International Research Journal of Engineering and Management Studies (IRJEMS)
Volume: 03 Issue: 05 | May -2019 ISSN: 2395-0126

A CLUSTERING BASED DEEP HYPERGRAPH MODEL FOR SENTIMENT CLASSIFICATION

Abinaya v[1], Mrs.K.Krishnakumari[2]
[1]*Student, Department of Computer Science and Engineering*
A.V.C. College of Engineering,
Mannampandal, Mayiladuthurai – 609305
abi.abinayav@gmail.com
[2] *Associate Professor, Department of Computer Science and Engineering*
A.V.C. College of Engineering,
Mannampandal, Mayiladuthurai – 609305
krishna.41999@gmail.com

Abstract - *In the existing sentiment classification system ignore the semantic correlation among different reviews, and causing ineffectiveness for sentiment classification. In this project we are using deep hyper-graph model based on the word embedding clustering to predict higher order relation among review samples. Improved hierarchical fast clustering algorithm is used to discover the semantic cliques, which provide label information for detecting precise semantic units in embedding space. Convolution neural networks are connected to extract the high-order textual and semantic features of reviews. Finally, the hyper-graph can be constructed based on high-order relations of samples for the sentiment classification of reviews.*

Key Words **Keywords: sentiment classification, word embedding clustering, hierarchical fast clustering, deep hyper-graph, Convolutional neural networks.**

1.INTRODUCTION

Sentiment classification is the task of text classification whose main aim is to classify a text according to the sentimental polarities of opinions it contains either positive or negative. Natural language processing (NLP) with deep learning is a useful combination. Using word2vec representations and embedding layers to train recurrent neural networks with outstanding performances in a wide variety of industries. Online reviews are playing a remarkable role for consumers in online purchasing decisions. In the environment of information overload today, valuable online reviews can help consumers make better decisions. In this the semantic correlation among the review can be identified using cluster based deep hypergraph model. The Amazon review dataset is used to calculate sentiment analysis.

1.2 DEEP HYPERGRAPH MODEL

A word embedding clustering-based deep hyper graph model is proposed for the sentiment analysis of online reviews. Hyper graph is used instead of using di-graph because it predict the higher order relation among the review in an effective manner compared with di-graph. While graph edges are pairs of nodes, hyper edge are arbitrary sets of nodes, and can therefore contain an arbitrary number of nodes. In sentiment analysis phase words are the vectors and the semantic relation exist between the set of words means the hyperedge cover those set of words.

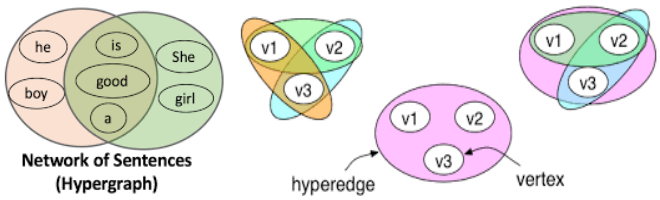


Figure 1 Hypergraph model

1.3 OVERVIEW

The main aim of the project is concerned with semantic correlation among different review in Amazon review dataset. A clustering based deep hypergraph model for sentiment classification is done through convolution neural network of word embedding data and deep hypergraph model. Word embedding is a set of feature engineering techniques that map sparse word vector into continuous space based on surrounding context word. The hierarchical fast clustering algorithm is used to discover word embedding clustering. After that the precise semantic unit are detected using the semantic cliques. There after the CNN is used to capture the feature vector of each text. Finally the deep hypergraph model is used to find the semantic polarity among the different review dataset.

