# BERT-based Financial Sentiment Index and LSTM-based Stock Return Predictability

**Joshua Zoen Git Hiew**
City University of Hong Kong
joshuahzg@yahoo.com.hk

**Xin Huang**
The Chinese University of Hong Kong
huangxin@se.cuhk.edu.hk

**Hao Mou**
DataStory
mouhao@datastory.com.cn

**Duan Li**[*]
City University of Hong Kong
dli226@cityu.edu.hk

**Qi Wu**
City University of Hong Kong
qiwu55@cityu.edu.hk

**Yabo Xu**
DataStory
arber@datastory.com.cn

## Abstract

Traditional sentiment construction in finance relies heavily on the dictionary-based approach, with a few exceptions using simple machine learning techniques such as Naive Bayes classifier. While the current literature has not yet invoked the rapid advancement in the natural language processing, we construct in this research a textual-based sentiment index using a novel model BERT recently developed by Google, especially for three actively trading individual stocks in Hong Kong market with hot discussion on Weibo.com. On the one hand, we demonstrate a significant enhancement of applying BERT in sentiment analysis when compared with existing models. On the other hand, by combining with the other two existing methods commonly used on building the sentiment index in the financial literature, i.e., option-implied and market-implied approaches, we propose a more general and comprehensive framework for financial sentiment analysis, and further provide convincing outcomes for the predictability of individual stock return for the above three stocks using LSTM (with a feature of a nonlinear mapping), in contrast to the dominating econometric methods in sentiment influence analysis that are all of a nature of linear regression.

## 1 Introduction

It is a common belief that investors' sentiment is one of the important driving sources behind the financial market movement. Although the classical financial theory hypothesizes that investors are rational, extensive studies have already revealed the significant influence of their *irrational* behavior, like optimistic or pessimistic sentiment (see Lee et al. (1991) and Baker and Wurgler (2006) among others). However, different research works adopt different sentiment measures. Sending questionnaires to investors is a very traditional way to collect public opinion upon the market environment and market trend. The obvious drawback of this approach is a low frequency on data acquisition, since a survey is usually conducted once a week, a month, or even a quarter. For instance, the sentiment proxy in Brown and Cliff (2005) is based on weekly survey data from the American Association of Individual Investors. Some quantitative methods are then proposed. For example, both

[*]Corresponding Author.

Baker and Wurgler (2006) and Chong et al. (2014) apply the principal component analysis (PCA) on a series of data set of selected market factors to extract the market-implied sentiment index, while Han (2008) measures the investor sentiment from option-implied information. These methods focus on finding a proxy of sentiment leading to an indirect measure, compared with those approaches that directly deal with sentimental texts from the internet, like Twitter (see Bollen et al., 2011), news or analysts' articles (see Chen et al., 2018). Recently, Kearney and Liu (2014) provide a comprehensive survey that summarizes different information sources, content analysis methods, and empirical models for the textual sentiment. As a conclusion, however, they suggest to extend the lexicons for textual content analysis, ignoring the rapid development in the natural language process (NLP). Although the sentiment analysis is a common research field in both machine learning (ML) and behavioral finance, there is still a big gap to integrate the research strength of these two.

In this paper, we construct a textual-based sentiment index by adopting the newly-devised NLP tool BERT from Devlin et al. (2018) to posts that are published on the Chinese social media, which represents the *first* attempt in the literature to apply this state-of-the-art learning model to the financial sentiment extraction. At this stage, our analysis focuses mainly on the *individual* stock level, by investigating three actively-trading listed companies in Hong Kong Stock Exchange (HKSE) in a pilot study, namely, Tencent (0700.HK), CCB (0939.HK), and Ping An (2318.HK), which all possess a sufficient exposure on Weibo.com. We then demonstrate a better performance of BERT on sentiment construction compared with the other well-known models such as Multichannel Convolutional Neural Network (CNN) of Kim (2014) and Transformer of Vaswani et al. (2017), among others. Through combining the BERT-based sentiment index with other two types of sentiment indices from the option-implied information and PCA on market data for the above three stocks, we next provide a deeper and more general financial sentiment analysis. More specifically, our BERT-based sentiment reflects more about individual investors' opinion, whereas the option-implied one followed by Han (2008) represents more about the institutions' attitude. We would like to see how these two counterparts influence the market, with the attendance of market-based index which is treated as an overall market sentiment. Finally, we address the stock return predictability by integrating sentiment indices from different information sources, through applying the powerful sequential neural network model Long Short-Term Memory (LSTM), parallel to the classical econometrics tool like Vector Autoregression (VAR). Note that, as Yan et al. (2018) point out, LSTM model on quantile regression outperforms those traditional time-series analysis tools that are commonly used in the financial literature.

