Combining Enterprise Knowledge Graph and News Sentiment Analysis for Stock Price Volatility Prediction

Jue Liu
School of Information,
Renmin University of China,
Beijing, 100872, China
liujue9918@ruc.edu.cn

Zhuocheng Lu
School of Information,
Renmin University of China,
Beijing, 100872, China
2015201938@ruc.edu.cn

Wei Du
School of Information,
Renmin University of China,
Beijing, 100872, China
ahduwei@ruc.edu.cn

Abstract

Many state of the art methods analyze sentiments in news to predict stock price. When predicting stock price movement, the correlation between stocks is a factor that can't be ignored because correlated stocks could cause co-movement. Traditional methods of measuring the correlation between stocks are mostly based on the similarity between corresponding stock price data, while ignoring the business relationships between companies, such as shareholding, cooperation and supply-customer relationships. To solve this problem, this paper proposes a new method to calculate the correlation by using the enterprise knowledge graph embedding that systematically considers various types of relationships between listed stocks. Further, we employ Gated Recurrent Unit (GRU) model to combine the correlated stocks' news sentiment, the focal stock's news sentiment and the focal stock's quantitative features to predict the focal stock's price movement. Results show that our method has an improvement of 8.1% compared with the traditional method.

1. Introduction

Stock price prediction has long been studied in financial field. The early articles on stock price prediction were mostly based on Fama's "Efficient Market Hypothesis" proposed in 1965 [1]. Fama believed that the stock price could reflect the impact of events that had occurred and events that had not occurred but were expected to affect stock price. This assumption also became the basis for the stock price prediction for many years to come.

Many state of the art methods analyze sentiments in news and comments on social media to predict stock price [23,24]. Besides, stock quantitative data is also an important factor to the prediction [10,25].

In practice, when considering the impact of public sentiments on stocks, the relevance between stocks is a factor that cannot be ignored. For example, the headline today is "Alibaba has completed its acquisition of XXX Company". Predictably, Alibaba's stock price will show an upward trend for a while. Naturally, companies that have certain links with Alibaba, such as their subsidiaries or companies with whom they have close relations, will also be affected by this favorable news and the corresponding stock price will rise. Therefore, when measuring the impact of a certain news on a company, it is necessary to consider not only the company mentioned in the news but also the company associated with it.

Therefore, in this paper we proposed a new method to measure the correlation between stocks—knowledge graph embedding. We first built the enterprise knowledge graph by crawling the news of each company, identifying named entities and extracting business relations. In the process of relation extraction, we chose the four most representative relations among the companies: shareholding, cooperation, management, and supply-customer. Afterwards, we used TransR model to translate the knowledge graph into vectors. Then we built the correlation matrix among stocks by calculating the Euclidean distance between the entity vectors. The advantage of this method is that it takes full account of the internal relevance between stocks in reality, avoiding the limitation of using only stock price data to measure relevance. Based on the constructed knowledge graph, relevant stocks can be identified.

By calculating the positive emotion score of each word in news, we structured daily news sentiment vectors for each stock. We used PB ratio, PE ratio, PS ratio, and PCF ratio to fully reflect the impact of quantitative features on stock price fluctuation. In terms of predicting model, we adopted the GRU model in consideration that the stock quantitative data and news sentiment vectors were both time series data. Then we took news sentiment vectors, correlation matrix and stock quantitative data as the input of GRU model and got the prediction results.

URI: https://hdl.handle.net/10125/59565
ISBN: 978-0-9981331-2-6

The rest of this paper is organized as follow: Section 2 summaries previous studies related to stock price prediction and sentiment analysis. Section 3 provides a review of the methodology used in this paper. Section 4 presents the process of news feature extraction. Section 5 gives an account of the acquisition of stock quantitative data. Section 6 describes in detail the use of the trans model to represent the entities and relations of the knowledge graph in the vector space. Section 7 describes GRU model. Section 8 provides the data we used and the experimental results. Section 9 summarizes the content of this paper and give a prospect of our future work.

2. Related work

With the advent of the Internet 2.0 era, more and more investors are beginning to use the emotion of investors on social media and the news about stocks to forecast stock price [2-4]. The concept of 'sentiment analysis' has been prosperous for a long time. 'Sentiment Analysis' refers to the use of natural language processing, text mining and computer linguistics to identify and extract subjective information from source material. Nuno Oliveira et al. used data from Twitter to extract emotion from microblogs and AAII, and used machine learning to forecast stock price [5]. Finally, using Diebold-Mariano to compare the model with the traditional autoregressive model. C.-Y. Chang et al. used the deep neural model to analyze the content of financial news to judge the economic value of the event to forecast the stock price [6]. Yue Zhang et al. used the Deep learning method to forecast the stock price [7]. They parsed news topics and extracted structured events. These extracted events were used as inputs to neural networks to forecast stock prices.