2. RELATED WORK

Laurence et al. [1] the author used two method deep computation model and back propagation to compress the parameters and to improve the training efficiency. Results demonstrate that the model can compress parameters greatly and improve the training efficiency significantly with a low classification accuracy drop.**Zhikui Chen et al. [2]** the author propose hypergraph model to capture textual information and sentimental information simultaneously for sentiment classification of reviews. The author used four domain datasets (books, DVD, electronics, kitchen) to evaluate the model. The main advantage is hypergraph gives the higher order relation among review sample. **zhang et al.[3]** Deep Pyramid Convolutional Neural Networks for Text Categorization, the author proposes a low-complexity word-level deep convolutional neural network (CNN) architecture for text categorization that can efficiently represent long range associations in text. The main advantage is this model is well performed for large training dataset. **Jin et al. [4]** the author propose a topic-level maximum entropy (TME) model for social emotion classification over short text. TME generates topic-level features by modeling latent topics, multiple emotion labels, and valence scored by numerous readers jointly. The main advantage of the model is it tackle data sparsity problem. **Richard et al. [5]** Better Word Representations with Recursive Neural Networks for Morphology, the author propose a novel model that is capable of building representations for morphologically complex words from their morphemes. The model gives syntactic and semantic information in clustering related words. **Wan et al. [6]** Implicit feature identification via hybrid association rule mining, the author used novel hybrid association rule mining to predict many association rules using several complementary algorithms. A few more parameters of the hybrid association rule mining so that it is a little hard to control in practical application **Zheng et al. [7]** Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification, the author used unsupervised dependency analysis-based approach to extract Appraisal Expression Patterns (AEPs) from customer reviews. It is a sentence-level model so that the accuracy is less. It having many useless AEP's so it leads to the model ineffectiveness **Corrado et al. [8]** the author propose two novel model architectures for computing continuous vector representations of words. The author used Google News corpus for training the word vectors.**Liang Liu et al. [9]** Movie-rating and review-summarization system in a mobile environment. The movie-rating information is based on the sentiment classification resultPLSA approach does not work well in product feature identification.

2.1 DISADVANTAGES OF EXISTING SYSTEM

- An early work only focus on textual representation and ignoring the semantic correlation among review
- Lexicon based methods show low level of reliability.
- In existing sentiment classification the accuracy level are not good compare with human judgment.
- The di-graph is used to predict the indirect and direct relation among the review sentence in the existing sentiment analysis process but it ignore higher order relation among review sample.

3 PROPOSED SYSTEM

In this paper, a deep hypergraph model based on the word embedding clustering is proposed for sentiment polarity classification of Amazon reviews. In this work, the proposed model will keep the composition of the texts as much as possible and acquire the semantic representations of each text. We introduce external knowledge by taking advantage of semantic units consisted of the pre-training word embedding representations and contextual information to improve classification performance for user reviews. Word2vec technique is used to form the word embedding data of amazon review dataset.

The word2vec has two main techniques are there-

- Continuous Bag of Word Model
- Skip-Gram Model

After that the improved hierarchical clustering algorithm is used to cluster the word embedding data based on their density peaks searching. Using these semantic cliques the meaningful semantic units are detected in an efficient way. The semantic cliques provide label information for detection of the latent semantic unit in embedding space. The text can then be represented as a matrix composed by embedding representations of semantic units. After getting the projected matrices, the deep CNN is trained to capture the feature and acquire the feature vector of each text.

Finally, the hypergraph model can be constructed based on feature vectors, where each text can be treated as a vertex and different polarity relations can be used to construct Hyper edge based on the similarity among texts.

3.1 ADVANTAGE OF PROPOSED SYSTEM

- A deep hypergraph model based on word embedding clustering, which can capture the high-level features and reflect the high-order relations among samples.
- Semantic units are detected with considering the central words, which maximally preserves original information of reviews for improving sentiment classification accuracy.

4 SYSTEM DESIGN

The amazon review dataset is used to analysis the sentiment polarity using the cluster based deep hypergraph model. In the dataset 80% is used for testing phase and the remaining part is used for test phase. The first step is to prepare the text corpus for learning the embedding by creating word tokens, removing punctuation, removing stop words etc. The second step is to form a word embedding data using pre-trained word2vec model. The word2vec algorithm processes documents sentence by sentence. The continuous bag of words and skip-gram model is used to find the semantic relation among the words in the sentence. The next step is to map embedding from the loaded word2vec model for each word to the vocabulary and create a matrix

with of word vectors. After the word embedding data is taken as input to the hierarchical fast clustering algorithm. It gives cluster of words in an embedding space.

