Machine Learning and Text Mining in Investor Sentiment

Tao Wang

School of Economics, Shanghai University, Jiading District, Shanghai, China

shoulder509@163.com

Keywords: Behavioral Finance; Sentiment Indexes; Machine Learning; Text Mining

Abstract: Investor sentiment, as a core concept in the field of behavioral finance, has been widely studied by scholars. Through scholars' long-term efforts, investor sentiment indexes have undergone a development process from direct-type sentiment indexes to indirect-type sentiment indexes. Although sentiment indices are currently able to better portray investor sentiment, they also have some shortcomings. This paper focuses on how machine learning and text mining techniques can identify investor sentiment. Firstly, it introduces how to transform text into data that can be used for computer analysis through methods such as word separation and vectorization, followed by how to use machine learning methods to mine investor sentiment. A large amount of practice shows that machine learning methods are better applied in the field of investor sentiment.

1. Introduction

In recent years, behavioral finance is growing rapidly because of its ability to explain many anomalies in the market. Investor sentiment, a core concept in the field of behavioral finance, has naturally gained the attention of scholars. In financial markets, investors' irrational behaviors abound, such as overreaction and underreaction, overconfidence and self-attribution, loss aversion and disposition effect, herding behavior and overtrading. All of these irrational behaviors are closely related to investor sentiment. Therefore, the correct identification of investor sentiment is of great significance, both for investment practice and theoretical research.

It seems too abstract to study investor sentiment directly. Therefore, scholars usually construct investor sentiment indices to quantify the changes in sentiment. After scholars' long-term efforts, the investor sentiment index has undergone the development process from direct-type sentiment index to indirect-type sentiment index, and now it can better portray the changes of investor sentiment. The direct sentiment index is a relatively simple index with the help of investors' judgment of the market trend in the future, including CCTV watch index and investor intelligence index, but the accuracy is not high. Scholars then developed indirect sentiment indices, which can be subdivided into single indirect sentiment indices and compound indirect sentiment indices. Single indirect sentiment indices are usually composed of indicators that reflect market trends, such as closed-end fund discount rate, turnover rate, number of new accounts and IPO first-day return. Baker and Wurgler (2006)[1] used the principal component analysis to calculate the sentiment index by using the market turnover rate, closed-end fund discount rate, issue number, first-day-of-market return, equity issuance as a percentage of securities issuance, and dividend yield are used to construct a composite indirect sentiment index. The weaknesses of the traditional method are, first, the market variables that are proxies for investor sentiment may not only reflect investor sentiment, but also the equilibrium outcome of the interaction between sentiment and other economic factors (Qiu and Welch, 2006; Da et al., 2014)[2, 3]; second, although the survey method directly measures investor sentiment, it is costly to implement, constructing sentiment indices lower frequency, and shorter time span.

With the development of machine learning and text mining techniques, scholars have proposed to quantify investor sentiment through these new methods. A large number of practices have shown that machine learning methods are better applied in the field of investor sentiment. Therefore, this paper also focuses on the application and development history of machine learning and text mining in this field.

2. Text Mining

2.1 Advantages and Challenges of Text Big Data

Thanks to the rapid development of the Internet and advances in computer technology, the application of textual big data in the fields of economics and finance is on the rise. Textual Big Data provides a new data source for measuring investor sentiment. On the one hand, as investors tend to post stock-related comments on online forums or make relevant searches, these textual data can directly reflect their views on the future of companies, their interpretation of the current state of the market, and information related to their investment decisions. On the other hand, these data are easily accessible, have a long time span and cover a large number of companies, which meet the need to study the relationship between sentiment and asset prices from different frequencies and levels.

