev1_interval_EOL	number of news items during the previous interval (source: Edgar Online)	integer
125	prev1_interval_MKW	number of news items during the previous interval (source: MarketWatch)	integer
126	prev1_interval_BREAKI	number of "Breaking News" news items during the previous interval	integer
127	prev2_interval_general	number of news items during the previous 2 intervals (general count)	integer
128	prev2_interval_RTRS	number of news items during the previous 2 intervals (source: PRN Reuters)	integer
129	prev2_interval_PRN	number of news items during the previous 2 intervals (source: PRN Newswire)	integer
130	prev2_interval_BSW	number of news items during the previous 2 intervals (source: Business Wire)	integer
131	prev2_interval_EOL	number of news items during the previous 2 intervals (source: Edgar Online)	integer
132	prev2_interval_MKW	number of news items during the previous 2 intervals (source: MarketWatch)	integer
133	prev2_interval_BREAKI	number of "Breaking News" news items during the previous 2 intervals	integer
134	prev3_interval_general	number of news items during the previous 3 intervals (general count)	integer
135	prev3_interval_RTRS	number of news items during the previous 3 intervals (source: PRN Reuters)	integer
136	prev3_interval_PRN	number of news items during the previous 3 intervals (source: PRN Newswire)	integer
137	prev3_interval_BSW	number of news items during the previous 3 intervals (source: Business Wire)	integer
138	prev3_interval_EOL	number of news items during the previous 3 intervals (source: Edgar Online)	integer
139	prev3_interval_MKW	number of news items during the previous 3 intervals (source: MarketWatch)	integer
140	prev3_interval_BREAKI	number of "Breaking News" news items during the previous 3 intervals	integer
141	prev4_interval_general	number of news items during the previous 4 intervals (general count)	integer
142	prev4_interval_RTRS	number of news items during the previous 4 intervals (source: PRN Reuters)	integer
143	prev4_interval_PRN	number of news items during the previous 4 intervals (source: PRN Newswire)	integer
144	prev4_interval_BSW	number of news items during the previous 4 intervals (source: Business Wire)	integer
145	prev4_interval_EOL	number of news items during the previous 4 intervals (source: Edgar Online)	integer
146	prev4_interval_MKW	number of news items during the previous 4 intervals (source: MarketWatch)	integer
147	prev4_interval_BREAKI	number of "Breaking News" news items during the previous 4 intervals	integer
148	end_prev_trading_day_	number of news items from the end of the previous trading day to the beginning of the current trading day (general count)	integer
149	end_prev_trading_day_	number of news items from the end of the previous trading day to the beginning of the current trading day (source: PRN Reuters)	integer
150	end_prev_trading_day_	number of news items from the end of the previous trading day to the beginning of the current trading day (source: PRN Newswire)	integer
151	end_prev_trading_day_	number of news items from the end of the previous trading day to the beginning of the current trading day (source: Business Wire)	integer
152	end_prev_trading_day_	number of news items from the end of the previous trading day to the beginning of the current trading day (source: Edgar Online)	integer
153	end_prev_trading_day_	number of news items from the end of the previous trading day to the beginning of the current trading day (source: MarketWatch)	integer
154	end_prev_trading_day_	number of "Breaking News" news items from the end of the previous trading day to the beginning of the current trading day	integer
155	trading_day_start_to_cu	number of news items from the end of the previous trading day to the beginning of the current trading day (general count)	integer
156	trading_day_start_to_cu	number of news items from the end of the previous trading day to the beginning of the current trading day (source: PRN Reuters)	integer
157	trading_day_start_to_cu	number of news items from the end of the previous trading day to the beginning of the current trading day (source: PRN Newswire)	integer
158	trading_day_start_to_cu	number of news items from the end of the previous trading day to the beginning of the current trading day (source: Business Wire)	integer
159	trading_day_start_to_cu	number of news items from the end of the previous trading day to the beginning of the current trading day (source: Edgar Online)	integer
160	trading_day_start_to_cu	number of news items from the end of the previous trading day to the beginning of the current trading day (source: MarketWatch)	integer
161	trading_day_start_to_cu	number of "Breaking News" news items from the end of the previous trading day to the beginning of the current trading day	integer

2.6 Bag of Words - Single (stemmed) Words

Variables (stemmed words) were selected according to the top ranked, 400 variables based on the Information Gain criteria.

Original, full variable set, from which these 400 variables were selected, included all possible (stemmed) words available in news items published during the first 3 months of available data. This time period corresponds to the time period of our first, sliding window, training set.

#	Field	Description	Type
162	Importantly	1 if the (stemmed) word appears during the last interval	binary
163	Auto	1 if the (stemmed) word appears during the last interval	binary
164	Attent	1 if the (stemmed) word appears during the last interval	binary
165	Rough	1 if the (stemmed) word appears during the last interval	binary
166	Prosecut	1 if the (stemmed) word appears during the last interval	binary
167	Deliveri	1 if the (stemmed) word appears during the last interval	binary
168	Attorney	1 if the (stemmed) word appears during the last interval	binary
169	Royal	1 if the (stemmed) word appears during the last interval	binary
170	Supplem	1 if the (stemmed) word appears during the last interval	binary
171	Experiment	1 if the (stemmed) word appears during the last interval	binary
172	Broader	1 if the (stemmed) word appears during the last interval	binary
173	Head	1 if the (stemmed) word appears during the last interval	binary
174	Collect	1 if the (stemmed) word appears during the last interval	binary
175	Phase	1 if the (stemmed) word appears during the last interval	binary
176	Own	1 if the (stemmed) word appears during the last interval	binary
177	Online	1 if the (stemmed) word appears during the last interval	binary
178	Elig	1 if the (stemmed) word appears during the last interval	binary
179	Attribute	1 if the (stemmed) word appears during the last interval	binary
180	Ground	1 if the (stemmed) word appears during the last interval	binary
181	Tag	1 if the (stemmed) word appears during the last interval	binary
182	Outperform	1 if the (stemmed) word appears during the last interval	binary
183	distract	1 if the (stemmed) word appears during the last	binary

		interval	
184	Basel	1 if the (stemmed) word appears during the last interval	binary
185	Denot	1 if the (stemmed) word appears during the last interval	binary
186	Reusable	1 if the (stemmed) word appears during the last interval	binary
187	Badg	1 if the (stemmed) word appears during the last interval	binary
188	First	1 if the (stemmed) word appears during the last interval	binary
189	Ingénue	1 if the (stemmed) word appears during the last interval	binary
190	Creative	1 if the (stemmed) word appears during the last interval	binary
191	Whole	1 if the (stemmed) word appears during the last interval	binary
192	Stop	1 if the (stemmed) word appears during the last interval	binary
193	Known	1 if the (stemmed) word appears during the last interval	binary
194	Congress	1 if the (stemmed) word appears during the last interval	binary
195	Flight	1 if the (stemmed) word appears during the last interval	binary
196	interoper	1 if the (stemmed) word appears during the last interval	binary
197	delight	1 if the (stemmed) word appears during the last interval	binary
198	trigger	1 if the (stemmed) word appears during the last interval	binary
199	bolster	1 if the (stemmed) word appears during the last interval	binary
200	stapl	1 if the (stemmed) word appears during the last interval	binary
201	calm	1 if the (stemmed) word appears during the last interval	binary
202	electron	1 if the (stemmed) word appears during the last interval	binary
203	steadili	1 if the (stemmed) word appears during the last interval	binary
204	out	1 if the (stemmed) word appears during the last interval	binary
205	exclud	1 if the (stemmed) word appears during the last interval	binary
206	hospit	1 if the (stemmed) word appears during the last interval	binary
207	band	1 if the (stemmed) word appears during the last interval	binary
208	simpli	1 if the (stemmed) word appears during the last interval	binary
209	clear	1 if the (stemmed) word appears during the last interval	binary
210	health	1 if the (stemmed) word appears during the last	binary

		interval	
211	succe	1 if the (stemmed) word appears during the last interval	binary
212	still	1 if the (stemmed) word appears during the last interval	binary
213	medic	1 if the (stemmed) word appears during the last interval	binary
214	term	1 if the (stemmed) word appears during the last interval	binary
215	oper	1 if the (stemmed) word appears during the last interval	binary
216	sheer	1 if the (stemmed) word appears during the last interval	binary
217	extend	1 if the (stemmed) word appears during the last interval	binary
218	pharmaci	1 if the (stemmed) word appears during the last interval	binary
219	product	1 if the (stemmed) word appears during the last interval	binary
220	highli	1 if the (stemmed) word appears during the last interval	binary
221	cite	1 if the (stemmed) word appears during the last interval	binary
222	copi	1 if the (stemmed) word appears during the last interval	binary
223	court	1 if the (stemmed) word appears during the last interval	binary
224	lifestyl	1 if the (stemmed) word appears during the last interval	binary
225	therapi	1 if the (stemmed) word appears during the last interval	binary
226	ramp	1 if the (stemmed) word appears during the last interval	binary
227	dealer	1 if the (stemmed) word appears during the last interval	binary
228	weaker	1 if the (stemmed) word appears during the last interval	binary
229	rank	1 if the (stemmed) word appears during the last interval	binary
230	smaller	1 if the (stemmed) word appears during the last interval	binary
231	treatment	1 if the (stemmed) word appears during the last interval	binary
232	converg	1 if the (stemmed) word appears during the last interval	binary
233	spectrum	1 if the (stemmed) word appears during the last interval	binary
234	rack	1 if the (stemmed) word appears during the last interval	binary
235	option	1 if the (stemmed) word appears during the last interval	binary
236	border	1 if the (stemmed) word appears during the last interval	binary
237	portabl	1 if the (stemmed) word appears during the last	binary

		interval	
238	kick	1 if the (stemmed) word appears during the last interval	binary
239	increment	1 if the (stemmed) word appears during the last interval	binary
240	scandal	1 if the (stemmed) word appears during the last interval	binary
241	ticket	1 if the (stemmed) word appears during the last interval	binary
242	offici	1 if the (stemmed) word appears during the last interval	binary
243	near	1 if the (stemmed) word appears during the last interval	binary
244	stabil	1 if the (stemmed) word appears during the last interval	binary
245	conduct	1 if the (stemmed) word appears during the last interval	binary
246	borrow	1 if the (stemmed) word appears during the last interval	binary
247	catalyst	1 if the (stemmed) word appears during the last interval	binary
248	fairli	1 if the (stemmed) word appears during the last interval	binary
249	billionair	1 if the (stemmed) word appears during the last interval	binary
250	proxi	1 if the (stemmed) word appears during the last interval	binary
251	incur	1 if the (stemmed) word appears during the last interval	binary
252	foremost	1 if the (stemmed) word appears during the last interval	binary
253	buyer	1 if the (stemmed) word appears during the last interval	binary
254	wholli	1 if the (stemmed) word appears during the last interval	binary
255	secret	1 if the (stemmed) word appears during the last interval	binary
256	conflict	1 if the (stemmed) word appears during the last interval	binary
257	overshadow	1 if the (stemmed) word appears during the last interval	binary
258	undergo	1 if the (stemmed) word appears during the last interval	binary
259	intens	1 if the (stemmed) word appears during the last interval	binary
260	commend	1 if the (stemmed) word appears during the last interval	binary
261	lacklust	1 if the (stemmed) word appears during the last interval	binary
262	develop	1 if the (stemmed) word appears during the last interval	binary
263	variat	1 if the (stemmed) word appears during the last interval	binary
264	claim	1 if the (stemmed) word appears during the last	binary

