Tales of sentiment driven tails

Jozef Baruník
Cathy Yi-Hsuan Chen
Wolfgang Karl Härdle

Institute of Economic Studies
Charles University in Prague
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin

http://ies.fsv.cuni.cz
http://lwb.wiwi.hu-berlin.de
“Forget the dot-com boom with its irrational exuberance and the real estate bubble that was supposed to be invincible: Current market sentiment eclipses all of that”

Jeff Cox, CNBC, March 1 2017
Sentiment moves market

John Maynard Keynes (1936): markets can fluctuate wildly under the influence of investors’ “animal spirits,” which move prices in a way unrelated to fundamentals.
Sentiment can cause mispricing

Fifty years later...


- Uninformed noise traders base their decisions on sentiment
  - Greater mispricing (Stambaugh et al., 2012)
  - Excess volatility (Dumas et al., 2009)
“Now, the question is no longer, as it was a few decades ago, whether investor sentiment affects stock prices, but rather how to measure investor sentiment and quantify its effects.”

(Baker and Wurgler, 2007)
News moves markets

- Baker and Wurgler (2007) investor sentiment affects securities whose valuations are highly subjective
- Large literature Huang et al. (2014), Da et al. (2015), Shefrin (2007+)
- Zhang et al. (2016) textual sentiment provides incremental information about future stock reactions
Is average enough?

- Sentiment affects cross section of returns or volatility
- Grand average is OK for expected payoffs
- Though...
  - bear vs. bull markets
  - extreme negative vs. positive returns
Is average man enough?

Contrarians vs. Trend followers

Tales of sentiment driven tails
We already know that we can measure sentiment…

but how to quantify its effect on prices?

Tales of sentiment driven tails
Contribution

- Step forward from classical asset pricing (EU based)
- Provide decision-theoretic foundations of pricing in quantiles
- Link sentiment with quantiles of the return distributions
- Nonlinear dynamic quantile asset pricing model
- Confirm empirically on Panel of 100 US stocks
Outline

1. Motivation ✓
2. Theoretical Framework
3. Data Collection
4. Sentiment Projection
5. Calibration of weighting function
6. Quantile Panel Regressions
7. Outlook

Tales of sentiment driven tails
Classical asset pricing

Investor maximizes utility subject to budget constraint. The FOC (Euler equation):

$$E_F [M \times (1 + R)] = 1,$$

(1)

where $M$ is a pricing kernel (PK), or stochastic discount factor (SDF), $R$ is the total return on a risky asset with physical distribution $F(R)$. 


Probability weighting

Decisions under risk are more sensitive to changes in probability of events at extremes, Tversky and Kahneman (1992). Polkovnichenko and Zhao (2013) use the rank-dependent expected utility (RDEU) $\mathcal{U}(R) = E_F[u(R)g\{F(R)\}]$ with PK

$$M = u'(R)g\{F(R)\},$$

where $g\{F(R)\} = G'\{F(R)\}$ is a probability weighting function.

Euler equation reads as:

$$E_F[u'(R)g\{F(R)\} (1 + R)] = 1.$$
Theoretical Framework

Figure: Probability weighting function \( G(v) = \exp\left\{ -(-\beta \log v)^\alpha \right\} \) with \( \alpha = 0.7 \) and \( \beta = 0.6 \)

Tales of sentiment driven tails
Figure: Probability weighting function \( G(v) = \exp\left\{ -(-\beta \log v)^\alpha \right\} = v^{0.6} \)
(\( \alpha = 1 \) and \( \beta = 0.6 \))

Tales of sentiment driven tails
A route towards quantile preferences

- $X$ is preferred to $Y$ if there exist utility function $U(.)$ such that

$$X \succeq Y \text{ iff } E_F[U(X)] \geq E_F[U(Y)] \quad (4)$$


$$X \succeq Y \text{ iff } Q_\tau[U(X)] \geq Q_\tau[U(Y)] \quad (5)$$

- Maximising lower quantile is more risk-averse than higher quantile (example of portfolio), de Castro et al. (2017)
Example

Utility function \( u(x) \overset{\text{def}}{=} x \)

\[
X = \begin{cases} 
10^7 & \text{with } p = 10^{-6} \\
-1 & \text{with } q = 1 - p
\end{cases} \quad Y = \begin{cases} 
10 & \text{with } p = 9/10 \\
-1 & \text{with } q = 1 - p
\end{cases}
\]

\( X \preceq_{E} Y \) since \( E[X] = 9 + 10^{-6} \) and \( E[Y] = 8 + 9/10 \)