The rest of our paper is organized as follows. We construct first our textual sentiment index by BERT and also by other NLP models for comparison in Section 2. We then carry out a general financial sentiment analysis based on three different information sources in Section 3. We address the stock return predictability issue at the individual level in Section 4. Finally we conclude our paper in Section 5.

## 2 Textual sentiment index construction

In this paper, we focus on the individual-level sentiment analysis and take three listed firms in HKSE to conduct our experiment. More precisely, we select Tencent (0700.HK), Ping An (2318.HK), and CCB (0939.HK) as our individual stocks, and grab posts related to these three companies, respectively, from Weibo, a popular social media in China, for the time period from January 1, 2016 to December 31, 2018 on a daily basis. In the following, we introduce our procedure of sentiment index construction and evaluations among different ML models.

### 2.1 Pre-processing work

The pre-processing work, after grabbing *all* the firm-specific data from Weibo during the above time period, consists of posts cleaning and labelling. To filter out noisy posts like advertisements or others that are published by *water army*[2], we also adopt a detection model[3] through labelling

[2] Internet water army, always sponsored by certain business entities, is a group of paid posters who post biased content for particular purposes and have flooded the social networks nowadays, as pointed out by Chen et al. (2013).

[3] A commercial software, launched by DataStory (www.datastory.com.cn), that is used to detect water army and has successfully served for more than 100 internet companies.

those jam information, at the same time when we label real sentimental posts about corresponding stocks. Note that in traditional sentiment analysis under supervised learning framework, one may label a piece of context by emotional words, like “happy” and “anger”. However, when it comes to financial texts, we prefer using the *polarity*, i.e., “positive”, “negative”, and “neutral” to label the data, as they could also represent the attitude of posters corresponding to bullish, bearish, and ambiguous markets, respectively. Furthermore, we adopt a *voting strategy*, similar as in Ribeiro et al. (2016), to enhance our accuracy of labelling. More precisely, each post, in each round out of six, is manually labelled by at least three experts and we only keep those answers that are agreed by at least two experts. Moreover, we also need to achieve consensus on those conflicted posts among us before the next round starts. Eventually, we end up with original 117,029 posts for three stocks in total from Weibo during the considered time period mentioned above and randomly label 10,165 ones (8.69%) that are ready to train and assess different ML models in the following.

### 2.2 Evaluation by BERT and other ML models and Comparison

Proposed by Devlin et al. (2018), BERT, as an open-source model[4], is *pre-trained* with massive datasets to encode bi-directional contexts through multi-layer transformers, and has been reported to achieve the state-of-the-art results in NLP downstream tasks. For instance, it completes the Stanford Sentiment Treebank (SST-2) task, as one of the General Language Understanding Evaluation (GLUE) benchmarks, with accuracy as high as 94.9%. In this paper, we rely on the Chinese version “BERT-Base, Chinese” to do the fine-tuning.

To tackle the sentiment analysis as a basic *text classification* task, there actually exist other genres of ML models. Formerly, researchers tend to use support-vector networks (SVM) (see Cortes and Vapnik, 1995) or ensemble methods (see Opitz and Maclin, 1999) to build a classification model, while simpler lexical-based approaches are always adopted in finance, as surveyed in Kearney and Liu (2014). With much more embedding methods and significant increase on computing power nowadays, deep learning methods are growingly dominating those statistical learning ones in all aspects of NLP. As a comparison with BERT, we mainly consider the other four famous models, i.e., the Recurrent Neural Network (RNN) based Bidirectional Long Short-Term Memory (BiLSTM) (see Hochreiter and Schmidhuber, 1997), the Multichannel Convolutional Neural Network (CNN) (see Kim, 2014), the CPU-efficient FastText (see Joulin et al., 2016) that is adopted by Facebook, and the Transformer with attention mechanism (see Vaswani et al., 2017). Note that since BiLSTM and Multichannel CNN are suggested to initialize with pre-trained word embedding like Shifted Positive pointwise mutual information (PMI) proposed by Levy and Goldberg (2014), we finally take the PMI-enhanced versions of them (see Li et al., 2018, for example).