		interval	
265	notifi	1 if the (stemmed) word appears during the last interval	binary
266	sue	1 if the (stemmed) word appears during the last interval	binary
267	violat	1 if the (stemmed) word appears during the last interval	binary
268	boardroom	1 if the (stemmed) word appears during the last interval	binary
269	testifi	1 if the (stemmed) word appears during the last interval	binary
270	volunt	1 if the (stemmed) word appears during the last interval	binary
271	farm	1 if the (stemmed) word appears during the last interval	binary
272	turbin	1 if the (stemmed) word appears during the last interval	binary
273	leap	1 if the (stemmed) word appears during the last interval	binary
274	exhibit	1 if the (stemmed) word appears during the last interval	binary
275	begin	1 if the (stemmed) word appears during the last interval	binary
276	hamper	1 if the (stemmed) word appears during the last interval	binary
277	cmo	1 if the (stemmed) word appears during the last interval	binary
278	absent	1 if the (stemmed) word appears during the last interval	binary
279	lethal	1 if the (stemmed) word appears during the last interval	binary
280	deni	1 if the (stemmed) word appears during the last interval	binary
281	underpin	1 if the (stemmed) word appears during the last interval	binary
282	silver	1 if the (stemmed) word appears during the last interval	binary
283	evolut	1 if the (stemmed) word appears during the last interval	binary
284	exploit	1 if the (stemmed) word appears during the last interval	binary
285	assist	1 if the (stemmed) word appears during the last interval	binary
286	decreas	1 if the (stemmed) word appears during the last interval	binary
287	monei	1 if the (stemmed) word appears during the last interval	binary
288	brokerag	1 if the (stemmed) word appears during the last interval	binary
289	appli	1 if the (stemmed) word appears during the last interval	binary
290	possibl	1 if the (stemmed) word appears during the last interval	binary
291	labor	1 if the (stemmed) word appears during the last	binary

		interval	
292	verifi	1 if the (stemmed) word appears during the last interval	binary
293	visibl	1 if the (stemmed) word appears during the last interval	binary
294	simul	1 if the (stemmed) word appears during the last interval	binary
295	centr	1 if the (stemmed) word appears during the last interval	binary
296	defect	1 if the (stemmed) word appears during the last interval	binary
297	ultra	1 if the (stemmed) word appears during the last interval	binary
298	injunct	1 if the (stemmed) word appears during the last interval	binary
299	shortli	1 if the (stemmed) word appears during the last interval	binary
300	strongli	1 if the (stemmed) word appears during the last interval	binary
301	precis	1 if the (stemmed) word appears during the last interval	binary
302	pocket	1 if the (stemmed) word appears during the last interval	binary
303	eager	1 if the (stemmed) word appears during the last interval	binary
304	figur	1 if the (stemmed) word appears during the last interval	binary
305	contest	1 if the (stemmed) word appears during the last interval	binary
306	competitor	1 if the (stemmed) word appears during the last interval	binary
307	coach	1 if the (stemmed) word appears during the last interval	binary
308	reloc	1 if the (stemmed) word appears during the last interval	binary
309	motiv	1 if the (stemmed) word appears during the last interval	binary
310	sharpli	1 if the (stemmed) word appears during the last interval	binary
311	ad	1 if the (stemmed) word appears during the last interval	binary
312	mark	1 if the (stemmed) word appears during the last interval	binary
313	abov	1 if the (stemmed) word appears during the last interval	binary
314	down	1 if the (stemmed) word appears during the last interval	binary
315	fluctuat	1 if the (stemmed) word appears during the last interval	binary
316	underscor	1 if the (stemmed) word appears during the last interval	binary
317	efficaci	1 if the (stemmed) word appears during the last interval	binary
318	hurd	1 if the (stemmed) word appears during the last	binary

		interval	
319	pentagon	1 if the (stemmed) word appears during the last interval	binary
320	magic	1 if the (stemmed) word appears during the last interval	binary
321	ebitda	1 if the (stemmed) word appears during the last interval	binary
322	peripher	1 if the (stemmed) word appears during the last interval	binary
323	spur	1 if the (stemmed) word appears during the last interval	binary
324	royc	1 if the (stemmed) word appears during the last interval	binary
325	stabl	1 if the (stemmed) word appears during the last interval	binary
326	face	1 if the (stemmed) word appears during the last interval	binary
327	viabl	1 if the (stemmed) word appears during the last interval	binary
328	rigor	1 if the (stemmed) word appears during the last interval	binary
329	hummer	1 if the (stemmed) word appears during the last interval	binary
330	incent	1 if the (stemmed) word appears during the last interval	binary
331	prize	1 if the (stemmed) word appears during the last interval	binary
332	intent	1 if the (stemmed) word appears during the last interval	binary
333	soft	1 if the (stemmed) word appears during the last interval	binary
334	tape	1 if the (stemmed) word appears during the last interval	binary
335	mobil	1 if the (stemmed) word appears during the last interval	binary
336	sever	1 if the (stemmed) word appears during the last interval	binary
337	food	1 if the (stemmed) word appears during the last interval	binary
338	chequ	1 if the (stemmed) word appears during the last interval	binary
339	payless	1 if the (stemmed) word appears during the last interval	binary
340	symptomat	1 if the (stemmed) word appears during the last interval	binary
341	buffer	1 if the (stemmed) word appears during the last interval	binary
342	dyer	1 if the (stemmed) word appears during the last interval	binary
343	lessor	1 if the (stemmed) word appears during the last interval	binary
344	iec	1 if the (stemmed) word appears during the last interval	binary
345	behaviour	1 if the (stemmed) word appears during the last	binary

		interval	
346	narrat	1 if the (stemmed) word appears during the last interval	binary
347	beasystem	1 if the (stemmed) word appears during the last interval	binary
348	ata	1 if the (stemmed) word appears during the last interval	binary
349	life	1 if the (stemmed) word appears during the last interval	binary
350	donat	1 if the (stemmed) word appears during the last interval	binary
351	pilot	1 if the (stemmed) word appears during the last interval	binary
352	issuer	1 if the (stemmed) word appears during the last interval	binary
353	opinion	1 if the (stemmed) word appears during the last interval	binary
354	period	1 if the (stemmed) word appears during the last interval	binary
355	month	1 if the (stemmed) word appears during the last interval	binary
356	resort	1 if the (stemmed) word appears during the last interval	binary
357	complaint	1 if the (stemmed) word appears during the last interval	binary
358	migrat	1 if the (stemmed) word appears during the last interval	binary
359	master	1 if the (stemmed) word appears during the last interval	binary
360	transpar	1 if the (stemmed) word appears during the last interval	binary
361	workplac	1 if the (stemmed) word appears during the last interval	binary
362	rumor	1 if the (stemmed) word appears during the last interval	binary
363	crisi	1 if the (stemmed) word appears during the last interval	binary
364	enemi	1 if the (stemmed) word appears during the last interval	binary
365	pave	1 if the (stemmed) word appears during the last interval	binary
366	time	1 if the (stemmed) word appears during the last interval	binary
367	suppli	1 if the (stemmed) word appears during the last interval	binary
368	telephon	1 if the (stemmed) word appears during the last interval	binary
369	unless	1 if the (stemmed) word appears during the last interval	binary
370	slower	1 if the (stemmed) word appears during the last interval	binary
371	aid	1 if the (stemmed) word appears during the last interval	binary
372	brought	1 if the (stemmed) word appears during the last	binary

		interval	
373	declar	1 if the (stemmed) word appears during the last interval	binary
374	search	1 if the (stemmed) word appears during the last interval	binary
375	bonu	1 if the (stemmed) word appears during the last interval	binary
376	municip	1 if the (stemmed) word appears during the last interval	binary
377	frequenc	1 if the (stemmed) word appears during the last interval	binary
378	repair	1 if the (stemmed) word appears during the last interval	binary
379	burden	1 if the (stemmed) word appears during the last interval	binary
380	victori	1 if the (stemmed) word appears during the last interval	binary
381	agenda	1 if the (stemmed) word appears during the last interval	binary
382	quiet	1 if the (stemmed) word appears during the last interval	binary
383	brain	1 if the (stemmed) word appears during the last interval	binary
384	gui	1 if the (stemmed) word appears during the last interval	binary
385	magnitud	1 if the (stemmed) word appears during the last interval	binary
386	thrust	1 if the (stemmed) word appears during the last interval	binary
387	premis	1 if the (stemmed) word appears during the last interval	binary
388	invalid	1 if the (stemmed) word appears during the last interval	binary
389	annual	1 if the (stemmed) word appears during the last interval	binary
390	billion	1 if the (stemmed) word appears during the last interval	binary
391	relianc	1 if the (stemmed) word appears during the last interval	binary
392	beyond	1 if the (stemmed) word appears during the last interval	binary
393	basic	1 if the (stemmed) word appears during the last interval	binary
394	output	1 if the (stemmed) word appears during the last interval	binary
395	sole	1 if the (stemmed) word appears during the last interval	binary
396	lift	1 if the (stemmed) word appears during the last interval	binary
397	specul	1 if the (stemmed) word appears during the last interval	binary
398	long	1 if the (stemmed) word appears during the last interval	binary
399	allianc	1 if the (stemmed) word appears during the last	binary

