Merging Naive Bayes and Causal Rules for Text Sentiment Analysis

Yongjian Luo*, Xiaohua Yang, Chunping Ouyang, Yaping Wan, Sixi He
School of Computer, University of South China, No.28, West Changsheng Road, Hengyang421000, China

*1007536712@qq.com

Abstract. Traditional machine learning sentiment analysis models are difficult to achieve good classification results from small sample data. This paper proposes to merging naive Bayes and causal rule (MNBA CR) for small sample data sentiment analysis scenarios. This model is based on the causal analysis theory, and introduce the causal inference algorithm into the field of text sentiment analysis. The causal inference algorithm extracts the causal rules of Chinese texts, and the causal rules can be used as the features of the naive Bayes algorithm to predict the sentiment polarity of small sample texts. In experiments, the model in this paper is evaluated on financial news datasets which have a small number for sample, and the results show that the proposed method achieves the best performance compared to the existing state-of-the-art models on the small sample data onto sentiment analysis.

Keywords: Sentiment analysis, Naive bayes, Causal rule, Machine learning, Small samples

1. Introduction

With the increasing prosperity of Internet technology, there’s a large amount of text information on the Internet, such as public opinions on hot events, shopping reviews, user experience of services, etc. These text information contains sentiment information, and the analysis of the sentiment categories are helpful for public opinion analysis, product sales and other applications. Therefore, how to predict the sentiment category of text information on the Internet has attracted more and more researchers’ attention.

Machine learning and deep learning has achieved great performance for text sentiment analysis. However, machine learning requires complex manual features design and feature extraction processes. Superior prediction results obtained by deep learning methods rely on big datasets, and the improvement on research accuracy had been limited. In fact, in some fields, there are situations where sufficient sample size is difficult to obtain because of the lack of knowledge, small amount of labeled data, and high cost to label data, such as social topic outbreak data in the field of public opinion monitoring, and online fraud data in the field of information security. Similarly, although published financial news texts data are of great amount, but they contain less effective information, and the financial news texts data after removing the short texts is too small, for getting a good result in sentiment categorizing of machine learning and deep learning methods. So the focus is on how to achieve better performance on sentiment classification task with a small number of data.
Causality is the most effective theory of event prediction, which can make full use of small sample data to solve classification problems. In view of the poor performance of machine learning and deep learning in small sample sentiment analysis, causal analysis is introduced into the field of sentiment analysis to solve the problem of low accuracy of sentiment analysis on small sample data. This paper proposes merging naive Bayes and causal rule (MNBACR) for text sentiment analysis. First, the model extracts causal rules for texts that have been marked with sentiment polarity, and counts the sentiment polarity of the causal rules. Such causal rules of the sentiment polarity represent the sentiment polarity between events in the text, and it is the reason why the text is judged as a certain sentiment polarity. Second, the causal rule is used as the classification feature of naive Bayes algorithm to predict the sentiment polarity of text. From the experimental results of datasets, we could find that the proposed model could achieve superior performance.

2. Related work

2.1. Text sentiment analysis
Text sentiment analysis is a vital task in natural language processing (NLP) [1]. The traditional sentiment analysis method mainly focused on sentiment knowledge, which calculates the sentiment polarity of the text according to certain rules by constructing a sentiment lexicon [2]. NTU evaluation lexicon [3] and SentiWordNet [4] are the representation of the sentiment lexicon. The sentiment lexicon provides sentiment scores for sentiment words, and the text to be classified obtains the sentiment tendency score of the text to determine the sentiment polarity of the text [5]. The quality of sentiment lexicons and judgment rules essentially determines the accuracy of text sentiment analysis. Both sentiment lexicon which requires prior knowledge and judgment rules which require manual design need a lot of manual work, so this type of the method is difficult to promote. Another type of sentiment analysis method is machine learning and deep learning. Traditional machine learning methods complete the sentiment analysis tasks by extracting a large number of meaningful features to train the classifier. Pang et al. [6] first applied machine learning to text sentiment analysis, and they use machine learning methods such as maximum entropy, naive Bayes (NB), and support vector machine (SVM) to predict the sentiment polarity of movie reviews. The experiment results show that the SVM achieves the best performance. Traditional machine learning methods required a lot of time and effort for feature selection. Deep learning maps words with word vectors, and then encodes the word vectors into semantic sentences to solve the sentiment analysis task. Long Short-Term Memory (LSTM) [7], Convolutional Neural Networks (CNN) [8], Attention [9] are commonly used deep learning models in the field of sentiment analysis. Although this kind of method has better predictive ability, it relies on utilizing the big data to train the model, and it is difficult to achieve great performance on small sample data.