Table 1: Comparison of performance across different models

| ML Model | Precision_micro | Recall_micro | F1_micro |
|---|---|---|---|
| **BERT** | **79.3** | **75.4** | **78.5** |
| Transformer + attention | 77.6 | 64.8 | 71.3 |
| PMI + Multichannel CNN | 75.9 | 60.6 | 64.3 |
| PMI + BiLSTM | 75.3 | 56.2 | 62.6 |
| FastText | 72.1 | 48.7 | 61.5 |

We split our labelled dataset described in Subsection 2.1 into a training set and test set by a ratio of 80% and 20%, and a 10-fold cross validation is performed on the training set for all models. Table 1 shows performance evaluation for all selected models that are trained by the same labelled Weibo-post dataset mentioned in Subsection 2.1. In order to avoid impact from imbalanced proportion for different categories of our labelled result (15% positive, 78% neutral, and 7% negative), we use *micro-average* method to calculate the precision, the recall, and an averaged F1 score, respectively, as the common criteria for model evaluation. From the table we can see that BERT is superior across all indicators in our training process, especially on its significantly dominating recall rate even with the presence of its better precision. The above outcome demonstrates its strong capability over the other ML models for financial sentimental texts classification in Chinese, leading to the *first* BERT-based financial sentiment index in the literature presented in the next subsection.

[4] See https://github.com/google-research/bert.

### 2.3 BERT-based sentiment index

We apply our fine-tuned BERT to all unlabelled posts filtered by our detection model and classify them into three categories of polarity. Note that we treat those posts that are published after the trading time (4 p.m. (GMT+8) for Hong Kong market) as the influence for the next trading day, and calculate a BERT-based sentiment value $BSI_t^i$ for stock $i$ on a trading-day basis through

$$BSI_t^i = \frac{Pos_t^i - Neg_t^i}{Pos_t^i + Neu_t^i + Neg_t^i} \tag{1}$$

where $Pos_t^i$, $Neu_t^i$ and $Neg_t^i$ are the number of positive, neutral and negative texts that are related to stock $i$ and outputted by BERT on the trading day $t$, respectively. Then, all $BSI_t^i$'s time-series data form our BERT-based financial sentiment index $BSI^i$ for stock $i$.

## 3 Financial sentiment analysis based on different information channels

Apart from the textual channel to extract financial sentiment from the social network by NLP techniques, there exist another two types of information sources that have been commonly utilized in finance community. One is the *risk-neutral implied skewness* based on option price (for example, Han, 2008), leading to the *option-implied* sentiment; another channel includes the *market data* (for example, Baker and Wurgler, 2006), resulting in the *market-implied* sentiment. In our invesigation we construct additional two sentiment indices and then take into consideration all three indices in hand to conduct a more general financial sentiment analysis.

### 3.1 Option-implied and market-implied financial sentiment

Han (2008) discovers the relationship between option volatility smile, risk-neutral skewness and market sentiment. He finds that when the market tends to be bearish (or bullish), the slope of option volatility smile becomes steeper (or flatter) and the risk-neutral skewness changes to be more negative (positive). Accordingly, Han (2008) proposes an option-implied sentiment proxy. Following his work, we construct the same sentiment index for our selected individual stocks using the implied skewness of their option information, respectively, denoted by $OSI^i$.

Market data offer a traditional source for extracting market sentiment. Baker and Wurgler (2006) identify a set of market data which they believe is driven by the investors' sentiment, and form an underlying proxy for such a data set. They apply the *principal component analysis* (PCA), which is sometimes considered as an unsupervised machine learning method, to extract this market-type sentiment. However, due to the low frequency of part of their selected market characteristics, such as the number of initial public offerings (IPO) within a month, Chong et al. (2014) consider another set of market data which could represent the investors' sentiment on a daily basis. Therefore, in order to be in line with our previous sentiment indices, we choose to follow the work in Chong et al. (2014) which concentrates on Hong Kong market as well and construct our market-implied sentiment index $MSI^i$ for each individual stock.[5]

### 3.2 Framework of financial sentiment analysis from three channels

In general, there are two types of market participators: *individuals* and *institutions*. It is reasonable to believe that these two groups express their sentiment in different ways. As we could imagine, the social media are more individual-oriented, and the institutional investors seldom express their attitude towards market directly in public. As remarked by Verma and Soydemir (2009), even a survey for institutions may contain biases since they could deviate heavily from what they published; it is obvious, however, that sophisticated investors like institutions constitute the majority in contributing to the derivatives market. As Easley et al. (1998) point out, the informed traders are more likely to trade in the option market rather than in the equity market. Therefore, we tend to interpret the sentiment extracted from the social media as *individual* investors' sentiment, while treat the

[5] All details of risk-neutral skewness and selected market characteristics that are used to construct the option-implied and market-implied sentiment indices, respectively, can be found in our supplementary material. Some graphical illustrations are also provided there.

option-implied one as *institutional* investors' attitude to the market. We expect these two to be significantly different. Finally, since the overall market is made up of both individuals and institutions, the market-type proxy could be interpreted as the sentiment for the whole market.

