Textual Sentiment, Option Information and Stock Return Predictability

Cathy Yi-Hsuan Chen
Matthias R. Fengler
Wolfgang Karl Härdle
Yanchu Liu

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
University of St. Gallen, Switzerland
Lingnan College, Sun Yat-sen University, China

http://lvb.wiwi.hu-berlin.de
http://www.mathstat.unisg.ch/
http://www.lingnan.sysu.edu.cn
Sentiment moves stock markets

- Growing evidence shows that textual sentiment provides incremental information about future stock returns.
  Confirmed at index levels as well as single-stock levels.
What about sentiment and options markets?


- Prediction power cannot be explained by ”rational” option pricing models.
Options market and stock market

- Dennis and Mayhew (2002), Xing et al. (2010): option data characteristics (skew, implied volatility) predict stock returns

Hypothesis:
private information about stocks can be best exploited via the option market because it’s easier to leverage and short-sell.

Therefore options market may lead stock markets in terms of price discovery.
Given sentiment predicts both stock returns and option data, is there still room for the private information hypothesis in option markets?

Maybe it’s all just a common sentiment factor that get’s internalized at different speed in the different markets.

Requires a joint study of

**Textual Sentiment, Option Information and Stock Return Predictability**
This research

- Extend Han’s (2008) ideas:
  - Study reaction of standard of single-stock options to news
  - Use language processing tools for sentiment construction

- Investigate influence of option market variables in presence of news sentiment (Xing et al.’s hypothesis)

- Study source of option markets predictability:
  Inside information? Internalized investor sentiment? Both?
**Current literature**

$B_t$ is sentiment, $OC_t$ an option market variable, $R_t$ a stock return.
This work

$B_t$ is sentiment, $OC_t$ an option market variable, $R_t$ a stock return
This work

$B_t$ is sentiment, $OC_t$ an option market variable, $R_t$ a stock return
Findings

- Our sentiment proxies predict single-stock option market variables
  - Both firm-specific sentiment and aggregate sentiment
  - Aggregate negative sentiment is a strong predictor

- Sentiment proxies predict single stock returns

- Asymmetry of informational relevance of news:
  - Overnight information more relevant than trading day information
  - Possibly due to a different thematic coverage and more complex topics.
Findings

- Option market variables remain relevant predictors of stock returns in presence of sentiment

  - Aggregate sentiment is a relevant factor for single stock returns
  - Option market variables where sentiment is partialled out remain significant predictors.
Outline

1. Motivation ✓
2. Data collection
3. Text analytics
4. Sentiment projection
5. Topic model
6. Panel regressions
7. Summary
Sentiment extraction from news data

There is a lot of news...
**Dimensions of news**

- **Source of news**
  - Official channels: government, federal reserve bank/central bank, financial institutions
  - Internet: blogs, social media, message boards

- **Type of news**
  - Scheduled vs. non-scheduled
  - Expected vs. unexpected
  - Event-specific vs. continuous news flows
Data

Sentiment variables: distilled from Nasdaq articles

- Terms of Service permit web scraping
- Currently > 580k articles between October 2009 and January 2017
- Data available at RDC
- Analysis is on data from 2012-2015
Number of articles per trading day

Black: # articles on a trading day; grey: # articles on weekend, holiday
Hourly distribution
In total we process

- 119,680 articles, out of which 6,600 articles (i.e., 5.51%) are posted on non-trading days (excluded)
- Out of 113,080 articles 50.26% are posted during trading hours and 49.74% during overnights.
Extracting sentiment from text

- Articles
- Scraping
- NLP
- Projection
- Sentiment

Nasdaq Articles

- URL
- Author
- Symbol
- Date
- Text

Token
Negation
POS
Lemmata

Unsupervised
BL
LM

Supervised
SM

Sentiment and Options
**Sentiment analysis**

Strategies:
- Lexica projection: positive, neutral and negative
- Machine learning: text classification

Based on:
- *Financial Sentiment Dictionary* (LM)
  Loughran and McDonald (JF, 2011)
- *Financial Phrase Bank* (LM)
  Malo et al. (2014)
Unsupervised projection

O gentle doves, O turtle-doves,
And all the birds that be,
The lentils that in ashes lie
Come and pick up for me!

The good must be put in the dish,
The bad you may eat if you wish.

Figure: Example of Text Numerisization

- Many texts are numerisized via lexical projection
- Goal: Accurate values for positive and negative sentiment
Lexicon-based sentiment

Consider sentence $i$ in some document, positive sentiment $Pos_i$, positive lexicon entries $W_j$ ($j = 1, \ldots, J$) and count frequency of those entries $w_j$:

$$Pos_i = n_i^{-1} \sum_{j=1}^{J} I(W_j \in L) w_j$$

with $n_i$: number of words in document $i$ (e.g. sentence)

Equivalent calculation of negative sentiment $Neg_i$
Sentence-level polarity

For sentence $i$, we compute the sentence-level polarity by:

$$
Pol_i = \begin{cases} 
1, & \text{if } Pos_i > Neg_i \\
0, & \text{if } Pos_i = Neg_i \\
-1, & \text{if } Pos_i < Neg_i 
\end{cases}
$$

Then, at the document level, we calculate,

$$
FP = n^{-1} \sum_{i=1}^{n} I(\text{Pol}_i = 1)
$$

$$
FN = n^{-1} \sum_{i=1}^{n} I(\text{Pol}_i = -1),
$$

where $n$ is the number of sentences in the document.
Supervised projection

- Training data: Financial Phrase Bank of Malo et al. (2014)
  - Sentence-level annotation of financial news
  - Manual annotation of 5,000 sentences by 16 annotators incorporates human knowledge
  - Example: “profit” with different semantic orientations
    - Neutral in “profit was 1 million”
    - Positive in “profit increased from last year”
Regularized linear models (RLM)