\( Q_{\tau}(X) \overset{\text{def}}{=} \inf\{\alpha \in \mathbb{R} : \mathbb{P}(X \leq \alpha) \geq \tau\} \)

\[
X \begin{cases} 
\equiv_{Q_{\tau}} Y & \text{for } \tau \leq 1/10 \\
\preceq_{Q_{\tau}} Y & \text{for } 1/10 < \tau \leq 1 - 10^{-6} \\
\preceq_{Q_{\tau}} Y & \text{for } \tau > 1 - 10^{-6}
\end{cases}
\]
A route towards a (dynamic) quantile model

Instead of classical preferences, look at an agent maximizing her stream of the future quantile utilities. For a given \( \tau \in (0,1) \), Euler equation reads:

\[
Q_\tau \left[ u'(R)g(v) (1+R) \right] = 1, \tag{6}
\]

where \( v = F(R) \),

\[
G(\cdot) : [0,1] \to [0,1] \text{ probability weighting fct and } g(\cdot) = G'(\cdot).
\]

Can we relate \( g(\cdot) \) to sentiment?

Tales of sentiment driven tails
Probability weighting function and sentiment

Prelec (1998) weighting function:

\[ G(v) = G(\alpha, \beta; v) = \exp\{-(-\beta \log v)^{\alpha}\} \]  \hspace{1cm} (7)

\(\alpha, \beta\) parameters govern the shape of \(G(\cdot)\).
Link sentiment $S_t$ to $\beta_t$:

$$\beta_t = \beta(S_t, \rho) = \exp\{-\rho(S_t^{-1} - 1)\} - 1$$

Figure: $\beta$ versus $S_t$ for $\rho = -0.1$ and $\rho = -0.05$
Fix $\alpha = 1$ to impose monotonicity and compute $v_t = (\text{rank } R_t)/n$

\[ G(v_t, S_t) = v_t^{\beta(S_t, \rho)} = v_t^{\exp\{-\rho(S_t^{-1}-1)\}-1} \]
\[ G(v_t, S_t) = \exp\{(\exp\{-\rho(S_t^{-1}-1)\} - 1) \log v_t\} \]
A dynamic quantile model with sentiments

Equation (6) is beneficial, since it can be log-linearized as for a general random variable $W$, $Q_\tau[\log(W)] = \log(Q_\tau[W])$. Hence

$$Q_\tau[u'(R_t)g(\nu_t, S_t) (1 + R_{t+1})] = 1 \quad (10)$$

considering power utility function:

$$Q_\tau[-\gamma \log(R_t) + \log\{g(\nu_t, S_t)\} + \log(1 + R_{t+1})] = 0. \quad (11)$$

One can estimate the parameter driving $g(\nu_t, S_t)$ with nonlinear quantile regression.

How to estimate sentiment $S_t$?
Data

- Panel of 100 most liquid constituents of S&P 500 stocks
- Sentiment variables: distilled from Nasdaq articles

Nasdaq Articles

- Terms of Service permit web scraping
- Currently > 580k articles between October 2009 and January 2017
- Data available at RDC
There is a lot of news...
Dimensions of News

- **Source of news**
  - Official channel: government, federal reserve bank/central bank, financial institutions
  - Internet: blog, social media, message board

- **Content of news: signal vs. noise**

- **Type of news**
  - Scheduled vs. non-scheduled
  - Expected vs. unexpected
  - Specific-event vs. continuous news flows
The Power of Words: Textual Analytics

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

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

Tales of sentiment driven tails
Supervised Projection

We use supervised projection (Zhang et al., JBES, 2016)

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

Articles → Scraping → NLP → Projection → Sentiment

- URL
- Author
- Symbol
- Date
- Text

- Token
- Negation
- POS
- Lemmata

- Unsupervised
  - BL
  - LM

- Supervised
  - SM

Tales of sentiment driven tails
Lexicon-based Sentiment

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

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

(12)

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

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

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

for sentence \( i \)

- Measure sentiment on sentence level
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 \(s(X) = \beta^\top X\) with \(\beta \in \mathbb{R}^p\)

Regularized training error:

\[
\frac{1}{n} \sum_{i=1}^{n} \underbrace{L\{y_i, s(X)\}}_{\text{Loss Function}} + \lambda \underbrace{R(\beta)}_{\text{Regularization Term}}
\]  

(14)

with hyperparameter \(\lambda \geq 0\)

Tales of sentiment driven tails
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

