algorithms

MDPI

Article

Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks

Yuhai Yu [1,2], Hongfei Lin [1,*,†], Jiana Meng [2,†] and Zhehuan Zhao [1]

1 School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China; yuyh@dlnu.edu.cn (Y.Y.); zhehuan@dlut.edu.cn (Z.Z.)
2 School of Computer Science & Engineering, Dalian Nationalities University, Dalian 116600, China; mengjn@dlnu.edu.cn
* Correspondence: hflin@dlut.edu.cn; Tel.: +86-0411-84706550
† These authors contributed equally to this work.

Academic Editor: Tom Burr
Received: 16 February 2016; Accepted: 1 June 2016; Published: 21 June 2016

Abstract: Sentiment analysis of online social media has attracted significant interest recently. Many studies have been performed, but most existing methods focus on either only textual content or only visual content. In this paper, we utilize deep learning models in a convolutional neural network (CNN) to analyze the sentiment in Chinese microblogs from both textual and visual content. We first train a CNN on top of pre-trained word vectors for textual sentiment analysis and employ a deep convolutional neural network (DNN) with generalized dropout for visual sentiment analysis. We then evaluate our sentiment prediction framework on a dataset collected from a famous Chinese social media network (Sina Weibo) that includes text and related images and demonstrate state-of-the-art results on this Chinese sentiment analysis benchmark.

Keywords: sentiment analysis; convolutional neural network; word vectors; microblog

1. Introduction

As the number of webcams has increased, more and more people enjoy posting their experiences in opinionated texts and images using audio or video and expressing their opinions about all sorts of events and subjects in online social networks. Currently, social networks such as Twitter and Sina Weibo have grown to be among the world's biggest information repositories. In this study, we focus on automatically detecting the sentiments expressed in these large-scale datasets. Extracting sentiment is both meaningful and a major challenge for many social media analytics tasks such as advertising or recommendations.

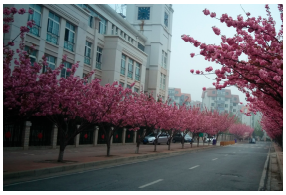
Users of microblogs often post one or several images in addition to text in their messages. Figure 1 shows several textual microblog messages with related images. In Figure 1a, both the text and the image carry an obvious sentiment; in Figure 1b, it is difficult to discern the sentiment from the image, but the text clearly shows a negative sentiment; and in Figure 1c, it is difficult to understand the elusive text; however, the blooming flowers in the image suggest a positive sentiment. Therefore, we can understand the sentiment in a message not only from its brief textual component but also from the visual content provided by the user.

Gorgeous!

(**a**)

Doing hard labor again!

(**b**)



"As long as the road is right, not afraid of distant future; if you believe it's worthwhile, do not care about vicissitudes."

(**c**)

Figure 1. Examples of textual microblog messages with related images from Sina Weibo. (**a**–**c**) are three types of micro-blogs.

