Ideas and steps about Ke et al. paper

-- Predicting Returns with Text Data

-- Zhongchen Wang

Main ideas

SESTM method (Sentiment Extraction via Screening and Topic Modeling)

A novel model-based approach: understand the sentimental structure of a text corpus without relying on pre-existing dictionaries.

Three main steps:

- Screening for Sentiment-Charged Words: isolates the most relevant terms from a very large vocabulary of terms via predictive correlation screening
- Learning Sentiment Topics: assigns term-specific sentiment weights using a supervised topic model
- Scoring New Articles: uses the estimated topic model to assign article-level sentiment scores via penalized maximum likelihood

Notation

notation	explanation	remark
n	Count of news articles	
m	A dictionary of m words	
$d_i \in \mathbb{R}_+^m$	the word (or phrase) counts of the i^{th} article in a vector	$d_{j,i}$ is the number of times word *j* occurs in article *i*
$D = [d_1, ..., d_n]$	$m \times n$ document-term matrix	
S, *N*	*S*: the index set of sentiment-charged words	*N*: the index set of sentiment-neutral words
$d_{[S],i}$	the column vector corresponding to the i^{th} column of $D_{[S],\cdot}$.	when $p_i = 1$, the article sentiment is maximally positive, and when $p_i = 0$, it is maximally negative
y_i	Associated stock return of article *i*	
$p_i \in [0,1]$	Sentiment score of article *i*	
s_i	Total count of sentiment-charged words in article *i*	
O_+, O_-	Probability distribution over words; O_+: Positive sentiment topic O_-: Negative sentiment topic	$\|S\|$ -vector of non-negative entries with unit l^1-norm
F, *T*	$F = \frac{1}{2}(O_+ + O_-)$, $T = \frac{1}{2}(O_+ - O_-)$	*F*: a vector of frequency *T*: a vector of tone

Dataset

Dataset
Dow Jones Newswires service, January 1984 - July 2017
Train the model using rolling window estimation.
Training and validation sample: January 1989 - January 2014: 15 year interval

  Training: first 10 years
  Validation: last 5 years

Out-of-sample Testing: subsequent one-year window
Roll the entire analysis forward by a year and re-train, iterate this procedure until exhaust the full sample, which amounts to estimating and validating the model 14 times.

Benchmark
A commercial vendor of financial news sentiment scores -- RavenPack

Data Pre-Processing

1	combining“chained”articles
	remove articles with more than one firm tag
	Track the date, exact timestamp, tagged firm ticker, headline, and body text of each article.
2	Timing choice
	Using ticker tags, we match each article with tagged firm’s market capitalization and adjusted daily close-to-close returns from CRSP.
	Without better guidance on timing choice, we train the model by matching articles published on day t (more specifically, between 4pm of day t - 1 and 4pm of day t) with the tagged firm’s three-day return from t - 1 to t + 1 (more specifically, from market close on day t - 2 to close on day t + 1)
3	Set out of sample analysis time duration: February 2014 - July 2017

Data Pre-Processing

4	Remove proper nouns
	Clean and structure news articles -- bag of words • Normalization 1. changing all words in the article to lower case letters 2. expanding contractions such as “haven’t” to “have not” 3. deleting numbers, punctuations, special symbols, and non-English words • Stemming and lemmatizing Group together the different forms of a word to analyze them as a single root word • Tokenization Splits each article into a list of words • Removes common stop words • Translate each article into a vector of word counts

Step 1 -- Screening for Sentiment-Charged Words

Strategy: Isolate the subset of sentiment-charged words, and then estimate a topic model to this subset alone (leaving the neutral words unmodeled)

A supervised approach that leverages the information in realized stock returns to screen for sentiment-charged words.
Intuitively, if a word frequently co-occurs in articles that are accompanied by positive returns, that word is likely to convey positive sentiment.