- Training data \((X_1, y_1) \ldots (X_n, y_n)\) with \(X_i \in \mathbb{R}^p\) and \(y_i \in \{-1, 1\}\)
- Linear scoring function \(s(X) = \beta^T X\) with \(\beta \in \mathbb{R}^p\)

Regularized training error:

\[
    n^{-1} \sum_{i=1}^{n} \left[ L\{y_i, s(X)\} + \lambda R(\beta) \right]
\]

with hyperparameter \(\lambda \geq 0\)
RLM estimation

- Optimize via Stochastic Gradient Descent
- 5-fold cross validation
- Oversampling
- Choice of: $L(\cdot), R(\cdot), \lambda, X$ ($n$-gram range, features) . . .
- Three categories: one vs. all sub-models
Model accuracy - polarity

Supervised Learning

- Chosen model: Hinge loss, L1 norm, $\lambda = 0.0001$, ...
- Mean accuracy (oversampling): 0.80
- Mean accuracy (normal sample): 0.82

Lexicon-based

- Mean accuracy BL: 0.58
- Mean accuracy LM: 0.64

So, we adopt the supervised learning methodology hereafter.
**Sentence-level and document-level polarity**

After training: Each document $i$ is split up into its sentences $j$ and the corresponding score is calculated.

Yields a predictor for the polarity of sentence $j$, $Pol_j$:

For each document, these scores are aggregated to

$$FP = n^{-1} \sum_{j=1}^{n} \mathbb{I}(Pol_j = 1)$$

$$FN = n^{-1} \sum_{j=1}^{n} \mathbb{I}(Pol_j = -1),$$

where $n$ is the number of sentences in the document.
Bullishness

\[ B = \log \left\{ \frac{1 + n^{-1} \sum_{j=1}^{n} I(Pol_j = 1)}{1 + n^{-1} \sum_{j=1}^{n} I(Pol_j = -1)} \right\} \]

by Antweiler and Frank (JF, 2004) with \( j = 1, \ldots, n \) sentences in document.

- \( B_{i,t} \) accounts for bullishness of company \( i \) on day \( t \)
- Consider \( BN_{i,t} = -I(B_{i,t} < 0)B_{i,t} \)
trading $B_{idx}$

SM bullishness index in trading hours

Date

2012-01-03  2013-01-02  2014-01-02  2015-01-02  2016-01-04
overnight $B_{idx}^{on}$
trading $BN_{idx}$

SM negative bullishness index in trading hours
overnight $BN_{idx}^{on}$
How do trading-day/overnight articles differ?

- Overnight information is more informative than trading-day information. Why?
- Uncover the thematic coverage of the alternate news archives using a statistical topic model.
Latent Dirichlet Allocation

LDA is a topic model suggested by Blei, Ng and Jordan (2003).

Structure:

- Documents are random mixtures over latent topics.
- A topic is a distribution over a fixed vocabulary (generated before the documents).
- A document may feature several topics.
**LDA: overnight archive**

<table>
<thead>
<tr>
<th>Topics</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dividends</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dividend</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scheduled</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>corporation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>september</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>june</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>march</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>november</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>august</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>trust</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>february</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>december</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>july</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>october</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Top 15 words</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dividend</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stock</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reasons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>focus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>great</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>investors</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>choice</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>jumps</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>session</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>growth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>momentum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rises</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>right</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>adds</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>moves</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>earnings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reasons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>follow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>history</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>indicator</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reaction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sensitive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>revenues</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beat</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>beats</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>surprise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>revenue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>misses</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>holdings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tale</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tape</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>continue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>higher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shares</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>focus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>march</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>surge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>strong</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>great</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>growth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>falls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>holdings</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fund</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>municipal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nuveen</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dividend</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>scheduled</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>date</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>high</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>new</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ahead</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>watch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>today</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>china</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dividend</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>growth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stocks</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>buy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>oil</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>higher</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>best</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>news</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>data</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lower</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ahead</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>watch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>today</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>china</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dividend</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>growth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Topics</td>
<td>Earnings</td>
<td>Equities</td>
<td>Asset mgmt</td>
<td>Econ. Outlook</td>
<td>Charts</td>
<td>Anal. Roundup</td>
<td>Sectors</td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>----------</td>
<td>------------</td>
<td>---------------</td>
<td>--------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td>3</td>
<td>Earnings</td>
<td>earnings</td>
<td>estimates</td>
<td>tale</td>
<td>fund</td>
<td>analyst</td>
<td>update</td>
</tr>
<tr>
<td>4</td>
<td>estimates</td>
<td>tape</td>
<td>follow</td>
<td>municipal</td>
<td>income</td>
<td>moving</td>
<td>sector</td>
</tr>
<tr>
<td>5</td>
<td>Asset</td>
<td>dividend</td>
<td>dividend</td>
<td>stocks</td>
<td>average</td>
<td>blog</td>
<td>energy</td>
</tr>
<tr>
<td>6</td>
<td>mgmt</td>
<td>ex</td>
<td>week</td>
<td>buy</td>
<td>moving</td>
<td>growth</td>
<td>health</td>
</tr>
<tr>
<td>7</td>
<td>Econ.</td>
<td>scheduled</td>
<td>bullish</td>
<td>oil</td>
<td>moving</td>
<td>new</td>
<td>care</td>
</tr>
<tr>
<td>8</td>
<td>Outlook</td>
<td>date</td>
<td>critical</td>
<td>higher</td>
<td>day</td>
<td>data</td>
<td>financial</td>
</tr>
<tr>
<td>9</td>
<td>Charts</td>
<td>data</td>
<td>energy</td>
<td>week</td>
<td>cross</td>
<td>energy</td>
<td>consumer</td>
</tr>
<tr>
<td>10</td>
<td>Anal.</td>
<td>critical</td>
<td>estimates</td>
<td>level</td>
<td>breaks</td>
<td>high</td>
<td>composite</td>
</tr>
<tr>
<td></td>
<td>Roundup</td>
<td>critical</td>
<td>revenues</td>
<td>low</td>
<td>key</td>
<td>high</td>
<td>closes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>critical</td>
<td>revenues</td>
<td>ahead</td>
<td>week</td>
<td>week</td>
<td>oil</td>
</tr>
<tr>
<td></td>
<td></td>
<td>critical</td>
<td>revenues</td>
<td>ahead</td>
<td>miss</td>
<td>miss</td>
<td>partners</td>
</tr>
<tr>
<td></td>
<td></td>
<td>critical</td>
<td>revenues</td>
<td>ahead</td>
<td>loss</td>
<td>loss</td>
<td>dis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>critical</td>
<td>revenues</td>
<td>ahead</td>
<td>roundup</td>
<td>revenues</td>
<td>dis</td>
</tr>
</tbody>
</table>
# LDA: trading-day archive

