The New Alchemy: Turning Words into Signals

Gideon Mann
gmann16@bloomberg.net
Bloomberg

- Founded in **1981**
- **320,000** subscribers
- Customers in **170 countries**
- Over **18,000 employees** in 192 locations
- **More News Reporters than** the New York Times + Washington Post + Chicago Tribune
- **Over 5,000 Engineers/Programmers**
Outline

Natural Documents

News that Moves the Market

Sentiment Signals
Amazon Financials

Item 6. Selected Consolidated Financial Data

The following selected consolidated financial data should be read in conjunction with the consolidated financial statements and the notes thereto in Item 5 of Part II, "Financial Statements and Supplementary Data," and the information contained in Item 7 of Part II, "Management’s Discussion and Analysis of Financial Condition and Results of Operations." Historical results are not necessarily indicative of future results.

Statements of Operations:

<table>
<thead>
<tr>
<th>Year Ended</th>
<th>2015</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net sales</td>
<td>$107,006</td>
<td>$88,988</td>
</tr>
<tr>
<td>Income from operations</td>
<td>$2,233</td>
<td>$178</td>
</tr>
<tr>
<td>Net income (loss)</td>
<td>$596</td>
<td>$(241)</td>
</tr>
<tr>
<td>Basic earnings per share (1)</td>
<td>$1.25</td>
<td>$(0.52)</td>
</tr>
<tr>
<td>Diluted earnings per share (1)</td>
<td>$1.25</td>
<td>$(0.52)</td>
</tr>
</tbody>
</table>

Weighted-average shares used in computation of earnings per share:

<table>
<thead>
<tr>
<th></th>
<th>2015</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>467</td>
<td>462</td>
</tr>
<tr>
<td>Diluted</td>
<td>477</td>
<td>462</td>
</tr>
</tbody>
</table>

Statements of Cash Flows:

<table>
<thead>
<tr>
<th>Year Ended</th>
<th>2015</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net cash provided by (used in) operating activities</td>
<td>$11,920</td>
<td>$6,842</td>
</tr>
</tbody>
</table>

Balance Sheets:

<table>
<thead>
<tr>
<th></th>
<th>2015</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total assets</td>
<td>$65,444</td>
<td>$54,505</td>
</tr>
<tr>
<td>Total long-term obligations</td>
<td>$18,181</td>
<td>$17,673</td>
</tr>
</tbody>
</table>

(1) For further discussion of earnings per share, see Item 8 of Part II, "Financial Statements and Supplementary Accounting Policies."
“Scatteract: Automated extraction of data from scatter plots” at ECML/PKDD 2017, with Mathieu Cliche, Dhruv Madeka, and Connie Yee.

https://github.com/bloomberg/scatteract
Outline

Understanding Natural Documents

News that Moves the Market

Sentiment Signals
## Market Moving News

### Options

**Date Range:** 01/29/18 - 02/05/18

**Sources:** All

<table>
<thead>
<tr>
<th>Wire</th>
<th>Time</th>
<th>Ticker</th>
<th>% Chg</th>
<th>Headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>BN</td>
<td>02/05</td>
<td>10:33</td>
<td>KEX US</td>
</tr>
<tr>
<td>11</td>
<td>BFW</td>
<td>02/05</td>
<td>10:07</td>
<td>NOW US</td>
</tr>
<tr>
<td>12</td>
<td>BFW</td>
<td>02/05</td>
<td>10:07</td>
<td>RHT US</td>
</tr>
<tr>
<td>13</td>
<td>BFW</td>
<td>02/05</td>
<td>09:47</td>
<td>BSX US</td>
</tr>
<tr>
<td>14</td>
<td>BN</td>
<td>02/05</td>
<td>09:47</td>
<td>KMX US</td>
</tr>
<tr>
<td>15</td>
<td>DBF</td>
<td>02/05</td>
<td>09:34</td>
<td>ELG GR</td>
</tr>
<tr>
<td>16</td>
<td>DBF</td>
<td>02/05</td>
<td>07:33</td>
<td>EIN3 GR</td>
</tr>
<tr>
<td>17</td>
<td>BN</td>
<td>02/05</td>
<td>07:01</td>
<td>GAW LN</td>
</tr>
<tr>
<td>18</td>
<td>BN</td>
<td>02/05</td>
<td>06:20</td>
<td>SPNO DC</td>
</tr>
<tr>
<td>19</td>
<td>BN</td>
<td>02/05</td>
<td>06:00</td>
<td>FABG SS</td>
</tr>
<tr>
<td>20</td>
<td>BN</td>
<td>02/05</td>
<td>04:57</td>
<td>CLAR IN</td>
</tr>
<tr>
<td>21</td>
<td>BN</td>
<td>02/05</td>
<td>04:36</td>
<td>IFCI IN</td>
</tr>
<tr>
<td>22</td>
<td>BN</td>
<td>02/05</td>
<td>03:48</td>
<td>TRGYO TI</td>
</tr>
<tr>
<td>23</td>
<td>BN</td>
<td>02/05</td>
<td>03:45</td>
<td>CPBI IN</td>
</tr>
<tr>
<td>24</td>
<td>BN</td>
<td>02/05</td>
<td>03:21</td>
<td>LELE IN</td>
</tr>
<tr>
<td>25</td>
<td>BN</td>
<td>02/05</td>
<td>03:08</td>
<td>WLDL IN</td>
</tr>
<tr>
<td>26</td>
<td>BN</td>
<td>02/05</td>
<td>02:52</td>
<td>EVRIN IN</td>
</tr>
</tbody>
</table>
Example
Introduction