		interval	
400	promis	1 if the (stemmed) word appears during the last interval	binary
401	larg	1 if the (stemmed) word appears during the last interval	binary
402	demand	1 if the (stemmed) word appears during the last interval	binary
403	proud	1 if the (stemmed) word appears during the last interval	binary
404	symptom	1 if the (stemmed) word appears during the last interval	binary
405	ethic	1 if the (stemmed) word appears during the last interval	binary
406	perspect	1 if the (stemmed) word appears during the last interval	binary
407	leagu	1 if the (stemmed) word appears during the last interval	binary
408	passion	1 if the (stemmed) word appears during the last interval	binary
409	newer	1 if the (stemmed) word appears during the last interval	binary
410	fundament	1 if the (stemmed) word appears during the last interval	binary
411	luxuri	1 if the (stemmed) word appears during the last interval	binary
412	former	1 if the (stemmed) word appears during the last interval	binary
413	measur	1 if the (stemmed) word appears during the last interval	binary
414	pack	1 if the (stemmed) word appears during the last interval	binary
415	resid	1 if the (stemmed) word appears during the last interval	binary
416	retail	1 if the (stemmed) word appears during the last interval	binary
417	fell	1 if the (stemmed) word appears during the last interval	binary
418	carefulli	1 if the (stemmed) word appears during the last interval	binary
419	creator	1 if the (stemmed) word appears during the last interval	binary
420	turnov	1 if the (stemmed) word appears during the last interval	binary
421	backbon	1 if the (stemmed) word appears during the last interval	binary
422	exclusion	1 if the (stemmed) word appears during the last interval	binary
423	imposs	1 if the (stemmed) word appears during the last interval	binary
424	stir	1 if the (stemmed) word appears during the last interval	binary
425	congratul	1 if the (stemmed) word appears during the last interval	binary
426	customari	1 if the (stemmed) word appears during the last	binary

		interval	
427	sanction	1 if the (stemmed) word appears during the last interval	binary
428	mirror	1 if the (stemmed) word appears during the last interval	binary
429	accredit	1 if the (stemmed) word appears during the last interval	binary
430	prosum	1 if the (stemmed) word appears during the last interval	binary
431	relai	1 if the (stemmed) word appears during the last interval	binary
432	overtak	1 if the (stemmed) word appears during the last interval	binary
433	account	1 if the (stemmed) word appears during the last interval	binary
434	immun	1 if the (stemmed) word appears during the last interval	binary
435	monopoli	1 if the (stemmed) word appears during the last interval	binary
436	congression	1 if the (stemmed) word appears during the last interval	binary
437	propel	1 if the (stemmed) word appears during the last interval	binary
438	kill	1 if the (stemmed) word appears during the last interval	binary
439	breadth	1 if the (stemmed) word appears during the last interval	binary
440	exactli	1 if the (stemmed) word appears during the last interval	binary
441	chose	1 if the (stemmed) word appears during the last interval	binary
442	unparallel	1 if the (stemmed) word appears during the last interval	binary
443	happi	1 if the (stemmed) word appears during the last interval	binary
444	notabl	1 if the (stemmed) word appears during the last interval	binary
445	thermal	1 if the (stemmed) word appears during the last interval	binary
446	genet	1 if the (stemmed) word appears during the last interval	binary
447	regularli	1 if the (stemmed) word appears during the last interval	binary
448	likelihood	1 if the (stemmed) word appears during the last interval	binary
449	ignor	1 if the (stemmed) word appears during the last interval	binary
450	cantor	1 if the (stemmed) word appears during the last interval	binary
451	instrument	1 if the (stemmed) word appears during the last interval	binary
452	closest	1 if the (stemmed) word appears during the last interval	binary
453	lighter	1 if the (stemmed) word appears during the last	binary

		interval	
454	invest	1 if the (stemmed) word appears during the last interval	binary
455	accur	1 if the (stemmed) word appears during the last interval	binary
456	soar	1 if the (stemmed) word appears during the last interval	binary
457	proactiv	1 if the (stemmed) word appears during the last interval	binary
458	nonprofit	1 if the (stemmed) word appears during the last interval	binary
459	insid	1 if the (stemmed) word appears during the last interval	binary
460	statem	1 if the (stemmed) word appears during the last interval	binary
461	worri	1 if the (stemmed) word appears during the last interval	binary
462	liquid	1 if the (stemmed) word appears during the last interval	binary
463	drive	1 if the (stemmed) word appears during the last interval	binary
464	rais	1 if the (stemmed) word appears during the last interval	binary
465	record	1 if the (stemmed) word appears during the last interval	binary
466	harbor	1 if the (stemmed) word appears during the last interval	binary
467	delta	1 if the (stemmed) word appears during the last interval	binary
468	lure	1 if the (stemmed) word appears during the last interval	binary
469	basi	1 if the (stemmed) word appears during the last interval	binary
470	assumpt	1 if the (stemmed) word appears during the last interval	binary
471	maxim	1 if the (stemmed) word appears during the last interval	binary
472	war	1 if the (stemmed) word appears during the last interval	binary
473	arriv	1 if the (stemmed) word appears during the last interval	binary
474	bet	1 if the (stemmed) word appears during the last interval	binary
475	risen	1 if the (stemmed) word appears during the last interval	binary
476	step	1 if the (stemmed) word appears during the last interval	binary
477	special	1 if the (stemmed) word appears during the last interval	binary
478	pretext	1 if the (stemmed) word appears during the last interval	binary
479	induc	1 if the (stemmed) word appears during the last interval	binary
480	protest	1 if the (stemmed) word appears during the last	binary

		interval	
481	yesterdai	1 if the (stemmed) word appears during the last interval	binary
482	compos	1 if the (stemmed) word appears during the last interval	binary
483	position	1 if the (stemmed) word appears during the last interval	binary
484	cut	1 if the (stemmed) word appears during the last interval	binary
485	predict	1 if the (stemmed) word appears during the last interval	binary
486	innov	1 if the (stemmed) word appears during the last interval	binary
487	servic	1 if the (stemmed) word appears during the last interval	binary
488	test	1 if the (stemmed) word appears during the last interval	binary
489	histor	1 if the (stemmed) word appears during the last interval	binary
490	track	1 if the (stemmed) word appears during the last interval	binary
491	logist	1 if the (stemmed) word appears during the last interval	binary
492	strateg	1 if the (stemmed) word appears during the last interval	binary
493	actual	1 if the (stemmed) word appears during the last interval	binary
494	outcom	1 if the (stemmed) word appears during the last interval	binary
495	clean	1 if the (stemmed) word appears during the last interval	binary
496	varianc	1 if the (stemmed) word appears during the last interval	binary
497	deliber	1 if the (stemmed) word appears during the last interval	binary
498	compliant	1 if the (stemmed) word appears during the last interval	binary
499	simultan	1 if the (stemmed) word appears during the last interval	binary
500	rescu	1 if the (stemmed) word appears during the last interval	binary
501	drove	1 if the (stemmed) word appears during the last interval	binary
502	occasion	1 if the (stemmed) word appears during the last interval	binary
503	verdict	1 if the (stemmed) word appears during the last interval	binary
504	underperform	1 if the (stemmed) word appears during the last interval	binary
505	contributor	1 if the (stemmed) word appears during the last interval	binary
506	defenc	1 if the (stemmed) word appears during the last interval	binary
507	boast	1 if the (stemmed) word appears during the last	binary

		interval	
508	retriev	1 if the (stemmed) word appears during the last interval	binary
509	opposit	1 if the (stemmed) word appears during the last interval	binary
510	intermedi	1 if the (stemmed) word appears during the last interval	binary
511	altitud	1 if the (stemmed) word appears during the last interval	binary
512	stood	1 if the (stemmed) word appears during the last interval	binary
513	reap	1 if the (stemmed) word appears during the last interval	binary
514	predecessor	1 if the (stemmed) word appears during the last interval	binary
515	contend	1 if the (stemmed) word appears during the last interval	binary
516	repeatedli	1 if the (stemmed) word appears during the last interval	binary
517	remind	1 if the (stemmed) word appears during the last interval	binary
518	vigor	1 if the (stemmed) word appears during the last interval	binary
519	twist	1 if the (stemmed) word appears during the last interval	binary
520	activ	1 if the (stemmed) word appears during the last interval	binary
521	neutral	1 if the (stemmed) word appears during the last interval	binary
522	util	1 if the (stemmed) word appears during the last interval	binary
523	sentim	1 if the (stemmed) word appears during the last interval	binary
524	failur	1 if the (stemmed) word appears during the last interval	binary
525	battl	1 if the (stemmed) word appears during the last interval	binary
526	economi	1 if the (stemmed) word appears during the last interval	binary
527	resist	1 if the (stemmed) word appears during the last interval	binary
528	slightli	1 if the (stemmed) word appears during the last interval	binary
529	locat	1 if the (stemmed) word appears during the last interval	binary
530	along	1 if the (stemmed) word appears during the last interval	binary
531	refer	1 if the (stemmed) word appears during the last interval	binary
532	benefit	1 if the (stemmed) word appears during the last interval	binary
533	control	1 if the (stemmed) word appears during the last interval	binary
534	caus	1 if the (stemmed) word appears during the last	binary

		interval	
535	secur	1 if the (stemmed) word appears during the last interval	binary
536	expens	1 if the (stemmed) word appears during the last interval	binary
537	Devot	1 if the (stemmed) word appears during the last interval	binary
538	Anyon	1 if the (stemmed) word appears during the last interval	binary
539	Choos	1 if the (stemmed) word appears during the last interval	binary
540	Try	1 if the (stemmed) word appears during the last interval	binary
541	Believ	1 if the (stemmed) word appears during the last interval	binary
542	Model	1 if the (stemmed) word appears during the last interval	binary
543	produc	1 if the (stemmed) word appears during the last interval	binary
544	interfer	1 if the (stemmed) word appears during the last interval	binary
545	Wear	1 if the (stemmed) word appears during the last interval	binary
546	Licenc	1 if the (stemmed) word appears during the last interval	binary
547	Advoc	1 if the (stemmed) word appears during the last interval	binary
548	Steep	1 if the (stemmed) word appears during the last interval	binary
549	Wrong	1 if the (stemmed) word appears during the last interval	binary
550	Stack	1 if the (stemmed) word appears during the last interval	binary
551	aftermarket	1 if the (stemmed) word appears during the last interval	binary
552	represent	1 if the (stemmed) word appears during the last interval	binary
553	Justify	1 if the (stemmed) word appears during the last interval	binary
554	Vacat	1 if the (stemmed) word appears during the last interval	binary
555	controversi	1 if the (stemmed) word appears during the last interval	binary
556	Elabor	1 if the (stemmed) word appears during the last interval	binary
557	acclaim	1 if the (stemmed) word appears during the last interval	binary
558	Upset	1 if the (stemmed) word appears during the last interval	binary
559	Bounc	1 if the (stemmed) word appears during the last interval	binary
560	Cfa	1 if the (stemmed) word appears during the last interval	binary
561	Afraid	1 if the (stemmed) word appears during the last	binary

		interval	

2.7 Bag of Words - Dual Adjacent (stemmed) Words

Variables (stemmed, 2 adjacent words) were selected according to the top ranked, 100 variables based on the Information Gain criteria.