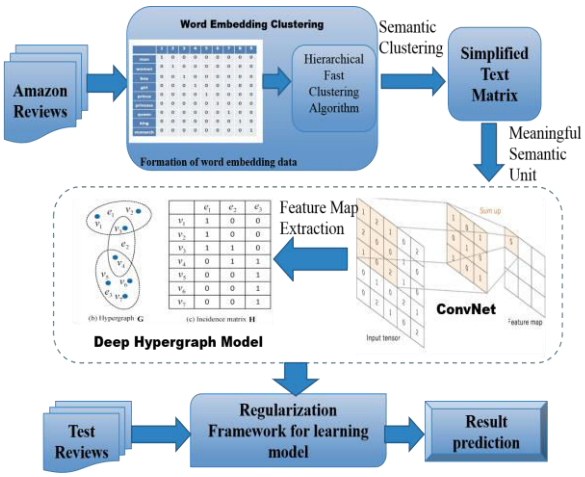


Figure 2 System architecture

The semantic cliques are used to discover the meaningful semantic units from the review dataset. Deep convolution neural network takes the sequence of embedding vectors as input and converts them to a compressed representation. The CNN layer capture the most meaningful global features with fixed size, and enable the output feature map meet the needs of hypergraph construction, a k-max pooling operation is used after the third convolution. This fixed sized global feature map can be then used to construct hypergraph model. The hypergraph model is constructed for sentiment classification, instead of softmax function used to get the probability distribution of output feature representations. Thus each review text can be represented as a vertex in hypergraph, and the similarity relationships among reviews can be treated as different hyperedge.

4.1 MODULES

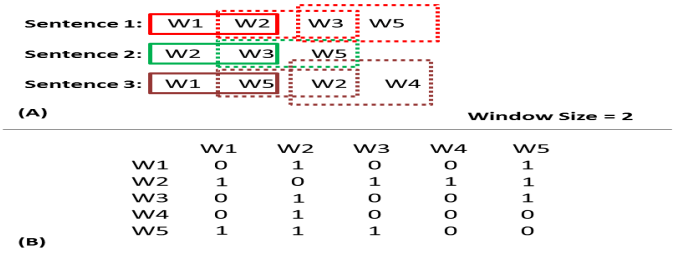
In this project there are five modules are used to classify the sentiment in the review dataset. The modules are,

1. Formation of word embedding data
2. Discovering semantic cliques
3. Semantic unit detection
4. Feature Extraction by CNN
5. Hypergraph construction

SAMPLE REVIEW:
Review 1: Very nice app
Review 2: This restaurant has best food.
Review 3: Most disgusting food I have ever had
Review 4: The game is great to pass time with it

4.1.1 FORMATION OF WORD EMBEDDING DATA

The formation of word embedding is an unsupervised learning based method. A word embedding is a technique that maps the words into a high dimensional continuous vector space where different words with a similar meaning have a similar vector representation. Word-vector representation means columns represent the dimensions of vector, row having input words, for each word column size vector is present. This Word-vector representation provides convenient properties for comparing words or phrases. Word2vec is not a single algorithm it is a combination of two techniques- continuous bag of words and skip gram model. Both of the techniques are shallow neural network which map word to the target word.



	W1	W2	W3	W4	W5
W1	0	1	0	0	1
W2	1	0	1	1	1
W3	0	1	0	0	1
W4	0	1	0	0	0
W5	1	1	1	0	0

Figure 3 Word Embedding

EXAMPLE:

Let's tokenize our sample and do usual cleaning steps like removing special character, stop words. So the dictionary may looks like –
[Very, nice', app, restaurant, best, food, disgusting, game, great, time]
Vocabulary size –10, Window size – 2

One hot encoding is used because it is the simple method to represent word in the vector form. Here 1 stands for the position where the word exist and 0 everywhere else.