Many researchers have contributed to sentiment analysis of textual content or visual content. Existing methods for sentiment analysis are divided into three main categories by Cambria [1]: knowledge-based techniques, statistical methods, and hybrid approaches. Knowledge-based techniques such as WordNet Affect [2], SentiWordNet [3], SenticNet [4] and AffectiveSpace [5], are popular because of their accessibility and economy, but their validity depends heavily on a comprehensive knowledge base and the knowledge representation. Statistical methods have been the mainstream natural language processing (NLP) direction in research since the late 1990s. But the performance of traditional statistical text classifiers depends on a sufficiently large text input [6]. Except for the bag-of-concepts and bag-of-narratives methods in [6], recent prominent deep learning methods [7–10] are also popular for their ability to analyze the sentiment of short texts by learning sentiment representations from a large corpus of labeled and unlabeled text. For example, Kim and dos Santos *et al.* [7,8] used a convolutional neural network (CNN) to extract sentence features and performed sentiment analysis of Twitter messages. Mesnil *et al.* [9] built an ensemble system to detect the polarity of a text document from a dataset of IMDB movie reviews. CNNs have also been applied to visual sentiment analysis. Chen *et al.* trained a deep CNN model called DeepSentiBank to classify visual sentiment concepts. Xu *et al.* [10] introduced a visual sentiment prediction framework that performs transfer learning from a pre-trained CNN with millions of parameters. Hybrid approaches [11–14] exploit both knowledge-based techniques and statistical methods to perform sentiment analysis of text or multimodal data. Cambria *et al.* [11,12] exploited an ensemble of SenticNet and deep learning methods to infer polarity from text. Chikersal [13,14] built a Twitter sentiment analysis system by combining a rule-based classifier with a supervised learning method. A vast literature focuses on sentiment analysis of multimodal content. Some researchers have focused on sentiment analysis of long texts assisted by visual sentiment analysis. Maynard *et al.* [15] centered on entity and event recognition of socially aware federated political archiving and socially contextualized broadcaster web archiving by combining opinion mining from text and multimedia. They extracted SIFT local features from images to analyze visual sentiment and assist in the sentiment analysis of text. Several researchers have performed video sentiment analysis by combining multimodal information. Rosas *et al.* [16] performed sentiment analysis of Spanish online videos. They integrated the following multimodal information to analyze sentiment: the bag-of-words representation of the video transcripts, the duration of smiles detected from video clips by a commercial software package, and the vocal intensity features of the audio track computed by an open source software project. Poria *et al.* [17–19] built models that used multimodal content including audio, visual, and textual information to harvest sentiments from videos posted to YouTube. They leveraged SenticNet 3.0 [4] and EmoSenticNet [20] for textual sentiment analysis and extracted facial expression features from video clips and vocal intensity features from the audio track. They used both feature- and decision-level fusion methods to merge the affective information extracted from these multiple modalities. Pereira *et al.* [21] presented an approach to perform sentiment analysis of news videos based on multimodal features extracted from closed captions, facial expressions and the participants' speeches. Some researchers have concentrated on

sentiment analysis of microblogs, which typically include a short text and one or more related images in each post. You *et al.* [22] and Cao *et al.* [23–25] employed both text and images to predict sentiment. The former focused mainly on detecting sentiment from English-language social media, while the latter established a dataset extracted from a Chinese social network (Sina Weibo) and provided a benchmark for combining visual and textual sentiment prediction from Chinese social media. In this paper, we focus on analyzing sentiment via both textual and visual content in the benchmark dataset mentioned above. We adopt deep convolutional neural networks (DNNs) with DropConnect [26] for visual sentiment analysis and train another CNN on word vectors using the short text for textual sentiment analysis and then analyze the complete sentiment by combining these two prediction results. Our textual results alone surpass the fusion results from [23] and we improve that performance even further by introducing the visual content component, thus achieving state of art performance on this benchmark.

The rest of this paper is organized as follows. In Section 2, we introduce the research status of textual and visual sentiment analysis. In Section 3, we introduce the architecture of our approach consisting of a textual model, a visual model and a fusion model. We conduct experiments to evaluate the performance of different models on sentiment analysis in Section 4. Finally, we summarize this article and mention possible future work in Section 5.

2. Related Work

The overwhelming majority of the researchers of natural language processing (NLP) address their single benchmark task by engineering intermediate representations. The task-specific features are effective, but they depend on the researchers' abundant linguistic knowledge. Collobert *et al.* [27] advocate a multilayer neural network architecture, which takes a sentence as input and extracts the features layer by layer. The first layer maps each word into a feature vector using a lookup table operation, starting from a random initialization. Subsequent layers extract local features from the sentence and feed a fixed-sized global feature vector by following standard neural network layers. They trained the neural networks by maximizing likelihood over large unlabeled data sets and then let the training algorithm discover useful representations for all types of NLP tasks. Mikolov *et al.* [28] introduced the CBOW (continuous bag-of-words) and Skip-gram models, which are efficient methods for learning high-quality word vectors from much larger unstructured text datasets. Mikolov *et al.* [29] created the tool *word2vec*, which provides an efficient implementation of the CBOW and Skip-gram architectures for computing word vectors. In addition, Mikolov *et al.* [30] demonstrated that the word vectors capture syntactic and semantic regularities in English language. Yoon Kim [7] trained a convolutional neural network on top of pre-trained word vectors and improved upon the state of the art for the task of sentence classification.