In order to have a more comprehensive understanding about the financial sentiment, it is natural to consider all three channels simultaneously. The equal-weighted sum as an overall sentiment index is the simplest but may cause information loss, since the three indices are not fully inter-independent, as shown by the correlation calculation given later in Table 2. Another possible *linear* combination of the three indices could be figured out by VAR when addressing the predictability issue. However, the most interesting mixture could be a *nonlinear* form through a neural network as discussed further in the next section.

## 4 Stock return predictability by sentiment indices

Investigation on how to predict the future stock return requires better time-series analysis tools and it still remains challenging. Verma and Soydemir (2009) study predicting ability of investors' sentiment, with the presence of other fundamental market factors like Fama-French three factors (see Fama and French, 1993), through the Vector Autoregression (VAR) as a basic model in econometrics. VAR, though simple and clear enough, only captures the linear relationship among different time-series data. In this paper, we propose to use Long Short-Term Memory (LSTM) model to analyze the predictability of different sentiment indices on stock return, as it could capture the nonlinear features that traditional VAR fails to include. Our testing results conclude that the LSTM model outperforms VAR in a yearly basis in terms of lower mean square error.

### 4.1 Basic statistics

The table below summarizes, for each individual stock, the correlation coefficients between any two quantities out of three different sentiment indices themselves. We also do the similar analysis for sentiment index at time $t$ with stock return at $t+1$, $r^i_{t+1}$.

Table 2: Correlation coefficient between two quantities for each individual stock, where $BSI^i$ stands for our BERT-based sentiment index for stock $i$, $OSI^i$ for option-implied sentiment index, and $MSI^i$ for market-implied one; $r^i_{t+1}$ represents stock return at $t+1$.

| Stock $i$ | Tencent (0700.HK) | CCB (0939.HK) | Ping An (2318.HK) |
|---|---|---|---|
| *Correlation between different sentiment indices for each stock* | | | |
| $BSI^i$ v.s. $OSI^i$ | 0.0347 | -0.0026 | -0.0448 |
| $BSI^i$ v.s. $MSI^i$ | -0.3442 | 0.1944 | 0.2024 |
| $OSI^i$ v.s. $MSI^i$ | -0.1776 | 0.1463 | 0.0116 |
| *Correlation between today's sentiment index and tomorrow's stock return* | | | |
| $BSI^i_t$ v.s. $r^i_{t+1}$ | -0.0205 | -0.0387 | 0.0710 |
| $OSI^i_t$ v.s. $r^i_{t+1}$ | -0.0052 | -0.0094 | -0.0327 |
| $MSI^i_t$ v.s. $r^i_{t+1}$ | -0.0304 | -0.0068 | 0.0337 |

From Table 2 we can see that there does not exist a persistent linear relationship across different quantities in the individual stock level. For instance, when we compare $BSI^i$ with $MSI^i$, it could have positive or negative correlations across different stocks, though relatively strong in magnitude. Given a certain individual, the relation between different pairs of sentiment indices looks borderline as well. As the simple prediction power check, all values of correlation coefficients seem low, which may indicate a hidden nonlinear affection of sentiment on future stock return.

### 4.2 Predictability of sentiment indices on stock return

In this subsection, we examine whether our BERT-based financial sentiment index and the other two could predict the market or not, especially on predicting the future stock return under the attendance of other classical risk factors that have been proved to have pricing power on stocks. Following the work by Verma and Soydemir (2009), we select eight fundamental factors as *control variables*, including one-month interest rate ($r_1$), economic risk premium defined by the difference between three-month and one-month interest rates ($r_3 - r_1$), inflation rate ($Inf$), the return on portfolio of winning stocks over past twelve months minus those losing stocks ($UMD$), the currency fluctuation of Hong Kong dollar ($HKD$), and the Fama-French three factors: the excess market portfolio return ($r_m - r_1$), the return on portfolio of small companies minus big ones ($SMB$), and the return on portfolio of high book to market value companies minus low book to market value ones ($HML$). The full time period in our experiment covers from January 1, 2016 to December 31, 2018. The individual stock return is defined by its *log return*, namely, $r_t^i = \log(S_t^i / S_{t-1}^i)$ where $S_t^i$ is the price for stock $i$ at time $t$. All time-series data are normalized to have mean 0 and variance 1.