Original, full variable set, from which these 100 variables were selected, included all possible (stemmed) 2 adjacent words available in news items published during the first 3 months of available data. This time period corresponds to the time period of our first, sliding window, training set.

Note: replacecompanynam -> replaces a company name

#	Field	Description	Type
562	previous_report	1 if the dual adjacent (stemmed) words appears during the last interval	binary
563	top_gainer	1 if the dual adjacent (stemmed) words appears during the last interval	binary
564	sell_prescript	1 if the dual adjacent (stemmed) words appears during the last interval	binary
565	contact_with	1 if the dual adjacent (stemmed) words appears during the last interval	binary
566	asset_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	binary
567	report_said	1 if the dual adjacent (stemmed) words appears during the last interval	binary
568	oil_futur	1 if the dual adjacent (stemmed) words appears during the last interval	binary
569	replacecompanynam_challeng	1 if the dual adjacent (stemmed) words appears during the last interval	binary
570	replacecompanynam_continu	1 if the dual adjacent (stemmed) words appears during the last interval	binary
571	fell_after	1 if the dual adjacent (stemmed) words appears during the last interval	binary
572	rais_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	binary
573	loss_from	1 if the dual adjacent (stemmed) words appears during the last interval	binary
574	it_revenu	1 if the dual adjacent (stemmed) words appears during the last interval	binary
575	can_lead	1 if the dual adjacent (stemmed) words appears during the last interval	binary
576	it_rose	1 if the dual adjacent (stemmed) words appears during the last interval	binary
577	busi_that	1 if the dual adjacent (stemmed) words appears during the last interval	binary
578	which_thei	1 if the dual adjacent (stemmed) words appears during the last interval	binary

579	share_jump	1 if the dual adjacent (stemmed) words appears during the last interval	binary
580	presid_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	binary
581	stock_rise	1 if the dual adjacent (stemmed) words appears during the last interval	binary
582	incom_tax	1 if the dual adjacent (stemmed) words appears during the last interval	binary
583	would_cut	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
584	oper_net	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
585	relat_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
586	cash_equival	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
587	incom_loss	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
588	program_detail	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
589	stock_purchas	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
590	quarterli_report	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
591	lead_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
592	stock_base	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
593	report_net	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
594	from_outperform	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
595	oper_system	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
596	team_with	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
597	benefit_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
598	provid_network	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
599	enabl_custom	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
600	product_develop	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
601	replacecompanynam_comput	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
602	replacecompanynam_advanc	1 if the dual adjacent (stemmed) words appears during the last interval	Binary

603	of_record	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
604	profit_from	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
605	win_new	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
606	board_member	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
607	of_competit	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
608	state_univers	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
609	invest_firm	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
610	after_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
611	high_at	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
612	replacecompanynam_patient	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
613	replacecompanynam_commun	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
614	which_provid	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
615	faster_than	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
616	base_compens	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
617	decreas_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
618	replacebrokeragenam_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
619	post_lower	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
620	replacecompanynam_establish	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
621	auto_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
622	quarter_sale	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
623	more_of	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
624	presid_ceo	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
625	cut_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
626	expens_net	1 if the dual adjacent (stemmed) words appears during the last interval	Binary

627	exclud_restructur	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
628	price_by	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
629	it_stock	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
630	product_will	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
631	annual_report	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
632	high_qualiti	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
633	dividend_of	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
634	includ_new	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
635	back_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
636	servic_of	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
637	percent_year	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
638	rank_no	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
639	chairman_of	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
640	with_percent	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
641	more_close	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
642	replacecompanynam_late	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
643	differ_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
644	replacecompanynam_profit	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
645	compet_with	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
646	inform_on	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
647	show_that	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
648	replacecompanynam_chairman	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
649	allianc_with	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
650	whose_share	1 if the dual adjacent (stemmed) words appears during the last interval	Binary

651	replacecompanynam_found	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
652	respons_team	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
653	between_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
654	technologi_product	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
655	increas_replacecompanynam	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
656	kick_off	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
657	market_with	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
658	coupl_with	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
659	develop_technologi	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
660	court_of	1 if the dual adjacent (stemmed) words appears during the last interval	Binary
661	order_aircraft	1 if the dual adjacent (stemmed) words appears during the last interval	Binary

2.8 Categories

Classification into the different categories is based on classifiers described in section 2.6 of the paper.

#	Field	Description	Type
662	Repeats Market Activity_cur	Repeats Market Activity- count items during last interval	integer
663	Analyst Upgrade_cur	Analyst Upgrade- count items during last interval	integer
664	Lawsuit_cur	Lawsuit- count items during last interval	integer
665	Regulatory Company Statements_cur	Regulatory Company Statements- count items during last interval	integer
666	Regulatory Activity_cur	Regulatory Activity- count items during last interval	integer
667	Order Imbalance buy_cur	Order Imbalance buy- count items during last interval	integer
668	Order Imbalance sell_cur	Order Imbalance sell- count items during last interval	integer
669	Analyst Downgrade_cur	Analyst Downgrade- count items during last interval	integer
670	Acquisition_cur	Acquisition- count items during last interval	integer
671	8k Filing_cur	8k Filing- count items during last interval	integer
672	Performance Outlook_cur	Performance Outlook- count items during last interval	integer
673	Earning or Revenues_cur	Earning or Revenues- count items during last interval	integer
674	Sales and Sales Contracts_cur	Sales and Sales Contracts- count items during last interval	integer
675	Executive HR issues_cur	Executive HR issues- count items during last interval	integer
676	Product or Service Announcement_cur	Product or Service Announcement- count items during last interval	integer
677	Joint Venture/Collaboration_cur	Joint Venture/Collaboration- count items during last interval	integer
678	Production and Shipping Issues_cur	Production and Shipping Issues- count items during last interval	integer
679	Repeats Market Activity_prev1	Repeats Market Activity- count items during 1 previous interval	integer
680	Analyst Upgrade_prev1	Analyst Upgrade- count items during 1 previous interval	integer
681	Lawsuit_prev1	Lawsuit- count items during 1 previous interval	integer

682	Regulatory Company Statements_prev1	Regulatory Company Statements- count items during 1 previous interval	integer
683	Regulatory Activity_prev1	Regulatory Activity- count items during 1 previous interval	integer
684	Order Imbalance buy_prev1	Order Imbalance buy- count items during 1 previous interval	integer
685	Order Imbalance sell_prev1	Order Imbalance sell- count items during 1 previous interval	integer
686	Analyst Downgrade_prev1	Analyst Downgrade- count items during 1 previous interval	integer
687	Acquisition_prev1	Acquisition- count items during 1 previous interval	integer
688	8k Filing_prev1	8k Filing- count items during 1 previous interval	integer
689	Performance Outlook_prev1	Performance Outlook- count items during 1 previous interval	integer
690	Earning or Revenues_prev1	Earning or Revenues- count items during 1 previous interval	integer
691	Sales and Sales Contracts_prev1	Sales and Sales Contracts- count items during 1 previous interval	integer
692	Executive HR issues_prev1	Executive HR issues- count items during 1 previous interval	integer
693	Product or Service Announcement_prev1	Product or Service Announcement- count items during 1 previous interval	integer
694	Joint Venture/Collaboration_prev1	Joint Venture/Collaboration- count items during 1 previous interval	integer
695	Production and Shipping Issues_prev1	Production and Shipping Issues- count items during 1 previous interval	integer
696	Repeats Market Activity_prev2	Repeats Market Activity- count items during 2 intervals before the current interval	integer
697	Analyst Upgrade_prev2	Analyst Upgrade- count items during 2 intervals before the current interval	integer
698	Lawsuit_prev2	Lawsuit- count items during 2 intervals before the current interval	integer
699	Regulatory Company Statements_prev2	Regulatory Company Statements- count items during 2 intervals before the current interval	integer
700	Regulatory Activity_prev2	Regulatory Activity- count items during 2 intervals before the current interval	integer
701	Order Imbalance buy_prev2	Order Imbalance buy- count items during 2 intervals before the current interval	integer
702	Order Imbalance sell_prev2	Order Imbalance sell- count items during 2 intervals before the current interval	integer
703	Analyst Downgrade_prev2	Analyst Downgrade- count items during 2 intervals before the current interval	integer
704	Acquisition_prev2	Acquisition- count items during 2 intervals before the current interval	integer
705	8k Filing_prev2	8k Filing- count items during 2 intervals before the current interval	integer

706	Performance Outlook_prev2	Performance Outlook- count items during 2 intervals before the current interval	integer
707	Earning or Revenues_prev2	Earning or Revenues- count items during 2 intervals before the current interval	integer
708	Sales and Sales Contracts_prev2	Sales and Sales Contracts- count items during 2 intervals before the current interval	integer
709	Executive HR issues_prev2	Executive HR issues- count items during 2 intervals before the current interval	integer
710	Product or Service Announcement_prev2	Product or Service Announcement- count items during 2 intervals before the current interval	integer
711	Joint Venture/Collaboration_prev2	Joint Venture/Collaboration- count items during 2 intervals before the current interval	integer
712	Production and Shipping Issues_prev2	Production and Shipping Issues- count items during 2 intervals before the current interval	integer
713	Repeats Market Activity_prev3	Repeats Market Activity- count items during 3 intervals before the current interval	integer
714	Analyst Upgrade_prev3	Analyst Upgrade- count items during 3 intervals before the current interval	integer
715	Lawsuit_prev3	Lawsuit- count items during 3 intervals before the current interval	integer
716	Regulatory Company Statements_prev3	Regulatory Company Statements- count items during 3 intervals before the current interval	integer
717	Regulatory Activity_prev3	Regulatory Activity- count items during 3 intervals before the current interval	integer
718	Order Imbalance buy_prev3	Order Imbalance buy- count items during 3 intervals before the current interval	integer
719	Order Imbalance sell_prev3	Order Imbalance sell- count items during 3 intervals before the current interval	integer
720	Analyst Downgrade_prev3	Analyst Downgrade- count items during 3 intervals before the current interval	integer
721	Acquisition_prev3	Acquisition- count items during 3 intervals before the current interval	integer
722	8k Filing_prev3	8k Filing- count items during 3 intervals before the current interval	integer
723	Performance Outlook_prev3	Performance Outlook- count items during 3 intervals before the current interval	integer
724	Earning or Revenues_prev3	Earning or Revenues- count items during 3 intervals before the current interval	integer
725	Sales and Sales Contracts_prev3	Sales and Sales Contracts- count items during 3 intervals before the current interval	integer
726	Executive HR issues_prev3	Executive HR issues- count items during 3 intervals before the current interval	integer
727	Product or Service Announcement_prev3	Product or Service Announcement- count items during 3 intervals before the current interval	integer
728	Joint Venture/Collaboration_prev3	Joint Venture/Collaboration- count items during 3 intervals before the current interval	integer
729	Production and Shipping Issues_prev3	Production and Shipping Issues- count items during 3 intervals before the current interval	integer