As a new data source, text big data has at least three characteristics. First, data sources are diversified. Compared with the traditional data mainly collected by government and institution-led, the publishing subjects of text big data include individuals (such as investors and consumers), enterprises, media, institutions and government-related functional departments; its specific forms are rich and diverse, such as tweets, microblogs, forum posts, consumer evaluations of products, WeChat public websites, annual reports of listed companies, recorded telephone scripts, job advertisements, company annual reports, quarterly reports, announcements, IPO prospectuses, analyst research reports, meeting minutes, speeches by influential political, economic and financial figures, various information released by central banks and other government agencies on a regular and irregular basis, and so on. Second, the volume of data is growing geometrically. Due to the cost of data collection, traditional data collection often requires the use of paper media, and the volume is small. With the transfer of text information from paper media to the Internet as a medium, the cost of text data collection and transmission has been significantly reduced, providing an application scenario for natural language processing methods (NLP) in the computer field. Third, the time frequency is high. Traditional data need to be systematically organized and arranged to collect, commonly used data in the economic and financial fields are mostly annual, quarterly, monthly and weekly data, and the availability of data with higher frequency is insufficient to meet the application needs of high-frequency data analysis in the economic and financial fields. In contrast, the frequency of textual big data can be as high as seconds, which provides a data basis for high-frequency research.

However, the use of unstructured textual big data has brought new challenges while broadening the field of empirical research in economics and finance. The core challenge of applying textual Big Data to economics and finance research is how to accurately and efficiently extract the needed information from texts and examine its ability to explain or predict the corresponding problems.

2.2 Text Splitting Techniques

Formally, a text is a string of letters and punctuation marks. If the text is broken down from large to small, it may yield chapters, chapters, sections, paragraphs, sentences, phrases, words and characters. The main difficulty and obstacle in natural language understanding is that the meaning of the same word (words) varies in different scenarios or contexts; also, due to the rich diversity of text, it is often necessary to deal with problems related to high-dimensional sparse matrices after conversion into data matrices. This section focuses on the word separation technique and word embedding technique to determine the base unit of text data, i.e., the method of converting words into vectors. After the above conversion, the unstructured text can be represented in the form of a matrix, where each row records different attribute information of the same individual, while the same column of data records information related to the same attribute of different individuals.

In the English environment, words are separated by spaces, so that words are realized as participles. The empirical application also expands individual words into phrases of length n, i.e., ngrams (n-words). Since Chinese characters are continuous sequences in Chinese, analyzing the text requires cutting the sequence of Chinese characters into words or phrases according to certain specifications, i.e. Chinese word segmentation. According to the segmentation principle, the existing word separation methods can be categorized into three types: string-based matching, comprehension-based and statistical-based. The string matching method matches the Chinese character string to be analyzed with a predefined dictionary entry, and if a string can be found in the dictionary, it is recorded as identifying a word. This method has the advantage of simplicity and speed, but ignores the problem of ambiguity. Comprehension-based word separation method performs syntactic and semantic analysis along with word separation to improve the processing of ambiguous words. Statistical-based word separation methods first use machine learning models to learn the patterns of words that have already been separated, and then realize the separation of unknown text, and commonly used methods include maximum probability word separation and maximum entropy word separation.

2.3 Techniques for Transforming Words into Vectors

What needs to be accomplished after the completion of word separation is how to further transform the text into a digital matrix. If a text is considered as a combination of words selected from all word banks, the main challenge of this transformation is often the problem of dimensionality reduction of the high-dimensional matrix of words. To understand this, we first need to introduce the One-Hot Representation.

2.3.1 One Hot Representation

The one hot representation is characterized by ignoring elements such as syntax and order, and treating text data as a collection of several independent words. First, a word list2 is constructed based on all the words appearing in the text, and each word is numbered 1, 2, 3..., N in order. Then, word j is represented by an N-dimensional vectorw_j After each word is converted into a vector, the text t can be transformed into a 1 × N vector by summing up the vectors of all wordsW_t , whereW_{tj} (j = 1, ..., N) is the frequency of the j word in the text t. If there are a total of t = 1, ..., T texts, the original text base Ψ can be transformed into a T × N numerical matrix after using the one hot representation.