Alex Krizhevsky *et al.* [31] demonstrated that a deep convolutional neural network (DNN) is capable of achieving record-breaking results for image classification. Borth *et al.* [32] constructed a large-scale Visual Sentiment Ontology (VSO) and built SentiBank, a visual concept detector library for visual sentiment analysis. Chen T. *et al.* [33] trained a DNN model named DeepSentiBank (SentiBank 2.0) on Caffe for visual sentiment concept classification. Xu [10] used a pre-trained DNN [31] on the ILSVRC-2012 dataset and transferred the learned parameters to the task of visual sentiment prediction.

However, research on combining visual and textual elements in sentiment prediction from microblogs is far behind the single-element research. You *et al.* [22] fine-tuned a CNN on a collection of images from Getty Image for visual sentiment analysis and trained a paragraph vector model for textual sentiment analysis. Cao *et al.* [24,25], and Wang *et al.* [23] established a dataset consisting of textual messages with related images extracted from Sina Weibo and performed sentiment analysis by combing the prediction results of using n-gram textual features and mid-level visual features [32]. Our work is greatly motivated by the work performed in [7,23,26]. We train a CNN on word vectors pre-trained using the *word2vec* tool to extract features from texts, train a DNN with DropConnect [26]

to learn visual features extracted from images, and then, make sentiment predictions by combining the textual and visual features.

3. Methods

This section describes the multiple deep convolutional neural networks (DNNs) architecture (see Figure 2). DNN_Text and DNN_Image indicate our textual and visual sentiment analysis model, respectively.

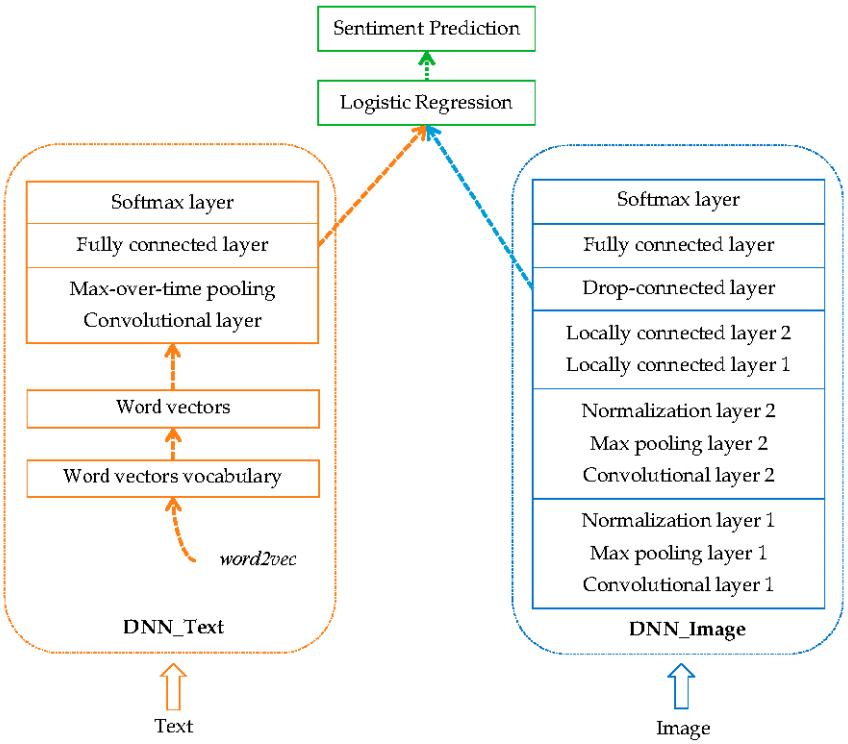


Figure 2. Architecture of our multiple deep convolutional neural networks approach.

3.1. Textual Features

Before pre-training the word vectors, we extract all texts from the Sina Weibo Dataset including both labeled and unlabeled data and segment them using two methods: One is splitting all texts into Chinese phrases using the Python package named jieba, and the second is splitting them into single Chinese characters—but preserving the English words. We split all texts using a delimiter of one blank space.

Based on these two different segmentation methods, we obtain two word vector vocabularies trained on all texts using the *word2vec* tool and all vectors have dimensionality of 300.