• Supervised learning setup
  o Training data: news stories + positive/negative labels
    ▪ Did this precipitate a price reaction?
    ▪ Positive labels are MMN events
  o Model will provide a probability of price/volume move given a headline

• Defining a positive MMN event
  o What constitutes a market move (anomaly)?
  o Aligning news headlines to that move (relevance)
Finding anomalies in the market
Finding price anomalies is hard
Finding Stories Aligned With Anomalies

• Find stories tagged with the ticker in the [-5min, +2min] window around the beginning of the anomaly

• +2 minutes because we might not have the breaking story, but might have similar stories
Annotation

- Stories are annotated by humans: whether the stories are related to the price move
- Limited to English, and no social media, for now
- Predominantly flash headlines

- Overall, of the one million examples, only about 2% are positive
- Negative examples are sampled
Negative Examples

<table>
<thead>
<tr>
<th>Wire</th>
<th>Time</th>
<th>Headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>110</td>
<td>01/19 02:08:30</td>
<td>*HS#FTA50RC1512D &lt;67241&gt; Notice of Occurrence of Mandatory</td>
</tr>
</tbody>
</table>
### Negative Examples

<table>
<thead>
<tr>
<th>Wire</th>
<th>Time</th>
<th>Headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>01/21 10:31:09</td>
<td>*ADTRAN EARNINGS CONFERENCE CALL BEGINS</td>
</tr>
</tbody>
</table>
Negative Examples

Wire | Time       | Headline                                           
-----|------------|----------------------------------------------------
10)  | BN 01/21 10:38:44 | *AOL CEO ENDS REMARKS AT INDUSTRY CONFERENCE
Why We Need Human Annotation

<table>
<thead>
<tr>
<th>Wire</th>
<th>Time</th>
<th>Headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>BFW</td>
<td>06/13 10:11:53</td>
<td>*AVIS BUDGET JUMPS AS MUCH AS 6.7%, TOUCHES HIGHEST SINCE MAY 23</td>
</tr>
</tbody>
</table>
Quiz: Is this Market Moving News?

*JAPAN POST PRESIDENT SAYS CO. WILL `TONE DOWN' ON M&AS; NIKKEI
Quiz: With context, should this be an MMN event?

Nomura Real Estate Holdings
The explanation (story minutes later...)

Japan Post Pres. Wants to ‘Tone Down’ M&As After Toll: Nikkei

By Go Onomitsu

(Bloomberg) -- Japan Post Holdings President Masatsugu Nagato says the company will “tone down” acquisitions both in Japan and overseas, Nikkei reports in an interview with the executive, reversing the company’s stance of proactively seeking takeovers.

