Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766213

HYBRID RANDOM FOREST AND CONVOLUTIONAL NEURAL NETWORK FOR DEEP LEARNING CROSS DOMAIN SENTIMENT CLASSIFICATION

[1]Mrs.V.Manimekalai, [2]Dr.S.Gomathi alias Rohini

[1]Research scholar,Sri Ramakrishna College of Arts & Science and Assistant Professor, Dr. N.G.P Arts & Science College, Coimbatore.

[2]Associate Professor, Dept. of Computer Application, Sri Ramakrishna College of Arts & Science, Coimbatore.

Abstract

Sentiment analysis becomes more popular in the research area. It assigns positive or negative polarity to an entity or items using various natural language processing methods, as well as predicts the high and poor performance of various sentiment classifiers. Our work focuses on sentiment analysis based on product reviews utilising novel text search algorithms. These reviews can be classified as positive or negative depending on specific factors in connection to a query based on phrases. We presented a hybrid strategy to identifying product reviews in this research. The results show that the proposed system approach outperforms these individual classifiers in this dataset. Cross-domain sentiment classification has drawn much attention in recent years.

Keywords: Sentiment Analysis, Random Forest, Convolutional Neural Network

I. INTRODUCTION

There are numerous uses for sentiment analysis. For example, businesses and organisations must get consumer and public feedback on their products and services. Individual consumers may also want to know what other people think about a product before purchasing it. Furthermore, they may be interested in hearing other people's perspectives on political candidates before making a vote decision in a political election. Furthermore, data is now widely and freely available on the Internet, but with no data quality guarantee, it is up to the readers to determine whether or not to believe it.

As a result, there is a considerable demand for enhanced fact checking methodologies and applications that contribute to data quality assurance. In particular, consumers using the Web are more frequently harmed/damaged by inconsistencies in information. Furthermore, establishing and implementing such fact-checking systems necessitates the use of robust sentiment analysis models. When fact checking systems are given by a general or universal model that can be trained once and then used to additional reviews, this work for fact checking detection can be completed. Certainly, this model should perform well in order to improve the accuracy of such fact-checking systems.

Deep learning-based sentiment categorization has advanced considerably in recent years. While these approaches provide excellent results, they usually require large-scale datasets containing sentiment polarity labels to train a viable sentiment classifier. Deep learning-based sentiment categorization has advanced considerably in recent years. While these approaches provide excellent results, they usually require large-scale datasets containing sentiment polarity labels to train a viable sentiment classifier.

Convolutional Neural Network (CNN) is a deep feed forward neural network class that uses local connection patterns to handle data in the form of numerous arrays. 21 CNNs achieved competitive state-of-the-art performance in a variety of classification tasks, including image and text classification. A CNN is made up of an input layer, M convolutional layers, H MLP hidden layers, and an output layer. The existence of the early convolutional layers, each consisting of a convolutional layer and a pooling layer, distinguishes it from MLP.

The convolutional layer's contribution to the CNN is the convolution operation itself, which is a type of sliding window function that performs a matrix product between the input and a filter matrix or vector, also known as a kernel or feature detector that is smaller than the size of the input matrix. Because this convolution procedure results in a sparser interaction in CNN, fewer parameters are evaluated, increasing computing efficiency. Another property that separates CNN from other neural networks is parameter sharing, which refers to the usage of the same parameter for more than one function in a model since each member of the kernel is employed at every input location. By implementing this parameter sharing, the layers also assume translation property equivariance.

The pooling layer, which further alters the convolutional layer output by substituting the values in certain locations with the summary statistics of the surrounding outputs, is another key component in a CNN. The max polling and average pooling functions are two well-known pooling functions. If the convolutional vector output c is partitioned into v rectangular sections, each comprised of e = cv components, then the pooling output is a vector of length v, with each element corresponding to the maximum or average of the e-th rectangle. In this article, we use a CNN with a convolution layer followed by another convolutional layer and an average pooling layer.