Step 1 -- Screening for Sentiment-Charged Words

1	calculates the frequency with which word *j* co-occurs with a positive return	
	$f_j = \frac{count\ of\ word\ j\ in\ articles\ with\ sgn(y) = +1}{k_j}$, $j = 1, ..., m$ $k_j = count\ of\ word\ j\ in\ all\ articles$ Variant: $f_j^* = \frac{count\ of\ articles\ including\ word\ j\ AND\ having\ sgn(y) = 1}{count\ of\ articles\ including\ word\ j}$	Can be viewed as a form of marginal screening statistics
2	Compare f_j with proper thresholds: α_+, α_-	Hyper-parameters
	Positive sentiment terms: $f_j > \hat{\pi} + \alpha_+$ Negative sentiment terms: $f_j < \hat{\pi} - \alpha_-$ $\hat{\pi}$: The fraction of articles tagged with a positive return in training sample	$f_j \approx \hat{\pi}$ for a sentiment neutral word
3	Third threshold: κ, to ensure minimal statistical accuracy of the f_j $k_j > \kappa$	Since some words may appear infrequently in the data sample
4	$\hat{S} = \{j : f_j >= \hat{\pi} + \alpha_+, or\ f_j <= \hat{\pi} - \alpha_-\} \cap \{j : k_j >= \kappa\}$	Estimate of the relevant wordlist: set *S*

Step 1 -- Screening for Sentiment-Charged Words

Choice of parameter

1	estimate a collection of SESTM models corresponding to a grid of tuning parameters	α_+, α_- :always set such that the number of words in each group (positive and negative) is either 25, 50, or 100 κ: 86%, 88%, 90%, 92%, and 94% quantiles of the count distribution each year λ: 1, 5, 10
2	• use all estimated models to score each news article in the validation sample • select the constellation of tuning parameter values that minimizes a loss function in the validation sample	l^1 -norm of the differences between estimated article sentiment scores and the corresponding standardized return ranks for all events in the validation sample. $\| \hat{p} - \hat{p}_{rank} \|_1$

Step 2 -- Learning Sentiment Topics

Assigns term-specific sentiment weights using a supervised topic model.

Each Newswire is associated with a stock return, and the return contains information about article sentiment. Hence, returns serve as training labels.

$O = [O_+, O_-]$	determines the data generating process of the counts of sentiment-charged words in each article
$h_i = d_{[S],i}/s_i$ Matrix form: $H = [h_1, h_2, ..., h_n]$	$\lvert S \rvert \times 1$ vector of word frequencies
$\mathbb{E}H = OW$ $W = \begin{bmatrix} p_1 & \dots & p_n \\ 1-p_1 & \dots & 1-p_n \end{bmatrix}$	estimate O via a regression of H on W

Step 2 -- Learning Sentiment Topics

Assigns term-specific sentiment weights using a supervised topic model.

Each Newswire is associated with a stock return, and the return contains information about article sentiment. Hence, returns serve as training labels.

1	Estimate *H*: $\hat{h}_i = d_{[\hat{S}],i}/\hat{s}_i$, $\hat{s}_i = \sum_{j\in\hat{S}} d_{j,i}$	Plugging in $\hat{S}$ from the screening step:
2	$\hat{p}_i = \frac{rank\ of\ y_i\ in\ \{y_l\}_{l=1}^{n}}{n}$	To estimate *W*, we use the standardized ranks of returns as sentiment scores for all articles in the training sample.
3	$\hat{O} = [\hat{h}_1, \hat{h}_2, ..., \hat{h}_n]\widehat{W}'(\widehat{W}\widehat{W}')^{-1}$	$\hat{O}$ may have negative entries. We set all negative entries of this matrix to zero and re-normalize each column to have a unit l^1-norm.

Step 3 -- Scoring New Articles:

To estimate the sentiment p for a new article that is not included from the training sample.

$$d_{[S]} \sim Multinomial(s, pO_+ + (1-p)O_-)$$

Given estimates $\hat{S}$ and $\hat{O}$, we can estimate p using maximum likelihood estimation (MLE).

$$\hat{p} = \arg\max_{p \in [0,1]} \{\hat{s}^{-1} \sum_{j \in \hat{S}} d_j \log\left(p\hat{O}_{+,j} + (1-p)\hat{O}_{-,j}\right) + \lambda \log(p(1-p))\}$$

Out-of-sample test period

- Estimate the sentiment scores of articles using the optimally tuned model determined from the validation sample.
- In the case a stock is mentioned in multiple news articles on the same day, we forecast the next-day return using the average sentiment score over the coincident articles.

Empirical Analysis

Alternative hypothesis: information in news text is not fully absorbed by market prices instantaneously, for reasons such as limits-to-arbitrage and rationally limited attention.