2.2. Application of causality in prediction
Causality is a core issue in data science [10]. It explains the direct interaction between variables, that is, it can reflect the more essential and internal connections between objective things. Therefore, it is possible to find that mastering the causal relationship between things is able to predict the result of things directly. Pearl et al. [11] introduces causality into the field of artificial intelligence to solve these problems. Yang et al. [12] combined the causal Bayesian network and collaborative filtering algorithm to build a prediction recommendation system. Shen et al. [13] predicted users' ad clicks behavior based on causal Bayesian network. Zhang et al. [14] used causal Bayesian network to predict online clothing purchase behavior and applied it to the recommendation system. Causal analysis can make full use of domain knowledge and sample data information to infer the causal relationship between variables, and has a great advantage in prediction. According to the advantages of causal inference in the field of prediction, this paper introduces causality into the field of sentiment analysis, and merging naive Bayes and causal rules to predict text sentiment.

3. Model
3.1. Keyword vectorization

The subject of the text is summarized with a set of keywords, and the TF-IDF algorithm is used to extract the keywords from the text. The principle of the TF-IDF algorithm is that if a word appears frequently in an article and appears less frequently in other articles, it is considered that the word can better represent the meaning of the current article. That is, the importance of a word is directly proportional to the number of times it appears in the document, and inversely proportional to the frequency of its appearance in the document in the corpus. TF-IDF is equal to the term frequency (TF) multiplied by the Inverse Document Frequency (IDF):

$$\text{TF-IDF} = \text{TF} \times \text{IDF}$$ (1)

where TF represents word frequency and IDF represents reverse document frequency:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$ (2)

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|}$$ (3)

The extraction of causal rules requires vectorization of keywords. The keywords are vectorized in an one-hot manner. The one-hot encoding steps can be summarized as follows.

1) Set up a dictionary consisting of n keywords \(KW_1, KW_2, \ldots, KW_n\), use keyword extraction algorithm to extract keywords from the text dataset, and get n keywords, these n keywords are the subject dictionary of the text datasets.

2) Vectorize each piece of text, and use a vector \(W_1, W_2, \ldots, W_n\) to represent each piece of text. For each piece of text, if a keyword appears in the piece of text, set to 1, otherwise set to 0.

3) The vector of the text dataset is formed into a m*n matrix, where m is the number of texts. After the above steps, the keyword vectorization matrix required to extract the causal rules are obtained. Keyword vectorization matrix is shown in Tab.1.

| keyword1 | Keyword2 | Keyword3 | …… | keywordn |
|----------|----------|----------|-----|----------|
| Text1    | 1        | 0        | 1   | ……       | 1        |
| Text2    | 0        | 1        | 1   | ……       | 0        |
| Text3    | 1        | 1        | 0   | ……       | 1        |
| ……      | ……      | ……      | …… | ……       | ……       |
| Textn    | 1        | 0        | 1   | ……       | 0        |

**Tab.1 Keyword vectorization**

3.2. Causal rules extraction

This paper uses the CCU causality algorithm to mine causal relationship, which is a constraint-based causality inference algorithm. It uses variable association, independence and conditional independence tests to limit the possible causal relationships between variables, and uses local searches methods to find causal relationships. The principle of the CCU causal algorithm is as follows: Suppose \(A, B,\) and \(C\) are three variables. If \(A\) is associated with \(C\), \(B\) is associated with \(C\), and \(A\) is independent of \(B\), but \(A\) is associated with \(B\) under the condition of \(C\), which means that \(A\) and \(B\) cause \(C\) without hidden variables and confusion in the case of variables. In the CCU causal algorithm, it is necessary to test the independence and conditional independence between variables. This paper uses mutual information (MI) and conditional mutual information (CMI) in information theory to test the independence and conditional independence of observation data respectively. The principle is to give a minimum mutual information threshold \(\varepsilon\), if \(MI(X,Y) < \varepsilon\), then \(X\) and \(Y\) are independent, if \(CMI(X,Y|W) < \varepsilon\), then \(X\) and \(Y\) are independent under the given \(W\) condition.

$$MI(X,Y) = \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}$$ (4)

$$CMI(X,Y|W) = \sum_{X,Y,W} P(X,Y,W) \log \frac{P(X,Y|W)}{P(X|W)P(Y|W)}$$ (5)
4.2. Experimental settings

and corresponding polarities, which are labeled with \{positive, negative\}. Analysis, and extracts the causal rule from the experimental data financial news text through the CCU

Tab.2, where MI and CMI represents independence and conditional independence respectively.

held by the company's shareholders. The CCU causal algorithm for extracting causal rules is shown in Tab.2, where MI and CMI represents independence and conditional independence respectively.