The above steps show that the one hot representation is simple to operate; however, the transformed matrix is often a high-dimensional sparse data matrix when the amount of data is large. This is due to the fact that the word vector dimension is determined by the number of words and that most words occur infrequently, so that the vast majority of elements in the vector corresponding to the text have zero values. In addition, the one hot representation may be ambiguous by ignoring the contextual structure.

There are two strategies to solve the problem that text data are high-dimensional sparse matrices. One is to take various measures to achieve dimensionality reduction for digitized text matrices, and a systematic summary of the corresponding dimensionality reduction methods has been made. Another idea is to use word embedding technique, which directly transforms words into low-dimensional vectors when they are converted into digitized matrices.

2.3.2 Word Embedding Technology

The word embedding technique is a model and technique that involves "embedding" a high-dimensional space of the number of all words in a continuous vector space of much lower dimension, i.e.$e_j = E \times W_j$ wheree_j denotes the word vector of the jth word mapped onto the real domain by the embedding matrix E, andW_j is the unique heat vector representation of the jth word. Since each element value of this vector can be a continuous value instead of just 0 or 1, thee_j the dimensionality ofN_e can be much lower than N.

The Word2Vec technique has been widely used in fields such as computational linguistics and performs well when combined with other statistical models for text analysis, but it has been relatively little used in the field of economics and finance (Gentzkow et al., 2019)[4]. In recent years, the method has also gradually gained importance. Scholars compared two methods, the one hot representation and Word2Vec, and found that using Word2Vec to represent text features can significantly improve the classification accuracy of text sentiment compared to the one hot representation.

3. Machine Learning

Depending on the existence of prior labeled training data, text-related problems in economics and finance can be analyzed using two types of methods, supervised learning or unsupervised learning. Among them, the main methods of unsupervised learning include lexicographical methods and topic classification models, while the classical methods of machine learning such as support vector machines and deep learning methods are more often used in the economic and financial fields in recent years belong to supervised learning.

Table 1. Machine Learning Methods

Types	Methods
Unsupervised Learning	Dictionary method, Topic classification model
Supervised Learning	Naive Bayes, Support Vector Machine

3.1 Unsupervised Learning

3.1.1 Dictionary Method

Dictionary method is a traditional method for analyzing text big data. The method starts from a predefined dictionary, and extracts text information by counting the number of occurrences of different categories of words in text data, combined with different weighting methods. The lexicon method is widely used in the field of economics and finance. A key aspect of using the lexicographical method is to select or construct a suitable dictionary, where a dictionary includes a specific lexicon, as well as a collection of specific words or phrases constructed by the author. When used properly, the lexicographical method is more capable of extracting information from a text, and this advantage is more obvious for short texts and applications where the logical relationships between words are weak. Therefore in practice, the lexical method can often be used as a benchmark method for big data analysis of texts.

3.1.2 Topic Classification Model

One application requirement in economics and finance is to classify text by topic without a prior annotation set. Since a text may have more than one topic, this type of classification problem differs from applications that classify a text into only one category according to a prior annotation set. A representative model for the topic classification problem is the implicit Dirichlet allocation (LDA) model proposed by Blei et al. (2003)[5], which is a probabilistic topic model.

3.2 Supervised Learning

3.2.1 Classical Supervised Machine Learning Methods

Classical machine learning methods include plain Bayes, support vector machines, decision trees, K-nearest neighbor algorithm, AdaBoost, maximum entropy method, etc. In text analysis in finance, the more commonly used traditional machine learning methods include Naive Bayes and Support Vector Machine (SVM).

Plain Bayesian algorithm is a supervised learning algorithm based on Bayesian theory. The common steps in dealing with text classification problems are as follows. First, the prior distribution (i.e., the prior probability that the text belongs to different categories) and the conditional probability distribution (i.e., the probability that a word appears in a given category) of the plain Bayesian classifier are obtained by learning the relationship between the words in the text and the categories to which they belong based on the training set. Second, using these probabilities, the conditional probability that the document belongs to different categories is computed based on the word features in the text, combined with the Bayesian conditional probability formula. Finally, the text is classified into the class with the maximum posterior probability according to the maximum posterior hypothesis.