- Nomura Real Estate, which was previously reported to be Japan Post takeover target, erased gains, falling as much as 5.9%

- NOTE: Nomura Holdings was in discussions to sell its stake of about one-third in the brokerage’s real estate unit to Japan Post, people familiar told Bloomberg News: Link

- NOTE: Japan Post in April said it will book a 400b yen writedown on its acquisition of Toll: Link

- Related story: Japan Post Is Still Interested in M&A After Toll Holdings Loss: Link
Logistic Regression Model

• One of the go-to methods for binary classification problems
  o Many features
  o Features are correlated…
  o Millions of samples (more complex methods are computationally intractable)

• Named after logistic (sigmoid) function:
  o \( L(z) = \frac{1}{1+e^{-z}} \)

• Based on a linear combination of ‘features’:
  o \( z(x) = \alpha + \beta \cdot x = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_d x_d \)

• Learning is finding the values for parameters:
  o \( \alpha, \beta_1, \beta_2, \beta_3, \ldots, \beta_d \)
Features

1-grams: Gold, Rallies, ...
2-grams: Gold Rallies, Rallies on, ...
3-grams: Gold Rallies on, Rallies on Weak, on Weak U.S., ...
1-skip-bigram: Gold on, Rallies Weak, on U.S., ...
...
2-skip-trigram: Gold Weak U.S., Rallies U.S. Retail, on Retail Sales, ...

m-skip-n-gram: Gold Rallies Weak U.S. Retail, Gold Rallies U.S. Retail Sales

<table>
<thead>
<tr>
<th>Topic (98)</th>
<th>Relevance</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>139900</td>
<td>(100) (100)</td>
<td>139900</td>
</tr>
<tr>
<td>2018/06/14</td>
<td>(100) (100)</td>
<td>2018/06/14</td>
</tr>
<tr>
<td>AGRKEY</td>
<td>(0) (99)</td>
<td>Key Agricultural...</td>
</tr>
<tr>
<td>ASIA</td>
<td>(0) (79)</td>
<td>Asia, Pacific Rim</td>
</tr>
<tr>
<td>BASIC</td>
<td>(0) (100)</td>
<td>Materials</td>
</tr>
<tr>
<td>BBPROFILE</td>
<td>(0) (100)</td>
<td>Bloomberg Client</td>
</tr>
<tr>
<td>BIZNEWS</td>
<td>(0) (100)</td>
<td>Bus, Eco, Govt...</td>
</tr>
<tr>
<td>BOJ</td>
<td>(79) (79)</td>
<td>Bank of Japan</td>
</tr>
<tr>
<td>BON</td>
<td>(0) (69)</td>
<td>Bonds</td>
</tr>
<tr>
<td>BONDWIRES</td>
<td>(100) (100)</td>
<td>Eligible Wires ...</td>
</tr>
<tr>
<td>BUSINESS</td>
<td>(0) (100)</td>
<td>Business News</td>
</tr>
<tr>
<td>CECO</td>
<td>(0) (100)</td>
<td>Country Economic...</td>
</tr>
</tbody>
</table>
Current Production System

• Logistic Regression:
  o Using about 24k features
  o Main features: skip-grams, topic codes…
  o (Tried many other features…)

• The pipeline:
  o POS tagging
  o name entity matching
  o skip-gram extraction, etc.

• Caveats:
  • Skip-grams are very sparse, making it difficult to reliably estimate their weights
  • This model needs careful feature selection
  • Many moving parts
Comparison between:

Logistic Regression

Next Version

Testing PR curve. AUC=0.0395512075459

Testing PR curve. AUC=0.06570717842
Outline

Natural Documents

News that Moves the Market

Sentiment Signals
## Social Velocity

<table>
<thead>
<tr>
<th>N</th>
<th>Time</th>
<th>Company Name</th>
<th>Price</th>
<th>%Chg</th>
<th>Δ AVAT</th>
<th>Option</th>
<th>Δ Volm</th>
<th>Snrmt</th>
<th>Headline</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>15:22</td>
<td>INDUSTREA ACQU</td>
<td>10.28</td>
<td>-39%</td>
<td>-99.5%</td>
<td></td>
<td></td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>11</td>
<td>15:21</td>
<td>MICRON TECH</td>
<td>39.60</td>
<td>-2.99%</td>
<td>3.9%</td>
<td>200676</td>
<td>2.6%</td>
<td></td>
<td>InvestorPlace: High-Level Measures for Q3...</td>
</tr>
<tr>
<td>12</td>
<td>15:21</td>
<td>GILEAD SCIENCES</td>
<td>79.00</td>
<td>-3.34%</td>
<td>-1.1%</td>
<td>55956</td>
<td>-7.6%</td>
<td></td>
<td>Merck: Lawyer Misdeeds Don’t Justify Kill...</td>
</tr>
<tr>
<td>13</td>
<td>15:20</td>
<td>PROCTER &amp; GAMBLE</td>
<td>81.75</td>
<td>-2.97%</td>
<td>72.1%</td>
<td>31540</td>
<td>-1.3%</td>
<td></td>
<td>Amazon, Netflix, P&amp;G Try to Make $5 Mill...</td>
</tr>
<tr>
<td>14</td>
<td>15:18</td>
<td>NVIDIA CORP</td>
<td>217.01</td>
<td>-7.07%</td>
<td>91.6%</td>
<td>117735</td>
<td>-18.9%</td>
<td></td>
<td>Nvidia Shares Fall 6.1%; Volume More Th...</td>
</tr>
<tr>
<td>15</td>
<td>15:15</td>
<td>ACADIA PHARMACEUTICALS</td>
<td>27.33</td>
<td>-5.63%</td>
<td>-15.2%</td>
<td>1305</td>
<td>-68.6%</td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>16</td>
<td>15:15</td>
<td>VERICEL CORP</td>
<td>6.70</td>
<td>-11.84%</td>
<td>16.7%</td>
<td>36</td>
<td>-35.9%</td>
<td></td>
<td>Vericel to Present at the BIO CEO &amp; Inv...</td>
</tr>
<tr>
<td>17</td>
<td>15:00</td>
<td>DOWDUPONT INC</td>
<td>67.51</td>
<td>-4.77%</td>
<td>117.6%</td>
<td>19459</td>
<td>49.8%</td>
<td></td>
<td>DowDuPont Down Most in 2 Years, Paces...</td>
</tr>
<tr>
<td>18</td>
<td>15:57</td>
<td>ETHAN ALLEN</td>
<td>23.80</td>
<td>-2.46%</td>
<td>-8.0%</td>
<td>157</td>
<td>115.0%</td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>19</td>
<td>14:56</td>
<td>BANK OF AMERICA</td>
<td>29.79</td>
<td>-6.74%</td>
<td>88.2%</td>
<td>570583</td>
<td>59.9%</td>
<td></td>
<td>Bank of America Cuts Gas Estimate as Pr...</td>
</tr>
<tr>
<td>20</td>
<td>14:48</td>
<td>VIKING THERAPEUTIC</td>
<td>5.52</td>
<td>-9.06%</td>
<td>228.2%</td>
<td>327</td>
<td>138.6%</td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>21</td>
<td>14:45</td>
<td>SPIRAX-SARCO ENG</td>
<td>5595</td>
<td>-8.9%</td>
<td>54.0%</td>
<td></td>
<td></td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>22</td>
<td>14:26</td>
<td>FORTINET INC</td>
<td>44.23</td>
<td>-1.58%</td>
<td>45.5%</td>
<td>5199</td>
<td>212.4%</td>
<td></td>
<td>PREVIEW CYBERSECURITY 4Q: High-Demand...</td>
</tr>
<tr>
<td>23</td>
<td>14:22</td>
<td>DRIPCAR INC</td>
<td>2.34</td>
<td>-8.95%</td>
<td>-11.9%</td>
<td></td>
<td></td>
<td></td>
<td>Dropcar, Inc.</td>
</tr>
<tr>
<td>24</td>
<td>14:18</td>
<td>STEELCASE INC-A</td>
<td>14.75</td>
<td>-3.91%</td>
<td>14.0%</td>
<td>25</td>
<td>-86.9%</td>
<td></td>
<td>Steelcase Inc</td>
</tr>
<tr>
<td>25</td>
<td>13:56</td>
<td>EXXON MOBIL CORP</td>
<td>79.52</td>
<td>-5.93%</td>
<td>153.3%</td>
<td>105014</td>
<td>148.3%</td>
<td></td>
<td>Big Oil Stocks Fail to Attract Buyers as S...</td>
</tr>
<tr>
<td>26</td>
<td>13:54</td>
<td>ADVANCE AUTO PAR</td>
<td>111.51</td>
<td>-2.12%</td>
<td>-4.2%</td>
<td>720</td>
<td>-54.0%</td>
<td></td>
<td>ADVANCE AUTO PARTS INC: 4 2018-02-01</td>
</tr>
<tr>
<td>27</td>
<td>13:49</td>
<td>HIMAX TECHNO-ADR</td>
<td>7.68</td>
<td>-10.76%</td>
<td>24.8%</td>
<td>11356</td>
<td>1.9%</td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>28</td>
<td>13:30</td>
<td>TABLEAU SOFTWARE</td>
<td>80.63</td>
<td>-4.02%</td>
<td>83.8%</td>
<td>5287</td>
<td>115.6%</td>
<td></td>
<td>Street Insider: First Analysis Remains Eq...</td>
</tr>
<tr>
<td>29</td>
<td>13:26</td>
<td>BITCOIN INVEST</td>
<td>9.80</td>
<td>18.33%</td>
<td>56.8%</td>
<td></td>
<td></td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>30</td>
<td>13:25</td>
<td>SOPHIES BIO INC</td>
<td>1.97</td>
<td>-7.94%</td>
<td>146.5%</td>
<td></td>
<td></td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>31</td>
<td>13:24</td>
<td>EVERSOURCE ENERG</td>
<td>58.79</td>
<td>-9.4%</td>
<td>-5.5%</td>
<td>31</td>
<td>-75.6%</td>
<td></td>
<td>No recent news headline</td>
</tr>
<tr>
<td>32</td>
<td>13:24</td>
<td>F5 NETWORKS</td>
<td>137.84</td>
<td>-4.36%</td>
<td>41.4%</td>
<td>2716</td>
<td>36.8%</td>
<td></td>
<td>F5 Networks Shares Down 4.5% Outpacing...</td>
</tr>
<tr>
<td>33</td>
<td>13:24</td>
<td>ALLERGAN PLC</td>
<td>166.52</td>
<td>-3.47%</td>
<td>33.8%</td>
<td>16317</td>
<td>40.6%</td>
<td></td>
<td>PREVIEW ALLERGAN 4Q: Migraine, Esmya...</td>
</tr>
</tbody>
</table>
Sentiment analysis