Next, we train the textual DNN model (DNN_Text, see Figure 2) on top of the pre-trained word vectors through stochastic gradient descent (SGD) over shuffled mini-batches with the Adadelta update rule [34]. The model randomly initializes unknown words. Then, while training the model, it keeps all the word vectors static and learns only the other parameters. Finally, we extract textual features with a length of 300 from the second-to-last layer of the textual model.

We use the hyperparameters similar to [7]: rectified linear units, filter windows of 3, 4, 5 with 100 feature maps each, dropout rate of 0.5, l_2 constraint of 3, mini-batch size of 50, adadelta decay of 0.95 and epochs of 25.

3.2. Visual Features

Before inputting the images into the visual DNN model (DNN_Image, see Figure 2), we resize them to a square of N × N pixels (where N = 32, 64, 128, *etc.*) and prepare a Python version of the dataset similar to the CIFAR-10 Python dataset.

Using a DNN with limited labeled data and billions of parameters can easily result in an overfitting problem. We utilize DropConnect to reduce overfitting when constructing our visual DNN model. DropConnect is a generalization of Dropout that randomly (e.g., by 50%) drops the weights rather than activating the full connected layer.

We use a feature extractor with two convolutional layers and two locally connected layers as described in [26,31,35]. The output of the convolutional layer is response-normalized and max-pooled separately by a normalization layer and a pooling layer. Two fully connected layers with ReLU activations are added between the softmax layer and the feature extractor (see Figure 2). The first fully connected layer is the DropConnect-layer, which has 128 output values. The second fully connected layer has only two or three output values. The softmax function is implemented at the final layer of the DNN used for classification.

We divide the training set into several batches used for different stages of training. To anneal the initial learning rate we choose a fixed multiplier for each stage of training. We report four numbers of epochs, such as 900-700-60-30 to define our schedule. We multiply the initial rate, such as 0.4 by 1 for the first and the second such number of epochs. Then we use a multiplier of 0.5 for the third number of epochs followed by 0.1 again for this third number of epochs. The fourth number of epochs is used for multipliers of 0.05, 0.01, 0.005, and 0.001 in that order, after which point we report our results. In addition, we choose different mini-batch sizes, such as 64-64-32-16 for different stages of training. The parameters throughout the model can be updated via stochastic gradient descent (SGD) by backpropagating gradients of the loss function. To balance the training time with the performance gains, we reduce the size of the images to 32 × 32 without rotation or scaling. Finally, we extract visual features with a length of 128 from the third-to-last layer of the trained visual sentiment analysis model.

3.3. Fusion

Because some messages have no accompanying images, in accord with reality, we record which microblogs include images before training and testing. When fusing the results, we adopt the following strategies: when an image exists in the current microblog message, we perform sentiment analysis of the message by fusing the visual and textual sentiment prediction; otherwise, we employ only the textual sentiment prediction.

We employ late fusion to analyze the performance of our model (see Figure 2). We use Logistic Regression to perform sentiment prediction of the text and any related images individually and, finally, fuse the probabilistic results using the average strategy and a weight, which is learned from the labeled data.

4. Experiments

In this section, we describe the baseline method, CBM [23], and compare it with our proposed method. Then we present the experimental results of our approaches as well as the baseline.

4.1. Dataset

Our experiments use the Sina Weibo dataset [23–25]. Sina-Weibo is a famous Chinese microblog company in Beijing. In [23], the authors chose the top ten hot topics from the Sina Weibo topic list, extracted the text and related images from the messages and built the dataset, which includes both labeled and unlabeled data. The labeled data consists of 6171 messages is (4196 positives, 1354 negatives and 621 neutrals), of which 5859 messages have one accompanying image. The unlabeled data includes large-scale messages stored in an ".sql" data table. This dataset covers a variety of topics including

weather events such as typhoons and smog, products such as iOS7 and Meizu MX3, and gossip about celebrities and films.

4.2. Baselines

Wang M. *et al.* [23] used the bag-of-words method to represent a microblog message as a vector in the CBM model, which consists of text features and image features. For text representation, they chose five basic properties such as the number of positive sentiment words from each message and selected the top N positive adjective-noun pairs (ANPs) and negative ANPs as image sentiment features. After preparing the message features, they applied some classifiers to perform sentiment prediction. In their experiments, Logistic Regression obtained the best performance. For convenience, we use CBM_Text, CBM_Image and CBM_Fusion as the baseline methods in this work, which respectively represent textual, visual and fusion methods in the CBM model.