Tales of sentiment driven tails
Bullishness

\[ B = \log \left\{ \frac{1 + n^{-1} \sum_{j=1}^{n} I(Pol_j = 1)}{1 + n^{-1} \sum_{j=1}^{n} I(Pol_j = -1)} \right\} \] (15)

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 \( BN_{i,t} = I(B_{i,t} < 0)B_{i,t} \)
Calibration of probability weighting functions

Estimate $\rho_\tau$ using nonlinear quantile regressions. Employ power utility $u(R) = R^{1-\gamma}/(1 - \gamma)$.

$$Q_\tau [-\gamma \log(R_t) + \log\{g(v_t, S_t)\} + \log(1 + R_{t+1})] = 0, \quad (16)$$

recall

$$g(v, S) = G'(1, \beta; v) = \beta v^{\beta-1},$$

$$\beta = \beta(S, \rho) = \exp\{-\rho(S^{-1} - 1)\} - 1.$$ 

where

$$g(v_t, S_t) = (\exp\{-\rho(1/S_t - 1)\} - 1) v_t^{\exp\{-\rho(1/S_t-1)\}-2}.$$
Calibration of probability weighting functions

Expect $\rho_\tau$ to differ across $\tau$ since sentiment distorts beliefs of a $\tau$-quantile preference maker.

**Figure:** Variation over firms for $\gamma = 20$
Calibration of probability weighting functions

Message: $\rho$ bigger for smaller $\tau$

Tales of sentiment driven tails
Higher values of $\rho_\tau$ in the left tail indicate that large negative sentiment is connected to higher overweighting of the PK.
Empirical Results: Pricing tails with Sentiment

- We propose a dynamic quantile asset pricing model

\[ Q_{\tau} \left[ \tilde{M}_t \times (1 + R_{t+1}) + 1 \right] = 0 \]

- with \( \tilde{M}_t = \exp(-\alpha_{\tau} - \beta_{S,\tau} S_t - FF_t^T \beta_{FF,\tau} - X_t^T \beta_{X_t,\tau}) \),

  FF=Fama French 5 factors

  \( X_t \) - control variables including idiosyncratic factors

- Factors are proxy for aggregate consumption

Tales of sentiment driven tails
Empirical Results: Pricing tails with Sentiment

After log-linearization, we arrive to a simple linear model

\[ Q_\tau \left[ \log(1 + R_{t+1}) - \alpha_\tau - \beta_{S,\tau} S_t - FF_t^\top \beta_{FF,\tau} - X_t^\top \beta_{X_t,\tau} \right] = 0 \] (17)

implying

\[ Q_\tau \left[ \log(1 + R_{t+1}) \right] = \alpha_\tau + \beta_{S,\tau} S_t + FF_t^\top \beta_{FF,\tau} + X_t^\top \beta_{X_t,\tau} \] (18)

with \textit{FF} \ Fama-French Factors

Tales of sentiment driven tails
Empirical Results: Sentiment as factor

- Aggregate market sentiment as possible risk factor.
- Control also for firm-specific sentiment and volatility.
- Negative sentiment captures “fear”, related to VIX (Da et al., 2015).
- Following high investor sentiment, aggregate returns are low (Baker and Wurgler, 2007).
- Overly optimistic beliefs about future cash flows is not justified by fundamentals.
A dynamic quantile model with sentiment

Linear asset pricing model

\[
Q_\tau(r_{i,t+1}) = \alpha_{i,\tau} + \beta_{1,\tau} B_{i,t} + \beta_{2,\tau} \sigma_{i,t} + \beta_{3,\tau} |BN_t| + FF_t^\top \beta_{FF,\tau} \tag{19}
\]

with \( \sigma_{i,t} \) Garman & Klass (1980) range-based volatility

- \(|BN_t|\) proxy for \( S_t \) (hence \( \beta_S \) from (18) is here \( \beta_3 \))
- \( B_{i,t} \) proxy for idiosyncratic sentiment
- \( \sigma_{i,t} \) proxy for volatility

\( B_{i,t}, \sigma_{i,t} \) control variables, contained in the matrix \( X \) in (18).
Eq (20) tests if sentiment prices quantiles of the excess asset returns.

- Coefficients capture marginal effects of pricing factors
- Coefficients varying across $\tau$ imply marginal effect
- Coefficients constant over $\tau$: EU works?
A dynamic quantile model with sentiment

Linear asset pricing model with Fama-French Factors

\[ Q_\tau(r_{i,t+1}) = \alpha_{i,\tau} + \beta_{1,\tau} B_{i,t} + \beta_{2,\tau} \sigma_{i,t} + \beta_{3,\tau} |BN_t| + FF_t^T \beta_{FF,\tau} \] (20)

with \( \sigma_{i,t} \) - Garman & Klass (1980) range-based volatility.

