Research on the Relevant Matching Method of Internet MediaInformation-Stock Assets Based on the Combination of Literal and Semantic

Xianzu Liu[1, *]

[1] School of Economics and Management, Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: njust_lxz@163.com

Abstract. Classify and match kinds of media information has been a problem to be solved. To solve the problem, the study finished the experiment of text matching, include the process of data collection and building the media information with stock asset matcher, based on the perspective of literal matching and semantic matching. The experimental data comes from financial news websites such as eastmoney.com and jrj.com. This study uses a small sample learning model to solve the problem of lacking labeled data and too much types of financial media information. Based on the Siamese network and Word2Vec model, the matcher performs well in matching tasks. This study will provide the basic way to support the research of impact of Internet media information on financial market asset prices from the perspective of salient effects.

Keywords: NLP; Few-Shot Learning; Media Information; Siamese Network.

1. Introduction

In the financial trading market, information plays a key role in guiding investors' behavioral. In particular, with the popularization of the Internet and the rapid development of new media, a lot of financial information has appeared, including stock bars, microblogs, blogs, forums and other platforms as the release medium. The emergence of this information, with its enhanced social interaction characteristics, is constantly changing the way that investors acquire, share, evaluate and utilize information. However, the booming of media information complicates the structure of Internet media information. So individuals or companies engaging in related research work need to spend more time to process the information. Compared with other media information, The difficulty of classifying and matching financial media information will be greatly increased.

With the help of the small sample learning algorithm in deep learning, the semantic information in the article-level article can be extracted, and the asset entity in the article can be captured from the article. Existing research lacks exploration in the field of Chinese data. The main reason is not only the lack of research materials, but also because the complexity of Chinese text data structure is much higher than that of English data. Compared with English data, there is no space between words in Chinese data, and the meaning of words is rich. These reasons make it difficult to recognize new words.

In this paper, I conduct research around Internet financial media information. First, I created a dataset of the Internet financial media information. Secondly, based on the literal information of the text, the matching of asset entities is realized by constructing a feature dictionary. Asset entities can be matched by rules. In order to capture the semantic signal of the asset entity in the article, a few-shot learning method is used. Finally combine the two methods to realize the matching of the financial media information. based on the semantic information of the text, a small sample learning method is used.

2. Research Design

First, this research used web crawlers to collect data. I should analyze the information distribution, characteristics of the websites and select appropriate crawler tools. Second, clean the data. The

purpose of which is to obtain high-quality data and modify the data into a format suitable for the research. Then the media information would be preprocessed. Including removing stop words and common words, word segmentation, training Word2Vec word vector model, building asset entity feature dictionary, vectoring text data, etc. These steps have two purposes, one is to construct a feature dictionary for extracting keywords, and the other is to process the text data into the format of the input of the siamese network text matching model. Next, the asset entity feature dictionarys can be used to implement the literal-based asset entity matching method. And the processed data can be used to train the semantic-based asset entity matching model through the few-shot method. Combining two methods, a asset entity matching model based on the combination of literal and semantic can be used to match asset entity.

3. Model Description

3.1. LSTM

Long short-term memory (LSTM)[1] is an artificial neural network used in the fields of artificial intelligence and deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. Such a recurrent neural network (RNN) can process not only single data points (such as images), but also entire sequences of data (such as speech or video). For example, LSTM is applicable to tasks such as unsegmented, connected handwriting recognition, speech recognition, machine translation, robot control, video games, and healthcare. LSTM has become the most cited neural network of the 20th century.

A common LSTM unit is composed of a cell, an input gate, an output gate[2] and a forget gate.[3] The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.

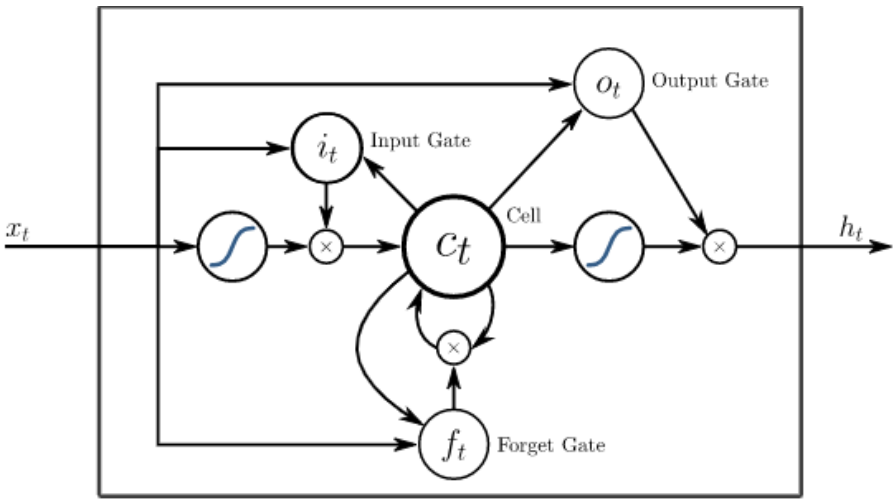


Fig. 1 The structure of LSTM

LSTM networks are well-suited to classifying, processing and making predictions based on time series data, since there can be lags of unknown duration between important events in a time series. LSTMs were developed to deal with the vanishing gradient problem[4] that can be encountered when training traditional RNNs. Relative insensitivity to gap length is an advantage of LSTM over RNNs, hidden Markov models and other sequence learning methods in numerous applications.