4.3. Experimental Results and Discussion

We evaluated our approach using both Two-Class (positive and negative) and Three-Class (positive, negative and neutral) evaluations similar to [23,25].

The experiment is performed with 10 fold cross-validation. We randomly split the dataset into training and testing data. For each such split, we train the model on the training data and assess the predictive accuracy using the testing data. We then average the results over the splits. Because the results would vary if we were to repeat the analysis with different random splits, we run all models on the same split.

4.3.1. Textual Sentiment Analysis

The results of our text based methods against baseline methods are listed in Table 1. While we expected some performance gains because of the use of pre-trained vectors, we were surprised at the magnitude of the gains. For example, in the Two-Class evaluation, the method of CNN_W2V_Char based only on the textual information performs remarkably well, giving competitive results of 81.1% and 74.8% compared to the textual results of 76% and 65% chieved by the CBM_Text model [23].

As shown in Table 1, segmenting the text messages into Chinese characters can obtain better performance (the enhancement amounts to approximately 2 percentage points) than into Chinese phrases. The learning features from word vectors of Chinese characters preserve more information for the CNN model because Chinese characters can express richer meanings than Chinese phrases. The relatively smaller number of Chinese characters greatly reduces the size of the word vector vocabulary and thus increases time and space efficiency when extracting features.

In this study, we have described two experiments with convolutional neural networks built on top of pre-trained vectors trained by *word2vec*. These results suggest that *word2vec* is an efficient and effective feature extractor when applied for the purpose of Chinese sentiment prediction.

Table 1. Accuracy of textual methods.

Type	CBM_Text [23]	DNN_W2V_Phrase	DNN_W2V_Char			
			Accuracy	Precision	Recall	F1
Two-Class	0.76	0.793	0.811	0.894	0.872	0.883
Three-Class	0.65	0.720	0.748	-	-	-

"CBM_Text [23]" is the baseline model. "DNN_W2V_Phrase"and "DNN_W2V_Char" are our models using two different parsing algorithms that either segment each message into Chinese phrases (DNN_W2V_Phrase) or into Chinese characters (DNN_W2V_Char). Each phrase or character is separated by a space.

4.3.2. Visual Sentiment Analysis

Table 2 shows the visual sentiment prediction results of the network described in Section 3.2. The baseline model, CBM_Image [23,32], uses a detector library based on a constructed ontology and

is aimed at establishing a mid-level representation to bridge the affective gap and perform well (below 0.725 and below 0.66) on the current benchmark. We trained our deep convolutional neural network to automatically learn from the raw input and leverage DropConnect [26] to effectively relieve the overfitting problem that would otherwise occur on such a small dataset. Table 2 shows that our deep learning visual method achieves better performance (0.763 and 0.688) and surpass the visual sentiment prediction baseline in both Two-Class way and Three-class way.

Table 2. Accuracy of visual methods.

Type	CBM_Image [23]	DNN_Image			
		Accuracy	Precision	Recall	F1
Two-Class	<0.725	0.763	0.955	0.747	0.838
Three-Class	<0.66	0.688	-	-	-

4.3.3. Multi-Modality Sentiment Analysis

From Tables 1 and 2 we can see that sentiment prediction from images is not as effective as sentiment prediction from text. Therefore, based on a grid search, we choose the best fusion result by assigning a greater weight to the textual information when fusing the predicted class probabilities for text and images ((0.62, 0.38) in the Two-Class and (0.66, 0.34) in the Three-Class evaluations). From the results shown in Figure 3 we find that the visual information is helpful for sentiment prediction (it improved the performance from 0.811 to 0.826 in Two-Class evaluation and from 0.748 to 0.765 in the Three-Class evaluation). The experiment demonstrates that visual information functions as an effective supplement to textual sentiment prediction, and that sentiment analysis via multi-modality information achieves better performance than considering any modality alone.

