A Method for Searching Emotional Tendencies by Expressions based on Convolutional Neural Networks

Yujia Xie

Senior High School of Hefei City, Anhui Province, China

YujiaXie.CS@gmail.com

Keywords: Natural Language Processing; Emotional Analysis; Orientation Analysis; Search Engines

Abstract. A search engine is becoming increasingly important given the massive influx of big data. In most cases, users get feedback based on the search theme. However, sometimes users need to search in terms of textual emotions. This thesis puts forward a regular emotional expression based on a regular expression, a method which searches emotional tendency in expressions via XieFCNN. At the same time, this thesis constructs a prototype of an emotional search engine, which enables searching in terms of themes and emotional tendencies. Experiments show that the prototype system can realize the use of expressions in the search of emotional tendencies.

INTRODUCTION

The data volume of the Internet showed an explosive growth in the past ten years. With the individualized developmental trend of the Internet, individuals cost less and less at information publishing. Meanwhile, it means that the phenomenon of information overload is getting increasingly serious. How search engines help users find required information efficiently has become an important issue. Traditional search engines usually return and search relevant contents based on matching keywords. However, just because of the growth of information individuality, users often prefer to search through emotional tendencies; however, traditional search engines are difficult to meet this requirement in most cases. Emotional tendency analysis, an emerging researching field in recent years, classifies the emotional information in texts according to tendencies. Currently, there are three mainstream methods, respectively based on dictionary, rules and statistical machine learning algorithm. The dictionary-based method uses attribute dictionary, emotional dictionary, degree dictionary and negative dictionary to judge the emotional tendencies of sentences comprehensively. The second method is based on rules. The third method is based on statistical machine learning algorithms, which deals with emotional tendency analysis as a clustering problem and obtains the emotional characteristics of sentences through the classifier. But most of the existing orientation algorithms are applied in commodity comments such restricted areas. In contrast, the environment of open domain is more complex, where the topics discussed are wider with more flexible words and more casual syntax. The method based on dictionary and rules run out of puff in such cases. If we use traditional machine learning algorithm, the artificial workload will be too huge to imagine. Under the background of deep learning and wide application, this thesis constructs a convolutional neural network XieFCNN which is applied in emotional tendency analysis. Experiments show the advantages of emotional tendency analysis that XieFCNN has in open domains. Further, the algorithm is encapsulated into a regular emotional tendency expression similar to regular expressions. At the same time, a search engine prototype aiming at emotions is established. The prototype reprocesses the results returned by traditional search engines and returns results according to the emotional tendencies of users. This prototype system contains information search, information extraction, information preprocessing, emotional tendency analysis, results display and other modules, which achieves the goal of returning results according to emotional tendencies. It proves the advantages of using deep neural network for emotional tendency analysis on the other hand, and provides a tool for emotional tendency search.

The structure of this thesis is as follows: the first section introduces the research background of

emotional analysis and in-depth study; the second section mainly introduces the emotion analysis method proposed in this thesis; the third quarter evaluates the feasibility of the method by setting up experiments; the fourth quarter wraps the invoking of XieFCNN to an expression and establishes a prototype system with the method; finally, the fifth section reviews and summarizes the main contributions of this thesis.

Research Background

Sentiment analysis technology can be divided into methods based on dictionary, rules and statistical machine learning algorithm, among which the learning method, based on emotional dictionary with supervised machine, is a hot research spot currently. Such method is mainly based on support vector machine, Bayes, maximum entropy model, etc. Combining all kinds of textual characteristics, texts are mapped as characteristic vectors for the training and clustering of models. By combining Naive Bayes with SVM, SidaWang [2] et al. gained good effects in multiple public data sets. Bollegala used the general characters of emotional expressions in different fields to build a relevant emotional dictionary and expand textual characteristics, so as to improve the effects on interdisciplinary classification of emotions.

While in the field of NLP, word is the basic element of texts. And One-hot Representation is one of the most common means of expression. But this method ignores the context relationship between words; neither will it provide the information carried by the word itself. In 2003, Bengio[3] et al. proposed to use neural network to build dualistic linguistic model, which could map words to a low-dimension real vector and then judge the semantic similarities by the distance between words. Andriy Mnih[4] et al. raised a hierarchic Log-Bilinear model to train linguistic models. Mikolov[5] realized the two linguistic models of CBOW and Skip-gram at word2vec. Then Socher, Johnson, Maas[6], Tang[7], Faruqiu[8] et al. built different linguistic models embedded with emotions respectively and obtained good effects at multiple public data sets.

At present work, the polarities and attributes of words will both affect the final polarity of articles, so pure sentiment analysis based on word embedding cannot achieve the optimal effect. While the methods based on rules and dictionary rely too heavily on specific corpus, so with the increase of rule and dictionary resources, textual characteristic dimension will rise in a linear fashion, which will cause dimension disaster and increase the cost of training, but reduce the generalization ability.