The remainder of this paper is organized as follows. Section II provides an overview of related work in the area, and Section III covers the methodology used for sentiment classification. In Section IV, we discuss the results and compared the performance factor in terms of accuracy. Section V offers the conclusions of this paper.

II. RELATED WORK

Hu. M. et al. [1] demonstrated the use of domain-invariant sentiment characteristics to enhance cross-domain sentiment classification by training an aspect detection task while removing domain-specific aspect features from the input phrase. They improve the separation of domain-invariant and domain-specific information by using two effective modules, the context allocation mechanism and the domain classifier, respectively. The experimental findings revealed the efficacy of the suggested strategy and, through visualisation, confirmed the rationale of the context allocation process.

To facilitate domain transfer, Li. T et al. [2] developed Contrastive Learning with Mutual Information Maximization. On two typical benchmarks, they exceed strong baselines and the preceding state-of-the-art technique. Furthermore, while back translation works well in this article, it would be interesting to investigate additional data augmentation approaches and conclude which ones best support contrastive learning for cross-domain sentiment categorization.

Because contrastive learning alone may not always aid domain transfer, we employ mutual information maximisation [2] to select discriminative characteristics that can best assist the final prediction. With mutual information maximisation, their model's prediction distribution will be more balanced, and the margin between classes on the target domain will be larger, making our model more resilient and allowing the same predictor to be optimum across domains. The notion is that the decision boundary learned in the source domain will be more likely to fall into the margin in the target domain. Experiments on two typical cross-domain sentiment classification benchmarks demonstrate the usefulness of the proposed technique, and ablation experiments are conducted to demonstrate that mutual knowledge aids cross-domain sentiment classification in a variety of domain situations.

Fei, R et al. [3] shown that the classic cross entropy loss function is enhanced using the LSTM and CNN models. *e LSTM-BO and CNN-BO models are developed in such a way that the enhanced model can more successfully fit the prediction error samples and avoid overfitting. Furthermore, the relevance of each word to the classification results is derived by analysing the effect of the input words on the final classification, which is integrated with the properties of the circulating neural network, and the W-RNN model is produced. *e model lends more weight to words having a greater emotional propensity and prevents emotional information loss. To test the efficacy of the three sentiment classification models, qualitative and quantitative sentiment analysis experiments were constructed using two types of datasets in Chinese and English. The experimental results show that the three models proposed in this paper improve the accuracy of text sentiment classification to a certain extent and also perform better in loss rate and time performance.

The HANP approach for CDSC task was described by Manshu, T et al [4]. For emotion categorization, the suggested HANP can pay greater attention to key words and phrases. The emotion dictionary match layer may additionally capture crucial pivots, non-pivots, and dis-pivots. The HANP can determine the meaning of a non-pivot or dis-pivot by searching for synonyms. Experiments on the Amazon review dataset demonstrate the efficacy of HANP. On this dataset, we achieve state-of-the-art accuracy.

For the cross-domain sentiment classification problem, Yin H et al [5] investigated a new capsule network with identifying transferable knowledge (CITK). They begin by extracting significant and consistent polar terms in two domains as a means of discovering transferrable information in order to increase classification accuracy as much as feasible. The common information is then embedded into a capsule network for improved sentiment categorization. Experiments on 12 source-target domain pairings reveal that our strategy outperformed earlier methods designed specifically for cross-domain sentiment categorization.

Manshu, T et al. [6] showed an end-to-end CDSC task model, CCHAN. For emotion categorization, the suggested CCHAN can pay more attention to key words and phrases. The efficacy of the auxiliary task cloze task, as well as the structure CNN they included in HAN, is also validated for the CDSC task. Because our model does not require labelled pivots as input, it may be easily applied to various cross-domain classification tasks. Experiments using Amazon review datasets demonstrated the efficacy of CCHAN. On these datasets, the author achieved state-of-the-art accuracy.