Sentiment analytics data is provided in two ways:

1. **Story-level sentiment**
   - Generated in real time upon the arrival of news stories or tweets
   - Score: 1 for positive, -1 for negative, and 0 for neutral
   - We generate three way confidences - interpreted as probability

2. **Company-level sentiment**
   - Weighted average of story-level sentiments – aggregated at daily and intra-day intervals
Binary classification can be viewed as a task of separating classes in feature space:

\[
w^T x + b = 0
\]

\[
w^T x + b > 0
\]

\[
w^T x + b < 0
\]

\[
f(x) = \text{sign}(w^T x + b)
\]
Classification margin

- Distance from example $x_i$ to the separator is $r = \frac{w^T x + b}{\|w\|}$
- Examples closest to the hyperplane are support vectors
- Margin $\rho$ of the separator is the distance between support vectors
Solving the Optimization Problem

Find \( w \) and \( b \) such that
\[
\Phi(w) = \frac{1}{2} w^T w \text{ is minimized and for all } \{(x_i, y_i)\}
\]
\[
y_i (w^T x_i + b) \geq 1
\]

- Need to optimize a *quadratic* function subject to *linear* constraints.

- Quadratic optimization problems are a well-known class of mathematical programming problems, and many (rather intricate) algorithms exist for solving them.

- The solution involves constructing a *dual problem* where a *Lagrange multiplier* \( \alpha_i \) is associated with every constraint in the primary problem:

Find \( \alpha_1 \ldots \alpha_N \) such that

\[
Q(\alpha) = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i^T x_j \text{ is maximized and}
\]

(1) \( \sum \alpha_i y_i = 0 \)

(2) \( \alpha_i \geq 0 \) for all \( \alpha_i \)
Sentiment aggregation methods

We take all stories tagged with a particular company:

1) Each news story or tweet is scored with "confidences" $C_+ , C_-, C_n$ for positive, negative and neutral sentiment respectively.

2) Story specific sentiment (polarity score): $S^i = C_+^i - C_-^i$

3) Sentiment Average:

$$\mu = \frac{\sum_{i=1}^{N} S^i}{N} = \frac{\sum_{i=1}^{N} (C_+^i - C_-^i)}{N} = \overline{C_+} - \overline{C_-}$$

1) Sentiment dispersion = Inter-story variance + story specific dispersion

Simplifies to: $\overline{C_+} + \overline{C_-} - \mu^2$
The daily sentiment score is constructed using story level sentiment data over the 24 hours period, and calculated around 9:05 AM every day.
Does market efficiently price in the sentiment information?
Strategy: Daily long-short

Does market efficiently price in the sentiment information?