### 4.2.1 VAR and LSTM modelling

We first employ VAR as a traditional time-series analysis tool to investigate the predictability of the sentiment on future stock return. More precisely, we consider the following model,

$$Y_t^i = A^i + \sum_{s=1}^{\ell} B_s^i Y_{t-s}^i + \boldsymbol{\epsilon}_t^i \tag{2}$$

where $Y_t^i$ is a column vector for stock $i$ consisting of variables which we believe could have an inter-temporal relationship (in our case it contains stock return and different sentiment indices with risky factors above), $A^i$ is a time-invariant constant term, $\ell$ is the look-backward length (which is set to be 2 here), taking into account the possible time lag effect of sentiment, $B_s^i$ is the matrix of coefficients for the $s$-lag vector $Y_{t-s}^i$, and $\boldsymbol{\epsilon}_t^i$ is the error term. Note that (2) can be written in an ordinary *linear regression* form

$$y_{tm}^i = a_m^i + \sum_{s=1}^{\ell} \sum_{n=1}^{N^i} b_{sn}^i y_{(t-s)n}^i + \epsilon_{tm}^i, \tag{3}$$

where $y_{tm}^i$ is our concerned component of $Y_t^i$ (that is stock return which we are going to predict) for stock $i$, and $y_{(t-s)n}^i$ is the $n$th-component of $N^i$-dimensional vector $Y_{t-s}^i$ with corresponding coefficient $b_{sn}^i$.

Despite VAR always acts as a benchmark forecasting model in finance, it requires strong model assumptions, like Gaussian white noise and dependence of predetermined variables. We emphasize that VAR is a linear prediction model as evidenced from (3). We now adopt LSTM as a powerful machine learning method in predicting future based on past information without assuming any noise form. Most importantly, LSTM could capture possible *nonlinear* features behind the time series. The hyperparameters of our LSTM model include the number of layers $L$ and the number of training epochs $E$. Moreover, we keep the same maximum time lag $\ell$ and set the hidden size to be the same number of independent variables of VAR above, in order to have a proper comparison. Note that, since a linear structure is actually a special case of a feedforward neural network when armed with a linear transformer, one should expect that LSTM performs at least as good as VAR.

### 4.2.2 Predictability testing results

In our experiments of each individual stock return prediction, *dates* within a calendar year are randomly distributed into the *training* set $D_{tr}^i$ and the *test* set $D_{te}^i$, with proportion 80% and 20%, respectively, which are shared for both VAR and LSTM models fitting in parallel. Namely, the stock return $r_t^i$ such that $t \in D_{tr}^i$ as output together with sentiment indices and all other factors at time $t-2$ and $t-1$ as inputs are used to train our models, and we choose to report mean square error (MSE) not only on $D_{te}^i$ but also on $D_{tr}^i$ as well as the whole year set $D_{wh}^i$. Note that we let three sentiment indices enter into inputs separately and also all together for different experiments in order to see whether the combination of different sentimental sources could enhance the prediction further or not. Besides, the reason why we start with the yearly basis testing is that the influence of sentiment, though may maintain for a while, cannot last for too long.

Table 3 lists the yearly-basis prediction accuracy of different sentiment indices and their mixture (with the presence of other factors mentioned above), in terms of MSE between real stock returns and predicted ones calculated on different sets of dates under both LSTM and VAR (in bracket) models, for three selected stocks, respectively. We can see that all MSE's calculated from LSTM are smaller than those from VAR in the table (including on test sets $D_{te}^{i}$'s), indicating that the yearly-basis stock return prediction from LSTM performs in general better than using traditional time-series tool VAR. In other words, investors' sentiment predicts the market more likely in a *nonlinear* fashion.

Table 3: Individual stock return prediction (of different sentiment indices and their mixture with presence of other factors) accuracy in terms of MSE (unit $\times 10^{-5}$) based on LSTM and VAR (in bracket) for 2016, 2017, and 2018, respectively; **bold** numbers show the best sentiment index as a predictor on a certain set of dates in a particular year; the case without using any sentiment index to predict the market is also examined.