3.2. Siamese Network

A siamese neural network (sometimes called a twin neural network) is an artificial neural network that uses the same weights while working in tandem on two different input vectors to compute comparable output vectors.[5,6,7] Often one of the output vectors is precomputed, thus forming a baseline against which the other output vector is compared. This is similar to

comparing fingerprints but can be described more technically as a distance function for locality-sensitive hashing.

Uses of similarity measures where a twin network might be used are such things as recognizing handwritten checks, automatic detection of faces in camera images, and matching queries with indexed documents. The perhaps most well-known application of twin networks are face recognition, where known images of people are precomputed and compared to an image from a turnstile or similar. It is not obvious at first, but there are two slightly different problems. One is recognizing a person among a large number of other persons, that is the facial recognition problem. DeepFace is an example of such a system.[8] In its most extreme form this is recognizing a single person at a train station or airport. The other is face verification, that is to verify whether the photo in a pass is the same as the person claiming he or she is the same person. The twin network might be the same, but the implementation can be quite different.

This research use LSTM and siamese neural network to build classifying predictions. The structure of the model is shown in Figure 2.

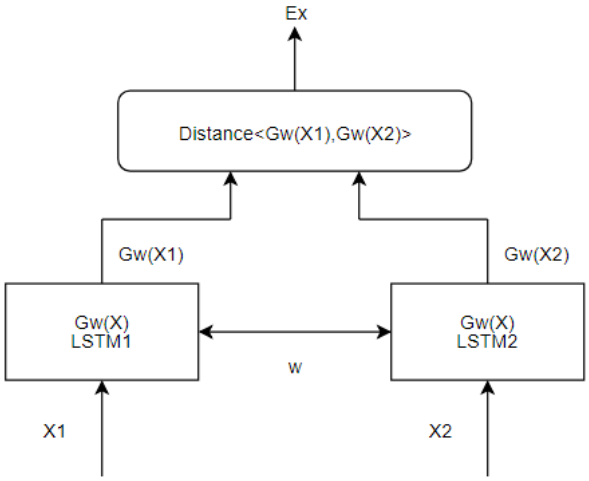


Fig. 2 The structure of Bi-LSTM

4. Data Description

4.1. Data Collection

Considered with investors' information collection habits, the data is selected from mainstream financial media websites. At present, the mainstream financial media websites include Oriental Fortune Network, Tencent Finance, Phoenix Finance, NetEase Finance and so on. Considering the availability of data, eastmoney.com is selected as the acquisition source of marked data and zgjrjw.com is selected as the acquisition source of unmarked historical news data.

A total of 75000 items were crawled. After eliminating useless and invalid data, I get 73,205 items. A total of more than 800,000 items were collected from zgjrjw.com. The date of items range from 2015 to 2020. The format of the data is shown in Table 1.

Table 1. The format of the data

Index	url	content	Stock Code
1	eastmoney.com	PingAn Bank Financial Report	000001

4.2. Data Processing

Data preprocessing is the most important part of natural language processing. Unlike English data, Chinese text does not have spaces between words, so the first processing is to split words on sentences. Data cleaning can be done either before word segmentation or after word segmentation. The main function of data cleaning is to remove useless words. Finally, the texts are converted into vectors by Doc2Vec model. The steps of data preprocessing are shown in Figure 3.

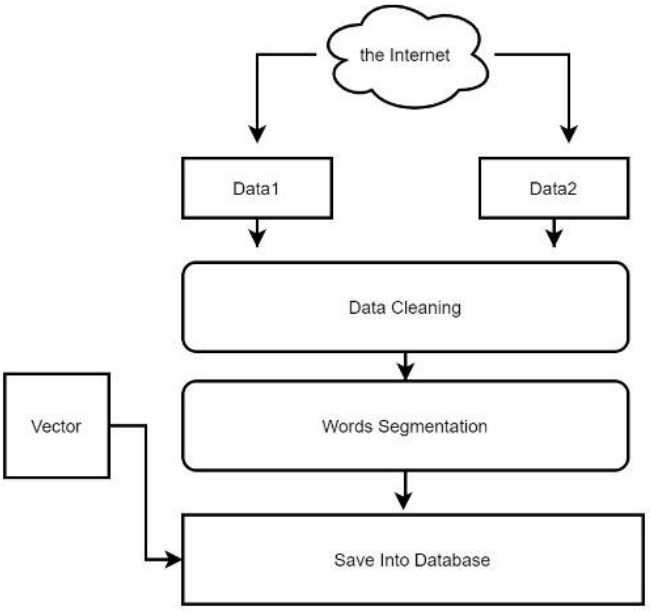


Fig. 3 The steps of data preprocessing

5. The procedure of experimental and result

5.1. Build characteristic dictionary of individual stock

In the experiment, Word2Vec is selected as the tool for converting words to vector. And I used genism to implement Word2Vec.Gensim is a free Python development package with Word2Vec model training functions, which can easily implement Word2Vec training tasks. Euclidean distance is used to calculate the similarity between words. The experiment used more than 800,000 pieces of corpus data to train the Word2Vec model to obtain the most relevant vocabularies of the stock codes and asset entities. The result of characteristic dictionaries of some stocks are shown as in table 2.