The form of the causal rule defined by the CCU causal inference algorithm is: "\(X, Y \rightarrow Z\)" the causal rule indicates that \(X\) and \(Y\) cause \(Z\) This paper applies this causal rule for the field of text sentiment analysis, and extracts the causal rule from the experimental data financial news text through the CCU causal inference algorithm. For example, a causal rules such as "shares, reduced holdings \(\rightarrow\) shareholder change", it means that the company's shareholders change because of the reduction of shares held by the company's shareholders. The CCU causal algorithm for extracting causal rules is shown in Tab.2, where MI and CMI represents independence and conditional independence respectively.

### 3.3. Merging naive bayes and causal rules

We extract causal rules for text, and then use causal rules as classification features of naive Bayes algorithm for text sentiment analysis.

Naive Bayes algorithm is a classification method based on Bayes' theorem and the assumption of independence of characteristic conditions.

\[
P(Y = c_k | X = x) = \frac{P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k)}{\sum_k P(Y = c_k) \prod_j P(X^{(j)} = x^{(j)} | Y = c_k)} k = 1, 2, \ldots k
\]

\[(6)\]

where \(x^{(j)}\) represents the feature, \(c_k\) represents the category.

Naive Bayes algorithm makes the assumption of conditional independence on conditional probability distribution. Naive Bayes assumes that all features are independent of each other, and the conditional independence assumption is defined as:

\[
P(X = x | Y = c_k) = P(X^{(1)} = x^{(1)}, \ldots, X^{(n)} = x^{(n)} | Y = c_k) = \prod_{j=1}^{n} P(X^{(j)} = x^{(j)} | Y = c_k)
\]

\[(7)\]

where \(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\) represents the feature, \(c_k\) represents the category.

In this paper, \(x^{(1)}, x^{(2)}, \ldots, x^{(n)}\) represents the causal rules and \(c_k\) represents the text sentiment category. Therefore, the naive Bayes algorithm is used to predict sentiment categories, if \(P(c_1 | x^{(1)}, x^{(2)}, \ldots, x^{(n)}) > P(c_0 | x^{(1)}, x^{(2)}, \ldots, x^{(n)})\), the sentiment category of the text belongs to \(c_1\), if \(P(c_1 | x^{(1)}, x^{(2)}, \ldots, x^{(n)}) < P(c_0 | x^{(1)}, x^{(2)}, \ldots, x^{(n)})\), the sentiment category of the text belongs to \(c_0\).

### 4. Experiment

4.1. Datasets

In order to evaluate the performance of proposed model on the small sample datasets, we conduct experiments on the financial news text dataset, and its source comes from the financial news text on Xueqiu.com. For the datasets, the training datasets, the positive samples and the negative samples were 500 row financial news texts respectively, the test datasets, the positive samples and the negative samples were 2500 row financial news texts respectively, each review contains a list of aspect terms and corresponding polarities, which are labeled with \{positive, negative\}.

4.2. Experimental settings
If the length of the financial news text is too short, it contains less effective information. Therefore, the minimum length of the financial news text are 100 characters, and the maximum number of keywords extracted from the text is 200. The thresholds for mutual information and conditional mutual information are both 0.001. The specific description of the parameters used in the experiment is shown in Tab.3.

| Parameter names                  | value | unit     |
|----------------------------------|-------|----------|
| Minimum text length              | 100   | character|
| Maximum number of keywords       | 200   | number   |
| Conditional independence threshold| 0.001 | -        |

Tab.3 Model parameter setting

4.3. Experimental result
We set up five sets of comparative experiments. Then we use the pre-trained word vector file to vectorize the segmented training set corpus. For traditional machine learning algorithms, the input is a N-dimensional vector, and use the average sum of the sentence vector. For CNN and RNN deep learning algorithms, the input is required to be N*M dimensional vectors, which are respectively searched and generated. The MNBACR model selects 200 keywords from the training set text for causal rule extraction. The 4 evaluation results of the five groups of comparative experiments on the financial news dataset are shown in Tab.4.