Myagmar et al. [7] demonstrated the cross domain sentiment classification problem using bidirectional contextualised Transformer language models from BERT and XLNet. BERT and XLNet both beat prior state-of-the-art algorithms for CDSC task due to their unsupervised pre-training tasks that use huge unlabeled datasets and their self-attention

Transformer mechanisms. XLNet outperforms BERT on all CDSC tasks when compared closely. XLNet is exceptionally effective at collecting context and delivers state-of-the-art results with just 50 fine-tune training samples, which is approximately 120 times less data than the prior top performing CDSC systems trained on. The improved prediction accuracy of XLNet is mostly due to its innovative pre-training aim, ability to capture long-term dependencies, and bigger pre-training dataset. XLNet consumes more resources than BERT, but it learns contextual data considerably faster and with fewer fine-tuning stages.

P. Zola et al. [8] investigated a unique cross-source cross-domain sentiment analysis method. Their goal is to easily classify the sentiment of distinct items (e.g., restaurant, hotel, book, music) by first fitting a sentiment classifier to easily collected labelled Web sources (such as Amazon and Trip Advisor) and then reusing such model to predict the sentiment of typically unlabeled social media reviews (from Facebook and Twitter). They used a three-step experimental methodology to test different modelling methods: balancing training methods – under sampling and oversampling; text pre-processing – stemming and POS tagging; and learning algorithms – Naive Bayes (NB), Support Vector Machine (SVM), deep Multilayer Perceptron (MLP), and Convolutional Neural Network (CNN).

Qu X et al [9] presented a unique strategy for achieving category-level alignment for sentiment analysis that makes use of multiple view classifiers. Unlike prior efforts, the author considers the decision boundary, properly categorising the 2504 target samples into the appropriate group. Experiments reveal that the suggested ACAN beats state-of-the-art approaches on the Amazon benchmark substantially. They would want to extend our solution to more domain adaptation jobs in the future, as well as examine more effective options for the generator regularizer.

Peng M et al [10] explored the role of domain-specific information in domain adaptation. In contrast to the majority of earlier techniques, which focused on domain invariant information, they demonstrated that domain specific information may also be beneficially utilised in the domain adaptation job with a modest quantity of in-domain labelled data. Specifically, the author suggested a unique approach based on the CMD metric for extracting domain invariant and domain specific features from target domain data at the same time. They co-trained these two distinct features using labelled data from the source domain and a modest quantity of labelled data from the target domain. Experiments with sentiment analysis showed the method's efficacy.

III. METHODOLOGY

The random forest is a member of a family of approaches that use the decision tree as an individual predictor and are based on the methods of Bagging, Randomizing Outputs, and Random Subspace Excluding Boosting. The random forest method is one of the greatest classification algorithms, capable of accurately classifying massive volumes of data. It is a classification and regression ensemble learning approach that generates a number of decision trees during training and offers the class that is the mode of the classes generated by individual trees.

Many classifiers are formed from smaller subsets of the input data in the random forest classification approach, and their individual findings are eventually combined based on a voting process to yield the desired output of the input data set. This ensemble learning technique has lately gained a lot of traction.

To begin, random forest is an ensemble learning approach that builds a number of decision trees with randomly chosen characteristics and predicts the class of a test instance by voting on the various trees. The concept of a margin – either side of a hyper plane that divides two classes — is essential to Support Vector Machine. Maximizing the margin and therefore establishing the greatest feasible distance between the separating hyper plane and the examples on either side of it has been shown to lower the upper bound on the predicted generalisation error. Because RF was unaffected by input parameters, we simply utilised the default settings for each classifier.

The trained classifiers produce scores ranging from 0 to 1, which are then converted to a binary state signalling 'negative' or 'positive.' The presence of an element is classified as positive (P) or negative (N) for each combination (N). The abbreviation TP stands for True Positives, which are the number of examples predicted positive that are actually positive, FP stands for False Positives, which are the number of examples predicted positive that are actually negative, TN stands for True Negatives, which are the number of examples predicted negative that are actually negative, and FN stands for False Negatives, which are the number of examples predicted negative that are actually positive.