There have been many researches on news sentiment analysis. Piger et al. analyzed the impact of positive and negative news on the future performance of films [17]. Chen et al. used sentiment analysis method to build sentiment features, and applied genetic algorithm to further screened the features [18]. Moreover, the news sentiment analysis is also widely used in the prediction of stock price fluctuation. X. Ding et al. adopted text extraction to transforms structured events into triples, which include agent predicates and objects[19]. Then triples were used to predict the stock price. B. Wang et al. translated textual data into vectors via bag-of-words model and used as an input into a time series model [20]. Results showed that the accuracy of stock price fluctuation prediction increased with the use of sentiment analysis. Yoshihara et al. built models to analyze the long-term and short-term impact of news temporal properties on stock price volatility [21].

When predicting stock price, besides news sentiment, many researches also take the correlation between companies into consideration. W.-J.Ma et al. proposed a coupled random walk model to measure the relevance between stocks [8]. P. Chen et al. regarded the stock price sequence as the Wiener process, and replaced the correlation between stocks with the correlation between the two Wiener processes [9]. Xi.Zhang et al. proposed to apply tensor decomposition to news and a correlation matrix built by analyzing stock data [10], which are inadequate to completely embody the complex relationships between the companies. Zhang et al. established an ECM (Energy Cascade Model) to analyze the trend of stock price volatility by analyzing the dynamic cooperation and competitive business network [22] but overlooked a variety of relationships such as shareholding and supply-customer. These methods only used the known time series data of stock price to measure their similarity only from the perspective of data. However, in practical issue, the relation between the companies is complicated. Cooperation, holding, supply and competition will make the stocks closely related. Therefore, the method of measuring correlation simply from the data without considering these realistic relevance seems too lopsided and hollow.

Whereas enterprise knowledge graph comprises a variety of relationships between companies, we adopted enterprise knowledge graph embedding to measure the correlation between stocks.

In recent years, Trans series models have been widely applied to the knowledge representation learning of the knowledge graph [14, 15, 16]. Compared with traditional training methods, Trans series models have simple parameters, are easy to train, and are not prone to overfitting problems. Boders et al. proposed TransE in 2013 [14]. Its basic idea is to treat the relation in the triples (head, relation, tail) as a translation from the entity head to the entity tail. That is $h+r\approx t$. However, when modeling 1-N, N-1, N-N relations, TransE has many problems. So Wang et al. proposed TransH in 2014 [15]. When it came to different types of relation, TransH enabled an entity with different expression. However, both TransE and TransH assumed that entities and relations were embedded in the same space R^k. However, an entity has many attributes. Different relations focus on different attributes of the entity. As a result, the following problems may arise: Some entities have a low degree of similarity, so their 'distance' in entity space is very far. However, these two entities are very similar in some specific aspects, which leads to their close 'distance' in the relation space. For example, in fact, the similarity between Obama and Trump is very low. But they are all presidents of America. Therefore, in the space of 'US president', their distance is very

close. So they will get high similarity after embedding them into the space, which is contrary to the facts. In order to solve this problem, Lin et al. proposed TransR [16]. TransR projected entities and relations into different spaces and performed translation in the corresponding relation space.

3. Model framework

This paper comprehensively uses the news sentiment, the correlation between stocks, and the stock quantitative data to forecast the rise and fall of stock prices. The main steps of our method is depicted in Figure 1.