## Topics and most frequent words

<table>
<thead>
<tr>
<th>Topics</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topics</strong></td>
<td>Press rel.</td>
<td>Earnings 1</td>
<td>Funds</td>
<td>Option trades</td>
<td>Charts</td>
<td>Sectors</td>
<td>Dividends</td>
</tr>
<tr>
<td>analyst</td>
<td>earnings</td>
<td>etf</td>
<td>options</td>
<td>average</td>
<td>update</td>
<td>stock</td>
<td>trading</td>
</tr>
<tr>
<td>blog</td>
<td>revenues</td>
<td>detected</td>
<td>trading</td>
<td>moving</td>
<td>moving</td>
<td>options</td>
<td>bullish</td>
</tr>
<tr>
<td>analysts</td>
<td>beat</td>
<td>big</td>
<td>using</td>
<td>day</td>
<td>moving</td>
<td>trading</td>
<td>day</td>
</tr>
<tr>
<td>blog</td>
<td>estimates</td>
<td>inflow</td>
<td>cross</td>
<td>cross</td>
<td>sector</td>
<td>sector</td>
<td>energy</td>
</tr>
<tr>
<td>highlights</td>
<td>beats</td>
<td>inflows</td>
<td>interesting</td>
<td>critical</td>
<td>financial</td>
<td>financial</td>
<td>energy</td>
</tr>
<tr>
<td>releases</td>
<td>miss</td>
<td>outflows</td>
<td>earn</td>
<td>makes</td>
<td>technology</td>
<td>technology</td>
<td>technology</td>
</tr>
<tr>
<td>press</td>
<td>season</td>
<td>notable</td>
<td>commit</td>
<td>key</td>
<td>consumer</td>
<td>consumer</td>
<td>consumer</td>
</tr>
<tr>
<td>energy</td>
<td>report</td>
<td>buy</td>
<td>buy</td>
<td>breaks</td>
<td>health</td>
<td>health</td>
<td>health</td>
</tr>
<tr>
<td>press</td>
<td>view</td>
<td>annualized</td>
<td>available</td>
<td>key</td>
<td>care</td>
<td>care</td>
<td>care</td>
</tr>
<tr>
<td>energy</td>
<td>beats</td>
<td>large</td>
<td>begin</td>
<td>market</td>
<td>mid</td>
<td>market</td>
<td>mid</td>
</tr>
<tr>
<td>group</td>
<td>estimates</td>
<td>noteworthy</td>
<td>begin</td>
<td>day</td>
<td>market</td>
<td>market</td>
<td>market</td>
</tr>
<tr>
<td>holdings</td>
<td>beats</td>
<td>alert</td>
<td>purchase</td>
<td>crossover</td>
<td>day</td>
<td>crossover</td>
<td>crossover</td>
</tr>
<tr>
<td>international</td>
<td>miss</td>
<td>experiences</td>
<td>october</td>
<td>alert</td>
<td>laggards</td>
<td>laggards</td>
<td>laggards</td>
</tr>
<tr>
<td>high</td>
<td>season</td>
<td>experiences</td>
<td>november</td>
<td>alert</td>
<td>oil</td>
<td>oil</td>
<td>oil</td>
</tr>
<tr>
<td>american</td>
<td>report</td>
<td>Ishares</td>
<td>january</td>
<td>option</td>
<td>corp</td>
<td>corp</td>
<td>corp</td>
</tr>
<tr>
<td>loss</td>
<td>view</td>
<td>etfs</td>
<td>november</td>
<td>morning</td>
<td>corp</td>
<td>corp</td>
<td>corp</td>
</tr>
<tr>
<td>week</td>
<td>store</td>
<td>spdr</td>
<td>morning</td>
<td>morning</td>
<td>corp</td>
<td>corp</td>
<td>corp</td>
</tr>
<tr>
<td>airlines</td>
<td>sales</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Top 15 words**

- analyst
- earnings
- etf
- detected
- big
- inflow
- inflows
- outflows
- notable
- large
- noteworthy
- alert
- experiences
- Ishares
- etfs
- spdr
### Topics and most frequent words