(20) tests if sentiment prices quantiles of the excess asset returns.

- Coefficients capture marginal effects of pricing factors
- Coefficients varying across \( \tau \) imply marginal effect
- Coefficients constant over \( \tau \): EU works?

Tales of sentiment driven tails
Results

Estimate (20) via QR

- Panel of 100 most liquid constituents of S&P 500 stocks
- 10 main sectors
- Check sentiments across $\tau$

Tales of sentiment driven tails
Quantile Panel Regressions

Results: Panel of 100 stocks

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20)

Further Graphics

Tales of sentiment driven tails
Results: Panel of 100 stocks

Figure: Estimates for $\beta_{i,\tau}$ together with box plots showing individual estimates with univariate individual $i=1,\ldots,100$ QR estimates
Empirical Results

- Tails are strongly influenced
- Sentiment and volatility effects similarly
- $\beta_\tau \neq 0$ for most of the $\tau$s
- Asymmetric impact of market sentiment
- Holds even after control for firm specific sentiment
- Increase in negative bullishness has positive effect on right tail, and negative effect on left tail
- Contrary to literature, factors explain daily data in quantiles
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Further Graphics

Tales of sentiment driven tails
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Tales of sentiment driven tails
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20)

Tales of sentiment driven tails
Results: Sectors

**Figure:** Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Tales of sentiment driven tails
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Tales of sentiment driven tails
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20)

Tales of sentiment driven tails
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) [Further Graphics]

Tales of sentiment driven tails
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20) Further Graphics

Tales of sentiment driven tails
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20)

Tales of sentiment driven tails
Results: Sectors

Figure: Estimates for $\beta_{i,\tau}$ from eq. (20) for $\tau \in (0, 1)$

Full estimates of eq. (20)

Tales of sentiment driven tails
Summary

- Tales of sentiment driven tails
- Dynamic quantile model for asset pricing with sentiment
- Investor sentiment distilled from public news with cross-section of future return’s quantiles.
Tales of sentiment driven tails

Jozef Baruník
Cathy Yi-Hsuan Chen
Wolfgang Karl Härdle

Institute of Economic Studies
Charles University in Prague
Ladislaus von Bortkiewicz Chair of Statistics
Humboldt-Universität zu Berlin

http://ies.fsv.cuni.cz
http://lvb.wiwi.hu-berlin.de
Antweiler, W. and Frank, M. Z.  
Is All That Talk Just Noise?  
J. Finance, 2004

Baker, M., and J. Wurgler.  
Investor sentiment and the cross-section of stock returns  
Journal of Finance, 2006

de Castro, L. I. and A. F. Galvao  
Dynamic quantile models of rational behavior  
2017

Da, Z., Engelberg J. and Gao, P.  
The Sum of All FEARS Investor Sentiment and Asset Prices  
Review of Financial Studies, 2015


Tales of sentiment driven tails
Keynes, J. M.  
*The general theory of employment, interest and money.*  
London: Macmillan, 1936

Koenker, R.  
*Quantile regression for longitudinal data.*  
Journal of Multivariate Analysis, 2004

Manski, C.F.  
*Ordinal utility models of decision making under uncertainty.*  
Theory and Decision, 1988

Polkovnichenko, V. and Zhao, F.  
*Probability weighting functions implied in options prices*  

Tales of sentiment driven tails
Prelec, D.
*The probability weighting function*
Econometrica, 1998

Rostek, M.
*Quantile maximization in decision theory.*
The Review of Economic Studies, 2010

Stambaugh, R.F., Yu, J.F., Yuan, Y.
*The short of it: Investor sentiment and anomalies.*

Tversky, A. and Kahneman, D.
*Advances in prospect theory: Cumulative representation of uncertainty*
Journal of Risk and uncertainty, 1992
Zhang, J., Chen C. Y., Härdle, W. K. and Bommes, E.
*Distillation of News into Analysis of Stock Reactions*
Appendix
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**

TXTMcDbm

Article source

Tales of sentiment driven tails
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

TXTMcDlm

Tales of sentiment driven tails
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**

  \[ L\{y, s(X)\} = \log(2)^{-1} \log[1 + \exp\{-s(X)y\}] \]  
  \hspace{1cm} (21)

- **Hinge: Support Vector Machines**

  \[ L\{y, s(X)\} = \max\{0, 1 - s(X)y\} \]  
  \hspace{1cm} (22)
Regularization Term