The classification metrics examined for sentiment analysis are Accuracy, Precision, Recall, and F-Measure, and these parameters are assessed using the suggested hybrid technique based on the computed positivity and negativity of reviews. The following formulae are used to evaluate the performance of classifiers: True positives should be reported.

It corresponds to:

$$TP\ Rate = \frac{TP}{TP + FN}$$

$$FP\ Rate = \frac{FP}{FP + TN}$$

It is thus the report between the number of positive instances classified well and the total number of elements which should be classified well.

Precision is the report between the number of the true positive and the sum of the true positives and the false positive. A value of 1 expresses the fact that all the positive classified examples were really:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

Accuracy is a common measure for the classification performance and it's proportional of correctly classified instances to the total number of instances, whereas the error rate uses incorrectly classified rather than correctly

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

IV. RESULT AND DISCUSSION

TABLE I PERFORMANCE COMPARISON OF ALGORITHMS

Algorithm name	Precision	Recall	Accuracy
Hybrid Random Forest and CNN	98.7	99.3	99
Convolutional Neural Network(CNN)	95.8	94.4	95
Random Forest	93.9	92.5	93

The above tabulation value depicts that hybrid random forest and convolutional neural network yields better performance than other two algorithms which are taken into comparison.

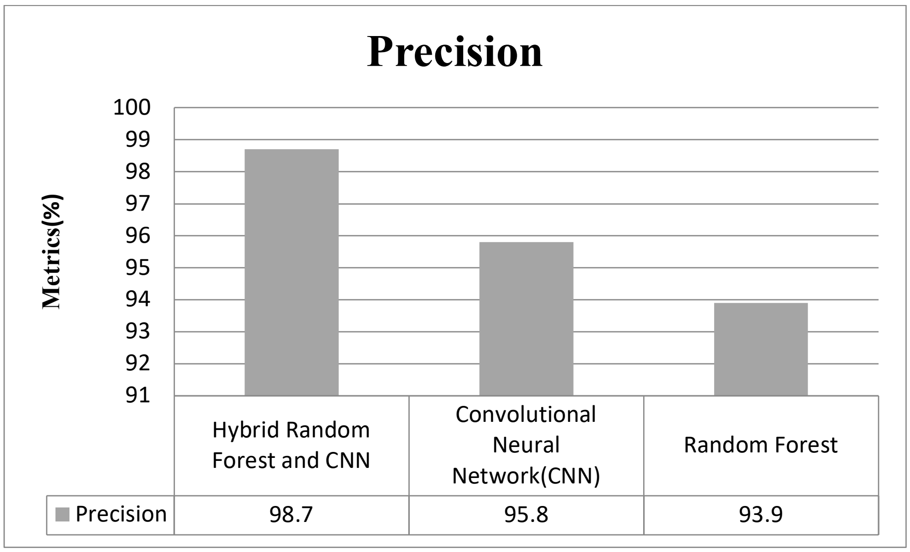

Fig 4.1 Comparison of Precision for Sentiment Classification

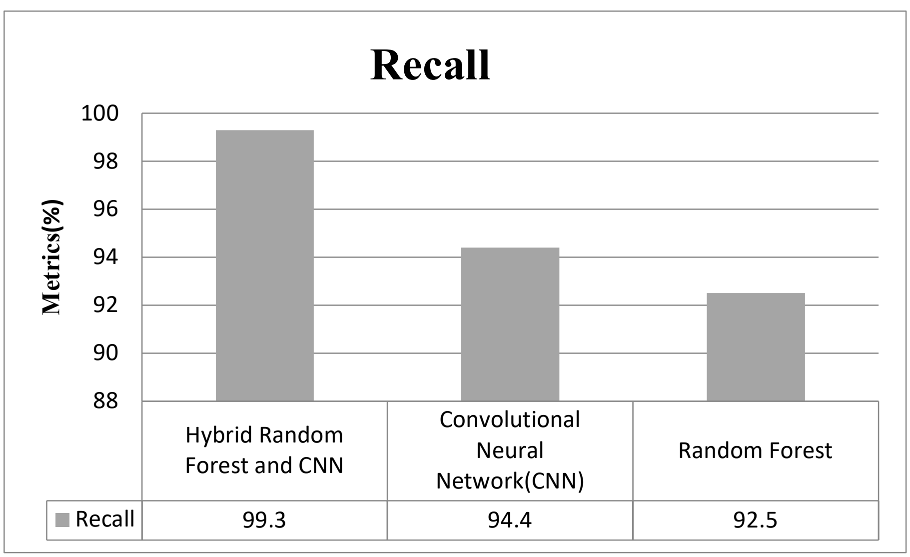

Fig 4.2 Comparison of Recall for Sentiment Classification