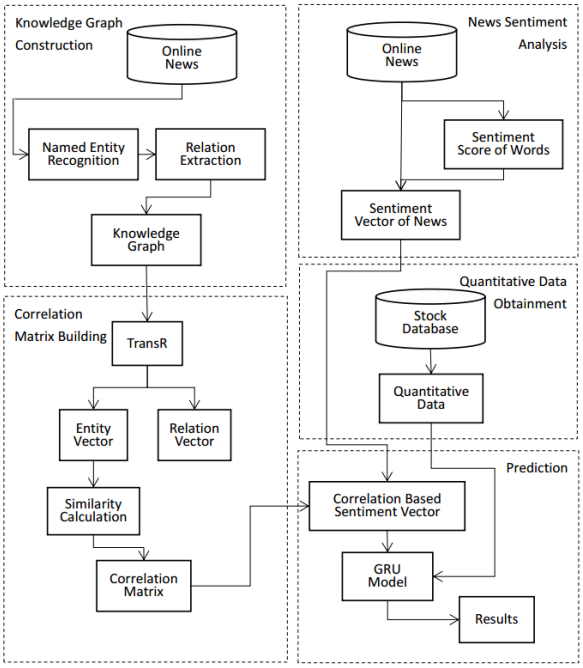

Figure 1. Model framework

This paper mainly adopted feature extraction method based on word pointwise mutual information while extracting news sentiment. We had artificially established two sets of positive and negative words. Then we used these two sets to calculate the positive sentiment score of words in the news. The sentiment of the news could be structured upon this score.

When choosing the stock quantitative feature, inspired by previous related studies [10, 11, 12], we chose four commonly used quantitative features of stock: PS, PE, PB, and PCF ratio. By obtaining the quantitative data of each stock over a period of time, we could obtain the quantitative data sequence of each stock for subsequent forecasting.

The innovation of this paper is that when considering the impact of news on stocks, it also considers the correlation between stocks. And when calculating the correlation matrix, starting from the existing knowledge graph, which is closer to the reality, can avoid the one-sidedness of the traditional methods that only use stock data to calculate the correlation. As a result, a news about stock A may affect the price of all the stocks related to stock A, which is more consistent with reality.

After obtaining the news sentiment sequences and quantitative data of the stock, we used them as the input of the GRU model. Then we performed model parameter learning and stock price forecasting. Compared with the traditional RNN model, the model not only inherits the characteristics that can effectively describe the feature influence on time series, but also adds the forgetting and saving mechanism. Besides the parameters are simpler, thus can acquire better forecast results.

3.1. Enterprise knowledge graph construction

We selected four most representative relationships between companies: shareholding, cooperation, management, and supply-customer.

In our knowledge graph, shareholding means holding a certain number of shares and not being directly engaged in production and operation business. Cooperation always accompanies by business activities in which different enterprises jointly develop products or markets through agreements or other means to share benefits in order to obtain overall advantages. Management is the acquisition of right of planning, organizing, coordinating, controlling and commanding in another company. And supply-customer refers to one enterprise offering goods or other things to another company.

In order to get the relationships of cooperation and supply-customer, we crawled related news from Sina[1], Tencent[2] and Phoenix News[3], which are the mainstream medias in China and can contain most significant news. The management and shareholding relationships between enterprises can be obtained from companies' annual reports. Then we transferred these PDF-formatted reports to a valid XML document.

After obtaining the raw textual data containing the four relationships from news and annual reports, we applied Named Entity Recognition (NER) and Relation Extraction (RE) to process the raw data. We used the

[1] http://www.sina.com.cn/

[2] http://news.qq.com/

[3] http://news.ifeng.com/

traditional rule-based named entity recognition to identify the company name. Then we used Baidu to search the company name plus stock and determine whether the company is listed by judging the search results. After we judged the direction and category of the relationship, we represented the relation as a form of three tuples (*h*, *r*, *t*). *h* and *t* represent head and tail entities (enterprises) respectively, *r* represents the correlation between the head and tail entities. For example, (Company A, shareholding, Company B) means that Company A holds Company B. After stored the triples into Neo4j database, we accomplished the construction of enterprise knowledge graph.

3.2. Feature extraction of news

3.2.1. Positive sentiment score calculation of a word. The impact of news on stocks falls into two categories: positive and negative. In order to analyze whether a news' impact on stock price is positive or negative, we first defined the concepts of positive and negative words. Positive words are words that are beneficial to the rise of stocks, such as ballooning. Negative words are words that are detrimental to the rise of stocks, such as bear market. We can construct the sentiment vector of the news by extracting the sentiment words in the news and analyzing the positive sentiment score of these words.

Suppose we already have a large enough news set N:

$$N = \{news_1, news_2, news_3, \dots, news_M\} \quad (1)$$

Each news can be seen as a sequence of words, that is:

$$news_i = w_1 w_2 \dots w_{n_i} \quad (2)$$

It is easy to see the probability of the word w appearing in the news $p(w) = \frac{|\{i|w \epsilon news_i\}|}{M}$.