- SVM and NB: The two models are classic models in sentiment prediction tasks, and have achieved good performance when predict the sentiment.
- CNN: Convolutional neural networks have achieved great performance in sentiment analysis, which are better than previous machine learning methods such as NB and SVM.
- LSTM: The LSTM model can learn what information should be stored and what information should be ignored in long-term memory. In addition, it can avoid the gradient explosion and gradient disappearance problems which frequently occur to CNN.

| Models   | Accuracy | Recall | Precision | F1-score |
|----------|----------|--------|-----------|----------|
| NB       | 65.8%    | 71.5%  | 64.2%     | 67.7%    |
| SVM      | 63.4%    | 79.6%  | 60.1%     | 68.5%    |
| LSTM     | 61.4%    | 82.7%  | 58.0%     | 68.2%    |
| CNN      | 58.1%    | 86.8%  | 55.1%     | 67.4%    |
| MNBACR   | 69.4%    | 84.0%  | 71.0%     | 77.0%    |

Tab.4 Experimental result

Fig.1 Model performance
Experimental results show that in sentiment prediction tasks, the MNBACR is better than the traditional naive Bayes model, SVM model, LSTM model and CNN model on the small sample datasets. The accuracy of the model increases with the number of keywords. As is show in the Fig.1, when there are 200 keywords, there are 9802 causal rules, and the accuracy rate reaches 69.4%. It shows that the causality we introduce is effective, and it played an important role in optimization. The experiment results also fully prove that MNBACR model is proposed in this paper is competitively in sentiment analysis tasks on small sample datasets.

5. Conclusion
To solve the problem of machine learning methods and deep learning methods which have low accuracy in the field of small sample sentiment analysis, this paper proposes MNBACR model to study the problem. The causal analysis theory about the field of sentiment analysis is introduced and the causality inference algorithm is used to extract the causal rules for the text, and then be utilized as the classification feature of the naive Bayes algorithm to determine the sentiment polarity of the text. Experimental results from datasets demonstrate that compared with other methods, the accuracy of our proposed model has been further improved, which effectively proves the feasibility of MNBACR model on sentiment analysis of small samples. But, the experimental results show that there is still a great deal of room for improvement. Therefore, the next step will focus on improved the classification effect of the MNBACR model on the sentiment classification of small samples of text.

Acknowledgment
This paper was funded by the National Defense Science and Technology Innovation Special Zone Project of the Central Military Commission (No.17-163-15-XJ-002-002-04), the Key Projects of the Hunan Provincial Department of Education(No.17A185).

References
[1] Ligong Y, Jian Z and Shiping T 2013 Survey of text sentiment analysis J. Journal of Computer Applications 33(6) pp1574-1607
[2] Ku L W, Wu T H and Lee L Y 2005 Construction of an evaluation corpus for opinion extraction C. NTCIR
[3] Ku L, Lo Y and Chen H 2007 Using polarity scores of words for sentence-level opinion extraction C. Proc. of NTCIR-6 workshop meeting pp316-322
[4] Esuli A and Sebastiani F 2006 Sentiwordnet: A publicly available lexical resource for opinion mining C. LREC 6 pp417-422
[5] Rao D and Ravichandran D 2009 Semi-supervised polarity lexicon induction C. Proc. of the 12th Conf. of the European Chapter of the ACL (EACL 2009) pp675-682
[6] Pang B, Lee L and Vaithyanathan S 2002 Thumbs up? Sentiment classification using machine learning techniques J. arXiv preprint 0205070
[7] Hochreiter S and Schmidhuber J 1997 Long short-term memory J. Neural computation 9(8) pp1735-1780
[8] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks C. Advances in neural information processing systems pp1097-1105
[9] Bahdanau D, Cho K and Bengio Y 2014 Neural machine translation by jointly learning to align and translate J. arXiv preprint 1409.0473
[10] Pearl J 2000 Models, reasoning and inference J. Cambridge, UK: Cambridge University Press
[11] Pearl J 2019 The seven tools of causal inference, with reflections on machine learning J. Communications of the ACM 62(3) pp54-60
[12] Yang X, Guo Y and Liu Y 2012 Bayesian-inference-based recommendation in online social networks J. IEEE Transactions on Parallel and Distributed Systems 24(4) pp642-651
[13] Shen S, Hu B and Chen W 2012 Personalized click model through collaborative filtering C. Proc. of the fifth ACM international conference on Web search and data mining pp323-332
[14] Xiaoyang Z 2016 Reseach on a big data-based causal inference recommended model for online garments among college students *D. Harbin Institute of Technology*