Textual Sentiment Analysis based on Convolutional Neural Network

In view of realizing textual sentiment analysis, the thesis puts forward a XieFCNN(Xie Feature Convolutional Neural Networks) for textual sentiment tendency analysis. This method firstly extracts the emotion sequence by the maximum matching method. Then the sequence is abstracted into vectors according to the emotional attributes and some grammatical features of the words in the sequence. And then emotional tendency weights will be gained with the sequence vectors. Meanwhile, part-of-speech dictionary is used to extract the emotion characteristic matrix in the full text; inputting the original characteristic matrix and emotional characteristic matrix in the convolution neural network, and then training in accordance with BP rules, a vector will be finally got. Multiplying each element in the vector by emotional tendency weight, and then entering the vector into SVM for further classification, we will get the emotional tendency of the text.

Emotional Characteristics Matrix Extraction based on the Maximum Matching Method and the of Part-of-speech Dictionary

According to view in the literature[9], when identifying the tendency of texts, if sequence fragments related to emotional expression can be extracted, it will help determine the emotional polarity of the text accurately. In order to extract specific emotional expression sequences in English texts, the thesis proposes a maximum matching extraction algorithm based on

part-of-speech dictionary, namely MMS. This algorithm has two core ideas, one of which is the maximum matching: starting from the left end of sentences, continuously matching the most striking emotional expression sequence until the end of the sentence. The second way is to judge the degrees of emotions through the part-of-speech dictionary: according to the lexical characteristics in table 1, extracting the number of "1" in current sequence shows the emotional polarity of the sequence. After finishing the matching, we extract the sequence $W_{[i,j]}$ which expresses emotions most strongly and map it to $\xi_{emotion}$ between [0, 1] by Sigmoid function, in which way emotional tendency weight of the sequence is expressed.

At the same time, according to the data in table 1, every word in the text is mapped to the k-dimension {0, 1} vector space, namely $w \in R^k$, where k represents the characteristics of the words themselves. The value of each dimension should be 0 or 1. The given sentence contains n w_i while $1 \le i \le n$, which forms a $n \times k$ emotional characteristic matrix.

characteristic name	yes	no
positive sentiment word or not	1	0
negative sentiment word or not	0	1
assertive word or not	1	0
negation word or not	0	1
adjective or not	1	0
degree adverb or not	1	0
noun or not	0	1
verb or not	0	1
adverb or not	0	1
punctuation or not	0	1

Tab.1 value table for lexical characteristics

Model	Precision+ Precision-	Recall+ Recall-	F-Measure+ F-Measure-
XieF	0.7852	0.7992	0.7921
CNN	0.7632	0.7401	0.7515
W2VCNN	0.7868 0.8442	0.8671 0.7190	0.8250 0.7766
NBSVM	0.7172 0.7547	0.8269 0.6204	0.7682 0.6810

The dictionary resources adopted in this experiment is HowNet. The word embedding is generated by the Skip-gram model For the NBSVM model, Unigram and Big- gram models are used for the construction of text embeddings.

Tab.2 performance contrast of models

XieFCNN convolution neural network model

Convolution nerve network, a kind of artificial neural network, has become the research hotspot in the fields of speech analysis and image recognition.On the basis of the convolution nerve network put forward by LeCun[10], the thesis introduces an emotional characteristic vector expressive method. Figure 1 is the structure chart of the XieFCNN.