Support vector machine is a supervised learning algorithm that can be used for both classification and regression analysis. The basic principle is that each text is first projected as a point in a high-dimensional space, and by finding a hyperplane, these points are partitioned according to their corresponding labels (e.g., positive, negative sentiment, etc.) so that the closest distance from each class of points to this hyperplane is maximized. The steps before using support vector machines for classification and regression analysis include first converting text into vectors using methods such as the unique heat representation or Word2Vec, then learning the relationship between text vectors and the categories they belong to based on the training set, then doing cross-validation on the model obtained from the training set, and finally using the best trained model to predict the classification of all texts.

3.2.2 Deep Learning Method

In text analysis, although classifiers such as SVM can handle certain nonlinearities, as linear classifiers, such methods often can only slice the input data into very simple regions, which also tend to lead to problems such as overfitting. With the increase of big data availability and the development of artificial intelligence software and hardware technologies, the powerful functions of deep learning methods in the field of natural language processing have gradually emerged. As a branch of machine learning, deep learning attempts to achieve goals such as classification by mimicking the neural networks of the human brain and using multiple processing layers composed of multiple nonlinear transformations to perform high-level abstraction of data. This type of approach can be used for both supervised and unsupervised learning, but it has not yet been widely used in the economic field, and its use in the financial field is mainly supervised learning.

Neural network is a machine learning model based on neural network imitating human brain to achieve artificial intelligence, which contains the structure of input layer, hidden layer and output layer, and can be used to deal with text classification problems. The principle is that the feature vector of the input layer reaches the output layer through the transformation of the hidden layer, and the classification result is obtained in the output layer, and the back propagation algorithm is usually used to train the neural network model.

Whether classical machine learning methods or emerging deep learning methods are used, supervised training requires two elements: high-quality labeled data as the training set and explicit model selection criteria. Since the quality of the training set directly affects the final information extraction effect, the cost of constructing the labeled data should be evaluated in advance for related studies. In terms of model selection criteria, the ideal model should not only avoid in-sample overfitting, but also have a good out-of-sample performance. A cross-validation approach is usually needed to evaluate the model: first, the annotated set is randomly divided into a training set, a validation set and a test set according to a certain ratio; then the model is trained on the training set and the model parameters are adjusted according to its performance on the validation set; finally, the model is applied to the test set to calculate the accuracy rate as a criterion to evaluate the out-of-sample performance of the model.

In summary, the selection of information extraction methods for text data requires comprehensive consideration of the source of text data, language environment, content length and the characteristics of the information to be extracted, as well as evaluation of the costs and benefits of each method. When conditions allow, both simple and complex methods can be considered, and the accuracy of information extraction can be improved by analyzing and comparing the differences between the two types of methods. Of course, the transparency and replicability of these methods need to be ensured when using complex methods. Finally, it is also important to note that the execution order of the two steps, data structured conversion and text data information extraction, needs to be determined by the specific problem and sometimes requires repeated attempts to find the best solution.

References

[1] Baker, M.,Wurgler, J. Investor sentiment and the cross-section of stock returns. Journal of Finance, 2006, 61 (4): 1645–1680.

[2] Qiu, L. and I. Welch, "Investor Sentiment Measures", National Bureau of Economic Research Working Paper, 2006.

[3] Da, Z., J. Engelberg, and P. Gao, "The sum of All FEARS Investor Sentiment and Asset Pricce", The Review of Financial Studies,2014,28(1),1-32.

[4] Gentzkow, M, B. T. Kelly, and M. Taddy, "Text as Data", Journal of Economic Literature, 2019, 57(3), 535-574.

[5] Blei, D.M., and J. D. Lafferty, "Dynamic Topic Models," Proceedings of the 23rd International Conference on Machine Learning. acm,2006,113-120.