| Stock $i$ | Tencent (0700.HK) | | | CCB (0939.HK) | | | Ping An (2318.HK) | | |
|---|---|---|---|---|---|---|---|---|---|
| MSE on | $D_{tr}^{0700}$ | $D_{te}^{0700}$ | $D_{wh}^{0700}$ | $D_{tr}^{0939}$ | $D_{te}^{0939}$ | $D_{wh}^{0939}$ | $D_{tr}^{2318}$ | $D_{te}^{2318}$ | $D_{wh}^{2318}$ |
| *For 2016* | | | | | | | | | |
| $BSI^i$ | 2.87 (4.04) | 2.91 (4.64) | 2.88 (4.16) | 2.51 (3.55) | 3.06 (**5.01**) | 2.62 (**3.84**) | 2.98 (4.47) | 4.17 (**6.66**) | 3.22 (4.90) |
| $OSI^i$ | 2.83 (3.99) | 2.97 (4.67) | 2.85 (4.13) | 2.56 (3.57) | 3.06 (5.09) | 2.66 (3.87) | 3.76 (4.64) | 4.44 (6.71) | 3.90 (5.05) |
| $MSI^i$ | 2.58 (4.00) | 2.52 (**4.62**) | 2.57 (4.12) | 2.56 (3.56) | 3.05 (5.06) | 2.66 (3.86) | 3.75 (4.62) | 4.59 (6.69) | 3.91 (5.03) |
| *Mixture* | **2.40** (**3.90**) | **2.34** (4.83) | **2.39** (**4.08**) | **2.17** (**3.52**) | **2.81** (5.26) | **2.29** (3.86) | **2.52** (**4.41**) | **3.28** (6.78) | **2.67** (**4.88**) |
| *No SI* | 3.88 (3.86) | 5.56 (5.53) | 4.21 (4.19) | 3.62 (3.60) | 4.96 (5.06) | 3.89 (3.89) | 4.77 (4.58) | 6.80 (7.26) | 5.17 (5.11) |
| *For 2017* | | | | | | | | | |
| $BSI^i$ | 2.12 (3.41) | 2.24 (5.32) | 2.14 (3.79) | 1.56 (2.58) | 1.52 (3.27) | 1.55 (2.72) | 2.57 (4.35) | **2.01** (**4.97**) | 2.46 (4.48) |
| $OSI^i$ | 2.07 (3.42) | **2.16** (5.33) | **2.09** (3.80) | 1.65 (2.60) | 1.52 (3.32) | 1.63 (2.74) | 2.80 (4.47) | 2.09 (4.97) | 2.66 (4.57) |
| $MSI^i$ | 2.32 (3.38) | 2.33 (**5.30**) | 2.32 (3.76) | 1.75 (2.57) | 1.52 (**3.30**) | 1.70 (**2.71**) | 2.97 (4.45) | 2.52 (5.04) | 2.88 (4.57) |
| *Mixture* | **2.05** (**3.32**) | 2.29 (5.34) | 2.10 (**3.72**) | **1.45** (**2.54**) | **1.46** (3.41) | **1.45** (2.72) | **2.48** (**4.30**) | 2.07 (5.07) | **2.40** (**4.46**) |
| *No SI* | 2.97 (3.60) | 4.45 (4.43) | 3.27 (3.77) | 2.29 (2.68) | 2.96 (2.98) | 2.42 (2.74) | 3.90 (4.51) | 5.30 (4.84) | 4.18 (4.58) |
| *For 2018* | | | | | | | | | |
| $BSI^i$ | 5.79 (8.89) | 5.52 (11.88) | 5.74 (9.49) | 2.99 (4.54) | 2.83 (5.10) | 2.96 (4.65) | 2.59 (4.51) | 2.23 (5.51) | 2.52 (4.71) |
| $OSI^i$ | 6.14 (8.84) | 4.85 (**11.33**) | 5.88 (9.34) | 3.34 (4.49) | 2.95 (**5.04**) | 3.27 (4.60) | 2.91 (4.48) | 2.75 (5.58) | 2.88 (4.70) |
| $MSI^i$ | 6.20 (8.95) | 5.67 (11.67) | 6.09 (9.49) | 2.85 (4.50) | 2.71 (5.06) | 2.83 (4.61) | 2.95 (4.42) | 2.54 (**5.43**) | 2.86 (4.62) |
| *Mixture* | **5.33** (**8.68**) | **4.40** (11.91) | **5.14** (**9.32**) | **2.58** (**4.43**) | **2.31** (5.13) | **2.53** (**4.57**) | **2.27** (**4.33**) | **2.04** (5.57) | **2.22** (**4.58**) |
| *No SI* | 7.20 (9.01) | 11.93 (12.17) | 8.14 (9.64) | 4.03 (4.42) | 5.93 (5.78) | 4.41 (4.69) | 3.79 (4.51) | 5.45 (5.81) | 4.12 (4.77) |

Another worth-mentioning observation is that the more complex combination of three sentiment indices leads to a better prediction, as almost all the mixtures have lower MSE's than the case when there is only one sentiment index added under LSTM setting (as we can find from the table that bold numbers, which stands for the *best* sentiment predictor within a certain set of dates, on LSTM positions always appear in the *Mixture* lines), while it is not always true for VAR model. This result confirms some complicated influence structure of sentiment on stock return in the individual level and suggests that the combination of different channels does help to improve the accuracy of the predicting power of financial sentiment. As a supplement, we also examine the simpler case that leverages *only* eight factors to predict the market without utilizing any sentiment index (denoted by *No SI*). As we could imagine, adding sentimental factors is indeed valuable under LSTM model, since the appearance of sentiment indices as predictors significantly reduces the predicting error, reflecting on lower MSE's in the Table 3.

