**Predicting Returns with Text Data**

1. **Introduction**

The previous researches on the sentiment analysis of financial documents concerns the use of pre-specified sentiment dictionaries. However, the sentiment scored derived from researches for other purposes might not be suitable in the financial context.

This paper uses machine learning techniques to understand the sentimental structure of a text corpus without relying on pre-existing dictionaries. It has three virtues: a) simplicity b) minimal computing power c) ability to generalize to other datasets.

The aim of the paper is to study the extent to which business news explains and predicts observed asset price variation. The dataset being used is the Dow Jones Newswires. The model is evaluated using back-testing and is shown to outperform similar strategies based on . The paper also studies the impacts of “stale” and “fresh” news and the assimilation of news for different stocks.

* **Procedures**

First: Isolate the most relevant features from a large vocabulary of terms using the bag-of-words representation. Second: Assign term-specific sentiment weight for the prediction task and taking into account the skewness in terms of frequencies. Third: Use the estimated topic model to assign an article-level sentiment score.

1. **Methodology**

To establish notation, we use to denote the number of time word occur in article , and denote the subset of columns from using . Each article is tagged with identifier of stock and we only study articles that correspond to a single stock, we label the return of the stock on the publication date as .

* 1. **Model Setup**

Assume that each article possesses a sentiment score , with corresponding to maximally positive sentiment and corresponding to maximally negative. Furthermore, we assume that serves as a sufficient statistic on the stock return.

* **Return Distribution**

For conditional return distribution, we assume that the higher the sentiment score, the higher the probability of realizing a positive return. In mathematical form

* **Sentiment Distribution**

We assume that the sentiment-charged words and sentiment-neutral words constitutes the entire dictionary, and that and are independent of each other. Assume that sentiment-charged words count are generated by mixture multinomial distribution

where is the total count of sentiment-charged words in article , is the probability distribution over words in maximally positive sentiment article, and describes the distribution over words in maximally negative sentiment article. A word is a “positive word” if the entry of is positive and is a “negative word” if the entry is negative.

For a given event , the distribution of sentiment word count and the distribution of returns are linked through the common parameters . The function is monotone increasing and therefore a higher maps into higher likelihood of positive returns, and thus high values in and low values in . The objective is to learn the mode parameters and .

A picture containing screenshot

Description automatically generated

* 1. **Sentiment Charged Words**

The screening procedure first calculates the frequency with which word co-occurs with a positive return, measured by

Then, we set an upper threshold and define all words having are positive sentiment terms. Likewise, any word satisfying for some lower threshold is deemed a negative sentiment term. Finally, we select a third threshold on the count of articles including by restricting analysis of words for which . The threshold are hyper-parameters that can be tuned via cross-validation.

A picture containing object

Description automatically generated

* 1. **Sentiment Distribution**

We have identified the relevant wordlist , we have the simplified problem of fitting two-topic model to the sentiment-charged counts. The multinomial distribution that relates word count to the sentiment score and the vectors by

If we gather the matrix into a single matrix , the relation above can be rewritten

Bases on this relation, we can estimate via a regression of on . To estimate , we use the standardized ranks of returns as sentiment scores for all articles in the training sample, more precisely

* 1. **Scoring New Articles**

The preceding steps construct estimators and , we can now use the estimators to estimate the sentiment score for a new article that is not included in the training sample. The value of can be estimated using the MLE from the multinomial distribution

* **Penalty**

In order to cope with the limited number of observations and the low signal-to-noise ratio inherent to return prediction, we add a penalty term in the likelihood function

1. **Empirical Analysis** 
   1. **Data and Pre-Processing**

**Dataset:** Dow Jones Newswires Machine Text Feed and Archive database (1989-2017)

**Process Dataset:** a) Select articles with one and only one firm tag b) Match each article with market capitalization and adjusted daily close-to-close return from CRSP c) Train the model by merging articles published between of day and of day with three-day return from close on day to close on day

Process Article: a) Normalization: a1) change to lower case a2) expand contractions a3) delete numbers, punctuations etc. b) stemming and lemmatizing c) tokenizing d) remove common words e) bag of words

* 1. **Return Predictions**

Train the model using rolling window estimation for which the first years are used for training and the last years for validation and subsequent year for out-of-sample testing. This amount to training the model times. In each training sample, we estimate a collection of SSESTM models corresponding to a grid of hyper-parameters.

The trading strategy is a zero-net-investment portfolio that, each day at the market open, buys the stocks with the most positive sentiment scores and shorts the stocks with the most negative sentiment score gathered in the following period:

A screenshot of a social media post

Description automatically generated

The cumulative one-day open-to-open trading strategy returns, based on out-of-sample sentiment forecast, is summarized below:

A close up of a map

Description automatically generated

Three basic facts emerge from the one-day forecast evaluation:

1. Equal-weighted portfolios substantially outperform their value-weighted counterparts.
2. Second, the long side of the trade outperforms the short side
3. Third, the SSESTM sentiment trading strategies have little exposure to standard aggregate risk factors
   1. **Most Impactful Words**

A close up of a piece of paper

Description automatically generated

* 1. **Fresh News and Stale News**

The previous analysis focuses on relating news sentiment on day to returns on day . In this section, we investigate the timing of price responses to news sentiment more precisely. In the following figure, we separately investigate the association between sentiment on day and returns on day , day and day .

A close up of a map

Description automatically generated

The day strategy quantifies the extent to which our sentiment score picks up on stale news. On average, prices respond strongly ahead of news in the sample, suggesting much of the daily news flow echoes previously reported news or is a new report of information already known to market participants.

* 1. **Price Assimilation**

We analyse trading strategies that trade in response to news sentiment with delays of positioning of and from 1 to 10 days following the announcement, for which the stocks are sold on the subsequent trading day.

A close up of a map

Description automatically generated

* 1. **Price Assimilation vs. Novelty**

In order to investigate the speed of assimilation without the influence of “stale” news, we construct measure of article novelty for firm on day with all articles about firm five trading days (denoted ) prior to day

The price assimilation of news based on article novelty is plotted below

A close up of a map

Description automatically generated

* 1. **Price Assimilation vs. Market Capitalization**

The figure below summarizes the price assimilation of news based on stock size

A close up of a map

Description automatically generated

* 1. **Price Assimilation vs. Stock Volatility**

The figure below summarizes the price assimilation of news based on stock price volatility

A close up of a map

Description automatically generated

* 1. **Comparison with Dictionary & RavenPack**

A close up of a map

Description automatically generated