Unique Words	One Hot Encoding	Unique Words	One Hot Encoding
Very	[1,0,0,0,0,0,0,0,0,0]	Food	[0,0,0,0,0,1,0,0,0,0]
Nice	[0,1,0,0,0,0,0,0,0,0]	Disgust	[0,0,0,0,0,0,1,0,0,0]
App	[0,0,1,0,0,0,0,0,0,0]	Game	[0,0,0,0,0,0,0,1,0,0]
Res	[0,0,0,1,0,0,0,0,0,0]	Great	[0,0,0,0,0,0,0,0,1,0]
Best	[0,0,0,0,1,0,0,0,0,0]	Time	[0,0,0,0,0,0,0,0,0,1]

Table 1 One-hot encoding

4.2.2 DISCOVERING SEMANTIC CLIQUES

The improved hierarchical fast clustering algorithm is used to cluster the word embedding data. The algorithm based on density peaks searching for semantic clustering of word embedding data. It is used to discover the semantic cliques, which provide label information for detection of the meaningful semantic unit in the Amazon review dataset. The algorithm first divides word embedding data into sub-datasets containing M samples, where M needs to be specified by the user. Then, clustering results of sub-datasets can be obtained by performing cluster process on each sub-dataset in parallel. Because the words in the same cluster are semantically close, a cluster can be represented by the clustering center as the sample for second clustering.

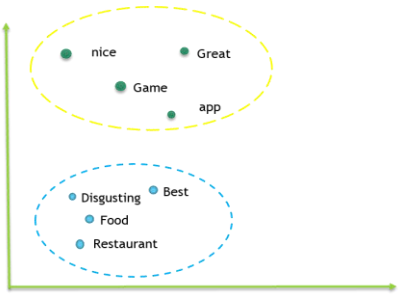


Figure 4 Semantic clustering

4.2.3 SEMANTIC UNIT DETECTION

The detection of the meaningful and precise semantic units is helpful for improving the model performance in an effective manner. In dense embedding space, semantically close words are likewise close in cosine distance. The cosine similarity is used to indicate the semantic similarity between words in Word2Vec. So computing cosine similarity to measure the similarity between semantic units and semantic clique's center for recognizing important and meaningful semantic units. A threshold τ is used as a constraint to fine-tune the detection of semantic units.

$$RM.M_{Win} = [su_1, su_2, \ldots . su_m]$$

Where RM is matrix representation of review text, M (win) is window matrix for word2vec model,

$$su_i = \sum_{j=1}^{s} RM_j^{win,i}$$

EXAMPLE:

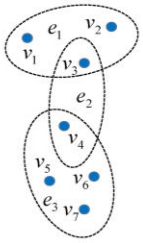
$$su_{food} = \sum (Food_2^{food}, Food_2^{restaurant}, Food_2^{disgusting}, Food_2^{best})$$

4.2.4 FEATURE EXTRACTION BY CNN

After detection for all semantic units of an input review text, review text can be represented as matrix constituted by the embedding representations of these semantic units, which is used as the input to CNN model. In this work, we construct a deep neural network for feature extraction with three convolutional layers to extract higher level features. After each convolution, the max-over-time pooling layer over the feature map is connected to capture the most useful local features for sentiment classification and reduce the size of input to the next convolutional layer, thereby reducing the complexity of the model. This fixed sized global feature map can be then used to construct hypergraph model.

4.2.5 HYPERGRAPH CONSTRUCTION

The feature representation of each input review text is obtained after the layer of feature extraction performed by operations convolution and max-pooling. Hypergraph model is constructed for sentiment classification, instead of softmax function used to get the probability distribution of output feature representations. Thus each review text can be represented as a vertex in hypergraph, and the similarity relationships among reviews can be treated as different hyperedge. The hypergraph G = (V, E,w), where V is a finite set of vertices, E is the hyperedge set and w is weight vector of the hyperedge.



(b) Hypergraph **G**

	e_1	e_2	e_3
v_1	1	0	0
v_2	1	0	0
v_3	1	1	0
v_4	0	1	1
v_5	0	0	1
v_6	0	0	1
v_7	0	0	1

(c) Incidence matrix **H**

Figure 5 Incidence matrix of hypergraph

5 RESULTS AND DISCUSSION

For the evaluation of the proposed methods, the three different training dataset are used. IMDB movie review dataset contain two labels namely Positive and Negative. The labels are encoded in the dataset: 0 is for negative and 1 for a positive review. The second one is airplane review dataset contains three labels namely positive, negative and neutral. Approximately 15000 airplane reviews are used to train the model. After that twitter dataset are used to analyse the sentiment of tweets. The results show that better results than conventional CNN can be obtained by the deep hyper graph model with concatenation of each word embeddings only. Since there is some relevance among reviews with same sentiment polarity syntactic or semantic correlation.