Trading strategy: It is a zeronet-investment portfolio

- each day buys the 50 stocks with the most positive sentiment scores
- shorts the 50 stocks with the most negative sentiment scores.

Two portfolio schemes: equal-weighted and value-weighted P16
Form portfolios only at the market open each day and exclude articles published between 9:00am and 9:30am EST

Empirical Analysis

- Fresh News and Stale News

$$Novelty_{i,t} = 1 - \max_{j \in \chi_{i,t}} \left(\frac{d_{i,t} \cdot d_j}{\parallel d_{i,t} \parallel \parallel d_j \parallel} \right)$$

For each article for firm i on day t, we calculate its cosine similarity with all articles about firm i on the five trading days prior to t (denoted by the set $\chi_{i,t}$)

- Stock volatility

$$\sigma_t = \sum_{i=0}^{\infty} (1-\delta)\delta^i u_{t-1-i}^2$$

Calculate idiosyncratic volatility from residuals of a market model using the preceding 250 daily return observations.
Estimate the conditional idiosyncratic volatility via exponential smoothing according to the formula above.
where u is the market model residual and δ is chosen so that the exponentially_x0002_weighted moving average has a center of mass $(\delta/(1-\delta))$ of 60 days.

Empirical Analysis

- Comparison Versus Dictionary Methods and RavenPack

Dictionary-based sentiment scoring: $\hat{p}_i^{LM}$
RavenPack News Analytics: $\hat{p}_i^{RP}$

Portfolio spanning test: For each sentiment-based trading strategy, regress its returns on the returns of each of the competing strategies, while also controlling for daily returns to the five Fama-French factors plus the UMD momentum factor.

If a trading strategy has a significant α after controlling for an alternative, it indicates that the underlying sentiment measure isolates predictive information that is not fully subsumed by the alternative. Likewise, the R^2 measures the extent to which trading strategies duplicate each other.

Empirical Analysis

- Practical asset management experiment -- taking into account trading costs

Novel trading strategy: EWCT (exponentially-weighted calendar time) portfolio
i) turns over (at most) a fixed proportion of the existing portfolio every period
ii) assigns weights to stocks that decay exponentially with the time since the stock was in the news.

<table>
<tr><td>1</td><td>Form an equal-weighted portfolio
● long the top N stocks in terms of news sentiment that day
● short N stocks with the most negative news sentiment</td></tr>
<tr><td>2</td><td>Parameter: γ the severity of the turnover constraint</td></tr>
<tr><td rowspan="2">3</td><td>Each subsequent day t, liquidate a fixed proportion γ of all existing positions, and reallocate that γ proportion to an equal-weighted long-short portfolio based on day t news.</td></tr>
<tr><td>$w_{i,t} = \frac{\gamma}{N} + (1-\gamma)w_{i,t-1}$: For a stock i in the long-side of the portfolio at day t - 1 and experience large positive sentiment news on day t
$w_{i,t} = (1-\gamma)w_{i,t-1}$: For a stock i in the long-side of the portfolio at day t - 1 but with no news on date t</td></tr>
</table>

Empirical Analysis

- Practical asset management experiment -- taking into account trading costs

- The turnover parameter simultaneously governs both the size of the weight spike at news arrival (the amount of portfolio reallocation) as well as the exponential decay rate for existing weights.
- The EWCT strategy guarantees daily turnover is never larger than γ.

1	Form an equal-weighted portfolio ● long the top N stocks in terms of news sentiment that day ● short N stocks with the most negative news sentiment
2	Parameter: γ the severity of the turnover constraint
3	Each subsequent day *t*, liquidate a fixed proportion γ of all existing positions, and reallocate that γ proportion to an equal-weighted long-short portfolio based on day *t* news.
	$w_{i,t} = \frac{\gamma}{N} + (1-\gamma)w_{i,t-1}$: For a stock *i* in the long-side of the portfolio at day *t* - 1 and experience large positive sentiment news on day *t* $w_{i,t} = (1-\gamma)w_{i,t-1}$: For a stock *i* in the long-side of the portfolio at day *t* - 1 but with no news on date *t*

Some typo

1. P5: We occasionally work with a subset of rows from D, where the indices of columns included in the subset are listed in the set S.
2. P17: count of words
3. P15 and P20: table 2, table 3, unit of return is confusing
4. P29: moving down