<table>
<thead>
<tr>
<th>3</th>
<th>Funds</th>
<th>ETF detected big inflows inflows outflows notable large noteworthy alert experiences ishares etfs spdr</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Option trades options trading using week interesting earn commit buy annualized available begin purchase october november</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Charts average moving day cross bullish notable critical makes breaks key crosses level crossover alert option</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Sectors update sector energy financial sector consumer health care mid market after noon day laggards oil</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Dividends stock reminder market preferred today series news ex cumulative dividend interesting corp round up market gains higher</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Equities stocks buy new strong oil mid sell etfs european adr s company international group systems technology holdings technologies</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Earnings 2 indicator earnings follow history reaction sensitive corp preferred dma dividend</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Share types shares cross earnings yield history series mark preferred dma dividend</td>
<td></td>
</tr>
</tbody>
</table>

---
Option markets’ reaction to sentiment

- Fixed-effect panel regression with IV

\[ OC_{it} = \alpha + \gamma_i + \beta_1 B_{it} + \beta_2^T X_{it} + \varepsilon_{it} \]  

- \( OC_{it} \in \{\text{Skew}_{it}, \text{IVol}_{it}, \text{OTM}_{it}\} \): option characteristic
- \( X_{it} \): the vector of control variables

More Information
Endogeneity

- Sentiment for single stocks and reaction in options market could be due to a common cause.
- Need to assert that NASDAQ news/articles are the only source of news.
- Idea:
  - Use lagged $B_{i,t-1}$, $B_{idx,t-1}$, $BN_{idx,t-1}$ as instruments
Panel regressions

**OCs and sentiment in trading hours**

- $B_i$
- $Skew$
- $OTM$
- $IVol$

**Table:** Significance codes
- [ ] 0.01
- [ ] 0.05
- [ ] 0.1
- [ ] 0.01
- [ ] 0.05
- [ ] 0.1

- IV regressions with constant, fixed effects, and FF1-5 factors
- instrument: $B_{i,t-1}$
- Blue (negative sign); Red (positive sign)
Panel regressions

OCs and sentiment in trading hours

\[
\begin{array}{ccc}
B_i & B_{idx} & BN_{idx} \\
Skew & \text{Blue} & \text{Red} \\
OTM & \text{Blue} & \text{Red} \\
IVol & \text{Blue} & \text{Red}
\end{array}
\]

Table: Significance codes

- 0.01
- 0.05
- 0.1
- 0.01
- 0.05
- 0.1

- IV regressions with constant, fixed effects, and FF1-5 factors
- instrument: \( B_{i,t-1}, B_{idx,t-1}, BN_{idx,t-1} \)
- Blue (negative sign); Red (positive sign)
Option markets’ reaction: summary

- Standard endogeneity tests (Durbin, Hausman-Wu) reject that $B_{it}$ is exogenous
- $Skew$, $IVol$ and $OTM$ react to investor sentiment
- Higher $B$ results in a flatter $Skew$, lower $OTM$ and $IVol$
- Higher $B_{idx}$ results in a flatter $Skew$, lower $OTM$ and $IVol$
- Higher $BN_{idx}$ results in a steeper $Skew$, higher $OTM$ and $IVol$
Panel regressions

Stock return predictability:
Option variables v.s. sentiment index

Pooled OLS regressions

\[ R_{i,t+1} = \alpha + \beta_1 OC_{it} + \beta_2 B_{i,t} + \beta_3 B_{idx,t} + \beta_4 B_{on, idx,t} + \beta_5 B_{on, i,t} + \beta_6 B_{on, idx,t} + \beta_7 B_{on, idx,t} + \beta_8 X_{it} + \epsilon_{it} \]

- Xing et al. (JFQA, 2010) only use \( OC_{it} \)
- Incremental predictability from sentiment index
Stock return predictability: 
Option variables

\[ S_{\text{Skew OTM IVol}} \]

\[ R_{i,t+1} \]

\[ R_{i,t+1} \]

\[ R_{i,t+1} \]

Table: Significance codes

- Blue (negative sign); Red (positive sign)

- Includes FF1-5, lagged return, idiosyncratic and market volatility
Stock return predictability: Option variables and sentiment

\[ R_{i,t+1}, R_{i,t+1}, R_{i,t+1} \]

Table: Significance codes

- Blue (negative sign); Red (positive sign)

- Includes FF1-5, lagged return, idiosyncratic and market volatility
Stock return predictability ctd

- Confirms Xing et al. (JFQA, 2010)’s results on the predictability of Skew
- Stock-specific sentiment insignificant
- Negative aggregate trading and overnight sentiment carry significant predictive content in presence of options market variables
- Aggregate overnight sentiment is a good predictor too.
Decompose option variables:  
Sentiment-related v.s. non-public part

Extract sentiment component from option market variables.

- Regress OC on sentiment and controls to get residuals:

\[
OC_{i,t} = \alpha + \theta^\top B_t + \beta^\top X_{i,t} + \epsilon_{OC,t}^i
\]

- \(\{Skew_{i,t}, Put_{i,t}, IV_{i,t}\} \subseteq OC_{i,t}\).
- \(B_t = (B_{i,t}, B_{idx,t}, BN_{idx,t}, B_{i, on}^{on}, B_{idx,t}^{on}, BN_{idx,t}^{on})^\top\).