730	Repeats Market Activity_prev4	Repeats Market Activity- count items during 4 intervals before the current interval	integer
731	Analyst Upgrade_prev4	Analyst Upgrade- count items during 4 intervals before the current interval	integer
732	Lawsuit_prev4	Lawsuit- count items during 4 intervals before the current interval	integer
733	Regulatory Company Statements_prev4	Regulatory Company Statements- count items during 4 intervals before the current interval	integer
734	Regulatory Activity_prev4	Regulatory Activity- count items during 4 intervals before the current interval	integer
735	Order Imbalance buy_prev4	Order Imbalance buy- count items during 4 intervals before the current interval	integer
736	Order Imbalance sell_prev4	Order Imbalance sell- count items during 4 intervals before the current interval	integer
737	Analyst Downgrade_prev4	Analyst Downgrade- count items during 4 intervals before the current interval	integer
738	Acquisition_prev4	Acquisition- count items during 4 intervals before the current interval	integer
739	8k Filing_prev4	8k Filing- count items during 4 intervals before the current interval	integer
740	Performance Outlook_prev4	Performance Outlook- count items during 4 intervals before the current interval	integer
741	Earning or Revenues_prev4	Earning or Revenues- count items during 4 intervals before the current interval	integer
742	Sales and Sales Contracts_prev4	Sales and Sales Contracts- count items during 4 intervals before the current interval	integer
743	Executive HR issues_prev4	Executive HR issues- count items during 4 intervals before the current interval	integer
744	Product or Service Announcement_prev4	Product or Service Announcement- count items during 4 intervals before the current interval	integer
745	Joint Venture/Collaboration_prev4	Joint Venture/Collaboration- count items during 4 intervals before the current interval	integer
746	Production and Shipping Issues_prev4	Production and Shipping Issues- count items during 4 intervals before the current interval	integer
747	Repeats Market Activity_trading_day_start_to_current_interval	Repeats Market Activity- count items from trading day start to current interval	integer
748	Analyst Upgrade_trading_day_start_to_current_interval	Analyst Upgrade- count items from trading day start to current interval	integer
749	Lawsuit_trading_day_start_to_current_interval	Lawsuit- count items from trading day start to current interval	integer
750	Regulatory Company Statements_trading_day_start_to_current_interval	Regulatory Company Statements- count items from trading day start to current interval	integer
751	Regulatory Activity_trading_day_start_to_current_interval	Regulatory Activity- count items from trading day start to current interval	integer
752	Order Imbalance buy_trading_day_start_to_current_interval	Order Imbalance buy- count items from trading day start to current interval	integer

753	Order Imbalance sell_trading_day_start_to_current_interval	Order Imbalance sell- count items from trading day start to current interval	integer
754	Analyst Downgrade_trading_day_start_to_current_interval	Analyst Downgrade- count items from trading day start to current interval	integer
755	Acquisition_trading_day_start_to_current_interval	Acquisition- count items from trading day start to current interval	integer
756	8k Filing_trading_day_start_to_current_interval	8k Filing- count items from trading day start to current interval	integer
757	Performance Outlook_trading_day_start_to_current_interval	Performance Outlook- count items from trading day start to current interval	integer
758	Earning or Revenues_trading_day_start_to_current_interval	Earning or Revenues- count items from trading day start to current interval	integer
759	Sales and Sales Contracts_trading_day_start_to_current_interval	Sales and Sales Contracts- count items from trading day start to current interval	integer
760	Executive HR issues_trading_day_start_to_current_interval	Executive HR issues- count items from trading day start to current interval	integer
761	Product or Service Announcement_trading_day_start_to_current_interval	Product or Service Announcement- count items from trading day start to current interval	integer
762	Joint Venture/Collaboration_trading_day_start_to_current_interval	Joint Venture/Collaboration- count items from trading day start to current interval	integer
763	Production and Shipping Issues_trading_day_start_to_current_interval	Production and Shipping Issues- count items from trading day start to current interval	integer
764	Repeats Market Activity_end_trading_day_to_next_day_start	Repeats Market Activity- count items from the end of previous trading day to current trading day	integer
765	Analyst Upgrade_end_trading_day_to_next_day_start	Analyst Upgrade- count items from the end of previous trading day to current trading day	integer
766	Lawsuit_end_trading_day_to_next_day_start	Lawsuit- count items from the end of previous trading day to current trading day	integer
767	Regulatory Company Statements_end_trading_day_to_next_day_start	Regulatory Company Statements- count items from the end of previous trading day to current trading day	integer
768	Regulatory Activity_end_trading_day_to_next_day_start	Regulatory Activity- count items from the end of previous trading day to current trading day	integer
769	Order Imbalance buy_end_trading_day_to_next_day_start	Order Imbalance buy- count items from the end of previous trading day to current trading day	integer
770	Order Imbalance sell_end_trading_day_to_next_day_start	Order Imbalance sell- count items from the end of previous trading day to current trading day	integer
771	Analyst Downgrade_end_trading_day_to_next_day_start	Analyst Downgrade- count items from the end of previous trading day to current trading day	integer
772	Acquisition_end_trading_day_to_next_day_start	Acquisition- count items from the end of previous trading day to current trading day	integer

773	8k Filing_end_trading_day_to_next_day_start	8k Filing- count items from the end of previous trading day to current trading day	integer
774	Performance Outlook_end_trading_day_to_next_day_start	Performance Outlook- count items from the end of previous trading day to current trading day	integer
775	Earning or Revenues_end_trading_day_to_next_day_start	Earning or Revenues- count items from the end of previous trading day to current trading day	integer
776	Sales and Sales Contracts_end_trading_day_to_next_day_start	Sales and Sales Contracts- count items from the end of previous trading day to current trading day	integer
777	Executive HR issues_end_trading_day_to_next_day_start	Executive HR issues- count items from the end of previous trading day to current trading day	integer
778	Product or Service Announcement_end_trading_day_to_next_day_start	Product or Service Announcement- count items from the end of previous trading day to current trading day	integer
779	Joint Venture/Collaboration_end_trading_day_to_next_day_start	Joint Venture/Collaboration- count items from the end of previous trading day to current trading day	integer
780	Production and Shipping Issues_end_trading_day_to_next_day_start	Production and Shipping Issues- count items from the end of previous trading day to current trading day	integer

2.9 Sentiment Scores

#	Field	Description	Type
781	INTRVAL_AV_SENT	average sentiment during last interval	decimal
782	PRV1_Av_Snt	average sentiment during the previous interval	decimal
783	PRV2_Av_Snt	average sentiment during 2 intervals before the current interval	decimal
784	PRV3_Av_Snt	average sentiment during 3 intervals before the current interval	decimal
785	PRV4_Av_Snt	average sentiment during 4 intervals before the current interval	decimal
786	DyStCuIntAvSe	average sentiment from day start	decimal
787	EnDyNxtDyStAvSe	average sentiment from the end of previous trading day to current trading	decimal
788	CmPRV1_Av_Snt	average sentiment during last 2 intervals	decimal
789	CmPRV2_Av_Snt	average sentiment during last 3 intervals	decimal
790	CmPRV3_Av_Snt	average sentiment during last 4 intervals	decimal
791	CmPRV4_Av_Snt	average sentiment during last 5 intervals	decimal

2.10 Calibrated Sentiment Scores

#	Field	Description	Type
794	INTRVAL_AV_CalibSENT	average calibrated sentiment during last interval	decimal
795	PRV1_Av_CalibSnt	average calibrated sentiment during the previous interval	decimal
796	PRV2_Av_CalibSnt	average calibrated sentiment during 2 intervals before the current interval	decimal
797	PRV3_Av_CalibSnt	average calibrated sentiment during 3 intervals before the current interval	decimal
798	PRV4_Av_CalibSnt	average calibrated sentiment during 4 intervals before the current interval	decimal
799	DyStCuIntAvCalibSe	average calibrated sentiment from day start	decimal
800	EnDyNxtDyStAvCalibSe	average calibrated sentiment from the end of previous trading day to curre	decimal
801	CmPRV1_Av_CalibSnt	average calibrated sentiment during last 2 intervals	decimal
802	CmPRV2_Av_CalibSnt	average calibrated sentiment during last 3 intervals	decimal
803	CmPRV3_Av_CalibSnt	average calibrated sentiment during last 4 intervals	decimal
804	CmPRV4_Av_CalibSnt	average calibrated sentiment during last 5 intervals	decimal

3. Initial Parameter Setup Evaluation

Since it was not apparent in advance which model parameters to use, we determined them based on an experimental study. This section details the results of various experimental setups which we used to select which setting to employ in the main research.

Several factors were evaluated:

1. Time intervals – the time interval of each data instance, (or alternatively, how much time elapses between each intraday prediction). We considered three options - 1 minute, 5 minutes, and 15 minutes intervals.
2. Prediction threshold for setting the dependent variable – whether 0.5% or 1.0% .
3. Prediction range: the amount of time we "hold" until we check whether stock returns exceed the prediction threshold, either 5 minutes, 15 minutes, 1 hour or until the end of the trading day.
4. Scoring method – whether we used single scoring mechanism (i.e., use a single "positive" algorithm predicting whether stock returns exceed a certain threshold) or a double scoring mechanism (see section 3)

The dataset for this experiment consisted of the first 3 months of data used in the main research, the first two months were used as a training set, and the remaining month as a validation set.

We used in this experimentation process the simple market data representation which was employed in the pilot study and the SLR forecasting algorithm (which is the dataset/algorithm combination that rendered one of the best results in pilot study). While confining ourselves to this data/algorithm combination could have potentially biased the main study results towards this combination, we believe

that the bias, if any, has a negligible effect. Indeed, our final modeling results show that other data representation/algorithm combinations noticeably outperformed this market data/SLR combination.

We present the results of the experiment in Table 3.1, 36 data setups in total. For each data setup we show simple statistics such as average returns, standard deviation, and number of trades recommended by the model. Due the multiple repetitions and our desire to keep this preliminary evaluation simple, we used simple performance measures similar to the ones used in the pilot study. We note that we ignored the transaction costs in this experiment.

We used standardized average returns (the ratio between average returns per trade and the standard deviation of the returns) as our primary performance measure. Two setups obtained the highest value of 0.17. From which we selected the one with the highest number of recommended trades (by a large difference) and better average returns. This yields the following parameters that we used later on in our study:

Time interval - 5 minutes

Prediction range – 1 hour

Prediction threshold – 1%

Double scoring mechanism

Table 3.1: Performance Evaluation for Determining Model Parameters.