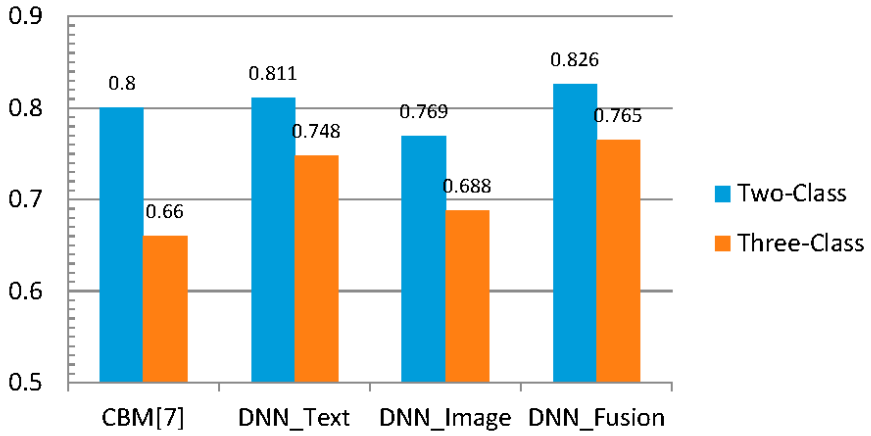


Figure 3. Sentiment prediction based on textual features, visual features and multi-modality features.

From Table 3 and Figure 3, we find our framework achieves better performance than CBM [23] in this benchmark not only in the Two-Class evaluation (0.826 *vs.* 0.80) but also in the Three-Class evaluation (0.765 *vs.* 0.66).

Our textual model utilizes large-scale unlabeled texts by employing unsupervised pre-training of word vectors to transfer ample semantic information for a simple CNN. This approach learns more discriminative features than traditional n-gram features. At the same time, our visual model extracts more abstract features via a deep and large neural network with billions of parameters and effectively relieves the overfitting problem by leveraging a regularized version of Dropout. Finally, we combine these representative features to analyze the sentiment expressed in samples. This combination inevitably leads toa great improvement in sentiment prediction.

Table 3. Accuracy of fusion methods.

Type	CBM_Fusion [23]	DNN_Fusion			
		Accuracy	Precision	Recall	F1
Two-Class	0.80	0.826	0.954	0.838	0.89
Three-Class	0.66	0.765	-	-	-

4.4. Error Analysis

From Figure 3, we can see that the accuracy of the single visual sentiment analysis model is lower than the fusion method, both in the Two-Class and Three-Class evaluations. We sum up two types of false prediction caused by introducing visual contents as follows (see Figure 4).

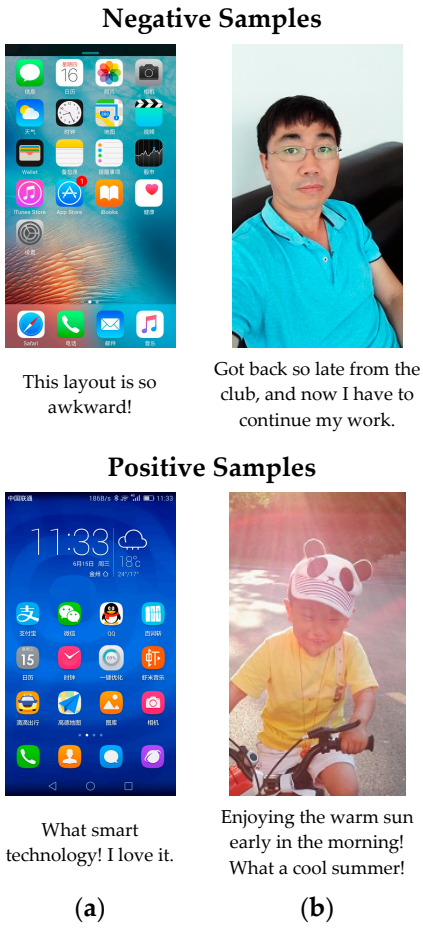


Figure 4. Examples of false prediction of negative and positive samples. Note: In example (**a**), smartphone screenshots do not carry explicit sentiment because different users of the same product have different views. The users are expressing opposite viewpoints about the smartphone layout using two similar pictures; in example (**b**), some users enjoy posting one portraiture no matter what their mood might be.