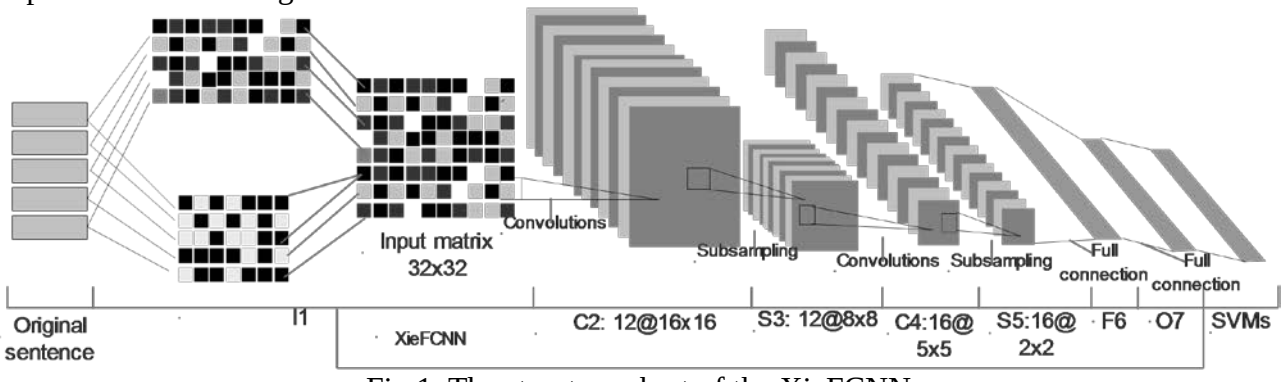

Fig.1. The structure chart of the XieFCNN

XieFCCN has six layers totally (including the input and output layers). In addition to the input and output layer, each layer contains connecting weights. The size of the input matrix is 32 x32, joint by textual characteristic matrix and emotional characteristic matrix, namely:

$$M_{input} = M_{word2vec} \oplus M_{emotion} \quad (1)$$

M_{input} represents the matrix which is finally entered in the convolution neural network; $M_{word2ec}$ is the textual characteristic matrix constructed by word embedding; and $M_{emotion}$ is the

emotional characteristic matrix, $\oplus$ meaning matrix splicing. For matrixes which are not long enough, zero matrixes will be stitched in the left in order to ensure that the significant characteristics can appear in the center of the highest-level detection area.

I1 is the input layer, mainly responsible for matrix splicing as mentioned above.

C2 layer is the convolutional layer, consisting of 12 FMs. Each neuron in FM is connected to the 5×5 border-upon domain of the input matrix. The size of FM is 16×16.There exist 144 trainable parameters and totally $144\times16\times16=36864$ linkages in the C1 layer. This layer does not take bias into consideration, namely:

$$c_{i,j}= f(w_i \bullet S_{j:j+h-1}) \tag{2}$$

$c_{i,j}$ represents the j th eigenvalue in the i th FM. $f(\bullet)$ is a convolutional kernel function; w_i is the filter, h is the sliding amount; $S_{j:j+h-1}$ means the partial characteristic matrix constructed by the j th line to $j+h-1$ th line. Therefore, FM is:

$$C_{i,j}=[c_{i,1},c_{i,2},...,c_{i,n-h+1}] \tag{3}$$

The S3 layer is a pooling layer with twelve 8×8 FMs. Each unit in FM connects to the 2×2 neighboring area of the FM in C2. Each unit in S3 inputs the sum of $2\times2=4$ multiplying by a trainable parameter, and then adds a trainable bias. Through the calculation by Sigmoid function, namely:

$$\sigma(w\bullet x+b)=\frac{1}{1+\exp(-\sum_j w_j x_j-b)} \tag{4}$$

x is the input and w is the weight of corresponding x. They share a bias b. As the 2×2 feeling area of each unit does not overlap, the FM size in the S3 layer is a quarter of that in the C2 layer where contains 12 trainable parameters and 768 connections.

C4 layer is a convolutional layer as well. This layer mainly extracts deep-seated characteristics by assembling different FMs. It convolutes sampling results in the S3 layer through sixteen 5×5 convolution kernels.

S5 layer is also a sub-sampling layer with sixteen 2×2 FMs. Its pooling process is similar to S3.

The F6 layer, fully linked with the S5 layer, is responsible for calculating the dot product between the input vector x_j and weight vector w_j.Adding the bias b, it passes a state generating unit j to Sigmoid function. The formula is similar to (3).

The O7 layer is the output layer. It outputs via *softmax* regression. According to the actual category labels of input data, it makes gradient update with BP algorithm, namely:

$$P(y\,|\,V,W_s,b)=softmax_y(W_s\bullet V+b) \tag{5}$$

Thereinto, $y\in\{1,-1\}$, $W_s\in R^{|V|}$, and b is a shared bias.

Then, each item in the vector is attached with an additional emotional weight, namely:

$$\hat{V}=[v_1\xi_{emotion},v_2\xi_{emotion},...,v_n\xi_{emotion}] \tag{6}$$

$v_{1:n}$ is an element in the original output vector.

Finally, $\hat{V}$ is entered into SVM for further classification and the emotional tendencies will be output eventually.