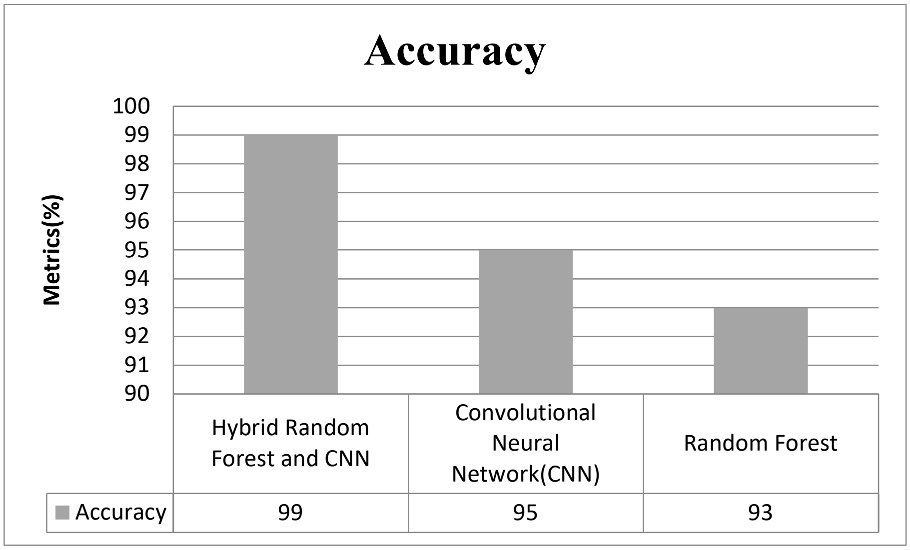

Fig 4.3 Comparison of Accuracy for Sentiment Classification

The above results comparison showed that performance metrics of precision, recall and accuracy of various algorithms. Hybrid random forest and convolutional neural network gives better performance.

V. CONCLUSION

We investigate the subject of aspect level cross-domain sentiment analysis in this paper and present a hybrid random forest and convolutional neural network for deep learning-based cross domain sentiment categorization. According to the testing results, Random Forest and Convolutional Neural Network appear to be superior to the other algorithms for product evaluations. Better results are obtained in the case of the hybrid classification approach utilised in this study because it takes use of the advantages of each of the distinct traditional classification methods.

REFERENCES

1. Hu, M., Wu, Y., Zhao, S., Guo, H., Cheng, R. and Su, Z., 2019. Domain-invariant feature distillation for cross-domain sentiment classification. *arXiv preprint arXiv:1908.09122*.
2. Li, T., Chen, X., Zhang, S., Dong, Z. and Keutzer, K., 2021, June. Cross-domain sentiment classification with contrastive learning and mutual information maximization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 8203-8207). IEEE.
3. Fei, R., Yao, Q., Zhu, Y., Xu, Q., Li, A., Wu, H. and Hu, B., 2020. Deep learning structure for cross-domain sentiment classification based on improved cross entropy and weight. *Scientific Programming*, *2020*.
4. Manshu, T. and Bing, W., 2019. Adding prior knowledge in hierarchical attention neural network for cross domain sentiment classification. *IEEE Access*, *7*, pp.32578-32588.