We note that the correlation between w and v is:

$$corr(w,v) = ln\frac{p(w,v)}{p(w)p(v)} = ln\frac{M*|\{i|w \epsilon news_i v \epsilon news_i\}|}{|\{i|w \epsilon news_i\}|*|\{i|v \epsilon news_i\}|} \quad (3)$$

Suppose we have obtained two sets of positive words and negative words, and they are denoted as

$$Pos = \{w_{p1}, w_{p2}, \dots, w_{pk}\}, \text{ and } Neg = \{w_{p1}, w_{p2}, \dots, w_{ps}\} \quad (4)$$

We can calculate the positive sentiment score of a word *w* based on its correlation with Pos and Neg:

$$pvalue(w) = \frac{1}{k}\sum_{v \epsilon Pos} corr(w,v) - \frac{1}{s}\sum_{v \epsilon Neg} corr(w,v) \quad (5)$$

3.2.2. News sentiment vector. We assume that the maximum value of the positive sentiment score of a word is Max and the minimum value is Min. By dividing the positive sentiment score into L intervals, and the *j*-th interval I_j is:

$$I_j = \left[Min + (j-1) * \frac{Max-Min}{L}, Min + j * \frac{Max-Min}{L}\right], j = 1,2, \dots, L \quad (6)$$

A stock's news at day i is denoted as $news_i = w_1 w_2 \dots w_{n_i}$, and its sentiment vector $f(news_i)$ is:

$$f(news_i) = (y_{i1}, y_{i2}, \dots, y_{iL}), \text{ where } y_{ij} = \frac{1}{n_i}|\{k|pvalue(w_k) \epsilon I_j\}| \quad (7)$$

The news sentiment vector can be viewed as a histogram approximation of the distribution of words' positive sentiment score. The number of histogram intervals is *L*. We can also adjust *L* to control the dimension of the sentiment vector, which will be the input of GRU model.

3.3. Quantitative features extraction

For the *i*-th stock, its quantitative features could be represented by a vector $x_i = (x_{i_1}, x_{i_2}, \dots, x_{i_k})$, where k represents the number of quantitative features, x_{i_k} represents the value of the *k*-th feature of the *i*-th stock.

According to related researches [12, 13], we chose the four most commonly used stock quantitative features: Price to Book Ratio (PB), Price Earnings Ratio (PE), Price to Sales Ratio (PS), Price Cash Flow Ratio (PCF). PE ratio is one of the most commonly used indicators to evaluate whether the stock price level is reasonable, and it reflects the ratio of market price per share to earnings per share (EPS). If a stock has a higher PE ratio, it illustrates that the market over-estimate its future earnings growth. Thus it indicates a lack of investment value and coming bubbles. At the same time, the difference for stock PE ratios in different industries is large, therefore, similar companies tend to have similar PE ratios. There are many similarities between PB and PE, and they all reflect the intrinsic value of a company. Normally companies with lower PB have higher investment value. For those stocks whose earnings per share changes greatly with the industry, the PB ratio is a more reliable indicator than the PE ratio, because the net assets per share of listed companies generally do not fluctuate too much regardless of the industry's prosperity. The PCF ratio is also a commonly used indicator for forecasting the value of stocks. It is the ratio of stock price to cash flow per share and can be used to evaluate the price and risk level of stocks. The

PCF ratio is a good indicator when evaluating stocks that have positive cash flow but are not profitable. The PS ratio is similar to the PB ratio, but it is relatively stable and not easily manipulated. In addition, it will not be negative. Therefore, for loss-making enterprises and insolvent enterprises, its value multiplier can also make sense.

3.4. Stock correlation matrix calculation

3.4.1. Knowledge graph embedding. The basic idea of TransR is shown in Figure 2. For each triples (*h*, *r*, *t*), the entity in the entity space is first projected to the relation r through M_r to get$h_r$andt_r, and then makes h_r+r≈t_r. A specific relation projection (colored circle) enables the head/tail entity to be close to each other when they do have this relation, and away from each other when they do not have this relation (colored triangle).

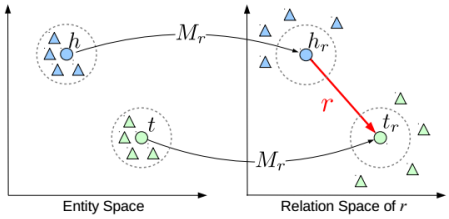

Figure 2. Sketch map of TransR

In TransR, for each triples *(h,r,t)*, entity embedding can be expressed as h, $t \in R^k$and relation embedding as $r \in R^d$. Here the embedded dimension of the entity and the relation can be different, i.e. there can be *k*≠*d*.