- \(\epsilon_{OC,t}^i\): residual term as a proxy for non-public information embedded in options data
Use residuals in the regression:

Pooled OLS regressions

\[ R_{i,t+1} = \alpha + \beta_1 \epsilon_{OC,t}^i + \beta_2 B_{i,t} + \beta_3 B_{idx,t} + \beta_4 B_{N_{idx,t}} + \beta_5 B_{on_{i,t}} + \beta_6 B_{on_{idx,t}} + \beta_7 B_{N_{on_{idx,t}}} + \beta_8^T X_{it} + \varepsilon_{it} \]
Panel regressions

**Stock return predictability: Option variables and sentiment**

\[ R_{i,t+1}, R_{i,t+1}, R_{i,t+1} \]

- \( \epsilon_{SKEW}^i \)
- \( \epsilon_{i OTM}^i \)
- \( \epsilon_{i IVol}^i \)
- \( B_i \)
- \( B_{idx} \)
- \( B_{N idx} \)
- \( B_{on}^i \)
- \( B_{on idx}^i \)
- \( B_{on N idx}^i \)

**Table:** Significance codes

- Blue (negative sign); Red (positive sign)

- Includes FF1-5, lagged return, idiosyncratic and market volatility
- Includes FF1-5, lagged return, idiosyncratic and market volatility
Source of the predictability ctd

- Sentiment-adjusted OCs remain significant
- Thus some information embedded in options markets data contains information other than sentiment
- Sentiment indices remain significant.
- Stock-specific bullishness not important.
Market consensus and stock returns

- data yield a cross section of firm-level sentiment measures
- observations are varying over time
- how does dispersion of sentiment affect stock returns?
  - low dispersion: cross-sectionally unequivocal sentiment
  - high dispersion: cross-sectionally differing sentiment
- implications unclear:
  - Miller (1977): dispersion could lead to negatively related returns if pessimists stay out of the market due to short sale constraints
  - Varian (1985); Cujean and Hasler (2016): investors demand compensation, e.g. due to adverse selection.
- measure dispersion by cross-sectional standard deviation and include in predictive regressions
Cross-section of $B_i$
Stock return predictability:
Option variables and sentiment

Table: Significance codes

- Includes FF1-5, lagged return, idiosyncratic and market volatility
- Blue (negative sign); Red (positive sign)
Market consensus and stock returns

- sentiment dispersion commands a high positive risk premium in the presence of market/ idiosyncratic volatility
- indeed sentiment dispersion and market volatility are only weakly correlated
- investors demand compensation for holding assets when sentiment is dispersed
- lends support to Varian (1985) / Cujian and Hasler (2016) among others
Trading

- Xing et al. (2010) show OC based trading strategies yield positive returns.
- Do OC strategies after partialling out sentiment do better?
- Strategy:
  - Group data of 97 firms into deciles according to OC / OC residuals
  - create long-short portfolios on the extreme deciles.
<table>
<thead>
<tr>
<th></th>
<th>Skew residual</th>
<th></th>
<th></th>
<th>Skew</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Long-Short</td>
<td>$FF_5$</td>
<td>$FF_3$</td>
<td>Long-Short</td>
<td>$FF_5$</td>
<td>$FF_3$</td>
</tr>
<tr>
<td>P value</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.004</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>Ann. Return</td>
<td>0.43</td>
<td>0.45</td>
<td>0.45</td>
<td>0.43</td>
<td>0.44</td>
<td>0.44</td>
</tr>
<tr>
<td>Daily Vol. (in bp)</td>
<td>86.25</td>
<td></td>
<td></td>
<td>92.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ann. Vol.</td>
<td>0.14</td>
<td></td>
<td></td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily Sharpe Ratio</td>
<td>0.17</td>
<td></td>
<td></td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ann. Sharpe Ratio</td>
<td>3.18</td>
<td></td>
<td></td>
<td>2.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IV residual</td>
<td></td>
<td></td>
<td>IV</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long-Short</td>
<td>$FF_5$</td>
<td>$FF_3$</td>
<td>Long-Short</td>
<td>$FF_5$</td>
<td>$FF_3$</td>
</tr>
<tr>
<td>Daily Return (in bp)</td>
<td>12.41</td>
<td>12.54</td>
<td>12.57</td>
<td>6.79</td>
<td>7.14</td>
<td>7.26</td>
</tr>
<tr>
<td>P value</td>
<td>0.009</td>
<td>0.010</td>
<td>0.010</td>
<td>0.181</td>
<td>0.121</td>
<td>0.141</td>
</tr>
<tr>
<td>Ann. Return</td>
<td>0.36</td>
<td>0.37</td>
<td>0.37</td>
<td>0.19</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>Daily Vol. (in bp)</td>
<td>88.67</td>
<td></td>
<td></td>
<td>99.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ann. Vol.</td>
<td>0.14</td>
<td></td>
<td></td>
<td>0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily Sharpe Ratio</td>
<td>0.14</td>
<td></td>
<td></td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ann. Sharpe Ratio</td>
<td>2.59</td>
<td></td>
<td></td>
<td>1.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Put residual</td>
<td></td>
<td></td>
<td>Put</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long-Short</td>
<td>$FF_5$</td>
<td>$FF_3$</td>
<td>Long-Short</td>
<td>$FF_5$</td>
<td>$FF_3$</td>
</tr>
<tr>
<td>Daily Return (in bp)</td>
<td>7.43</td>
<td>7.86</td>
<td>7.70</td>
<td>6.52</td>
<td>6.92</td>
<td>6.87</td>
</tr>
<tr>
<td>P value</td>
<td>0.098</td>
<td>0.090</td>
<td>0.098</td>
<td>0.178</td>
<td>0.118</td>
<td>0.140</td>
</tr>
<tr>
<td>Ann. Return</td>
<td>0.20</td>
<td>0.22</td>
<td>0.21</td>
<td>0.18</td>
<td>0.19</td>
<td>0.19</td>
</tr>
<tr>
<td>Daily Vol. (in bp)</td>
<td>85.66</td>
<td></td>
<td></td>
<td>94.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ann. Vol.</td>
<td>0.14</td>
<td></td>
<td></td>
<td>0.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily Sharpe Ratio</td>
<td>0.09</td>
<td></td>
<td></td>
<td>0.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ann. Sharpe Ratio</td>
<td>1.51</td>
<td></td>
<td></td>
<td>1.19</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