5. Yin, H., Liu, P., Zhu, Z., Li, W. and Wang, Q., 2019. Capsule network with identifying transferable knowledge for cross-domain sentiment classification. IEEE Access, 7, pp.153171-153182.
6. Manshu, T. and Xuemin, Z., 2019. CCHAN: An end to end model for cross domain sentiment classification. *IEEE Access*, *7*, pp.50232-50239.
7. Myagmar, B., Li, J. and Kimura, S., 2019. Cross-domain sentiment classification with bidirectional contextualized transformer language models. IEEE Access, 7, pp.163219-163230.
8. Zola, P., Cortez, P., Ragno, C. and Brentari, E., 2019. Social media Cross-Source and cross-domain sentiment classification. International Journal of Information Technology & Decision Making, 18(05), pp.1469-1499.
9. Qu, X., Zou, Z., Cheng, Y., Yang, Y. and Zhou, P., 2019, June. Adversarial category alignment network for cross-domain sentiment classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 2496-2508).
10. Peng, M., Zhang, Q., Jiang, Y.G. and Huang, X.J., 2018, July. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2505-2513).
11. Hai, Z.; Cong, G.; Chang, K.; Liu, W.; Cheng, P. Coarse-to-fine review selection via supervised joint aspect and sentiment model. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval; ACM: Berlin/Heidelberg, Germany, 2014; pp. 617–626.
12. Bajpai, R.; Hazarika, D.; Singh, K.; Gorantla, S.; Cambria, E.; Zimmerman, R. Aspect-Sentiment Embeddings for Company Profiling and Employee Opinion Mining. arXiv 2019, arXiv:1902.08342.
13. Sahu, T.P.; Ahuja, S. Sentiment analysis of movie reviews: A study on feature selection & classification algorithms. In Proceedings of the 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), Durgapur, India, 23–25 January 2016; pp. 1–6.
14. Sachan, D.S.; Zaheer, M.; Salakhutdinov, R. Revisiting LSTM Networks for Semi-Supervised Text Classification via Mixed Objective Function. Proc. AAAI Conf. Artif. Intell. 2019, 6940–6948.
15. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing—Volume 10 (EMNLP '02); Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 79–86.
16. Salvetti, F.; Lewis, S.; Reichenbach, C. Automatic Opinion Polarity Classification of Movie Reviews. Colo. Res. Linguist. 2004,
17. Dong, L.; Wei, F.; Liu, S.; Zhou, M.; Xu, K. A statistical parsing framework for sentiment classification. Comput. Linguist. 2015, 41, 293–336.
18. Kim, Y. Convolutional neural networks for sentence classification. arXiv 2014, arXiv:1408.5882.

19. Socher, R.; Pennington, J.; Huang, E.H.; Ng, A.Y.; Manning, C.D. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 151–161.
20. Tsutsumi, K.; Shimada, K.; Endo, T. Movie review classification based on a multiple classifier. In Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation, Seoul, Korea, 1–3 November 2007; pp. 481–488.
21. LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. Handb. Brain Theory Neural Netw. 1995, 3361, 1995.
22. Klambauer, G.; Unterthiner, T.; Mayr, A.; Hochreiter, S. Self-normalizing neural networks. In Advances in Neural Information Processing Systems; Curran Associates Inc: Red Hook, NY, USA, 2017; pp. 971–980.
23. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving neural networks by preventing co-adaptation of feature detectors. arXiv 2012, arXiv:1207.0580.
24. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989, 2, 359–366.