For each relation *r*, we define a projection matrix $M_r \in R^{k\times d}$that projects the entity space into relation space. The projected entity vector is $h_r = hM_r, t_r = tM_r$ the corresponding loss function is

$$f_r(h,t) = |h_r + r - t_r|^2 \quad (8)$$

In practice, we add restrictions on h, r, t and the mapping matrix, that is,

$$\forall h,r,t \text{ has } \| h \|_2 \leq 1, \| r \|_2 \leq 1, \| t \|_2 \leq 1, \| hM_r \|_2 \leq 1, \| tM_r \|_2 \leq 1 \quad (9)$$

We define the following score function in training:

$$L = \sum_{(h,r,t)\epsilon S} \sum_{(h',r',t')\epsilon S'} max(0, f_r(h,t) + \gamma - f_r(h',t')) \quad (10)$$

Where γ is the margin and *S* is the correct triples, S' is an incorrect triples. The learning process is implemented by the stochastic gradient descent. The specific algorithm is shown in Algorithm 1.

Algorithm 1. TransR

Input Training Set S = $\{(h,r,t)\}$, entities and relations, sets E and R, and margin γ, entity embeddings dim K, relation embeddings dim d, Projection matrix M_r

initialize r ← uniform$(-\frac{6}{\sqrt{d}}, \frac{6}{\sqrt{d}})$ for each r ∈ R
  r ← r/|r| for each r ∈ R
  e ← uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity e ∈ E
loop
  h_r ← hM_r, t_r ← tM_r for each h, r
  h_r ← $h_r/|h_r|$, t_r ← $t_r/|t_r|$ for each h_r, t_r
  e ← e/|e| for each entity e ∈ E
  S_{batch} ← sample(S,b) // sample a minibatch of size b
  T_{batch} ← ∅ // initialize the set of pairs of triplets
  for (h, r, t) ∈ S_{batch} **do**
    (h′, r, t′) ← $sample(S'_{(h,r,t)})$ // sample a corrupted triplet
    T_{batch} ← T_{batch} ∪ {((h, r, t))}
  end for
  Update embeddings w.r.t $\sum_{((h,r,t),(h',r,t'))\in T_{batch}} \nabla[f_r(h,t) + \gamma - f_r(h',t')]_+$
end loop

In this way, the entities (enterprise) in the enterprise knowledge graph are embedded in the *k*-dimensional hyperplane, and the relations are embedded in the *d*-dimensional hyperplane. *K* and *d* can be selected according to the practical experience.

3.4.2. Correlation matrix calculation. According to the principle of TransR, the distance of *k*-dimensional vectors between similar entities is close. As we had represented each company in the knowledge graph as a *k*-dimensional vector in 3.3.1, we measured the similarity between stocks by calculating Euclidean distance. Noted that the similarity between the i-th stock and the *j*-th stock was a_{ij}, the vector corresponding to the *i*-th stock is $z_i = (z_{i1}, z_{i2}, \dots, z_{ik})$, and the vector corresponding to the *j*-th stock is $z_j = (z_{j1}, z_{j2}, \dots, z_{jk})$. We have:

$$a_{ij} = \frac{1}{1+d(z_i,z_j)} \quad (11)$$

Then we can get the stock correlation matrix$(a_{ij})_{n\times n}$, where *n* is the number of stocks.

3.5. GRU model

For stock price prediction, both quantitative features and news information are time series data. Gated Recurrent Unit (GRU) [31] can well characterize the continuous effects of features over time. GRU is a variant of Long Short-Term Memory (LSTM) [30], and LSTM is a special kind of Recurrent Neural Network (RNN) that is capable of learning long-term dependencies. RNN has an advantage in dealing with time series data because RNN can retain the previous information instead of only based on current information. There are three gates in each cell in LSTM, forget gate, which decides what information we're going to throw away, input gate, which determines what new information we're going to store in the cell state, and output gate, which resolves the final results. Information attenuates every time it passes through

forget gate. Therefore, the closer the distance between the previous information and the current one, the larger the weight of previous information is, which conform with the circumstance when considering news' impact on stock market. GRU merges LSTM's forget gate and input gate into update gate and therefore is more efficient and useful with less variants than LSTM. So we use the GRU model as our prediction model.