The above conclusion can also be found graphically in the following figures, taking *Mixture* as the sentimental predictor along the dates of test set in 2018 as an example (while the other sentiment indices under the rest of years behave similarly). The first row of Figure 1 shows real stock return movement (blue line) versus *Mixture*-predicted ones (under attendance of those considered factors) from both LSTM (in red) and VAR (in black) for three stocks, respectively. It is apparent that LSTM prediction is much closer to the real return than VAR prediction for any individual stock. We also display the comparison between whether we utilize sentiment as predictor or not in LSTM model, as exhibited by the second row in the figure. We can still see the dominating performance of *Mixture* (in red) over *No SI* (in black) on predicting the real return fluctuation (in blue) for each stock. Note that since we randomly assign dates into test set, the curves in figures are accomplished through connecting real or predicted return data points on dates of test set in a correct time order.

As for the predictability testing for the whole time period from 2016 to 2018, we find that LSTM performs worse than VAR on the test sets for all three stocks, even with the same hyperparameters used for yearly-based testing, as shown by the underlined numbers in Table 4. This is possibly because the ML model is over-fitted. To see this, let us design another trivial model that always predicts zero return for any stock, and then we still calculate the MSE between real stock return and constant zero, the result of which is summarized in Table 5. Since the prediction accuracy is not far away from those of LSTM and VAR, it seems that any attempt to predict the future stock return over a *longer* time period is in vain, as if the model is learning a random signal with zero mean.

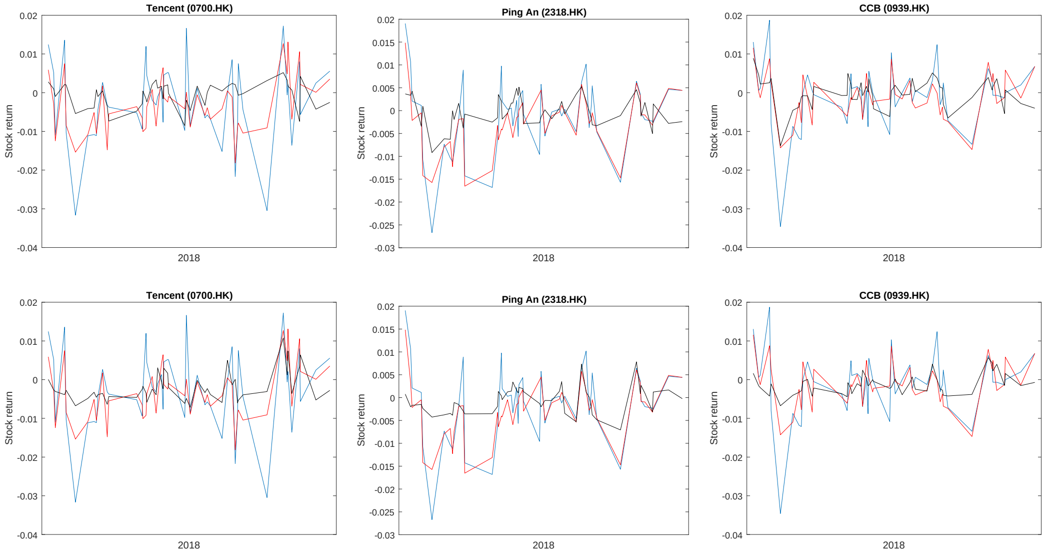

Figure 1: Real stock return movement (in blue) versus predicted ones from different models on test set of 2018 for each individual stock. The first row displays prediction of *Mixture* (with presence of other factors) from LSTM (in red) and VAR (in black), while the second row exhibits prediction of *Mixture* (in red) and *No SI* (in black) for LSTM model only; among all figures *Mixture* under LSTM setting performs best on prediction of stock return in the sense of closer distance to the real one.

Table 4: Individual stock return prediction (of different sentiment indices and their mixture with the presence of other factors) accuracy in terms of MSE (unit: $\times 10^{-5}$) based on LSTM and VAR (in bracket) models for the complete time period from 2016 to 2018; underline numbers show where VAR outperforms LSTM.