Table 2. The result of characteristic dictionaries of some stocks

Index	Company	Characteristic1	Characteristic2
1	000651	midea	Dong Mingzhu
2	000100	TCL	Huawei
...	...	...	...

5.2. Model Training

The network is implemented by Keras module package, setting batch processing 512, word vectors' dimension 300, and training rounds 20. The structure of the model is shown in figure 4.

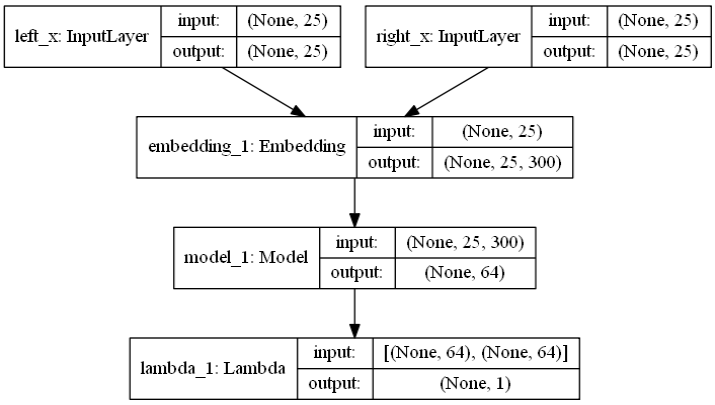


Fig. 4 The structure of training model

5.3. Construction of classifier prediction

First, the feature dictionaries are used to build the asset entities extract model based on literal matching and the twin network text matching model is used to build the asset entities classifier prediction based on semantics LSTM model. Finally, the two methods are combined to build the asset entities matching model.

5.4. Results of Comparison

In order to verify the model, 10 pieces of data in the database are randomly selected as the test set. A classifier based on small sample learning and a method based on a feature dictionary are used to extract asset entities from the test set, and a comparative analysis is made. See Table 3 and Table 4 for the specific comparison. Mark 1 represents that the model detects that the article is related to the stock.

Table 3. Results based on Bi-LSTM matching

Index	000001	000002	000063	000069	000333
1	0	1	1	1	1
2	0	0	1	0	1
3	0	0	1	0	1
4	0	0	1	0	1
5	0	0	1	0	1
6	0	0	1	0	1
7	0	0	1	0	1
8	0	0	1	0	1
9	0	0	1	0	1
10	0	1	1	1	1

Table 4. Results of characteristic dictionaries

Index	000001	000002	000063	000069	000333
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	1	0	0	0	0
5	0	1	0	0	0
6	0	0	1	0	0
7	0	0	0	1	0
8	0	0	0	0	0
9	0	0	0	0	0
10	0	0	0	0	0

The asset entities matching model based on characteristic dictionaries may not capture the semantic information between the text and the asset entities. The model based on Bi-LSTM can capture the semantic information between articles and asset entities, and excessive matching problem may happen. The reason for this over matching phenomenon may be that the amount of asset entities in the database is insufficient.

6. Summary

From the comparison results of the experiments, we can find that the asset entities matching model based on characteristic dictionaries and asset entities matching model based on Bi-LSTM have their own advantages and disadvantages. Literal matching method is fast and accurate, but they cannot capture the semantic information of texts. Semantic matching method based Bi-LSTM can capture potential asset entities information, but its efficiency is slow and results may be over matched. Combining these two methods, we can not only capture the asset entity signal literally, but also capture the potential asset entity information semantically, but the efficiency is still too low.

References

[1] Sepp Hochreiter, Jürgen Schmidhuber. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780.

[2] Hochreiter, Sepp, Schmidhuber. LSTM can solve hard long time lag problems. Advances in Neural Information Processing Systems, 1996.

[3] Felix A. Gers, Jürgen Schmidhuber, Fred Cummins. Learning to Forget: Continual Prediction with LSTM. Neural Computation, 2000, 12(10): 2451-2471.

[4] Hochreiter, Sepp. Untersuchungen zu dynamischen neuronalen Netzen. Technical University Munich, Institute of Computer Science, 1991.

[5] Chicco, Davide. Siamese neural networks: an overview. Artificial Neural Networks, Methods in Molecular Biology, 2020, vol. 2190 (3rd ed.)

[6] Bromley, Jane, Guyon, et al. Signature verification using a Siamese time delay neural network. Advances in Neural Information Processing Systems, 1994, 6: 737–744.

[7] Chopra, Hadsell, LeCun. Learning a similarity metric discriminatively, with application to face verification. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, 1: 539–546.

[8] Taigman, Yang, Ranzato, et al. DeepFace: Closing the Gap to Human-Level Performance in Face Verification. IEEE Conference on Computer Vision and Pattern Recognition, 2014, 1701–1708.