The input layer of this model contains two parts. First is the daily news sentiment vector of each related stock. We denote the i-th day's news sentiment sequence of the j-th stock as $f_j(news_i)$. If we had selected n stocks to predict their movements, we arranged the news sentiment vectors for the n stocks in the i-th day by column, and we would get the news sentiment matrix for the i-th day as $B = [f_1(news_i), f_2(news_i), \dots, f_n(news_i)]'$. We could calculate the correlation based sentiment matrix of the i-th day by multiplying the correlation matrix $A_{n\times n}$ with the news sentiment matrix as $C = A \times B$. The j-th row of C was denoted as $g_j(news_i)$ which shows the news sentiment vector of the i-th day of the j-th stock after considering the correlation between stocks. We take it as GRU input. Another input for the GRU model is the vector that denotes stock quantitative features.

We used a concatenation of sequences of news sentiment and quantitative features as the input of GRU model. Our GRU model builds on the framework provided by Keras[4] and the model is make up of five layers. The specific structure of the model is shown in Figure 3.

The first layer is 'recurrent.GRU' with the activation function 'relu'. The layer updates its parameters while constantly entering the sequences of news sentiment and quantitative features of M days. With this layer, the previous news and quantitative features can have an impact on the later day's prediction, and can retain the timing characteristics of the sequences. The second and fourth layers are Dropout layers, and are used to randomly inactivate neurons and prevent over fitting. Experiments shows that the accuracy of prediction can be effectively improved with the dropout layers. The third and fifth layers are Dense layers, and are used to fuse the information of news and quantitative data and get the final prediction. The output of the GRU model is a M-dimension vector consist of '0' and '1' ($\{0,1\}^M$) for indicating the rise and fall of the stock price.

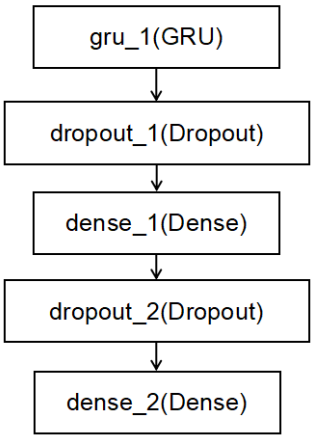

Figure 3. Structure of GRU model

4. Experiment

4.1. Dataset description

Our datasets can be divided into three parts: stock-related news, enterprise knowledge graph, and stock quantitative data. We crawled news related to each stock from Sina Finance[5] from January 2017 to May 2018. Then we obtained quantitative data of China A-share from UQER[6] from January 2017 to May 2018. Since the stock market closes on weekends and holidays, we counted news of the close day on the first open day after the close in order to match the stock quantitative data and the news sentiment sequence by the number of days. We integrated news from Sina[7], Tencent[8], and Phoenix News [9], and extracted triples with named entity recognition and relation extraction. After storing the triples into Neo4j database, we accomplished the construction of enterprise knowledge graph. Our knowledge graph contains 3,113 companies, from which we selected China's top 500 companies and A-share listed companies as our research objects. Statistical information of the datasets is shown in Table 1.

Table 1. Dataset Description

Trading day period	From 2017-01-01 to 2018-05-15
The number of trading days	331

[4] https://keras.io/
[5] http://finance.sina.com.cn/
[6] https://uqer.io/
[7] http://sina.com.cn/
[8] http://news.qq.com/
[9] http://news.ifeng.com/

The number of companies in our knowledge graph	3113
The number of stocks	357
The number of news	182500

We used Jieba[10] as our text segmentation tool. After the segmentation of all the news in our dataset, we sorted the words according to their frequency and selected the top 10% ones. Then we manually selected our positive and negative words from the top 10% ones. The words set selected in the experiment is shown in Table 2.

Table 2. Positive and Negative Words Set

Positive Words	优 增长 崛起 提升 推进 回暖 涨停 发展 高速 增资 扩股 收购 领先 优势 标杆 升级 稳健 涨幅 利润 上市 收益 增加 稳定 提高 增值 盈利 增持 (excellent, increase, rise, improve, advance, upturn, limit-up, develop, rapid, raise, expansion, purchase, lead, superiority, benchmarking, promotion, steady, inflate, profit, list, earnings, add, stable, enhance, increment, gain, increase holdings)
Negative Words	亏损 减少 下滑 新低 下跌 受贿 缺陷 下降 违约 虚假 误导性 亏损额 外债 减持 涉嫌 退市 债务 负债 违反 侵权 (loss, decline, slash, new low, fall, bribery, defect, descend, default, false, misleading, deficit, foreign debt, reduce holdings, suspected, delist, debt, liabilities, violate, tort)