| Stock $i$ | Tencent (0700.HK) | | | CCB (0939.HK) | | | Ping An (2318.HK) | | |
|---|---|---|---|---|---|---|---|---|---|
| MSE on | training set | test set | whole set | training set | test set | whole set | training set | test set | whole set |
| *From 2016 to 2018* | | | | | | | | | |
| $BSI^i$ | 5.05 (5.96) | 7.23 (6.54) | 5.49 (6.07) | 3.32 (3.97) | 4.83 (4.49) | 3.62 (4.08) | 4.45 (5.18) | 5.64 (5.23) | 4.69 (5.19) |
| $OSI^i$ | 4.71 (5.95) | 7.61 (6.53) | 5.29 (6.07) | 3.54 (3.97) | 4.70 (4.46) | 3.78 (4.06) | 4.67 (5.16) | 5.56 (5.21) | 4.85 (5.17) |
| $MSI^i$ | 5.10 (5.95) | 7.42 (6.59) | 5.57 (6.08) | 3.43 (3.97) | 4.70 (4.48) | 3.68 (4.08) | 4.69 (5.16) | 5.74 (5.22) | 4.90 (5.18) |
| Mixture | 4.52 (5.92) | 8.00 (6.61) | 5.21 (6.06) | 3.29 (3.96) | 5.09 (4.52) | 3.65 (4.07) | 4.36 (5.13) | 5.93 (5.23) | 4.67 (5.15) |

Table 5: Individual stock return prediction accuracy in terms of MSE based on a trivial model that always predicts constant zero stock return, for the complete time period from 2016 to 2018.

| Stock $i$ | Tencent (0700.HK) | CCB (0939.HK) | Ping An (2318.HK) |
|---|---|---|---|
| MSE (unit: $\times 10^{-5}$) on | | | |
| Training set | 6.42 | 4.15 | 5.52 |
| Test set | 6.35 | 4.35 | 5.16 |
| Whole set | 6.40 | 4.19 | 5.45 |

## 5 Summary

In this paper, we construct a textual financial sentiment index for three stocks that are listed in the Hong Kong Stock Exchange and have hot discussion on Weibo.com, using the state-of-the-art NLP model BERT developed by Google recently, as the first BERT-based sentiment index in the literature, to the best of our knowledge. We also demonstrate the dominating feature of our financial sentiment classification result over other existing deep learning methods in terms of precision, recall, and averaged F1 score. Apart from textual channel to extract investors' sentiment, the traditional approaches utilize the option data from derivatives markets and stock market data directly, resulting in option-implied and market-implied sentiment indices, respectively. Based on these three different information channels, we propose a more comprehensive framework for financial sentiment analysis by interpreting the textual sentiment as individual investors' emotion, and the option-implied one as institutional investors' opinion, while the market-implied one as overall attitude of all market participants. We also discuss the predictability of sentiment on stock return in the individual level. Rather than using traditional econometric methods like VAR, we adopt LSTM as an ML tool in order to capture the possible nonlinearity of sentiment impact on stock return. It turns out that LSTM performs better than VAR on prediction for a yearly basis in the sense of lower MSE's. However, when it comes to a longer time period, ML models seem easily over-fitted. How to deal with this issue could be our future research direction, other than extension of our BERT-based sentiment construction to a market-level analysis.

## References

Baker, M. and Wurgler, J. (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61(4).

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Brown, G. and Cliff, M. (2005). Investor sentiment and asset valuation. *Journal of Business*, 78(2).

Chen, C., Fengler, M. R., Härdle, W. K., and Liu, Y. (2018). Textual sentiment, option characteristics, and stock return predictability, available at ssrn: https://ssrn.com/abstract=3210585.

Chen, C., Wu, K., Srinivasan, V., and Zhang, X. (2013). Battling the internet water army: Detection of hidden paid posters. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, pages 116–120. IEEE.

Chong, T. T.-L., Cao, B., and Wong, W.-K. (2014). A new principal-component approach to measure the investor sentiment, available at ssrn: https://ssrn.com/abstract=2631910.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. In *Machine Learning*, pages 273–297.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Easley, D., O'Hara, M., and Srinivas, P. S. (1998). Option volume and stock prices: Evidence on where informed traders trade. *Journal of Finance*, 53(2).

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

Han, B. (2008). Investor sentiment and option prices. *The Review of Financial Studies*, 21(1).

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kearney, C. and Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Lee, C. M. C., Shleifer, A., and Thaler, R. H. (1991). Investor sentiment and the closed-end fund puzzle. *Journal of Finance*, 46(1).

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185.

Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., and Du, X. (2018). Analogical reasoning on chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*.

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., and Benevenuto, F. (2016). Sentibench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Verma, R. and Soydemir, G. (2009). The impact of individual and institutional investor sentiment on the market price of risk. *The Quarterly Review of Economics and Finance*, 49(3).

Yan, X., Zhang, W., Ma, L., Liu, W., and Wu, Q. (2018). Parsimonious quantile regression of financial asset tail dynamics via sequential learning. In *Advances in Neural Information Processing Systems*, pages 1575–1585.