We also found two cases of false prediction that even CNN on top of word vectors cannot handle well, as follows.

- First, emerging Chinese cyberspeak—the shorthand language used on the Internet, increases the difficulty of understanding text, especially when the intent of the symbols differs from their literal meaning.
- Second, film review fragments out of context make textual sentiment prediction more difficult.

5. Conclusions

Sentiment analysis based on multimodality content is an interesting and challenging task. In this work, we leverage recently developed deep learning models to extract both textual and visual features to analyze the sentiment expressed in Chinese microblogs. On this Chinese benchmark, the experimental results demonstrate that our proposed model outperforms state-of-art sentiment analysis models that use only textual or only visual content for sentiment analysis. Our results provide further support for the use of unsupervised pre-training word vectors as robust features for NLP tasks such as textual sentiment analysis. The proposed model based upon CNN can learn higher level representations of both message text and related images. In future work, adding facial expression recognition, character recognition and compound image detection and separation could further improve the performance of visual prediction, giving it a larger contribution in multi-modality sentiment analysis.

Acknowledgments: This research was supported by the National Natural Science Foundation of China (No. 61272373, No. 61202254 and No. 71303031) and the Fundamental Research Funds for the Central Universities (No. DC13010313 and No. DC201502030202).

Author Contributions: Yuhai Yu designed and wrote the paper; Hongfei Lin supervised the work; Jiana Meng conceived and designed the experiments; Yuhai Yu and Zhehuan Zhao performed the experiments; and Yuhai Yu analyzed the data. All authors have read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional neural network
DNN	Deep convolutional neural network
VSO	Visual Sentiment Ontology
ANPs	Adjective-Noun Pairs
NLP	Natural language processing

References

1. Cambria, E. Affective computing and sentiment analysis. *IEEE Intell. Syst.* **2016**, *31*, 102–107. [CrossRef]
2. Strapparava, C.; Valitutti, A. WordNet Affect: An Affective Extension of WordNet. In Proceedings of the LREC, Lisbon, Portugal, 26–28 May 2004.
3. Esuli, A.; Sebastiani, F. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the LREC; Citeseer: Genoa, Italy, 2006.
4. Cambria, E.; Olsher, D.; Rajagopal, D. SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec, QC, Canada, 27–31 July 2014.
5. Cambria, E.; Fu, J.; Bisio, F.; Poria, S. AffectiveSpace 2: Enabling Affective Intuition for Concept-Level Sentiment Analysis. In Proceedings of the AAAI, Austin, TX, USA, 25–30 January 2015.
6. Cambria, E.; White, B. Jumping NLP curves: A review of natural language processing research [review article]. *IEEE Comp. Intell. Mag.* **2014**, *9*, 48–57. [CrossRef]
7. Kim, Y. Convolutional neural networks for sentence classification. 2014, arXiv preprint. arXiv:1408.5882.
8. dos Santos, C.N.; Gatti, M. Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of the 25th International Conference on Computational Linguistics (COLING), Dublin, Ireland, 23–29 August 2014.
9. Mesnil, G.; Mikolov, T.; Ranzato, M.; Bengio, Y. Ensemble of Generative and Discriminative Techniques for Sentiment Analysis of Movie Reviews. 2014, arXiv preprint. arXiv:1412.5335.
10. Xu, C.; Cetintas, S.; Lee, K.-C.; Li, L.-J. Visual Sentiment Prediction with Deep Convolutional Neural Networks. 2014, arXiv preprint. arXiv:1411.5731.
11. Cambria, E.; Hussain, A. *Sentic Computing: A Common-Sense-Based Framework For Concept-Level Sentiment Analysis*; Springer: New York, NY, USA, 2015; Volume 1.