- L2 norm

\[ R(\beta) = 2^{-1} \sum_{i=1}^{p} \beta_i^2 \]  \hspace{1cm} (23)

- L1 norm

\[ R(\beta) = \sum_{i=1}^{p} |\beta_i| \]  \hspace{1cm} (24)
Appendix

**RLM Example**

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

\[ y = (1, -1) \quad (25) \]

\[ X = \begin{pmatrix}
    1 & 2 \\
    1 & 1 \\
    1 & 1 \\
    1 & 0 \\
    0 & 1 \\
  \end{pmatrix} \quad (26) \]
**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

![Diagram of 3-fold Cross Validation](image.png)

**Figure**: 3-fold Cross Validation

Tales of sentiment driven tails
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 = $\frac{FP}{FP + TN}$
- 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

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

(27)

- Iteratively update

\[ \theta_i = \theta_{i-1} - \eta \frac{\partial L_i(\theta)}{\partial \theta} \]  

(28)
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, $x_t \sim N(5, 1)$

$\eta \in \{1/t, \ 1/1000, \ 1/1500, \ 1/2000, \ 1/2500\}$

Tales of sentiment driven tails
# Evaluation Supervised Learning

<table>
<thead>
<tr>
<th>True</th>
<th>Pred</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>1,983</td>
<td>298</td>
<td>254</td>
<td></td>
<td>2,535</td>
</tr>
<tr>
<td>0</td>
<td>96</td>
<td>2,134</td>
<td>305</td>
<td></td>
<td>2,535</td>
</tr>
<tr>
<td>1</td>
<td>105</td>
<td>469</td>
<td>1,961</td>
<td></td>
<td>2,535</td>
</tr>
<tr>
<td>Total</td>
<td>2,184</td>
<td>2,901</td>
<td>2,520</td>
<td></td>
<td>7,605</td>
</tr>
</tbody>
</table>

**Table:** Confusion Matrix - Supervised Learning with Oversampling

Tales of sentiment driven tails
## Abbreviations

<table>
<thead>
<tr>
<th>Sector</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consumer Discretionary</td>
<td>CD</td>
</tr>
<tr>
<td>Consumer Staples</td>
<td>CS</td>
</tr>
<tr>
<td>Energy</td>
<td>EN</td>
</tr>
<tr>
<td>Financials</td>
<td>FI</td>
</tr>
<tr>
<td>Health Care</td>
<td>HC</td>
</tr>
<tr>
<td>Industrials</td>
<td>IN</td>
</tr>
<tr>
<td>Information Technology</td>
<td>IT</td>
</tr>
<tr>
<td>Materials</td>
<td>MA</td>
</tr>
<tr>
<td>Telecommunication</td>
<td>TE</td>
</tr>
<tr>
<td>Utilities</td>
<td>UT</td>
</tr>
</tbody>
</table>

**Table:** Sector Abbreviations
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 cont.

*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
Garman & Klass range-based volatility

\[
\sigma_{i,t} = 0.511(u - d)^2 - 0.019\{c(u + d) - 2ud\} - 0.838c^2 \tag{29}
\]

with \(u = \log(P_{i,t}^H) - \log(P_{i,t}^L)\)

\(d = \log(P_{i,t}^L) - \log(P_{i,t}^O)\)

\(c = \log(P_{i,t}^C) - \log(P_{i,t}^O)\),

where the \(P_{i,t}^H, P_{i,t}^L, P_{i,t}^O, P_{i,t}^C\) are the daily highest, lowest, opening and closing stock prices.
Results: Panel of 100 stocks

Tales of sentiment driven tails
Results: Sectors

Tales of sentiment driven tails
Results: Sectors

Tales of sentiment driven tails
Results: Sectors

**Sentiment**

**Volatility**

**|SM_neg| index**

**FF_1**

**FF_2**

**FF_3**

**FF_4**

**FF_5**

Tales of sentiment driven tails
Results: Sectors

Tales of sentiment driven tails
Results: Sectors

Industrials

Sentiment

Volatility

|SM_neg| index

FF_1

FF_2

FF_3

FF_4

FF_5

Tales of sentiment driven tails
Results: Sectors

Tales of sentiment driven tails
Results: Sectors

Tales of sentiment driven tails
Results: Sectors

Tales of sentiment driven tails
Appendix

Results: Sectors

Tales of sentiment driven tails
Results: Sectors

Telecommunication Services

Sentiment

Volatility

|SM_neg| index

FF_1

FF_2

FF_3

FF_4

FF_5

Tales of sentiment driven tails