Table 2 Confusion Matrix

	Predicted NO	Predicted YES
Actual NO	True Negative 11106	False Positive 2316
Actual YES	False Negative 1394	True Positive 10184

The word embeddings clustering by H-CFS algorithm can effectively assist the model with detecting meaningful latent semantic units and characterizing the original text accurately. Hyper graph learning can capture the high-order relations among samples, which has contributions to the classification accuracy. Especially, there is syntactic or semantic correlation among user reviews with same sentiment polarity.

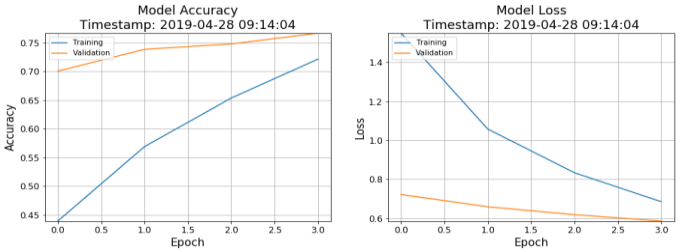


Figure 6 Accuracy and Loss of model

	PRECISION	RECALL	F1-SCORE	SUPPORT
Negative	0.89	0.83	0.86	13422
Positive	0.81	0.88	0.85	11578
Avg_Total	0.85	0.85	0.85	25000

Table 3: Precision, Recall, F1- Score

6. CONCLUSION

In this paper, a deep hyper graph scheme is proposed for online reviews modeling and sentiment classification. One property of our presented model is to construct hyper graph using feature representation of review to detect high-order relations among different reviews. Another property of the model is to use an improved hierarchical clustering algorithm to discover semantic cliques used for detecting precise semantic units as the supervision information. The three different dataset is used to perform the proposed model. The better result is achieved for all the three dataset using the clustering based deep hypergraph model. In the future work, the fusing multi-modal features and task specific embedding learning are employed to improve performance of the classification model. The proposed deep computation model will be further validated on low-end devices using real datasets in terms of CPU usage and memory usage.

REFERENCES

[1] Zhang, Q., Yang, L. T., Chen, Z., & Li, P. (2017). An improved deep computation model based on canonical polyadic decomposition. IEEE Transactions on Systems, Man, and Cybernetics: Systems, (99), 1-10.
[2] Chen, Z., Lu, F., Yuan, X., & Zhong, F. (2018). TCMHG: Topic-based cross-modal hypergraph learning for online service recommendations. IEEE Access, 6, 24856-24865.
[3] Johnson, R., & Zhang, T. (2017, July). Deep pyramid convolutional neural networks for text categorization. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 562-570).
[4] Rao, Y., Xie, H., Li, J., Jin, F., Wang, F. L., & Li, Q. (2016). Social emotion classification of short text via topic-level maximum entropy model. Information & Management, 53(8), 978-986.
[5] Luong, T., Socher, R., & Manning, C. (2013). Better word representations with recursive neural networks for morphology. In Proceedings of the Seventeenth Conference on Computational Natural Language Learning (pp. 104-113).
[6] Wang, W., Xu, H., & Wan, W. (2013). Implicit feature identification via hybrid association rule mining. Expert Systems with Applications, 40(9), 3518-3531.
[7] Zheng, X., Lin, Z., Wang, X., Lin, K. J., & Song, M. (2014). Incorporating appraisal expression patterns into topic modeling for aspect and sentiment word identification. Knowledge-Based Systems, 61, 29-47.
[8][Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
[9][Yang, H. L., & Lin, Q. F. (2013, July). Sentiment analysis in multi-scenarios: Using evolution strategies for optimization. In 2013 International Conference on Machine Learning and Cybernetics (Vol. 3, pp. 1230-1233). IEEE.
[10] Liu, C. L., Hsaio, W. H., Lee, C. H., Lu, G. C., & Jou, E. (2012). Movie rating and review summarization in mobile environment. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(3), 397-407.