12. Cambria, E.; Poria, S.; Bisio, F.; Bajpai, R.; Chaturvedi, L. The CLSA model: A novel framework for concept-level sentiment analysis. In *Computational Linguistics and Intelligent Text Processing*, Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14–20 April 2015; Springer: New York, NY, USA, 2015; pp. 3–22.
13. Chikersal, P.; Poria, S.; Cambria, E. SeNTU: Sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In Proceedings of the International Workshop on Semantic Evaluation, (Semeval 2015), Denver, CO, USA, 31 May–5 June 2015.
14. Chikersal, P.; Poria, S.; Cambria, E.; Gelbukh, A.; Siong, C.E. Modelling public sentiment in Twitter: Using linguistic patterns to enhance supervised learning. In *Computational Linguistics and Intelligent Text Processing*, Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14–20 April 2015; Springer: New York, NY, USA, 2015; pp. 49–65.
15. Maynard, D.; Dupplaw, D.; Hare, J. Multimodal sentiment analysis of social media. In Proceedings of the BCS SGAI Workshop on Social Media Analysis, Cambridge, UK, 10 December 2013.
16. Rosas, V.P.; Mihalcea, R.; Morency, L.-P. Multimodal sentiment analysis of spanish online videos. *IEEE Intell. Syst.* **2013**, *28*, 38–45. [CrossRef]
17. Poria, S.; Cambria, E.; Howard, N.; Huang, G.-B.; Hussain, A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* **2016**, *174*, 50–59. [CrossRef]
18. Poria, S.; Cambria, E.; Hussain, A.; Huang, G.-B. Towards an intelligent framework for multimodal affective data analysis. *Neural Net.* **2015**, *63*, 104–116. [CrossRef] [PubMed]
19. Poria, S.; Cambria, E.; Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the EMNLP, Lisbon, Portugal, 17–21 September 2015.
20. Poria, S.; Huang, G.; Gelbukh, A.; Cambria, E.; Hussain, A.W.; Huang, B. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowl. Based Syst.* **2014**, *69*, 108–123. [CrossRef]
21. Pereira, M.H.; Pádua, F.L.; Pereira, A.; Benevenuto, F.; Dalip, D.H. Fusing Audio, Textual and Visual Features for Sentiment Analysis of News Videos. In Proceedings of the International AAAI Conference on Web and Social Media, Cologne, Germany, 17–20 May 2016.
22. You, Q.; Luo, J.; Jin, H.; Yang, J. Joint Visual-Textual Sentiment Analysis with Deep Neural Networks. In Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, Brisbane, Australia, 26–30 Octorber 2015; ACM: New York, NY, USA, 2015.
23. Wang, M.; Cao, D.; Li, L.; Li, S.; Ji, R. Microblog Sentiment Analysis Based on Cross-media Bag-of-words Model. In Proceedings of the International Conference on Internet Multimedia Computing and Service, Xiamen, China, 10–12 July 2014; ACM: New York, NY, USA, 2014.
24. Cao, D.; Ji, R.; Lin, D.; Li, S. Visual sentiment topic model based microblog image sentiment analysis. *Multimed. Tools Appl.* **2014**, *73*, 1–14. [CrossRef]
25. Cao, D.; Ji, R.; Lin, D.; Li, S. A cross-media public sentiment analysis system for microblog. *Multimed. Syst.* **2014**, *71*, 1–8. [CrossRef]
26. Wan, L.; Zeiler, M.; Zhang, S.; Cun, Y.L.; Fergus, R. Regularization of neural networks using dropconnect. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), Atlanta, GA, USA, 16–21 June 2013.
27. Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; Kuksa, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **2011**, *12*, 2493–2537.
28. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. 2013, arXiv preprint. arXiv:1301.3781.
29. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems 26, Carson, NV, USA, 5–10 December 2013.
30. Mikolov, T.; Yih, W.-T.; Zweig, G. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of the HLT-NAACL, Atlanta, GA, USA, 9–14 June 2013.
31. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems 25, Carson, NV, USA, 3–6 December 2012.

32. Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM international conference on Multimedia, Barcelona, Spain, 21–25 Octorber 2013; ACM: New York, NY, USA, 2013.
33. Chen, T.; Borth, D.; Darrell, T.; Chang, S.F. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. 2014, arXiv preprint. arXiv:1410.8586.
34. Zeiler, M.D. ADADELTA: An adaptive learning rate method. 2012, arXiv preprint. arXiv:1212.5701.
35. Yu, Y.; Lin, H.; Yu, Q.; Meng, J.; Zhao, Z.; Li, Y.; Zuo, L. Modality classification for medical images using multiple deep convolutional neural networks. *J. Comput. Inf. Syst.* **2015**, *11*, 5403–5413.