• Strategy:
  o Each day before market open, rank all stocks by their daily sentiment

  o Three variations:
    1. HML 1/3 : Long top 1/3 and short bottom 1/3 stocks
    2. HML 5% : Long top 5% and short bottom 5% stocks
    3. Prop : Positions proportional to the difference of sentiment scores from its cross-sectional mean

  o Create portfolio at market open and close out at market close

We backtest the strategy for Russell 2000 stocks from Jan-01-2015 to Jul-31-2016
The portfolio daily return can be computed as the following,

\[ Ret_j = \sum_{i \in \text{Long}_j} w_{ij}^{\text{Long}} \left( \frac{P_{ij}^{\text{close}}}{P_{ij}^{\text{open}}} - 1 \right) - \sum_{i \in \text{Short}_j} w_{ij}^{\text{Short}} \left( \frac{P_{ij}^{\text{close}}}{P_{ij}^{\text{open}}} - 1 \right) \]

Where
- \( Ret_j \) is portfolio return on day \( j \);
- \( P_{ij}^{\text{close}} \) is the close price of stock \( i \) on day \( j \);
- \( P_{ij}^{\text{open}} \) is the open price of stock \( i \) on day \( j \);
- \( \text{Long}_j \) is the basket of stocks to long on day \( j \);
- \( w_{ij}^{\text{Long}} \) is the weight of stock \( i \) in \( \text{Long}_j \);
- \( \text{Short}_j \) is the basket of stocks to short on day \( j \);
- \( w_{ij}^{\text{Short}} \) is the weight of stock \( i \) in \( \text{Short}_j \);

For HML portfolio, \( w_{ij}^{\text{Long}} = \frac{1}{\text{# of Stocks in Long}_j} \), \( w_{ij}^{\text{Short}} = \frac{1}{\text{# of Stocks in Short}_j} \)

For proportional portfolio, \( w_{ij}^{\text{Long}} = \frac{SS_{ij}^{\text{Long}} - \mu_j}{\sum_{i \in \text{Long}_j} (SS_{ij}^{\text{Long}} - \mu_j)} \), \( w_{ij}^{\text{Short}} = \frac{SS_{ij}^{\text{Short}} - \mu_j}{\sum_{i \in \text{Short}_j} (SS_{ij}^{\text{Short}} - \mu_j)} \)

\( \mu_j \) is the cross-sectional mean of the company-level sentiment on day \( j \);
- \( SS_{ij}^{\text{Long}} \) is the sentiment score of stock \( i \) in \( \text{Long}_j \) on day \( j \);
- \( SS_{ij}^{\text{Short}} \) is the sentiment score of stock \( i \) in \( \text{Short}_j \) on day \( j \)
Sentiment Dispersion: To calculate the overall dispersion, we need to track two components.

- INTRA STORY VARIANCE: Variance within one story due to classification uncertainty

The variance of the sentiment per story is defined as the variance of the trinomial probability distribution

$$Var^i = E[(X^i - S^i)^2] = C_+^i (1 - S^i)^2 + C_-^i (-1 - S^i)^2 + C_n^i (0 - S^i)^2$$

Simplifies to

$$Var^i = C_+^i + C_-^i - S^i^2 = C_+^i + C_-^i - (C_+^i - C_-^i)^2$$

- INTER STORY VARIANCE: Sample variance of the different stories [i.e., $SampleVar(S^i)$]

- TOTAL VARIANCE: We expect it to be $\overline{C_+} + \overline{C_-} - \mu^2$ since we have trinomial distribution for the overall aggregation as well. Turns out we can show that sum of two components $SampleVar(S^i) + \frac{\sum_{i=1}^{N} Var^i}{N}$ indeed equates to $\overline{C_+} + \overline{C_-} - \mu^2$
Russell 2000 stocks

Equity curve of Sentiment portfolios

- HML 1/3
- HML 1/3 + dispersion
- Prop
- Prop + dispersion
- Index ETF

Time periods: Feb 2015 to Feb 2016
Russell 2000 – combining news and social
Russell 2000 – with transaction cost
THANK YOU