- We connect investor sentiment distilled from public news with equity and equity options markets.
- Options markets react to firm-level sentiment and aggregate sentiment.
- Relevance of inside information in option data after partialling out sentiment information from option data.
- Negative bullishness indices are important regressors in predictive regressions.
- Market consensus carries a positive risk premium.
- OC residual-based trading strategies slightly outperform pure OC based strategies.
- Results robust to lexicon projection techniques.
Textual Sentiment, Option Information and Stock Return Predictability

Cathy Yi-Hsuan Chen
Matthias R. Fengler
Wolfgang Karl Härdle
Yanchu Liu

Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin
University of St. Gallen, Switzerland
Lingnan College, Sun Yat-sen University, China

http://lvb.wiwi.hu-berlin.de
http://www.mathstat.unisg.ch/
http://www.lingnan.sysu.edu.cn
Antweiler, W. and Frank, M. Z.  
*Is All That Talk Just Noise?*  
J. Finance, 2004

Dennis, P. and S. Mayhew.  
*Risk-Neutral Skewness: Evidence from Stock Options*  
J. Financial Quant. Anal., 2002

Fama, E. and K. French.  
*A Five-Factor Asset Pricing Model*  
J. Financial Econom., 2015

Garman, M. and Klass, M.  
*On the Estimation of Security Price Volatilities from Historical Data*  
J. Bus., 1980
Han, B.
Investor Sentiment and Option Prices
Rev. Financial Stud., 2008

Härdle, W. K. and Lee, Y. J. and Schäfer D. and Yeh Y. R.
Variable Selection and Oversampling in the Use of Smooth Support Vector Machines for Predicting the Default Risk of Companies
J. Forecast., 2009

Hu, M. and Liu, B.
Mining and Summarizing Customer Reviews
10th ACM SIGKDD, 2004

Loughran, T. and McDonald, B.
When is a liability not a liability?
J. Finance, 2011
Malo, Pekka and Sinha, Ankur and Korhonen, Pekka and Wallenius, Jyrki and Takala, Pyry
*Good debt or bad debt*

Wilson, T. and Wiebe, J. and Hoffmann, P.
*Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis*
HLT-EMNLP, 2005

Xing, Y., X. Zhang and R. Zhao.
*What does the Individual Option Volatility Smirk Tell Us about Future Equity Returns*
J. Financial Quant. Anal., 2010

Zhang, J., Chen C. Y., Härdle, W. K. and Bommes, E.
*Distillation of News into Analysis of Stock Reactions*
Appendix
Correlation - Positive Sentiment

Figure: Monthly correlation between positive sentiment: BL and LM, BL and MPQA, LM and MPQA. Source: Zhang et al. (2016)
Correlation - Negative Sentiment

Figure: Monthly correlation between negative sentiment: BL and LM, BL and MPQA, LM and MPQA. Source: Zhang et al. (2016)
Tagging Example - BL

... McDonald’s has an obesity **problem** that continues to get **worse**. And that’s nothing to do with the food itself, but rather the huge menus that can now double as medieval fortification. For perspective, the chain’s menu has grown 70% since 2007. And while more offerings might seem **like** a **good** thing, large menus result in **slower** service and more flare-ups between franchisees and the corporation. **Bloated** menus raise inventory costs for smaller franchisees and **lead** to lower profit margins. The McDonald’s corporate franchise fee is based upon sales instead of profits, making it a smaller **concern** for the company overall. ...

3 **positive words** and 5 **negative words**
Tagging Example - LM

... McDonald’s has an obesity problem that continues to get worse. And that’s nothing to do with the food itself, but rather the huge menus that can now double as medieval fortification. For perspective, the chain’s menu has grown 70% since 2007. And while more offerings might seem like a good thing, large menus result in slower service and more flare-ups between franchisees and the corporation. Bloated menus raise inventory costs for smaller franchisees and lead to lower profit margins. The McDonald’s corporate franchise fee is based upon sales instead of profits, making it a smaller concern for the company overall. ...

1 positive word and 4 negative words
Web Scraping

- Databases to buy?
- Automatically extract information from web pages
- Transform unstructured data (HTML) to structured data
- Use HTML tree structure to parse web page
- Legal issues
  - Websites protected by copyright law
  - Prohibition of web scraping possible
  - Comply to Terms of Service (TOS)
Natural Language Processing (NLP)

- Text is unstructured data with implicit structure
  - Text, sentences, words, characters
  - Nouns, verbs, adjectives, ..
  - Grammar
- Transform implicit text structure into explicit structure
- Reduce text variation for further analysis
- Python Natural Language Toolkit (NLTK)
- TXTnlp
Tokenization

- **String**
  
  "McDonald’s has its work cut out for it. Not only are sales falling in the U.S., but the company is now experiencing problems abroad."

- **Sentences**
  
  "McDonald’s has its work cut out for it.",
  "Not only are sales falling in the U.S., but the company is now experiencing problems abroad."