Setup Number	Time Intervals	Prediction Range	Predicted threshold	Scoring Method	Average Returns Per Trade	Standard Deviation	Number of Trades	Average Returns / Standard Deviation
1	1 min	1 hour	Increase above 0.5%	Single	0.09%	0.99%	13945	0.09
2	1 min	1 hour	Increase above 0.5%	Double	0.14%	0.92%	5384	0.15
3	1 min	15 min	Increase above 0.5%	Single	-0.10%	1.08%	316	-0.09
4	1 min	15 min	Increase above 0.5%	Double	-0.02%	1.10%	220	-0.02
5	1 min	5 min	Increase above 0.5%	Single	-0.01%	0.60%	295	-0.02
6	1 min	5 min	Increase above 0.5%	Double	0.06%	0.57%	104	0.11
7	1 min	1 hour	Increase above 1%	Single	0.12%	1.05%	9997	0.11
8	1 min	1 hour	Increase above 1%	Double	0.15%	0.98%	4204	0.15
9	1 min	15 min	Increase above 1%	Single	-0.09%	0.98%	512	-0.09
10	1 min	15 min	Increase above 1%	Double	-0.05%	0.96%	441	-0.05
11	1 min	5 min	Increase above 1%	Single	0.01%	0.56%	236	0.02
12	1 min	5 min	Increase above 1%	Double	0.09%	0.52%	113	0.17
13	5 min	1 hour	Increase above 0.5%	Single	0.10%	1.02%	2522	0.10
14	5 min	1 hour	Increase above 0.5%	Double	0.14%	0.93%	1183	0.15
15	5 min	15 min	Increase above 0.5%	Single	-0.09%	0.90%	147	-0.10
16	5 min	15 min	Increase above 0.5%	Double	0.10%	0.85%	83	0.12
17	5 min	5 min	Increase above 0.5%	Single	-0.05%	0.55%	154	-0.09
18	5 min	5 min	Increase above 0.5%	Double	-0.21%	0.38%	23	-0.55
19	5 min	1 hour	Increase above 1%	Single	0.10%	1.02%	2513	0.10
20	5 min	1 hour	Increase above 1%	Double	0.16%	0.93%	1479	0.17
21	5 min	15 min	Increase above 1%	Single	-0.10%	0.81%	191	-0.12
22	5 min	15 min	Increase above 1%	Double	-0.03%	0.64%	87	-0.05
23	5 min	5 min	Increase above 1%	Single	-0.02%	0.62%	29	-0.03
24	5 min	5 min	Increase above 1%	Double	0.04%	0.55%	28	0.07
25	15 min	1 hour	Increase above 0.5%	Single	0.05%	0.96%	1567	0.05
26	15 min	1 hour	Increase above 0.5%	Double	0.09%	0.83%	670	0.11
27	15 min	15 min	Increase above 0.5%	Single	-0.02%	0.63%	775	-0.03
28	15 min	15 min	Increase above 0.5%	Double	0.01%	0.52%	98	0.02
29	15 min	end of trading day	Increase above 0.5%	Single	0.02%	1.25%	3688	0.02
30	15 min	end of trading day	Increase above 0.5%	Double	0.04%	1.22%	3425	0.03
31	15 min	1 hour	Increase above 1%	Single	0.07%	1.05%	1040	0.07
32	15 min	1 hour	Increase above 1%	Double	0.09%	0.99%	614	0.09
33	15 min	15 min	Increase above 1%	Single	-0.11%	0.67%	115	-0.16
34	15 min	15 min	Increase above 1%	Double	-0.02%	0.64%	52	-0.03
35	15 min	end of trading day	Increase above 1%	Single	0.05%	1.47%	2032	0.03
36	15 min	end of trading day	Increase above 1%	Double	0.05%	1.45%	1723	0.03

4. Simulation Results - 100 Shares per Trade

In this section we present the results of the simulation process when trading 100 shares each time the simulation "carries out" a trade. (Instead of the previously detailed, $5,000 worth of stocks per trade).

Table 4.1: Sharpe Measure (for $100K Increments in Investment Funds)

Data Representation	Algorithm	100000	200000	300000	400000	500000
Market	SLR	-1.10	-1.00	-0.92	-0.83	-0.85
	NN	0.43	0.76	0.32	0.02	-0.06
	GA	-0.89	-0.62	-0.59	-0.76	-0.95
Market, Simple News Count	SLR	-0.56	-0.35	-0.19	-0.37	-0.59
	NN	0.47	0.86	0.32	-0.07	-0.26
	GA	-1.41	-0.90	-0.71	-0.77	-0.96
Market, Simple News Count, Categories	SLR	-1.45	-0.89	-0.90	-1.51	-2.01
	NN	1.37	1.29	1.07	1.19	1.03
	GA	-1.45	-0.89	-0.90	-1.51	-2.01
Market, Simple News Count, Categories, Sentiment	SLR	-1.06	-1.10	-0.82	-0.69	-0.48
	NN	1.56	1.63	1.25	1.24	1.19
	GA	-0.59	-0.24	-0.58	-0.78	-1.00
Market, Simple News Count, Categories, Calibrated Sentiment	SLR	-1.11	-1.10	-0.74	-0.60	-0.55
	NN	1.59	1.79	1.47	1.29	1.18
	GA	-1.71	-1.12	-0.99	-0.90	-1.41

Table 4.2: Returns (for $100K Increments in Investment Funds)

Data Representation	Algorithm	100000	200000	300000	400000	500000
Market	SLR	-1.24%	0.22%	0.98%	1.48%	1.68%
	NN	5.55%	6.08%	4.22%	3.39%	3.22%
	GA	-0.55%	1.43%	1.81%	1.70%	1.62%
Market, Simple News Count	SLR	1.31%	2.41%	2.94%	2.68%	2.42%
	NN	5.69%	6.46%	4.24%	3.17%	2.85%
	GA	-2.36%	0.60%	1.50%	1.65%	1.58%
Market, Simple News Count, Categories	SLR	-2.13%	0.74%	1.23%	0.60%	0.32%
	NN	10.75%	8.12%	6.57%	6.44%	5.71%
	GA	-2.13%	0.74%	1.23%	0.60%	0.32%
Market, Simple News Count, Categories, Sentiment	SLR	-1.88%	-0.69%	0.93%	1.61%	2.25%
	NN	12.04%	9.64%	7.20%	6.63%	6.07%
	GA	0.37%	2.41%	1.64%	1.59%	1.52%
Market, Simple News Count, Categories, Calibrated Sentiment	SLR	-1.84%	-0.56%	1.21%	1.84%	2.11%
	NN	11.92%	10.30%	7.91%	6.76%	6.04%
	GA	1.93%	1.21%	1.12%	1.24%	1.15%

5. Classifier Performance

Category	Best Performing Model	Precision	Recall	F-Measure
Regulatory Company Statements	Rule Based Classifier	92.3%	75.0%	82.8%
Lawsuit	Top 500 Variables, SMO	100.0%	68.0%	81.0%
Order Imbalance-Buy Side	Rule Based Classifier	100.0%	100.0%	100.0%
Order Imbalance-Sell Side	Rule Based Classifier	100.0%	100.0%	100.0%
Analyst Upgrade	Top 500 Variables, SMO	91.7%	78.6%	84.6%
Analyst Downgrade	Rule Based Classifier	100.0%	71.4%	83.3%
Acquisition	Rule Based Classifier	88.9%	53.3%	66.7%
Regulatory Activity	Top 30 Variables, SMO	91.7%	64.7%	75.9%
8k form Filing	Rule Based Classifier	100.0%	100.0%	100.0%
Performance Outlook	Rule Based Classifier	86.2%	80.6%	83.3%
Earning or Revenues	Rule Based Classifier	97.1%	77.3%	86.1%
Product or Service Announcement	Top 50 Variables, SMO	85.0%	50.0%	63.0%
Sales and Sales Contracts	Rule Based Classifier	82.6%	50.0%	62.3%
Joint Venture and Collaboration	Top 100 Variables, SMO	77.3%	45.9%	57.6%
Executive Human Resources Issues	Rule Based Classifier	100.0%	61.5%	76.2%
Production and Shipping Issues	Top 50 Variables, Naïve-Bayes	81.8%	52.9%	64.3%
Repeats Market Activity	Top 500 Variables, SMO	97.0%	94.1%	95.5%

6. Dual -- Long and Short Trading Strategy

We present here the results for a variant of the recommendation rules which support both long and short selling. Overall system design remains the same as before, except for modification in the initiation of trading signals and in the simulation procedure.

6.1 Initiation of Trading Signals

Given the probability estimates emerging from the modeling module for each data instance *i*:

(a.) $ScrPos_i$ - the probability that stock returns will increase by more than 1% by the end of the hour;

(b.) $ScrNeg_i$ - the probability that stock returns will decrease by more than 1% by the end of the hour;

We issue a "buy" trading signal if they satisfy the condition:

($ScrPos_i >= ThrPos$) and ($ScrNeg_i < ThrNeg$)

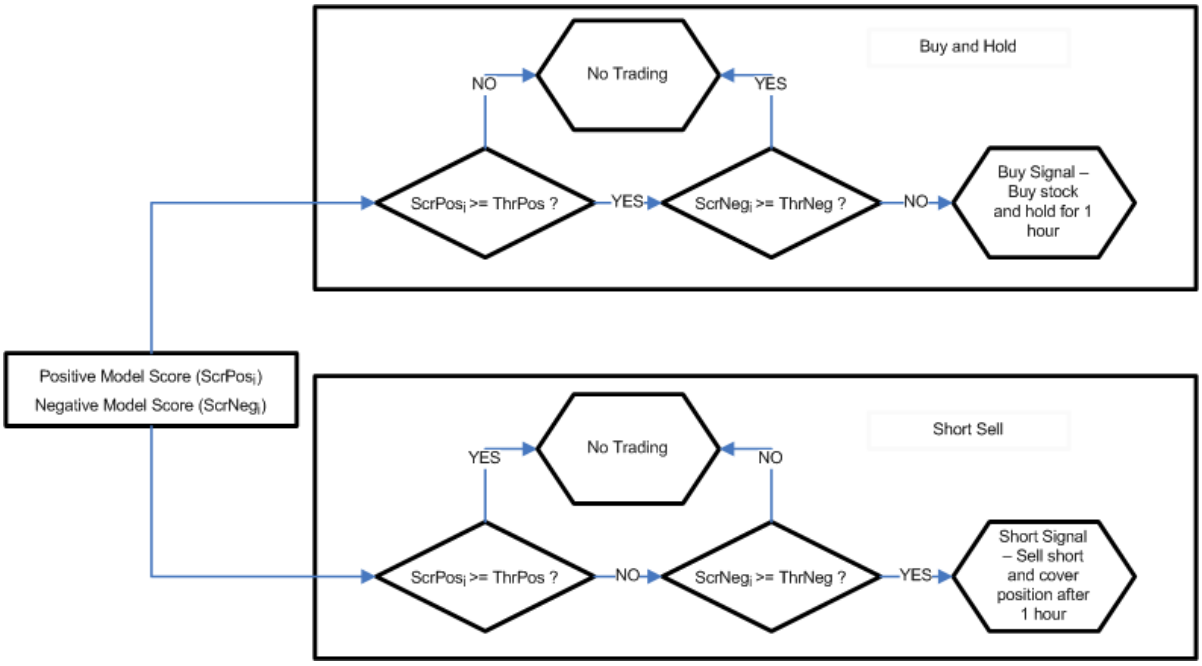
We issue a "short sell" trading signal if they satisfy the condition:

($ScrPos_i < ThrPos$) and ($ScrNeg_i >= ThrNeg$)

Where the *ThrPos* and *ThrNeg* are pre-specified threshold values for the "positive" and "negative" models, respectively.[1] See graphical illustration in figure 6.1

[1] We run a separate optimization process for determining ThrPos and *ThrNeg* values for the short sell.