4.2. Results

In order to test the effect of using the knowledge graph to evaluate the relevance between stocks and the function of the GRU neural network model mentioned in the paper, we designed several groups of comparative experiments. In Method 1 (denoted as Quan+News+GRU), the news sentiment sequences of each stock was directly taken as the input of a GRU network. In Method 2 (denoted as Quan+News+CSS+GRU), we used stock data to calculate the stock Correlation Matrix. Inspired by the previous work [10, 28, 29], we apply a Coupled Stock Similarity (CSS) measure to calculate the correlation between stocks. The detail of the method can be referred to literature [28]. Method 3 (denoted as Quan+News+Corr+GRU) is our proposed method. The accuracy of prediction is measured by $\frac{numberofdayscorrectlypredicted}{totalnumberoftradingdays} \times 100\%$. The results are shown in Table 3.

Table 3. Results

Method	Quan+News+GRU	Quan+News+CSS+GRU	Quan+News+Corr+GRU
Mean Accuracy	51.3%	55.6%	59.4%
Worst Accuracy	41.7%	48.2%	51.0%
Best Accuracy	63.1%	67.7%	79.2%

Table 4 shows the prediction accuracy of five stocks randomly selected from our dataset. We found that for all the five stocks, Method 3's performance is evidently better than Method 1 and Method 2.

Table 4. Accuracy of Several Stocks

Stock Code	Stock Name	Quan+News+GRU	Quan+News+Corr+GRU	Quan+News+Corr+GRU
601668.SH	China State Construction Engineering Co., Ltd.	54.4%	56.1%	58.6%
601669.SH	Power Construction Corporation of China, Ltd.	57.2%	62.7%	64.4%
601600.SH	Aluminum Corporation Of China, Ltd.	58.2%	66.4%	71.3%
000100.SZ	TCL Corporation	56.9%	58.8%	61.0%
600027.SH	Huadian Power International Co., Ltd.	53.3%	54.0%	55.2%

[10] https://pypi.org/project/jieba/

From Table 4, we can see that the mean accuracy of Method 1 in the experimental set (357 stocks for 87 days) was 51.3%. While in Method 2, the accuracy was improved to 55.6%, which shows that using stock data to calculate the correlation between stocks can improve the accuracy of news sentiment sequences in stock price forecasting. Method 3 also takes correlation into consideration as Method 2. But with the using of enterprise knowledge graph, the accuracy of Method 3 was 3.8% higher than that of Method 2, which shows that our enterprise knowledge graph has an advantage over stock data when measuring correlation between stocks.

5. Summary and Future Work

This paper proposes a stock price prediction model based on knowledge graph and news sentiment analysis. Specifically, GRU model is employed to combine the stock quantitative data and news sentiment to make prediction. In the experiment, we find that our method can improve the accuracy of 8.1% with the use of enterprise knowledge graph.

In our future work, we plan to predict stock price movements only for stocks having strong correlation with each other in the knowledge graph. For the knowledge graph that includes a large number of enterprises, it will be a heavy workload for TransR to embed such a large graph. In addition, the stock correlation matrix will become high dimensional and may cause sparsity which can reduce the accuracy of prediction. Therefore, future work will apply clustering algorithm to knowledge graph to divide strongly correlated enterprises into a community, and then use embedding method within the communities. This will not only save computing time, but also solve the problem of sparsity in correlation matrix.

6. References

[1] Fama E F. The behavior of stock-market prices[J]. The journal of Business, 1965, 38(1): 34-105.

[2] J. Bollen,H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.

[3] T.H. Nguyen, K. Shirai, Topic modeling based sentiment analysis on social media for stock market prediction, in: Proceedings of the 53rd Annural Meeting of the Association for Computational Linguistics (ACL-15), 2015.

[4] R. Feldman, B. Rosenfeld, R. Bar-Haim, M. Fresko, The stock sonar-sentiment analysis of stocks based on a hybrid approach, in: Proceedings of the Twenty-Third IAAI Conference, 2011.

[5] Nuno Oliveira, Paulo Cortez, Nelson Areal, The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices, Expert Systems with Applications, Volume 73, 2017, Pages 125-144, ISSN 0957-4174.

[6] C.-Y. Chang, Y. Zhang, Z. Teng, Z. Bozanic, B. Ke, Measuring the information content of financial news, in: Proceedings of the 26th International Conference on Computational Linguistics (COLING'16), 2016, pp. 3216–3225.

[7] Ding X, Zhang Y, Liu T, et al. Using Structured Events to Predict Stock Price Movement: An Empirical Investigation[C]//EMNLP. 2014: 1415-1425.