- **Words**
  
  "McDonald”, "’s”, "has”, "its”, "work”, "cut”, "out”..."
Negation Handling

- “not good” ≠ “good”
- Reverse polarity of word if negation word is nearby
- Negation words
  - "n’t", "not", "never", "no", "neither", "nor", "none"
Part of Speech Tagging (POS)

- Grammatical tagging of words
  - dogs - noun, plural (NNS)
  - saw - verb, past tense (VBD) or noun, singular (NN)

- Penn Treebank POS tags

- Stochastic model or rule-based
Lemmatization

- Determine canonical form of word
  - dogs - dog
  - saw (verb) - see and saw (noun) - saw

- Reduces dimension of text
- Takes POS into account
  - Porter stemmer: saw (verb and noun) - saw
Loss Functions for Classification

- **Logistic: Logit**

  \[ L(y, s(X)) = \log(2)^{-1} \log[1 + \exp(-s(X)y)] \]  
  \[ (4) \]

- **Hinge: Support Vector Machines**

  \[ L(y, s(X)) = \max\{0, 1 - s(X)y\} \]  
  \[ (5) \]
Regularization Term

- **L2 norm**

\[
R(\beta) = 2^{-1} \sum_{i=1}^{p} \beta_i^2
\]  

(6)

- **L1 norm**

\[
R(\beta) = \sum_{i=1}^{p} |\beta_i|
\]  

(7)
RLM Example

Sentence 1: “The profit of Apple increased.”
Sentence 2: “The profit of the company decreased.”

\[ y = (1, -1) \] (8)
\[ X = \begin{pmatrix} x_1 & x_2 \\ \text{the} & 1 \\ \text{profit} & 2 \\ \text{of} & 1 \\ \text{Apple} & 1 \\ \text{increased} & 1 \\ \text{company} & 0 \\ \text{decreased} & 0 \end{pmatrix} \] (9)
$k$-fold Cross Validation (CV)

- Partition data into $k$ complementary subsets
- No loss of information as in conventional validation
- Stratified CV: equally distributed response variable in each fold

**Figure:** 3-fold Cross Validation
Oversampling

- Härdle (2009) Trade-off between Type I and Type 2 error in classification
- Balance size of neutral sentences and ones with polarity in sample
- Duplicate sentences within folds of stratified cross validation until the sample is balanced
Classification Error Rates

- Type I error rate $= \frac{FP}{FP + TP}$
- Type II error rate $= \frac{FN}{FN + TP}$
- Total error rate $= \frac{FN + FP}{TP + TN + FP + FN}$

with TP as true positive, TN as true negative, FP as false positive and FN as false negative.
Stochastic Gradient Descent (SGD)

- Approximately minimize loss function

\[ L(\theta) = \sum_{i=1}^{n} L_i(\theta) \]  \hspace{1cm} (10)

- Iteratively update

\[ \theta_i = \theta_{i-1} - \eta \frac{\partial L_i(\theta)}{\partial \theta} \]  \hspace{1cm} (11)
SGD Algorithm

1. Choose learning rate $\eta$
2. Shuffle data
3. For $i = 1, \ldots, n$, do:
   $$\theta_i = \theta_{i-1} - \eta \frac{\partial L_i(\theta)}{\partial \theta}$$

Repeat 2 and 3 until approximate minimum obtained.
SGD Example

$X \sim N(\mu, \sigma)$ and $x_1, \ldots, x_n$ as randomly drawn sample

$$
\min_{\theta} \ n^{-1} \sum_{i=1}^{n} (\theta - x_i)^2
$$

Update step

$$
\theta_i = \theta_{i-1} - 2\eta(\theta_{i-1} - x_i)
$$

Optimal gain

Set $2\eta = 1/i$ and obtain $\theta_n = \bar{x}$ with $\bar{x}$ as sample mean.
SGD Example ctd

Figure: Estimate Mean via SGD, $x_t \sim N(5, 1)$

$\eta \in \{1/t, 1/1000, 1/1500, 1/2000, 1/2500\}$
## Evaluation Supervised Learning

<table>
<thead>
<tr>
<th></th>
<th>Pred</th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td></td>
<td>1992</td>
<td>289</td>
<td>254</td>
<td>2,535</td>
</tr>
<tr>
<td>0</td>
<td>96</td>
<td></td>
<td>2134</td>
<td>305</td>
<td>2,535</td>
</tr>
<tr>
<td>1</td>
<td>105</td>
<td>469</td>
<td></td>
<td>1,961</td>
<td>2,535</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2,193</td>
<td>2,892</td>
<td>2,520</td>
<td></td>
<td>7,605</td>
</tr>
<tr>
<td>Precision</td>
<td>0.91</td>
<td>0.74</td>
<td>0.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recall</td>
<td>0.78</td>
<td>0.84</td>
<td>0.77</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Confusion Matrix - Supervised Learning with Oversampling
## Evaluation Unsupervised Learning

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True</strong></td>
<td><strong>-1</strong></td>
<td><strong>0</strong></td>
<td><strong>1</strong></td>
<td><strong>Total</strong></td>
<td></td>
</tr>
<tr>
<td><strong>-1</strong></td>
<td>213</td>
<td>289</td>
<td>12</td>
<td>514</td>
<td></td>
</tr>
<tr>
<td><strong>0</strong></td>
<td>200</td>
<td><strong>2,187</strong></td>
<td>148</td>
<td>2,535</td>
<td></td>
</tr>
<tr>
<td><strong>1</strong></td>
<td>111</td>
<td>772</td>
<td><strong>285</strong></td>
<td>1,168</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>524</td>
<td>3,248</td>
<td>445</td>
<td>4,217</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Precision</strong></th>
<th><strong>Recall</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.41</td>
<td>0.41</td>
</tr>
<tr>
<td>0.67</td>
<td>0.86</td>
</tr>
<tr>
<td>0.64</td>
<td>0.24</td>
</tr>
</tbody>
</table>

*Table: Confusion Matrix - Lexicon Projection*
LDA – details

Assumed process of generating a document:

1. Choose number of words $N$ (randomly, deterministically).

2. Draw a distribution over $K$ topics:
   \[ \theta \sim \text{Dir}(\alpha) \]

3. For each of the $N$ words $w_n$:
   3.1 Choose a topic from $z_n \sim \text{M}(\theta)$
   3.2 Choose a word from $p(w_n|z_n, \beta)$, a multinomial probability conditional on topic $z_n$ parametrized by
   \[ \beta = [\beta_{ij}] = p(w^j = 1|z^i = 1) \]
Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

where

\( p(\mathbf{w} | \alpha, \beta) = \int p(\boldsymbol{\theta} | \alpha) \left( N \prod_{n=1}^{\mathbf{N}} \sum_{\mathbf{z}_{\mathbf{N}}} \right) p(\mathbf{z}_{\mathbf{N}} | \boldsymbol{\theta}) p(\mathbf{w}_{\mathbf{N}} | \mathbf{z}_{\mathbf{N}}, \beta) ) \,
\end{align}

Finally, taking the product of the marginal probabilities of single documents, we obtain the probability of a corpus:

\[
p(\mathbf{D} | \alpha, \beta) = M \prod_{d=1}^{\mathbf{M}} \int p(\boldsymbol{\theta}_d | \alpha) \left( N_d \prod_{n=1}^{\mathbf{N}_d} \sum_{\mathbf{z}_{d\mathbf{N}}} \right) p(\mathbf{z}_{d\mathbf{N}} | \boldsymbol{\theta}_d) p(\mathbf{w}_{d\mathbf{N}} | \mathbf{z}_{d\mathbf{N}}, \beta) ) \, d\theta_d.
\]

The LDA model is represented as a probabilistic graphical model in Figure 1. As the figure makes clear, there are three levels to the LDA representation. The parameters \( \alpha \) and \( \beta \) are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables \( \boldsymbol{\theta}_d \) are document-level variables, sampled once per document. Finally, the variables \( \mathbf{z}_{d\mathbf{N}} \) and \( \mathbf{w}_{d\mathbf{N}} \) are word-level variables and are sampled once for each word in each document.

It is important to distinguish LDA from a simple Dirichlet-multinomial clustering model. A classical clustering model would involve a two-level model in which a Dirichlet is sampled once for a corpus, a multinomial clustering variable is selected once for each document in the corpus, and a set of words are selected for the document conditional on the cluster variable. As with many clustering models, such a model restricts a document to being associated with a single topic. LDA, on the other hand, involves three levels, and notably the topic node is sampled repeatedly within the document. Under this model, documents can be associated with multiple topics.

Structures similar to that shown in Figure 1 are often studied in Bayesian statistical modeling, where they are referred to as hierarchical models (Gelman et al., 1995), or more precisely as conditionally independent hierarchical models (Kass and Steffey, 1989). Such models are also often referred to as parametric empirical Bayes models, a term that refers not only to a particular model structure, but also to the methods used for estimating parameters in the model (Morris, 1983). Indeed, as we discuss in Section 5, we adopt the empirical Bayes approach to estimating parameters such as \( \alpha \) and \( \beta \) in simple implementations of LDA, but we also consider fuller Bayesian approaches as well.

Source: Blei et al. (2003)
Inference

- The estimation problem is to find the hidden topic structure over the set of documents given observed words.
- Need to approximate the posterior distribution, i.e., the conditional distribution of topics, topic proportions, and topic assignments given observed words.
- Posterior computation is achieved by Gibbs sampling, see Blei et al. (2012) for details.
A plot of Skew

Figure: Skew of Apple Inc. in the sample period
Control Variables

\( Ret_{it} \) - Stock \( i \)'s contemporaneous return
\( Volu_{it} \) - Stock \( i \)'s trading volume
\( OC_{it} \) - option characteristics of stock \( i \)
\( VIX_t \) - CBOE VIX

and Fama-French 5 factors (Fama and French (JFE, 2015))
Fama-French 5 factors

*FF1* - the Mkt factor: excess return on the market index

*FF2* - the SMB factor: (Small Minus Big) the average return on the nine small-stock portfolios minus that on the nine big-stock portfolios.

*FF3* - the HML factor: (High Minus Low) the average return on the two value-stock portfolios minus that on the two growth-stock portfolios.
Fama-French 5 factors ctd

*FF4* - the RMW factor: (Robust Minus Weak) the average return on the two robust operating profitability portfolios minus that on the two weak operating profitability portfolios

*FF5* - the CMA factor: (Conservative Minus Aggressive) the average return on the two conservative investment portfolios minus that on the two aggressive investment portfolios
VIX

- Implied volatility
- Measures market expectation of S&P 500
- Calculated by Chicago Board Options Exchange (CBOE)
- Measures 30-day expected volatility
- Calculated with put and call options with more than 23 days and less than 37 days to expiration
Variables Definitions

- **Skew**: difference between volume-weighted average of implied volatilities (IVs) of OTMP and ATMC:

  \[ SKEW_{it} = IV_{it}^{OTMP} - IV_{it}^{ATMC} \]

  **Example**

- **OTMP**: a put with moneyness between 0.8 and 0.95
- **ATMC**: a call with moneyness between 0.95 and 1.05
- **Moneyness**: ratio of the strike price to the stock price
- **Use delta as moneyness**
Variables Definitions ctd

- **IVol**: volume-weighted average of IVs of all the ATM options
- **OTM**: volume-weighted average of prices of OTM put options (moneyness between 0.8 and 0.95) *relative* to stock price
- **$B$**: degree of bullishness defined in (4), positive (negative) value implies positive (negative) sentiment
- **$BN = -I(B < 0)B$**, indicating negative sentiment