Figure 6.1 Graphical Illustration of the Trading Rules



6.2 Simulation

When conducting the trading simulation we use the following additional assumptions:

1. For the purpose of being conservative, in case of a short sell recommendation we consider the value of the short sell as "funds at risk" and omit these funds from our balance until the position is "covered".
2. We assume that the calculated NBBO-based bid and ask prices are also valid for short sells. However, we note that in reality, due to the need to borrow a stock from the broker, actual prices may vary.

6.3 Results

Results for the trading simulation which utilizes both long and short strategies are presented in tables 6.1 and 6.2. In table 6.3 we extract the results for the *best-performing algorithm* for both returns and Sharpe ratio measures. The superiority of the NN algorithm is clearly evident in table 6.3. This finding is similar to the one in the main body of the paper (section 6). Another similar finding is that the more advanced the textual data representation, the better the result obtained by the NN model.

Table 6.1: Returns over the Validation Months as a Function of Investment Levels

Data Representation	Algorithm	$100K	$200K	$300K	$400K	$500K
Market	SLR	3.4%	3.3%	3.3%	3.2%	3.1%
	NN	8.7%	5.5%	4.3%	3.5%	3.2%
	GA	-6.8%	-3.1%	-0.1%	2.1%	2.9%
Market, Simple News Count	SLR	6.0%	6.6%	7.8%	7.6%	7.1%
	NN	8.3%	7.1%	4.9%	4.0%	3.5%
	GA	-7.4%	-0.1%	2.1%	2.7%	2.1%
Market, Simple News Count, Categories	SLR	-0.2%	1.4%	4.5%	5.9%	6.6%
	NN	-5.2%	2.8%	5.2%	6.8%	7.4%
	GA	-5.7%	0.4%	1.2%	0.7%	0.5%
Market, Simple News Count, Categories, Sentiment	SLR	3.6%	4.4%	6.7%	7.5%	8.0%
	NN	-4.9%	3.9%	7.2%	8.1%	8.0%
	GA	2.3%	4.6%	5.2%	4.9%	4.9%
Market, Simple News Count, Categories, Calibrated Sentiment	SLR	6.6%	6.7%	6.3%	6.9%	7.1%
	NN	-6.1%	5.1%	7.6%	8.9%	9.0%
	GA	4.9%	2.9%	3.0%	2.8%	2.4%

Columns show returns over the validation months as a function of initial investment ($100k - $500K)

Table 6.2: Sharpe Measure Over the Validation Months as a Function of Investment Levels

Data Representation	Algorithm	$100K	$200K	$300K	$400K	$500K
Market	SLR	0.05	0.00	-0.01	-0.07	-0.12
	NN	1.08	0.52	0.30	0.05	-0.04
	GA	-1.87	-1.59	-0.91	-0.33	-0.11
Market, Simple News Count	SLR	0.47	0.67	1.04	1.13	1.15
	NN	1.03	0.88	0.46	0.24	0.05
	GA	-1.94	-0.84	-0.38	-0.22	-0.48
Market, Simple News Count, Categories	SLR	-0.45	-0.32	0.26	0.59	0.76
	NN	-1.40	-0.09	0.46	0.96	1.19
	GA	-1.88	-0.78	-0.76	-1.19	-1.59
Market, Simple News Count, Categories, Sentiment	SLR	0.08	0.23	0.69	0.94	1.10
	NN	-1.29	0.15	0.90	1.18	1.27
	GA	-0.15	0.33	0.61	0.59	0.66
Market, Simple News Count, Categories, Calibrated Sentiment	SLR	0.51	0.59	0.62	0.84	0.99
	NN	-1.63	0.40	1.01	1.44	1.57
	GA	0.35	-0.10	-0.12	-0.25	-0.46

Columns show Sharpe ratios over the validation months as a function of initial investment ($100k - $500K)
For convenience, Sharpe ratios are presented as annualized values.

Table 6.3: Returns and Sharpe Measure for the Best-Performing Algorithm

	Returns					Sharpe				
Data Representation	$100K	$200K	$300K	$400K	$500K	$100K	$200K	$300K	$400K	$500K
Market	8.7%	5.5%	4.3%	3.5%	3.2%	1.08	0.52	0.30	0.05	-0.04
Market, News Count	8.3%	7.1%	7.8%	7.6%	7.1%	1.03	0.88	1.04	1.13	1.15
Market, News Count, Categories	-0.2%	2.8%	5.2%	6.8%	7.4%	-0.45	-0.09	0.46	0.96	1.19
Market, News Count, Categories, Sentiment	3.6%	4.6%	7.2%	8.1%	8.0%	0.08	0.33	0.90	1.18	1.27
Market, News Count, Categories, Calibrated Sent	6.6%	6.7%	7.6%	8.9%	9.0%	0.51	0.59	1.01	1.44	1.57

Best Performing Algorithm
SLR
NN
GA

7. Data Aspects

Real-life trading systems involve much larger datasets than the dataset used in this study. Nevertheless, an import aspect of our modeling and recommendation modules is that they allow for almost complete parallelization of the real time scoring components as well as offline modeling. This, combined with the fact that we 5-minites interval trading, rather than milliseconds, allows expanding our methodology to support much larger sets of data.

For instance, instead of using a single process or machine to monitor news and market data for thousands of companies, as well as score them to derive trading signals – it is possible to run multiple processes or machines, each responsible for a smaller subset of companies. The only component that is required to use an aggregation of data for multiple companies is the t-test designed to monitor the stability of the trading process. However, this component has a very low load.

Offline modeling can also be split into separate tasks. (e.g., build separate models for groups of companies with similar characteristics). However, offline modeling performance is not a concern due to the fact that the model is updated monthly, so that there should be sufficient time to run the update.

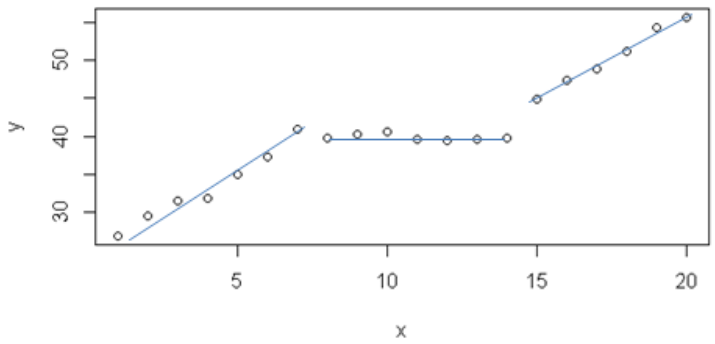
Last, we provide below figures regarding our disk space usage. These figures are for un-optimized and unzipped data for a period of 11.5 months. For 72 S&P500 companies for which there are affluent trades, quotes and news. From these figures it is evident that even using a single strong modern workstation, disk capacity will not be an issue even if we increase the number of companies by 2 orders of magnitudes.

- SAS dataset for 11.5 months ~ 35 gigabytes.
- Raw Market Data (text files unzipped) ~ 130 gigabytes
- News items ~50 Megabytes.

8. Piecewise Linear Segmentation

Piecewise linear segmentation is a set of methods for representing continuous time series data in a form of linear segments to obtain a compact representation of the time series while still maintaining a good approximation of the original data. See figure 8.1 for a graphical illustration

Figure 8.1 Graphical Illustration of a piecewise linear representation



There are several additional aspects that make this method appealing for representing financial time series data. First, it is intuitive as the output provides linear segments which can be visually reviewed. Second, the slope of the different segments captures the extent that the trend is either positive or negative. Third, the duration of each segment provides important information regarding the recency of the trend.

In this study we use the popular "sliding window" implementation of the piecewise linear representation as it is suitable for online calculation allowing one to determine the PLR segment immediately when new data points are added. This is in contrast to the bottom up and top down approaches which require an access to the entire time series in order to divide it into segments.

Specifically we use the sum of squares errors (SSE) as the criteria for "opening" a new segment. For each new data point D*i*, we check whether to consider it as a part of the last segment. This is done by running an OLS linear regression model which includes the data points of the last segment and the current data

point (i.e. data points D*i-n*... D*i-1,* D*i*). We then check whether the resulting SSE exceeds a certain threshold. If the SSE value does not exceed the threshold, we add the data point to the last segment. If it does, we "close" the segment and the new data point is considered as the first data point of a new segment.[2] Lastly, we take the latter 5 linear segment created by the PLR algorithm and use their coefficient and durations as explanatory variables for the prediction models.

Notes:

1. In the process of calculating the PLR coefficient we regard our data points as continuous between consecutive trading days.

2. Contrary to plain market data and technical analysis indicators which are calculated over pre-determined duration of time (e.g., over the last 2 time intervals), the duration of history used for the PLR data representation is not pre-determined because it is based on the duration of the 5 last linear segments, whose duration, in turn, is determined by the PLR algorithm.

[2] Determining this threshold value involves a tradeoff analysis. Reducing its value would improve accuracy, but would shorten the segment and increase the number of segments. Increasing the threshold value would harm accuracy but would enable us to use fewer and longer segments. In our case, we set the threshold value at 0.5. This value was based on subjective visual inspection of the fit between PLR based lines that were plotted against the original data points, (multiple time series data over a sample of 30 days taken from the first 3 month of training set data). We also tested the performance of our prediction models, over the first 3 months of training data using market data and different SSE thresholds values from 0.3-0.7 at 0.1 intervals, with the 0.5 threshold obtaining best results.