[8] W.-J. Ma, C.-K. Hu, R.E. Amritkar, Stochastic dynamical model for stock-stock correlations, Phys. Rev. E 70 (2004) 026101.

[9] P. Chen, Modelling the Stochastic Correlation, KTH, Mathematical Statistics,2016 Master's thesis.

[10] Xi Zhang, Yunjia Zhang, Senzhang Wang, Yuntao Yao, Binxing Fang, Philip S. Yu, Improving stock market prediction via heterogeneous information fusion, Knowledge-Based Systems, Volume 143, 2018, Pages 236-247, ISSN 0950-7051.

[11] Q. Li, L. Jiang, P. Li, H. Chen, Tensor-based learning for predicting stock movements, in: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15), 2015, pp. 1784–1790.

[12] E.F. Fama, K.R. French, Common risk factors in the returns on stocks and bonds,J. Financ. Econ. 33 (1) (1993) 3–56.

[13] Bordes, A.; Glorot, X.; Weston, J.; and Bengio, Y. 2014. A semantic matching energy function for learning with multirelational data. *Machine Learning* 94(2):233–259.

[14] Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of NIPS*, 2787–2795.

[15] Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In Proceedings of AAAI, 1112–1119.

[16] LIN, Y.; LIU, Z.; SUN, M.; LIU, Y.; ZHU, X.. Learning Entity and Relation Embeddings for Knowledge Graph Completion. AAAI Conference on Artificial Intelligence, North America, feb. 2015.Available at: Date accessed: 01 Jun. 2018.

[17] A.K.Davis, J.M. Piger, L.M. Sedor, Beyond the Numbers: An Analysis of Optimistic and Pessimistic Language in Earnings Press Releases, Federal Reserve Bank of St. Louis Working Paper Series, 2006.

[18] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Trans. Inform. Syst. (TOIS) 26 (2008) 12.

[19] X. Ding, Y. Zhang, T. Liu, J. Duan, Using structured events to predict stock price movement: An empirical investigation, in: Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP-14), 2014, pp. 1415–1425.

[20] B. Wang, H. Huang, X. Wang, A novel text mining approach to financial time series forecasting, Neurocomputing 83 (2012) 136–145.

[21] A. Yoshihara, K. Fujikawa, K. Seki, K. Uehara, Predicting stock market trends by recurrent deep neural networks, in: Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Springer, 2014, pp. 759–769.
[22] Zhang, W., Li, C., Ye, Y., Li, W., and Ngai, E. W. 2015. Dynamic business network analysis for correlated stock price movement prediction, IEEE Intelligent Systems, 30(2), pp.26-33.
[23] Adam Atkins, Mahesan Niranjan, Enrico Gerding, Financial news predicts stock market volatility better than close price, The Journal of Finance and Data Science, Volume 4, Issue 2, 2018, Pages 120-137, ISSN 2405-9188.
[24] Yu-Chen Wei, Yang-Cheng Lu, Jen-Nan Chen, Yen-Ju Hsu, Informativeness of the market news sentiment in the Taiwan stock market, The North American Journal of Economics and Finance, Volume 39, 2017, Pages 158-181, ISSN 1062-9408.
[25] Yu-Chen Wei, Yang-Cheng Lu, Jen-Nan Chen, Yen-Ju Hsu, Informativeness of the market news sentiment in the Taiwan stock market, The North American Journal of Economics and Finance,
Volume 39, 2017, Pages 158-181, ISSN 1062-9408.
[26] Mourad Gridach, Character-level neural network for biomedical named entity recognition, Journal of Biomedical Informatics, Volume 70, 2017, Pages 85-91, ISSN 1532-0464.
[27] Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, Bo Xu, Joint entity and relation extraction based on a hybrid neural network, Neurocomputing, Volume 257, 2017, Pages 59-66, ISSN 0925-2312.
[28] C. Wang, L. Cao, M. Wang, J. Li, W. Wei, Y. Ou, Coupled nominal similarity in unsupervised learning, in: Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM-11), ACM, 2011, pp. 973–978.
[29] F. Li, G. Xu, L. Cao, Coupled matrix factorization within non-iid context, in: Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2015, pp. 707–719.
[30] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Neural Computation, Vol. 9, No. 8, 1997, pp. 1735-1780.
[31] Cho, Kyunghyun & van Merriënboer, Bart & Bahdanau, Dzmitry & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. 10.3115/v1/W14-4012.