The Experimental Data

The experimental data applied in this thesis is the ModApte version of Reuters - 21578, which evaluates the effectiveness of the proposed method.Reuters - 21578 is divided into training set and testing set with totally 90 categories. There are 7769 training documents and 3019 testing documents. The experiment designed by this thesis is to compare the performance of XieFCNN model and that of word-vector CNN model proposed by Kim based on word2vec training (marked

as W2VCNN) as well as the NBSVM model proposed by Sida Wang.This set of experiments compares the performance of XieFCNN with that of other models. Table 2 lists the experimental comparison results on Reuters-21578 data set. Seen from the experimental results of table 2, the performance of XieFCNN model is close to that of word-vector W2VCNN which is based on word2vec. But the dimension of XieFCNN model is much lower than word2vec which has 50 or even hundreds of dimensions. It reduces the complexity of the model, speeds up the training model and guarantees the performance as well. Meanwhile, XieFCNN model is based on dictionary, so it is more adept at extracting implicit characteristics of texts with stronger generalization ability. Compared with traditional NBSVM based on domain knowledge, XieFCNN not only have advantages in model training, but also has great advantages on performance. Known from the experimental results in table 2, on F1 - Measure which identifies positive emotions, XieFCNN is 8.67% higher than NBSVM; while in the recognition of negative emotions, XieFCNN is 1.12% higher than NBSVM.

The Emotional Expression Prototype System

The preceding part of the thesis elaborates the application of XieFCNN in the analysis of emotional tendentiousness. However, not all the users can freely use neural network for classification. Even for computer professionals, if not majoring in neural network, it is also difficult for them to use the neural network. So the thesis proposes to call and wrap XieFCNN to an expression similar to regular expressions, and the symbol agreement is shown in table 3.The parsing engine uses limited automaton to realize function matching.

Expressions	Function declaration	Examples	Expressions	Function declaration	Examples
\e*	Matching any emotional tendencies	GitHub\e*	\e:d	Matching negative emotional tendencies	financial crisis\e:d
\e:a	Matching positive emotional tendencies	Microsoft\e:a	\e(pattern)	Matching pattern emotional tendencies	financial crisis\e (very bad)
\e:n	Matching neutral emotional tendencies	New York\e:n	\e[x]	Matching information whose emotional tendency value closes to x	New York\e[0]

Tab.3 The symbol agreement

And the emotional search engine prototype system Search. Emo is established by the use of above expression with Lucene. The system mainly consists of the following modules: information searching module, information extraction module, information preprocessing module, emotional calculation module and the parsing module for expression query. Finally, the system sequences the data according to positive, negative and neutral emotional tendencies.

Conclusion

This thesis proposes a textual emotional tendency analysis model based on convolution neural network. Through the use of the inherent characteristics of words, it maps texts to a low-dimension characteristic matrix. On the basis of guaranteeing the emotional tendentiousness recognition performance of texts, it reduces the complexity of the convolution neural network model and

accelerates the model training. In addition, the thesis also puts forward the maximum matching sequence algorithm which makes a preliminary analysis on the emotional tendencies of sentences. It can grade the emotional tendencies of sentences promptly. At the same time, the thesis conceives a searching method for emotional expressions and builds a search engine prototype system on the basis, which verifies the reliability of this method. However, the model in this thesis relies too much on manual annotation dictionary, so the quantity and quality of dictionaries will directly affect the classification effects of models. Therefore, how to make use of machine for calculation learning to make the model supplement and perfect dictionaries automatically is our next key job.

References

[1] Feng Shi, Fu Yongcnen, Yang Feng,et al.Blog Sentiment Orientation Analysis Based on Dependency Parsing. Journal of Computer Research and Development.2012:49(11):2395-2406.

[2] Wang S.,Manning C. D Baselines and bigrams: Simplegood sentiment and topic classiIication [C]//Proceedings of the ACL.2012: 90-94.

[3]Bengio Y., Ducharme R.,Ducharme R.,Vincent P. et al,A neural probabilistic language model[J]. The Journal of Machine Learning Research,2003,3,1137-1155.

[4] Mnih A.,Hinton G.E.,A scalable hierarchical distribued language model[G]//Proceedings of the NIPS.2009:1081-1088.

[5] Mikolov T.,Chen K.,Corrado G.,et al. Efficient estimation of word representations in vector space[J].Computing Research Repository,2013: 1301:3781.

[6] Maas A.L.,Daly R.E.,Pham P.T.,et al.Learning word vectors for sentiment analysis[C]//Proceeding of the ACL. 2011:142-150.

[7] Tang D.,Wei F.,Yang.,et al.Learning sentiment-speciIic word embedding for twitter sentiment classiIication[C]//Proceedings of the ACL.2014:1555-1565.

[8] Faruqui M.,Dodge J.,Jauhar S.K et al.,RetroIitting word vectors to semantic lexicons [J]. Computing Research Repository, 2014: 1441-4166.

[9] CHEN Zhao.,XU Ruifeng.,GUI Lin.,LU Qin.Combining Convolutional Neural Networks and Word Sentiment Sequence Features for Chinese Text Sentiment Analysis[J].Journal of Chinese Information Processing.2015:29 (6) 172-178.

[10] LeCun Y.,Bottou L,Bengio Y., et al. Gradient-based learning applied to document recognition[C]//Proceedings of the IEEE. 1998, 86(11):2278-2324.