9. Textual Data Representation in Related Data Mining Studies

In table 9.1 we summarize the various textual data representations used in forecasting methods in related data-mining studies.

Table 9.1 Features of Related Data-Mining Studies

Reference	Textual Data Representation	Forecasting Method	Predicted
[20] (Fawcett & Provost)	Words and bi grams after stemming and stop list	Use a method called "DC1", developed by the authors.	Stock prices
[22] (Fung et al.)	Pre-defined key words, following identification of news items for "relevant" stocks	SVM	Stock trend
[23] (Fung et al.)	TF-IDF weighted Words, following document reduction using a statistical test	SVM	Stock trend
[32] (Lavrenko et al.)	Bag of Words	"Language Models" similar to Naïve Bayes	Stock trend
[39] (Macskassy et al.)	Words	Naive Bayes and TF IDF ("Rocchio algorithm")	Whether stock returns will exceed one std
[41] (Mittermayer)	Stemmed Words, feature reduction using TF-IDF weighting (1000 top words). Feature vector are cosine normalized	SVM	Three categories (up down no change)
[42] (Mittermayer and Knolmayer)	Words using WDF-IDF weighting using a custom thesaurus. Actual features are selected using multiple different feature selection scoring procedures.	Rocchio algorithm, KNN, SVM	Good, Bad, Neutral Categories
[46] (Robertson & Geva)	Stemmed Words. Actual features are selected using with multiple different feature selection scoring procedures.	C4.5, SVM	Shocks in Volatility
[49] (Schumaker & Chen)	Bag of Words, Noun Phrases, Named Entities	SMO	Prices
[50] (Schumaker & Chen)	Proper Nouns	SVM	Returns
[53] (Thomas)	Document count, categories based on manually determined rules (regular expressions)	Genetic programming and manually crafted rules	Sharpe Measure (trading rules)

[54] (Thomas & Sycara)	Word count, Number of words, Number of posts.	Maximal Entropy classifier and Genetic Algorithm	Returns (trading rules)
[58] (Wuthrich et al.)	Stemmed key words, from a single word and up to 5 words combination. Words are weighted using "category discrimination factor"	Rule based method, KNN and Regression	Stock Index Returns

10. Logistic Regression

Logistic regression models are at the forefront of discrete choice models. Most common is the binary model, where the dependent variable, Y_i, is a simple yes/no, which is coded as 0/1: 0 – for "no" (e.g., no increase in stock price), 1 - for "yes" (positive increase in the stock price). There are several ways to formulate the logistic regression model. We follow here the random utility approach which is more in line with consumer theory (Ben Akiva, 1987) according to which there is an underlying "utility" for each customer *i*, Y_i^* , defined by the linear relationship:

$$Y_i^* = \beta' X_i + \varepsilon_i \qquad (1)$$

Where:

X_i - Vector of explanatory variables, or predictors, for observation *i*

β - Vector of coefficients, estimated based on real observations.

ε_i - Random disturbance, or residual, of observation *i*, and there exist $E(\varepsilon_i)=0$

The utility is a latent variable exhibiting the benefits that the customer derives by making the choice (e.g., purchasing a product). But in practice, Y_i^* is not observable. Instead, one observes the response variable Y_i, which is related to the latent variable Y_i* by:

$$Y_i = \begin{cases} 1 & \text{if } Yi^* > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Namely, if the utility derived from the decision is positive, the customer will take the choice, otherwise s/he will decline it.

From (1) and (2), we obtain:

$$Prob(Y_i = 1) = Prob(Y_i^* = \beta' X_i + \varepsilon_i > 0) = Prob(\varepsilon_i > -\beta' X_i) = 1 - F(-\beta' X_i) \qquad (3)$$

Yielding, for symmetrical distribution of ε_i around zero:

$$Prob\left(Y_i = 1\right) = F\left(\beta' X_i\right) \quad \text{and} \quad Prob(Y_i = 0) = F(-\beta' X_i)$$

Where $F(\cdot)$ denotes the CDF of the disturbance ε_i .

The parameters β's are estimated by the method of maximum likelihood. In case the distribution of ε_i is logistic, we obtain the logistic regression model (also referred to as the logit model) with closed-form formula to calculate the choice probabilities [50]:

$$Prob(Y_i = 1) = \frac{1}{1 + exp(-\hat{\beta}'X)} \quad \text{and} \quad Prob(Y_i = 0) = \frac{1}{1 + exp(\hat{\beta}'X)}$$

Where $\hat{\beta}$ is the MLE (Maximum likelihood estimate) of $\tilde{\beta}$

An alternative assumption is that ε_i is normally distributed. The resulting model in this case is referred to as the probit model. This model is more complicated to estimate because the cumulative normal variable does not have a closed-form solution. But fortunately, the cumulative normal distribution and the logistic distribution are very close to one another, and consequently, the resulting probability estimates are similar. Thus, for all practical purposes, one can use the more convenient and more efficient logit model instead of the probit model.

The logistic regression that we used in our study involves a stepwise procedure, conducted as part of the inter-set analysis of GainSmarts, to select the most influential predictors in the model (hence the name SLR for Stepwise Logistic Regression). In the most elaborate case, the stepwise regression involves a series of steps where at each step a variable, or a group of variables, is introduced or eliminated from the model based on F-tests. The process ends when no individual or a group of predictors are worth eliminating or introducing to the model.

11. Alternative Market Data Model

One of the important findings of our study is that adding textual news-based data to a market data benchmark model, improves predictive accuracy. In this section we evaluated our benchmark market data model against an alternative (benchmark) market data model suggested by Schöneburg (1990)

Schöneburg evaluates several NN stock prediction models and obtains best results using a back-propagation Neural Network. The model uses different variable representations, a different forecasting horizon (one day ahead), and a different sliding window period for the training set.

Schöneburg's paper does not detail how to transform the predictions it generates into trading rules. Therefore, we evaluated several setups using different thresholds, and experimented with a buy and hold strategy as well as with a dual long and short trading strategy. We also tested several NN architectures. The evaluations, which took into account transaction costs, were conducted using the same simulation procedure as reported in section 5 in the main body of this paper.

Best results were obtained with the NN model with one layer of hidden nodes, equal number of hidden nodes and inputs nodes, using a buy and hold strategy, and initiating trades when predicted returns per trade were expected to exceed 1%. As can be seen in Table 11.1, even this setup yields lower performance figures compared with our market data benchmark model. (Market data, using NN and a buy and hold strategy. Original results are reported in section 6, in the body of the paper. For convenience, results are repeated in Table 11.2).

Table 11.1 Modeling Performance using "Best Implementation" (Schöneburg, 1990)

	$100K	$200K	$300K	$400K	$500K
Returns	-5.58%	-3.91%	-2.50%	-1.87%	-1.50%
Sharpe Ratio	-3.58	-1.27	-1.51	-1.81	-2.11

Columns show returns and Sharpe ratio as a function of initial investment ($100k - $500K)
We use the same 8.5 months that we previously used as a validation period.

Table 11.2 Modeling Performance of Our Benchmark Model (using NN and Market Data)

	$100K	$200K	$300K	$400K	$500K
Returns	5.44%	4.58%	3.48%	2.90%	2.89%
Sharpe Ratio	0.36	0.31	0.06	-0.15	-0.19

Columns show returns and Sharpe ratio as a function of initial investment ($100k - $500K)
This table is provided for comparison purposes. It repeats results already presented in Tables 1, 2 in the main body of the paper.

References

T. G. Andersen, "Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility", The Journal of Finance , Vol. 51, No. 1 (Mar., 1996), pp. 169-204

W. Antweiler, M. Z. Frank, "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards", The Journal of Finance, vol. 59, pp. 1259-1294, 2004.

M. Ben-Akiva, S. Lerman; Discrete choice analysis. MIT Press, Cambridge, Massachusetts (1985).

J. A. Busse, T. C. Green, "Market Efficiency in Real Time", Journal of Financial Economics, vol. 65, pp. 415-437, 2002.

V. Dhar, D. Chou, "A Comparison of Nonlinear Methods for Predicting Earnings Surprises and Returns", Neural Networks, IEEE Transactions on, vol. 12, pp. 907-921, 2001.

T. Fawcett, F. Provost, Activity monitoring: Noticing interesting changes in behavior, in: Proceedings of the International Conference on Knowledge Discovery and Data Mining, 1999.

G.P.C. Fung, J.X. Yu, L. Wai, Stock prediction: Integrating text mining approach using real-time news, in: Proceedings of the IEEE International Conference on Computational Intelligence for Financial Engineering, 2003.

G.P.C. Fung, J.X. Yu, H. Luz, The predicting power of textual information on financial markets, IEEE Intelligent Informatics Bulletin, 5 (2005).

V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, J. Allan, Mining of concurrent text and time series, in: Proceeding of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Text Mining, 2000, pp. 37-44.

S.A. Macskassy, H. Hirsh, F.J. Provost, R. Sankaranarayanan, V. Dhar, Intelligent information triage, Research and Development. (2001) 318-326.

M.A. Mittermayer, Forecasting intraday stock price trends with text mining techniques, in: Proceedings, Annual Hawaii International Conference on System Sciences, 2004.

M.A. Mittermayer, G.F. Knolmayer, Newscats: A news categorization and trading system, in: Proceedings of the Sixth International Conference on Data Mining, 2006.

R. J. Rendleman, C. P. Jones, H. A. Latan, "Empirical Anomalies Based on Unexpected Earnings and the Importance of Risk Adjustments", Journal of Financial Economics, vol. 10, pp. 269-87, 1982.

C. Robertson, S. Geva, R.C. Wolff, Can the content of public news be used to forecast abnormal stock market behaviour?, in: Proceedings of ICDM, 2007.

E. Schöneburg, 1990. Stock price prediction using neural networks: A project report. Neurocomputing. 2(1) 17-27.

R.P. Schumaker, H. Chen, Textual analysis of stock market prediction using financial news articles, in: Proceeding of the 12th Americas Conference on Information Systems, 2006.

R.P. Schumaker, H. Chen, Evaluating a news-aware quantitative trader: The effect of momentum and contrarian stock selection strategies, Journal of the American Society for Information Science and Technology. 59 (2008) 247-255.

J. D. Thomas, News and Trading Rules, PhD Thesis: Carnegie Mellon University, 2003.

J.D. Thomas, K. Sycara, Integrating genetic algorithms and text learning for financial prediction, in: Data Mining with Evolutionary Algorithms, 2000, pp. 72-75.

B. Wuthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, J. Zhang, Daily stock market forecast from textual Web data, in: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, 1998.